eISSN 2300-262X
ISSN 0137-5075

POLISH ACADEMY OF SCIENCES
INSTITUTE OF FUNDAMENTAL TECHNOLOGICAL RESEARCH
COMMITTEE ON ACOUSTICS

ARCHIVES of ACOUSTICS

QUARTERLY

Vol. 51, No. 3, 2026

WARSAW

PAN
POLISH ACADEMY OF SCIENCES
INSTITUTE OF FUNDAMENTAL TECHNOLOGICAL RESEARCH
COMMITTEE ON ACOUSTICS

ARCHIVES of ACOUSTICS

QUARTERLY, Vol. 51, No. 3, 2026

Research Papers

N. Okur, C. Saricam, *Development of sustainable lightweight acoustic composites from denim, flax, and hemp fibers* 429

Z. Laly, R. Panneton, N. Atalla, S. Ghinet, *Modeling and investigation of a multi-resonant metamaterial sound insulator* 443

S. Arawade, J. Piechowicz, *Broadband vibration-isolating metamaterial for MRI-induced vibrations* 459

M. Meissner, *Examination of low-frequency sound decay in two connected enclosed spaces having different sound absorption on boundary surfaces* 477

C. Qian, D. Wang, H. Wang, F. Tang, M. Lin, J. Xu, *Low noise optimization of automotive air conditioning based on response surface methodology* 495

G. Rusinek, A.P. Pietrzak, *The impact of the VR interface on the accuracy and ergonomics of spatial hearing tests* 507

J. Yang, H. Liu, L. Shi, L. Gan, H. Nishizaki, C.S. Leow, *Adversarial audio inpainting with selective state space model, efficient attention, and large-scale pre-trained model* 515

S.P. Sakthidevi, C. Divya, *Audio strips network (ASNet) and amalgamated audio features (A2F): a synergistic approach for audio source separation* 533

T. Łukaszewicz, D. Kania, *Musical key detection using a triple composite signature of fifths* 549

M. Kaur, S. Kumar, V. Sharma, *Dispersion characteristics of shear horizontal-waves in a rotating micro-structured plate with restrained boundaries* 561

Technical Notes

H. Daniszewski, K. Szklanny, P. Wrzeciono, *Implementation of a COVAREP-integrated application for voice analysis using LLMs* 581

O. Deeb, S. Mahmoud, L. Saleh, A. Jafar, O. Al Dakkak, I. Chouaib, *From speech to underwater acoustics: a transfer learning framework for real-time passive diver detection using keyword spotting models* 601

Editorial Board

Editor-in-Chief: Nowicki Andrzej (Institute of Fundamental Technological Research PAS, Poland)
Deputy Editor-in-Chief: Gambin Barbara (Institute of Fundamental Technological Research PAS, Poland)

Associate Editors

General linear acoustics and physical acoustics
Ambroziński Łukasz (AGH University of Krakow, Poland)
Rdzanek Wojciech P. (University of Rzeszów, Poland)
Snakowska Anna (AGH University of Krakow, Poland)
Szemela Krzysztof (University of Rzeszów, Poland)

Architectural acoustics
Kamisiński Tadeusz (AGH University of Krakow, Poland)
Meissner Mirosław (Institute of Fundamental Technological Research PAS, Poland)
Pilch Adam (AGH University of Krakow, Poland)

Musical acoustics and psychological acoustics
Miśkiewicz Andrzej (The Fryderyk Chopin University of Music, Poland)
Preis Anna (Adam Mickiewicz University, Poland)
Wicher Andrzej (Adam Mickiewicz University, Poland)

Underwater acoustics and nonlinear acoustics
Marszal Jacek (Gdańsk University of Technology, Poland)

Speech, computational acoustics, and signal processing
Drgas Szymon (Poznan University of Technology, Poland)
Kłaczyński Maciej (AGH University of Krakow, Poland)
Kociński Jędrzej (Adam Mickiewicz University, Poland)

Ultrasonics, transducers, and instrumentation
Gambin Barbara (Institute of Fundamental Technological Research PAS, Poland)
Opieliński Krzysztof (Wrocław University of Science and Technology, Poland)
Tasinkiewicz Jurij (Institute of Fundamental Technological Research PAS, Poland)

Sonochemistry
Dzida Marzena (University of Silesia in Katowice, Poland)

Electroacoustics
Żera Jan (Warsaw University of Technology, Poland)

Vibroacoustics, noise control and environmental acoustics
Adamczyk Jan Andrzej (Central Institute for Labor Protection – National Research Institute, Poland)
Klekot Grzegorz (Warsaw University of Technology, Poland)
Kompała Janusz (Central Mining Institute, Poland)
Leniowska Lucyna (University of Rzeszów, Poland)
Piechowicz Janusz (AGH University of Krakow, Poland)
Pleban Dariusz (Central Institute for Labor Protection – National Research Institute, Poland)

Journal Managing Editor: Jezierska Eliza (Institute of Fundamental Technological Research PAS, Poland)

Advisory Editorial Board

Chairman: Tortoli Piero (University of Florence, Italy)
Kozaczka Eugeniusz (Polish Academy of Sciences, Poland)

Batko Wojciech (AGH University of Krakow, Poland)
Blauert Jens (Ruhr University, Germany)
Bradley David (The Pennsylvania State University, USA)
Crocker Malcolm J. (Auburn University, USA)
Dobrucki Andrzej (Wrocław University of Science and Technology, Poland)
Hansen Colin (University of Adelaide, Australia)
Hess Wolfgang (University of Bonn, Germany)
Leighton Tim G. (University of Southampton, UK)
Lewin Peter A. (Drexel University, USA)
Maffei Luigi (Second University of Naples SUN, Italy)
Pustelny Tadeusz (Silesian University of Technology, Poland)
Serebryany Andrey (P.P. Shirshov Institute of Oceanology, Russia)
Sundberg Johan (Royal Institute of Technology, Sweden)
Śliwiński Antoni (University of Gdańsk, Poland)
Tittmann Bernhard R. (The Pennsylvania State University, USA)
Vorländer Michael (Institute of Technical Acoustics, RWTH Aachen, Germany)

Polish Academy of Sciences
Institute of Fundamental Technological Research PAS
Committee on Acoustics PAS

Editorial Office
Institute of Fundamental Technological Research PAS, Pawińskiego 5B, 02-106 Warsaw, Poland
akustyka@ippt.pan.pl https://acoustics.ippt.pan.pl

Indexed in BazTech, Science Citation Index-Expanded (Web of Science Core Collection), ICI Journal Master List, Scopus, PBN – Polska Bibliografia Naukowa, Directory of Open Access Journals (DOAJ)
Recognised by The International Institute of Acoustics and Vibration (IIAV)

Edition co-sponsored by the Ministry of Science and Higher Education

PUBLISHED IN POLAND
Typesetting in LaTeX: Jezierska Katarzyna (Institute of Fundamental Technological Research PAS, Poland)

Archives of Acoustics Vol. 51, No. 3, pp. 429–442 (2026), https://doi.org/10.24423/archacoust.2026.4349

Research Paper

Development of Sustainable Lightweight Acoustic Composites from Denim, Flax, and Hemp Fibers

Nazan Okur*, Canan Saricam

Textile Engineering, Istanbul Technical University
Istanbul, Türkiye

*Corresponding Author: okurn@itu.edu.tr

Received September 26, 2025; revised April 5, 2026; accepted April 8, 2026; available online April 22, 2026; version of record July 13, 2026; published issue September 23, 2026.

The textile and clothing industries pose a serious problem in terms of environmental sustainability as they are among the sectors that produce the large amounts of waste worldwide. In this context, recycling both pre-consumer and post-consumer clothing waste into composite materials is of great importance in terms of both waste management and materials science. In this study, composite panels were developed using waste denim fabrics with different natural fiber compositions, recycled flax fibers derived from waste linen fabrics, and virgin hemp fibers as reinforcement materials. Polypropylene (PP)/polyethylene (PE) nonwoven interlining waste served as the matrix material, and the panels were fabricated by hot pressing at two different reinforcement-to-matrix ratios. The acoustic properties were evaluated through measurements of the sound absorption coefficient (SAC), noise reduction coefficient (NRC), and sound transmission loss (STL). In the case of denim waste, the results showed that flax- or hemp-blended samples provided better sound absorption than 100 % cotton samples, especially at a low matrix ratio. For the rest, while pure flax samples absorbed sound well, pure hemp samples showed good sound barrier properties with high STL.

Keywords: recycling, denim waste, hemp, flax, linen fabric, cotton, composite, sound absorption, sound insulation.

Copyright © 2026 The Author(s).
This work is licensed under the Creative Commons Attribution 4.0 International CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

Notations

NFC – natural fiber composite,
NRC – noise reduction coefficient,
PE – polyethylene,
PLA – polylactic acid,
PP – polypropylene,
SAC – sound absorption coefficient,
STL – sound transmission loss.

1. Introduction

Environmental concerns and awareness have led to the use of natural fibers in composite materials to achieve sustainability in several industrial applications. Although synthetic fibers meet the requirements of many applications due to their higher tensile strength, natural fibers possess higher specific tensile strength due to their low density, making them competitive candidates (Lambrache, 2022). In addition to their environmentally friendliness, light weight, high specific strength, and in most cases, low cost, natural fibers also have advantages such as low thermal conductivity and high sound absorption. Due to these features, they are intensively used as reinforcement in construction, building, and insulation materials (Pichardo *et al.*, 2018; Ahmad *et al.*, 2015; Takagi, 2019).

The use of natural fibers as reinforcement material in composites developed for noise control, which is important for human health and comfort, has drawn the attention of many researchers seeking to develop sustainable and environmentally friendly materials alternative to silica-based and petroleum-based materials such as glass fiber, rock wool, mineral wool, and expanded polystyrene, which are commonly used in acoustic applications (Hassani *et al.*, 2021). By using natural fibers such as jute, ramie, flax, kenaf, coir, abaca, cabuya, totora, and palm fiber, lightweight composite materials with satisfactory acoustic properties were obtained (Hassan *et al.*, 2021; Pawar *et al.*, 2022; Das *et al.*, 2022; Majumder *et al.*, 2023; Bravo-Moncayo *et al.*, 2024).

Natural fiber composite (NFC) used in acoustic and other applications can be produced from virgin fibers as well as waste and recycled fibers, further supporting sustainability. Therefore, the promising results obtained in the development of NFCs for acoustic applications encouraged researchers to investigate the acoustic characteristics of waste fiber-reinforced composites. In the literature, there are studies using recycled natural fibers, cotton fly, coconut/coir husk, sugarcane waste, industrial banana fiber waste, and industrial waste tea fiber as reinforcing materials in the development of composites for acoustic applications (Ersoy, Küçük, 2009; Putra *et al.*, 2013; Othmani *et al.*, 2016; Prabhu *et al.*, 2019; Hassan *et al.*, 2020; Gokulkumar *et al.*, 2021; Santoni *et al.*, 2021; Mondal *et al.*, 2022).

In these studies, fibers obtained from plant wastes found in certain geographical regions were generally used. In other words, agrowastes that are generally disposed of through throwing away, in situ burning, and composting were used. In some of these studies, hybrid reinforcement materials were used, including synthetic fibers such as E-glass, were used in the developed composites (Prabhu *et al.*, 2019; Gokulkumar *et al.*, 2021). As matrix materials, thermoset epoxy resin, polyurethane and polyester binder, and urea formaldehyde/ammonium chloride were used.

On the other hand, population growth and the corresponding increase in production increased the amount of waste generated in the textile and apparel industry, making it a critical issue (Sezgin *et al.*, 2021). Thus, in addition to agrowaste, it is important to investigate the use of wastes from textile products made of natural fibers in this field. Textile waste can be classified as production, pre-consumer, and post-consumer waste (Yalcin-Enis *et al.*, 2019). Recycling and reusing the waste generated in all these stages not only lessen the amount of waste but also reduce the burden on the environment. Because of the need to address global challenges in terms of sustainability and promotion of circular textile and fashion systems, the adoption of recycling and reuse of textile waste that replaces virgin fibers in the production of textiles, garments, accessories, and textile-reinforced composite materials has become more important (Pichardo *et al.*, 2018). According to 2023 data, the global production volume of virgin cotton fiber and other plant-based fibers was 24.4 million tons and 6.7 million tons, respectively. Cotton and non-cotton plant-based fibers accounted for about 29 % of the total global fiber production volume in 2023; however, the market share of recycled cotton was estimated at around 1 % of total cotton production. On the other hand, the share of recycled fibers is expected to grow significantly in the coming years to achieve a significant reduction in environmental pollution (Textile Exchange, 2024). For this to happen, new ideas, research, and development are needed for the use of waste and recycled materials from design to end use. At this context, transforming textile and clothing wastes into composite materials is an effective strategy to reduce environmental impacts and recycle industrial waste.

In the last few years, studies have been conducted for evaluating the acoustic and/or other insulation properties of composites reinforced with natural fibers sourced from apparel production or post-consumer waste. Among these, denim fabric waste, particularly 100 % cotton denim widely used in the fashion industry, has been extensively utilized. Recycled cotton denim fibers have been used as reinforcement in composite panels produced via hot pressing, with polypropylene (PP) fibers, ground PP derived from food and cleaning containers, and polyethylene (PE) granules obtained from bottle caps serving as matrix materials (Sezgin *et al.*, 2021; Kucukali-Ozturk *et al.*, 2022; Oztemur *et al.*, 2023). Similarly, Mago *et al.* (2024) used hydraulic compression molding to produce panels reinforced with denim shoddy and matrixed with recycled PP from cosmetic containers. Raj *et al.* (2020) investigated recycled denim shoddy and waste jute fibers as sustainable sound-absorbing materials, demonstrating that both exhibit comparable or superior acoustic performance to glass wool, with denim shoddy achieving up to tenfold and jute fiber up to fivefold cost advantages. Islam *et al.* (2022) used 100 % recycled cotton denim

with Sorona or polylactic acid (PLA) binder fibers to develop acoustic insulation panels, while Hassani *et al.* (2021) created phenolic resin-bonded composites using an air-laid nonwoven processes. Although denim is primarily made from cotton, blends with other fibers such as elastane (for comfort) and polyester (for durability and cost reduction) are also common. Jucienė *et al.* (2022) produced interior panels using cotton/PES denim waste and confirmed the potential of textile waste as the main component in acoustic applications.

Recent sustainability trends have also led to the inclusion of plant-based fibers like hemp and flax in denim fabrics. While most studies have focused on 100 % cotton denim or undefined 'denim shoddy,' there is limited information on the effect of fiber composition. However, evaluating the performance of NFCs made from denim waste of varied fiber contents is crucial for a sustainable design approach within the textile value chain. Therefore, this present study aims to develop 100 % recycled composite panels using denim waste containing different natural fibers: cotton, hemp, and flax, as reinforcement, and nonwoven interlining waste as the matrix. Since flax and hemp fibers are generally used in denim fabrics at low proportions, recycled flax fibers obtained from waste linen fabrics and virgin hemp fibers were also employed in this study in order to achieve higher blend ratios and to allow direct comparison with pure flax and pure hemp composites. Two different reinforcement-to-matrix ratios were used in the production of composite panels, their acoustic performance was tested, and their potential use in buildings and vehicles as insulation materials was evaluated. Investigating recycling alternatives for denim, a timeless fashion product, particularly with diverse natural fiber compositions, is expected to offer valuable contributions to sustainability.

2. Experimental part

2.1. Materials

Fabric wastes made of natural fibers were used as fiber reinforcement material in the composite panels. Cotton fiber was obtained from waste denim fabric, and flax fiber was obtained from waste linen fabric. Hemp fiber was supplied in virgin form by Karacasu Tekstil. These fibers were used in the production of composite panels containing pure fibers and 50/50 blends. In addition, 80/20 cotton/flax, and 80/20 cotton/hemp blends were obtained from waste denim fabrics. PP/PE (80/20) spunbonded nonwoven interlining waste was used as the matrix material in the composite panels. Figure 1 shows the fibers and matrix material used in the study. The areal density of the nonwoven interlining was measured to be approximately 20 g/m^2. The average densities of cotton, flax, hemp, PP, and PE fibers were approximated in accordance with the literature (Lu *et al.*, 2012; Yan *et al.*, 2014; Rohen *et al.*, 2017; Unterweger, 2014; Çolakoğlu, 2006).

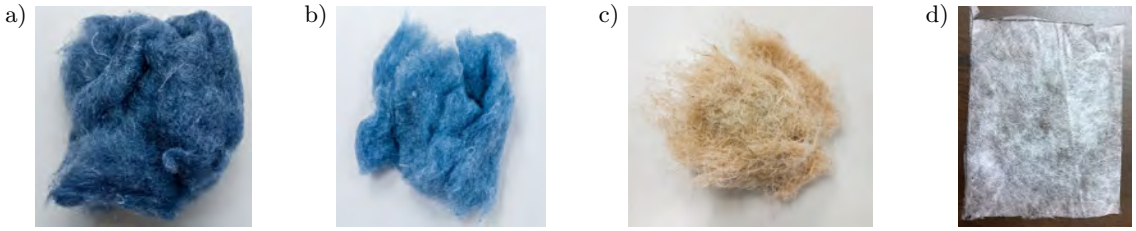


Fig. 1. Fibers and matrix material: a) cotton, b) flax, c) hemp, d) PP/PE nonwoven.

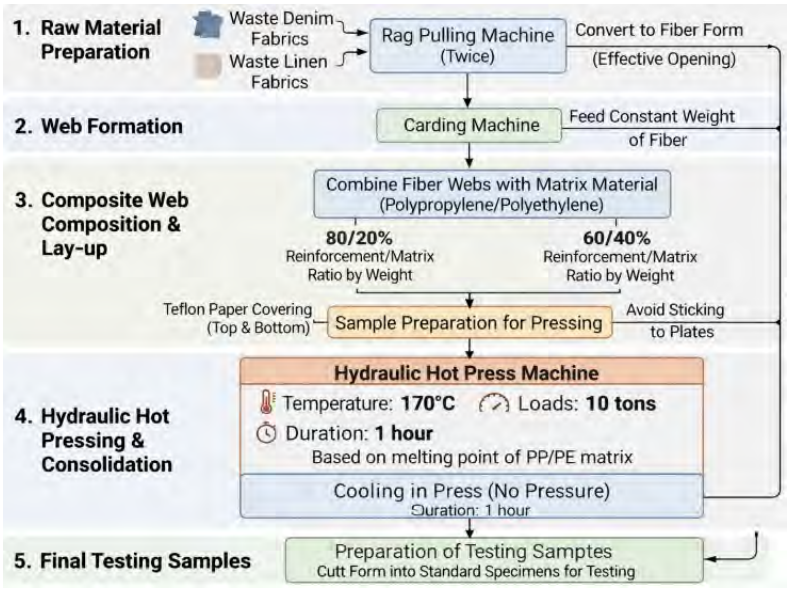
The characteristics of the materials used in the experiments are shown in Table 1. Due to the mechanical recycling process (rag pulling), the fibers exhibited a non-uniform length distribution. The length of the hemp fibers was reported as approximately 36 mm according to the supplier's specifications.

Table 1. Characteristics of the materials used.

Material type	Source	Fiber type	Form	Density [g/cm^3]
Waste denim fabric	Post-consumer textile waste	Cotton	Recycled (mechanically opened)	~1.51
Waste linen fabric	Post-consumer textile waste	Flax	Recycled (mechanically opened)	~1.50
Hemp fiber	Karacasu Tekstil (Türkiye)	Hemp	Virgin	~1.40
PP/PE spunbonded nonwoven	Industrial interlining waste	PP/PE (80/20)	Nonwoven	~0.91–0.97

2.2. Production of composite panels

Waste denim fabrics and waste linen fabric were converted into fiber form by using a rag-pulling machine twice for effective opening. The fibers were then transferred to a carding machine to obtain fiber webs by feeding a constant mass of fibers. The fiber webs were then combined with the matrix material at reinforcement-to-matrix ratios of 80 %/20 % and 60 %/40 %. A series of composite panels were produced by a hydraulic hot press machine. The webs were placed on the bottom plate of hot press machine, with Teflon paper covering the bottom and top surfaces of the samples to prevent the sticking of matrix material onto the plates. The composite panels were pressed at 170 °C, based on the melting temperature of the PP/PE matrix material, for 1 h under a load of 10 tons. The samples were then cooled for 1 h in the press without applying pressure. The production steps of the composite panels are presented in Fig. 2.



a) b) c) d)

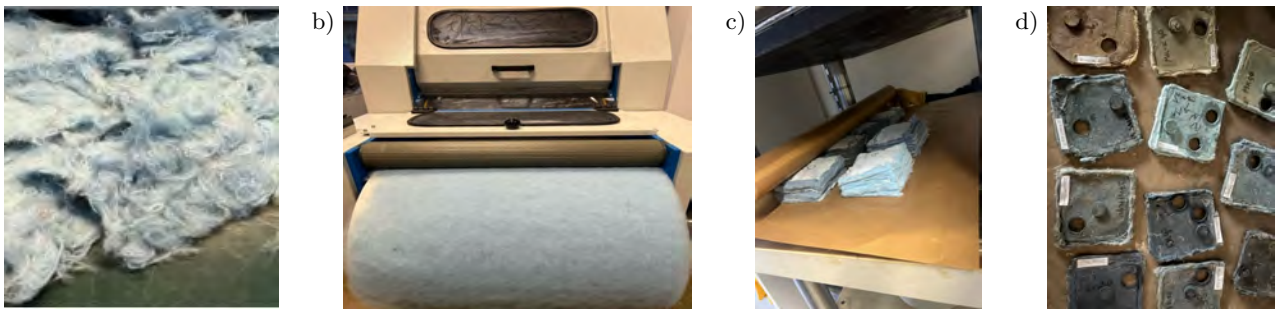


Fig. 2. Production steps of the composite panels: a) recycled fibers, b) carding process, c) samples for hot pressing, d) preparation of test samples.

2.3. Testing methods

The thickness of the composite panels was measured using a digital caliper. Areal densities were calculated by dividing the measured mass [g] by the measured area of the sample [cm^2]. To calculate the bulk density, the areal density was divided by the sample thickness [cm]. The porosity of the samples was calculated according to ASTM D2734-16 (2016) standard by using:

$$\text{Porosity } (\Phi) = \frac{\rho_{ct} - \rho_e}{\rho_{ct}} \times 100, \tag{1}$$

where ρ_{ct} is the theoretical density, and ρ_e is the experimental density (bulk density) of the composite panels. The theoretical density of the composite panels was calculated as the weighted average fiber density of the sample.

The tortuosity of the composite panels, which is the ratio of the actual flow path length to the straight distance between the ends of the flow path, was calculated by using the empirical formula based on porosity (Fatima, Mohanty, 2011):

$$\text{Tortuosity}\,(\tau) = 1 + \frac{1-\Phi}{2\Phi}. \tag{2}$$

The sound absorption coefficient (SAC) and sound transmission loss (STL) of the samples were measured by an impedance tube (TestSENS acoustic performance measurement system) developed by BIAS Engineering (Türkiye) shown in Fig. 3. The impedance tube was selected because it allows measurements on small samples. Circular samples with a diameter of 29 mm were prepared, the sound absorption test was conducted according to ASTM E1050-19 (2019) standard using a two-microphone impedance measurement tube, whereas the STL test was conducted according to ASTM E2611-17 (2019) standard using a four-microphone impedance measurement tube at frequencies from 200 Hz to 6400 Hz at room temperature. Schematic diagrams of impedance tube with two microphones for SAC measurements, and with four microphones for STL measurements are shown in Fig. 4 and Fig. 5, respectively.

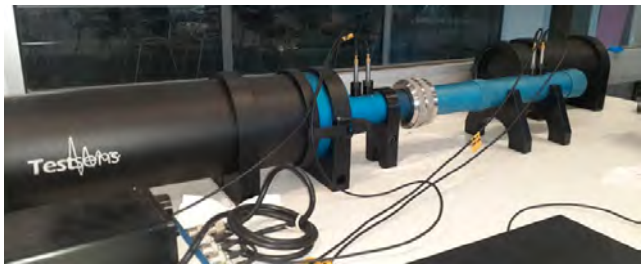


Fig. 3. Impedance tube used for SAC and STL measurements.

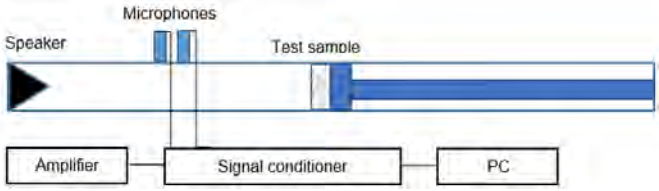


Fig. 4. Schematic diagram of the impedance tube with two microphones for SAC measurements.

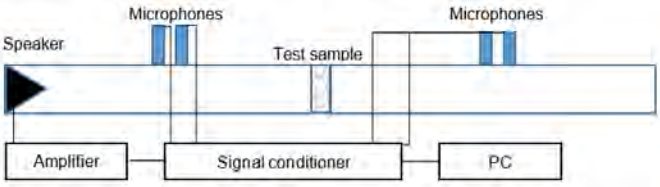


Fig. 5. Schematic diagram of the impedance tube with four microphones for STL measurements.

The SAC and STL of the samples were calculated by the measurement system according to the following equations:

$$\text{SAC}(\alpha) = \frac{\text{energy of sound absorbed by the surface}}{\text{total sound energy incident on the surface}}, \tag{3}$$

$$\text{STL}\,(\text{db}) = 10\log 10\frac{\text{power of the incident wave energy}}{\text{power of the transmitted wave energy}}. \tag{4}$$

The noise reduction coefficient (NRC), which enables the absorption performance to be expressed as a single quantitative value, was calculated as the arithmetic mean of the SAC values at frequencies of (250, 500, 1000, 2000, 3000, and 3500) Hz (see Eq. (5)). It corresponds to the ability of a material to absorb broad-spectrum sound, and since textile materials are porous absorbers and provide higher sound absorption at medium to high frequencies, the SAC values at 3000 Hz and 3500 Hz were included in the calculation of NRC (Lou *et al.*, 2005;

Hassanzadeh *et al.*, 2014; Selver, 2019; Sağlam *et al.*, 2022). The calculated NRC values were rounded to the nearest 0.05 according to ASTM C423-17 (2017) standard:

$$\text{NRC} = \frac{\alpha 250 + \alpha 500 + \alpha 1000 + \alpha 2000 + \alpha 3000 + \alpha 3500}{6}. \tag{5}$$

3. Results and discussion

The physical properties of the composite samples produced with different fiber and matrix ratios are shown in Table 2. In particular, the 100 % hemp panels exhibited the highest bulk density and the lowest porosity values, while lower density and higher porosity were obtained for pure flax and flax-blended samples. Notably, the thinnest panels were those composed entirely of hemp.

Table 2. Characteristics of the composite panels.

Sample code	Composition	Fiber ratio	Matrix ratio	Sample thickness [mm]	Sample weight [g]	Areal density [g/cm^2]	Bulk density [g/cm^3]	Weighted average density [g/cm^3]	Porosity (Φ)	Tortuosity (τ)
Cotton_20	100 % cotton	80	20	4.34	1.674	0.254	0.584	1.975	0.704	1.104
Flax_20	100 % flax			5.56	1.407	0.213	0.383	1.967	0.805	1.078
Hemp_20	100 % hemp			3.90	2.309	0.350	0.897	1.887	0.525	1.125
50Co50Flax_20	50 % cotton/ 50 % flax			5.12	1.202	0.182	0.356	1.971	0.820	1.074
50Co50Hemp_20	50 % cotton/ 50 % hemp			4.74	1.392	0.211	0.445	1.931	0.770	1.089
50Hemp50Flax_20	50 % hemp/ 50 % flax			4.58	1.343	0.203	0.444	1.927	0.770	1.089
80Co20Flax_20	80 % cotton/ 20 % flax			5.48	1.181	0.179	0.326	1.973	0.835	1.069
80Co20Hemp_20	80 % cotton/ 20 % hemp			5.17	1.307	0.198	0.383	1.957	0.804	1.079
Cotton_40	100 % cotton	60	40	4.94	2.195	0.332	0.673	1.712	0.607	1.119
Flax_40	100 % flax			4.82	1.655	0.251	0.520	1.706	0.695	1.106
Hemp_40	100 % hemp			3.91	2.021	0.306	0.783	1.646	0.524	1.125
50Co50Flax_40	50 % cotton/ 50 % flax			5.23	1.968	0.298	0.570	1.709	0.666	1.111
50Co50Hemp_40	50 % cotton/ 50 % hemp			4.55	1.606	0.243	0.535	1.679	0.681	1.109
50Hemp50Flax_40	50 % hemp/ 50 % flax			5.28	1.890	0.286	0.542	1.676	0.676	1.109
80Co20Flax_40	80 % cotton/ 20 % flax			5.37	1.607	0.243	0.453	1.710	0.735	1.097
80Co20Hemp_40	80 % cotton/ 20 % hemp			5.65	1.730	0.262	0.464	1.698	0.727	1.099

In the present study, hemp fiber was used in its virgin raw fiber form in the pure hemp and 50/50 blended composites. This means it was not recycled from fabrics that had undergone some finishing treatments. The intrinsic stiffness of natural fibers plays a decisive role in determining the porosity of the resulting composites. As fiber stiffness increases, the fiber elastic recovery ability decreases and they deform less when compressed by the matrix, which results in higher composite density and lower porosity. Fibers with a high elastic modulus, such as hemp (~23.5 GPa to 90 GPa), form more rigid and compact reinforcement networks, which facilitates closer fiber packing and consequently leads to lower porosity levels in the composite. In contrast, cotton fibers, with a significantly lower stiffness (~5.5 GPa to 12.6 GPa), are more flexible and deformable, often resulting in irregular packing and higher porosity. Although flax fibers also have a high stiffness (~27.6 GPa to 103 GPa), comparable to that of hemp, the flax fibers used in the pure flax and 50/50 blended composites were recycled from fabric. Due to the chemical treatments in textile processing, their stiffness had decreased. This resulted in higher porosity and lower bulk density in these samples (Yan *et al.*, 2014).

Significant differences were generally observed in thickness, density, and porosity values between the samples containing 80 % and 60 % fiber. While lower bulk density and higher porosity were observed in the 80 % fiber samples, bulk density increased and porosity decreased, increasing the matrix amount in the 60 % fiber samples. In addition, the tortuosity values increased slightly with increasing matrix content and varied between 1.07 and 1.12.

Table 3 provides the first SAC peak for each sample, along with the corresponding frequency. It also indicates the frequency range over which the SAC exceeded 0.5 – a value commonly considered the threshold for effective sound absorption (Khair *et al.*, 2016). The SAC curves of the samples are shown in Fig. 6, Fig. 7, and Fig. 8 over the frequency range from 200 Hz to 6400 Hz.

Table 3. Maximum SAC values of the composite panels.

Sample code	Frequency range [Hz] of $\alpha > 0.5$	First peak SAC (α)	First peak SAC frequency [Hz]
Cotton_20	–	0.418	5054
Flax_20	1382–6400	0.531	1674
Hemp_20	–	0.227	4330
50Co50Flax_20	1684–6400	0.726	4636
50Co50Hemp_20	1740–6400	0.722	2916
50Hemp50Flax_20	2306–6400	0.569	5686
80Co20Flax_20	1716–6400	0.577	2158
80Co20Hemp_20	1570–6400	0.569	1764
Cotton_40	–	0.346	3112
Flax_40	1440–6400	0.693	2336
Hemp_40	–	0.171	5752
50Co50Flax_40	3604–6400	0.735	5886
50Co50Hemp_40	2806–6400	0.635	2712
50Hemp50Flax_40	2809–6400	0.542	5146
80Co20Flax_40	2004–6400	0.501	2848
80Co20Hemp_40	[illegible]	[illegible]	2694

AoA 4349 / Rec. 23.09.2025 / Rev. 5.04.2026 / Acc. 7.04.2026 / Proof. 22.04.2026 //
archacoust.2026.4349 / Research Paper

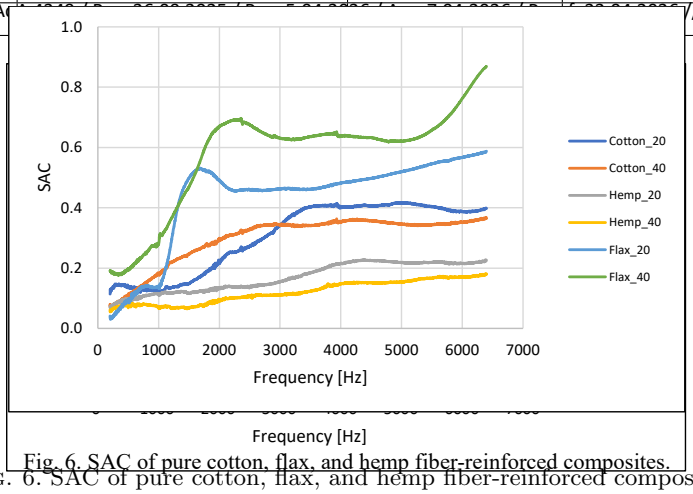


Fig. 6. SAC of pure cotton, flax, and hemp fiber-reinforced composites.

Fig. 6. SAC of pure cotton, flax, and hemp fiber-reinforced composites.

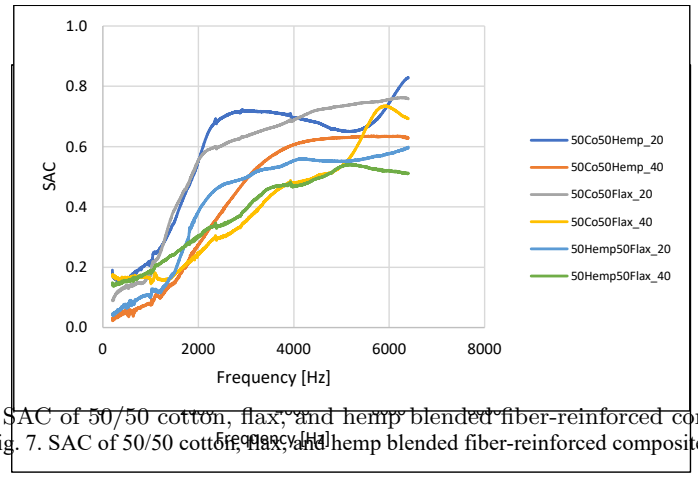


Fig. 7. SAC of 50/50 cotton, flax, and hemp blended fiber-reinforced composites.

Fig. 7. SAC of 50/50 cotton, flax, and hemp blended fiber-reinforced composites.

In the 50/50 blended samples, the combination of cotton with flax or hemp considerably improved sound absorption compared with pure cotton, with 50Co50Flax_20 and 50Co50Hemp_20 achieving maximum SAC values of 0.726 and 0.722, respectively. Lower peak SAC values were obtained for the 80/20 blended composites; however, the peak values

In the 50/50 blended samples, the combination of cotton with flax or hemp considerably improved sound absorption compared with pure cotton, with 50Co50Flax_20 and 50Co50Hemp_20 achieving maximum SAC values of 0.726 and 0.722, respectively. Lower

AoA 4349 / Rec. 26.09.2025 / Rev. 5.04.2026 / Acc. 7.04.2026 / Proof. 22.04.2026 //
archacoust.2026.4349 / Research Paper

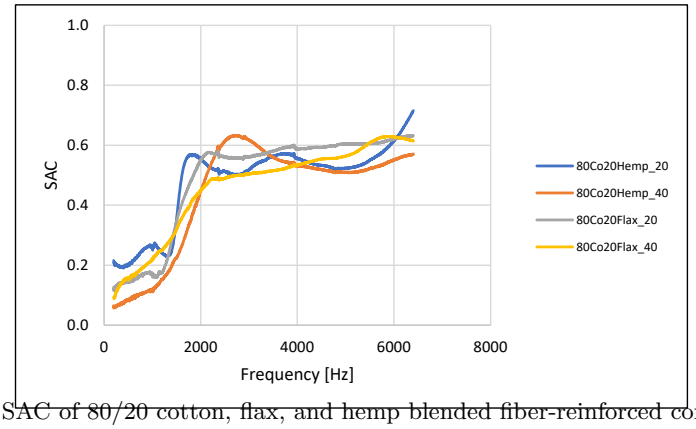


FIG. 8. SAC of 80/20 cotton, flax, and hemp blended fiber-reinforced composites.
Fig. 8. SAC of 80/20 cotton, flax, and hemp blended fiber-reinforced composites.

The results highlight that the sound absorption performance of the composite panels was found to be strongly dependent on both the fiber type and the matrix ratio. Among the pure fiber composites, the flax panels exhibited the highest performance, with Flax_20 achieving a maximum SAC of 0.531 and Flax_40 reaching 0.693. Pure cotton panels showed moderate performance, while hemp panels displayed the lowest values, particularly when the matrix ratio was increased from 20 % to 40 %, where the maximum SAC decreased from 0.227 to 0.171. The reason for this was the fact that these panels were the thinnest and had the lowest porosity, as mentioned before. Regarding the fiber type, pure flax samples, and in terms of matrix ratio, samples with a low matrix content exhibited SAC values exceeding 0.5 across a broader frequency range.

In the 50/50 blended samples, the combination of cotton with flax or hemp considerably improved sound absorption compared with pure cotton, with 50Co50Flax_20 and 50Co50Hemp_20 achieving maximum SAC values of 0.726 and 0.722, respectively. Lower peak SAC values were obtained for the 80/20 blended composites; however, the peak values were recorded at lower frequencies, especially in the samples with higher fiber content. In general, lower SAC values were recorded at higher matrix ratios, which was expected due to the decreased porosity.

Table 4. SACs of the composite panels in the range from 250 Hz to 3500 Hz with the corresponding NRC values.

Sample code	SAC						NRC
	250 Hz	500 Hz	1000 Hz	2000 Hz	3150 Hz		
Cotton_20	0.133	0.134	0.125	0.217	0.343	0.403	0.25
Flax_20	0.040	0.084	0.133	0.491	0.465	0.462	0.30
Hemp_20	0.077	0.089	0.112	0.131	0.154	0.185	0.10
50Co50Flax_20	0.103	0.135	0.179	0.554	0.634	0.661	0.40
50Co50Hemp_20	0.151	0.212	0.349	0.549	0.718	0.715	0.40
50Hemp50Flax_20	0.047	0.063	0.103	0.386	0.498	0.528	0.25
80Co20Flax_20	0.129	0.142	0.170	0.3471	0.362	0.581	0.35
80Co20Hemp_20	0.200	0.203	0.258	0.552	0.518	0.564	0.40
Cotton_40	0.076	0.105	0.181	0.292	0.343	0.340	0.20
Flax_40	0.183	0.152	0.228	0.668	0.693	0.671	0.45
Hemp_40	0.064	0.068	0.068	0.090	0.111	0.124	0.10
50Co50Flax_40	0.168	0.158	0.158	0.240	0.333	0.435	0.25
50Co50Hemp_40	0.032	0.044	0.076	0.274	0.491	0.565	0.25
50Hemp50Flax_40	0.140	0.156	0.188	0.301	0.392	0.461	0.25
80Co20Flax_40	0.108	0.158	0.217	0.448	0.504	0.545	0.30
80Co20Hemp_40	0.061	0.084	0.117	0.444	0.617	0.560	0.30

Table 4 presents the SAC values of the composite panels between 250 Hz and 3500 Hz together with the corresponding NRC values. The 100 % flax composites (Flax_20 and Flax_40) and the 50Co50Hemp_20, 50Co50Flax_20, and 80Co20Hemp_20 blended samples exhibited better sound absorption performance, with NRC values reaching 0.40–0.45. In contrast, the NRC values of 100% cotton and 100 % hemp samples remained quite low (0.10–0.25). The fiber-to-matrix ratio was also an important factor, and the samples containing 60 % fiber generally exhibited lower SACs, with performance weakening particularly at mid-frequencies (1000 Hz to 2000 Hz). An NRC of 0.40 or higher indicates effective sound absorption, while values above 0.20 classify a material as acoustic. Higher NRC values reflect better absorption (SAMUEL *et al.*, 2023). In this context, all samples except the pure hemp ones can be classified as acoustic materials; however, the Flax_20, Flax_40, 50Co50Hemp_ 20, 50Co50Flax_20, and 80Co20Hemp_20 samples, in particular, may be regarded as effective sound-absorbing materials.

Table 5 shows the STL values of the composite panels with different fiber compositions and matrix ratios across the frequency range between 250 Hz and 6300 Hz. The corresponding STL graphs as a function of frequency are shown in Fig. 9, Fig. 10, and Fig. 11.

All samples displayed a gradual increase in transmission loss with rising frequency, particularly above 4000 Hz, confirming the enhanced barrier effect of these composites at higher frequencies. Pure hemp samples exhibited the highest STL values for both matrix ratios. This can be attributed mainly to the high bulk densities of these samples. According to the mass law, as the mass per unit area of a material increases, STL rises logarithmically. Since bulk density contributes directly to the mass of the material, an increase in bulk density generally leads to higher STL values (ARJUNAN *et al.*, 2021).

The STL results of the pure and blended fiber composites revealed a clear dependence on both the fiber type and the matrix ratio. For the pure fiber composites (Fig. 9), an increase in the matrix content from 20 % to 40 % generally improved the STL values, in agreement with previous findings in the literature (SEZGIN *et al.*, 2022).

Table 4. SACs of the composite panels in the range from 250 Hz to 3500 Hz with the corresponding NRC values.

Sample code	SAC						NRC
	250 Hz	500 Hz	1000 Hz	2000 Hz	3000 Hz	3500 Hz	
Cotton_20	0.133	0.134	0.125	0.217	0.343	0.403	0.25
Flax_20	0.040	0.084	0.133	0.491	0.463	0.462	0.30
Hemp_20	0.077	0.089	0.112	0.131	0.154	0.185	0.10
50Co50Flax_20	0.103	0.135	0.179	0.554	0.634	0.661	0.40
50Co50Hemp_20	0.161	0.154	0.212	0.549	0.718	0.715	0.40
50Hemp50Flax_20	0.047	0.063	0.103	0.386	0.498	0.528	0.25
80Co20Flax_20	0.129	0.142	0.170	0.547	0.562	0.581	0.35
80Co20Hemp_20	0.200	0.203	0.258	0.552	0.518	0.564	0.40
Cotton_40	0.076	0.105	0.181	0.292	0.343	0.340	0.20
Flax_40	0.183	0.192	0.281	0.668	0.631	0.636	0.45
Hemp_40	0.064	0.068	0.068	0.090	0.111	0.124	0.10
50Co50Flax_40	0.168	0.158	0.158	0.240	0.353	0.435	0.25
50Co50Hemp_40	0.032	0.044	0.076	0.274	0.491	0.565	0.25
50Hemp50Flax_40	0.140	0.156	0.188	0.301	0.392	0.461	0.25
80Co20Flax_40	0.108	0.150	0.217	0.448	0.504	0.515	0.30
80Co20Hemp_40	0.061	0.084	0.117	0.444	0.617	0.560	0.30

Table 5. STL values of the composite panels at different frequencies.

Sample code	250 Hz	315 Hz	400 Hz	500 Hz	630 Hz	800 Hz	1000 Hz	1250 Hz	1600 Hz	2000 Hz	2500 Hz	3150 Hz	4000 Hz	5000 Hz	6300 Hz
Cotton_20	5.22	4.29	5.64	5.53	6.78	7.03	7.21	9.54	11.00	12.46	13.87	15.35	17.47	19.68	22.72
Flax_20	6.86	7.45	8.18	8.99	9.98	11.32	10.99	9.67	11.41	13.68	15.97	18.06	21.86	24.52	27.11
Hemp_20	17.77	16.95	17.76	17.84	18.57	19.10	19.44	21.21	22.15	22.75	22.05	23.18	25.12	26.55	29.35
50Co50Flax_20	4.88	4.59	5.46	5.75	6.34	7.32	6.04	7.95	9.50	11.11	12.52	14.61	16.57	17.89	19.30
50Co50Hemp_20	4.33	3.53	5.07	5.02	6.55	4.58	4.61	5.66	9.87	9.72	11.47	15.60	12.96	15.54	20.31
50Hemp50Flax_20	14.45	14.81	15.27	15.72	16.16	16.66	17.56	17.62	14.02	15.73	17.59	18.78	19.55	21.49	23.63
80Co20Flax_20	6.40	5.41	6.82	6.96	8.34	9.01	4.91	8.31	9.80	10.94	12.18	13.10	14.92	17.09	18.63
80Co20Hemp_20	13.00	11.38	12.48	12.45	13.43	13.35	8.09	11.44	13.31	15.15	16.12	17.19	20.51	22.61	26.21
Cotton_40	10.48	9.20	10.56	10.55	11.65	11.99	11.33	13.83	14.66	14.98	16.30	17.86	20.09	22.57	24.69
Flax_40	4.63	3.71	4.91	5.00	6.06	5.33	4.27	7.07	8.13	9.27	11.08	13.07	15.45	17.44	20.99
Hemp_40	32.36	31.25	31.93	31.39	33.40	33.33	33.38	33.09	33.98	34.61	35.49	34.31	32.68	31.80	34.84
50Co50Flax_40	28.94	26.63	27.33	27.02	27.22	27.56	26.71	26.74	27.05	26.81	27.86	27.22	26.15	28.70	35.96
50Co50Hemp_40	18.58	18.68	18.77	18.84	18.82	18.97	19.28	19.24	19.29	19.69	19.99	19.83	19.16	20.23	21.91
50Hemp50Flax_40	13.09	11.49	12.71	12.54	13.55	13.94	12.88	14.82	14.87	14.88	16.07	17.16	17.98	19.32	21.66
80Co20Flax_40	13.94	13.30	14.08	14.20	14.94	15.17	14.87	15.99	15.68	15.85	16.70	17.85	19.31	21.07	24.25
80Co20Hemp_40	5.54	4.12	5.27	5.28	6.40	6.62	6.37	8.76	10.07	11.20	12.20	13.69	15.69	18.02	20.57

previous findings in the literature (Sezgin *et al.*, 2022). This effect was most pronounced in the frequency range, followed by the cotton- and flax-based composites. This can be attributed to hemp-based samples, where Hemp_40 consistently exhibited the highest STL across the tested the higher bulk density of the hemp fiber composites, contributing to improved STL(Berardi, Iannace, 2015).

frequency range, followed by the cotton- and flax-based composites. This can be attributed to
the higher bulk density of the hemp fiber composites, contributing to improved STL(Berardi,
Iannace, 2

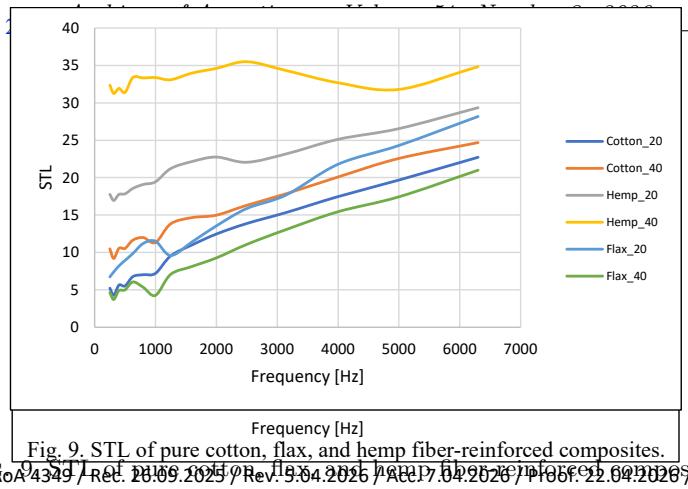


Fig. 9. STL of pure cotton, flax, and hemp fiber-reinforced composites.

Fig. 9. STL of pure cotton, flax, and hemp fiber-reinforced composites.

AoA 4349 / Rec. 26.09.2025 / Rev. 5.04.2026 / Acc. 7.04.2026 / Proof. 22.04.2026 //
archacoust.2026.4349 / Research Paper

For the 50/50 blended composites (Fig. 10), the positive influence of the higher matrix ratio was also evident. In particular, 50Co50Flax_40 showed superior STL performance compared with the other blends. However, the 80/20 blended composites (Fig. 11) demonstrated a different trend. The 80Co20Hemp_20 sample outperformed its 40 % counterpart, while 80Co20Flax_40 yielded higher STL values than its 20% counterpart. Nevertheless, the overall STL levels of the 80/20 blends were generally lower compared to those of the pure fiber and 50/50 blended composites, suggesting that higher cotton content diminishes the acoustic barrier efficiency of the composites. Furthermore, the surface roughness caused by cracks and cavities on the surface of flax and hemp fibers may have increased the friction between the fiber surface and the sound waves, causing an increase in the STL in these samples (Sezgin *et al.*, 2022).

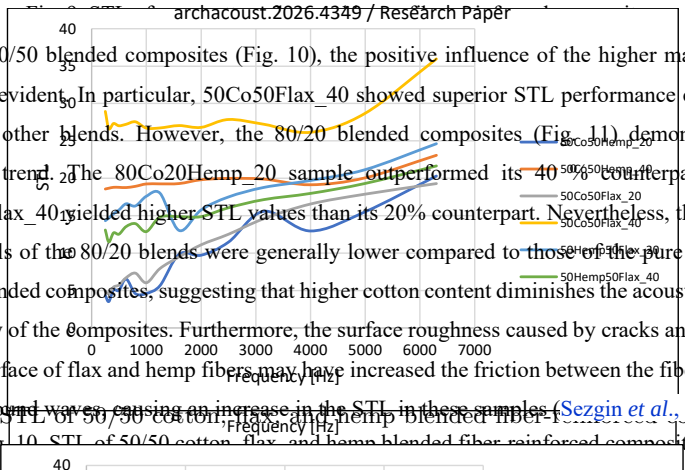


Fig. 10. STL of 50/50 cotton, flax, and hemp blended fiber-reinforced composites.

Fig. 10. STL of 50/50 cotton, flax, and hemp blended fiber-reinforced composites.

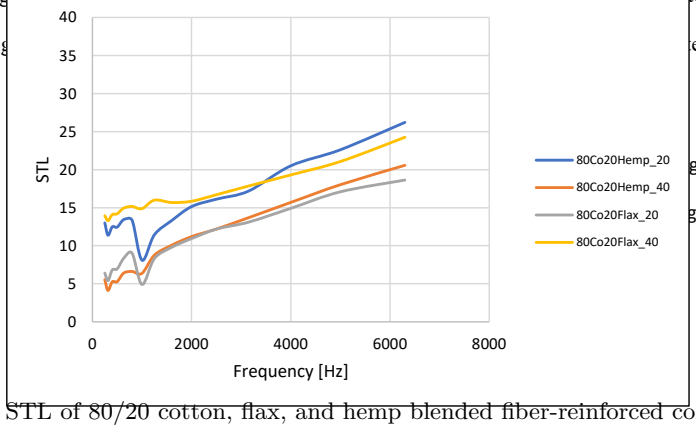


Fig. 11. STL of 80/20 cotton, flax, and hemp blended fiber-reinforced composites.

Fig. 11. STL of 80/20 cotton, flax, and hemp blended fiber-reinforced composites.

This effect was most pronounced in the hemp-based samples, where Hemp_40 consistently exhibited the highest STL across the tested frequency range, followed by the cotton- and flax-based composites. This can be attributed to the higher bulk density of the hemp fiber composites, contributing to improved STL (Berardi, Iannace, 2015).

For the 50/50 blended composites (Fig. 10), the positive influence of the higher matrix ratio was also evident. In particular, 50Co50Flax_40 showed superior STL performance compared with the other blends. However, the 80/20 blended composites (Fig. 11) demonstrated a different trend. The 80Co20Hemp_20 sample outperformed its 40 % counterpart, while 80Co20Flax_40 yielded higher STL values than its 20 % counterpart. Nevertheless, the overall STL levels of the 80/20 blends were generally lower compared to those of the pure fiber and 50/50 blended composites, suggesting that higher cotton content diminishes the acoustic barrier efficiency of the composites. Furthermore, the surface roughness caused by cracks and cavities on the surface of flax and hemp fibers may have increased the friction between the fiber surface and the sound waves, causing an increase in the STL in these samples (Sezgin *et al.*, 2022).

4. Conclusion

The present study demonstrated that denim waste and recycled natural fibers can be effectively transformed into lightweight and sustainable acoustic composites, offering a viable alternative to conventional petroleum-based materials. The results clearly showed that both fiber type and reinforcement matrix ratio significantly influence the acoustic performance. Specifically, blended composites based on denim waste exhibited higher sound absorption coefficients and NRC values compared to the 100 % cotton samples, confirming the positive contribution of flax and hemp fibers. Flax-based composites showed strong sound absorption

4. Conclusion

The present study demonstrated that denim waste and recycled natural fibers can be effectively transformed into lightweight and sustainable acoustic composites, offering a viable alternative to conventional petroleum-based materials. The results clearly showed that both fiber type and reinforcement/matrix ratio significantly influence the acoustic performance.

Specifically, blended composites based on denim waste exhibited higher sound absorption coefficients and NRC values compared to the 100 % cotton samples, confirming the positive contribution of flax and hemp fibers. Flax-based composites showed strong sound absorption performance, whereas hemp-based panels provided superior STL due to their higher bulk density and more compact structure.

Increasing the reinforcement content (i.e., reducing the matrix ratio) improved the sound absorption performance by increasing porosity and enhancing the complexity of the sound propagation paths, thereby promoting viscous and thermal dissipation mechanisms.

In terms of practical applications, such as interior acoustic panels for buildings or vehicle cabins, composite blends containing 50 % cotton/50 % flax or 50 % cotton/50 % hemp with a low matrix ratio (20 %) are recommended. These configurations achieved a high NRC of 0.40, making them effective for broad-spectrum sound absorption. For applications requiring a sound barrier effect, such as wall partitions, 100 % hemp composites with a higher matrix ratio (40 %) are the most practical recommendation due to their STL and higher bulk density.

Beyond their acoustic performance, the use of pre- and post-consumer textile wastes contributes directly to environmental sustainability by diverting valuable resources from landfills and supporting circular economy principles. Although cotton remains the primary fiber used in denim production, growing sustainability concerns and evolving consumer expectations have encouraged the incorporation of flax and hemp into denim products. The results of this study indicate that these natural fiber blends not only perform well during use but also offer advantages in terms of end-of-life waste management. Consequently, these emerging trends support sustainability across production, consumption, and post-use phases.

Future research should focus on evaluating durability under varying environmental conditions to better understand long-term performance. In addition, although the fibers in the present study were distributed in a relatively random manner, future studies should investigate controlled fiber orientation and alignment, as these factors may significantly influence acoustic performance.

Moreover, the use of bio-based matrices such as PLA should be explored to develop fully biodegradable composites and further reduce environmental impact. Finally, the incorporation of suitable additives to enhance functional properties, such as moisture resistance and fire retardancy, should be investigated to improve the applicability of these composites in construction and automotive applications.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interests

The authors declare that there are no known competing financial interests or personal relationships that could have influenced the work described in this paper.

Authors' contributions

Nazan Okur: conceptualization, methodology, material procurement, data interpretation and analysis, and writing the original manuscript. Canan Saricam: conceptualization, material procurement and preparation, experimental testing. All authors reviewed and approved the final manuscript.

Acknowledgments

The authors would like to thank Burak Kadir CECELI and Zilan OZTURK for their kind contribution to the sample preparation and the experiments carried out as part of their Senior Design Project. The authors also wish to express their thanks to Karacasu Tekstil for providing the hemp fibers.

References

1. Ahmad F., Choi H.S., Park M.K. (2015), A review: natural fiber composites selection in view of mechanical, light weight, and economic properties, *Macromolecular Materials and Engineering*, **300**(1): 10–24, https://doi.org/10.1002/mame.201400089.
2. Arjunan A., Baroutaji A., Latif A. (2021), Acoustic behaviour of 3D printed titanium perforated panels, *Results in Engineering*, **11**: 100252, https://doi.org/10.1016/j.rineng.2021.100252.
3. ASTM C423-17 (2017), *Standard test method for sound absorption and sound absorption coefficients by the reverberation room method*, ATM International.
4. ASTM D2734-16 (2016), *Standard test methods for void content of reinforced plastics*, ASTM International.
5. ASTM E1050-19 (2019), *Standard test method for impedance and absorption of acoustical materials using a tube, two microphones and a digital frequency analysis system*, ASTM International.
6. ASTM E2611-17 (2019), *Standard test method for normal incidence determination of porous material acoustical properties based on the transfer matrix method*, ASTM International.
7. Berardi U., Iannace G. (2015), Acoustic characterization of natural fibers for sound absorption applications, *Building and Environment*, **94**(Part 2): 840–852, https://doi.org/10.1016/j.buildenv.2015.05.029.
8. Bravo-Moncayo L., Argotti-Gómez M., Jara O., Puyana-Romero V., Ciaburro G., Guerrero V.H. (2024), Thermo-acoustic properties of four natural fibers, *Musa textilis*, *Furcraea andina*, *Cocos nucifera*, and *Schoenoplectus californicus*, for building applications, *Buildings*, **14**(8): 2265, https://doi.org/10.3390/buildings14082265.
9. Çolakoğlu M. (2006), Damping and vibration analysis of polyethylene fiber composite under varied temperature, *Turkish Journal of Engineering and Environmental Sciences*, **30**(6): 351–357.
10. Das P.P., Chaudhary V., Ahmad F., Manral A., Gupta S., Gupta P. (2022), Acoustic performance of natural fiber reinforced polymer composites: influencing factors, future scope, challenges, and applications, *Polymer Composites*, **43**(3): 1221–1237, https://doi.org/10.1002/pc.26455.
11. Ersoy S., Küçük H. (2009), Investigation of industrial tea-leaf-fibre waste material for its sound absorption properties, *Applied Acoustics*, **70**(1): 215–220, https://doi.org/10.1016/j.apacoust.2007.12.005.
12. Fatima S., Mohanty A.R. (2011), Acoustical and fire-retardant properties of jute composite materials, *Applied Acoustics*, **72**(2–3): 108–114, https://doi.org/10.1016/j.apacoust.2010.10.005.
13. Gokulkumar S., Thyla P.R., Prabhu L., Sathish S. (2021), Characterization and comparative analysis on mechanical and acoustical properties of camellia sinensis/ananas comosus/glass fiber hybrid polymer composites, *Journal of Natural Fibers*, **18**(7): 978–994, https://doi.org/10.1080/15440478.2019.1675215.
14. Hassan T. *et al.* (2020), Acoustic, mechanical and thermal properties of green composites reinforced with natural fibers waste, *Polymers*, **12**(3): 654, https://doi.org/10.3390/polym12030654.
15. Hassan T. *et al.* (2021), Factors affecting acoustic properties of natural-fiber-based materials and composites: a review, *Textiles*, **1**(1): 55–85, https://doi.org/10.3390/textiles1010005.
16. Hassani P., Soltani P., Ghane M., Zarrebini M. (2021), Porous resin-bonded recycled denim composite as an efficient sound-absorbing material, *Applied Acoustics*, **173**: 107710, https://doi.org/10.1016/j.apacoust.2020.107710.
17. Hassanzadeh S., Hasani H., Zarrebini M. (2014), Analysis and prediction of the noise reduction coefficient of lightly-needled Estabragh/polypropylene nonwovens using simplex lattice design, *The Journal of The Textile Institute*, **105**(3): 256–263, https://doi.org/10.1080/00405000.2013.835903.
18. Islam S., El Messiry M., Sikdar P.P., Seylar J., Bhat G. (2022), Microstructure and performance characteristics of acoustic insulation materials from post-consumer recycled denim fabrics, *Journal of Industrial Textiles*, **51**(4_suppl): 6001S–6027S, https://doi.org/10.1177/1528083720940746.

19. Juciene M., Dobilaitė V., Albrektas D., Bliūdžius R. (2022), Investigation and evaluation of the performance of interior finishing panels made from denim textile waste, *Textile Research Journal*, **92**(23–24): 4666–4677, https://doi.org/10.1177/00405175221109636.
20. Khair F.A., Putra A., Nor M.J., Selamat M.Z. (2016), Analysis of sound absorption of hollow tube absorbers, *International Journal of Automotive and Mechanical Engineering*, **13**(2): 3492–3502, https://doi.org/10.15282/ijame.13.2.2016.16.0288.
21. Kucukali-Ozturk M., Yalçın Enış I., Sezgin H. (2022), Development of 100% recycled thermoplastic composites for sound insulated acoustic panels, *Materials Science Forum*, **1053**: 352–357, https://doi.org/10.4028/p-q9e23m.
22. Lambrache N., Renagi O., Olaru L., N'Drelan B. (2022), Composite materials with natural fibers, [in:] *Fiber-Reinforced Plastics*, Masuelli M.A. [Ed.], https://doi.org/10.5772/intechopen.101818.
23. Lou C.-W., Lin J.-H., Su K.-H. (2005), Recycling polyester and polypropylene nonwoven selvages to produce functional sound absorption composites, *Textile Research Journal*, **75**(5): 390–394, https://doi.org/10.1177/0040517505054178.
24. Lu N., Swan R.H. Jr., Ferguson I. (2012), Composition, structure, and mechanical properties of hemp fiber reinforced composite with recycled high-density polyethylene matrix, *Journal of Composite Materials*, **46**(16): 1915–1924, https://doi.org/10.1177/0021998311427778.
25. Mago J., Jaish S., Negi A., Bolton J.S., Fatima S. (2024), Investigating the acoustic performance of sustainable composites utilizing recycled polypropylene and denim shoddy: fabrication, characterization, and theoretical modelling, [in:] *INTER-NOISE and NOISE-CON Congress and Conference Proceedings,* (7): 4017–4028, https://doi.org/10.3397/IN_2024_3403.
26. Majumder A., Achenza M., Mastino C.C., Baccoli R., Frattolillo A. (2023), Thermo-acoustic building insulation materials fabricated with recycled fibers – jute, wool and loofah, *Energy and Buildings*, **293**: 113211, https://doi.org/10.1016/j.enbuild.2023.113211.
27. Mondal M.S., Sarkar J., Hasan N., Mehedi H.A., Dutta P. (2022), Development of recycled natural fiber based composite material by hand lay-up process and analysis of its acoustic & physical properties, *Journal of Natural Fibers*, **19**(16): 14898–14908, https://doi.org/10.1080/15440478.2022.2069193.
28. Othmani C. *et al.* (2016), Experimental and theoretical investigation of the acoustic performance of sugarcane wastes based material, *Applied Acoustics*, **109**: 90–96, https://doi.org/10.1016/j.apacoust.2016.02.005.
29. Oztemur J., Ozdemir S., Sezgin H., Yalcin-Enis İ. (2023), Manufacturing sound absorption panel from textile waste: a design approach, [in:] *International Conference – Smart Textiles and Emerging Technologies*, **2**(1): 40–49, https://doi.org/10.61135/stet2024.
30. Pawar K.S., Bagha A.K., Bahl S., Nandan D. (2022), Significant applications of composite and natural materials for vibration and noise control: a review, *Advancement in Materials, Manufacturing and Energy Engineering*, **2**: 211–220, https://doi.org/10.1007/978-981-16-8341-1_17.
31. Pichardo P.P., Martínez-Barrera G., Martínez-López M., Ureña-Núñez F., Ávila-Córdoba L.I. (2018), Waste and recycled textiles as reinforcements of building materials, [in:], *Natural and Artificial Fiber Reinforced Composites as Renewable Sources*, Gururaja M.N. [Ed.], pp. 89–106, https://doi.org/10.5772/intechopen.70620.
32. Prabhu L., Krishnaraj V., Gokulkumar S., Sathish S., Ramesh M. (2019), Mechanical, chemical and acoustical behavior of sisal – tea waste – glass fiber reinforced epoxy based hybrid polymer composites, *Materials Today: Proceedings*, **16**(Part 2): 653–660, https://doi.org/10.1016/j.matpr.2019.05.142.
33. Putra A., Abdullah Y., Efendy H., Farid W.M., Ayob M.R., Py M.S. (2013), Utilizing sugarcane wasted fibers as a sustainable acoustic absorber, *Procedia Engineering*, **53**: 632–638, https://doi.org/10.1016/j.proeng.2013.02.081.
34. Raj M., Fatima S., Tandon N. (2020), Recycled materials as a potential replacement to synthetic sound absorbers: a study on denim shoddy and waste jute fibers, *Applied Acoustics*, **159**: 107070, https://doi.org/10.1016/j.apacoust.2019.107070.
35. Rohen L.A. *et al.* (2017), Hemp fiber density using the pycnometry technique, [in:] *Characterization of Minerals, Metals, and Materials*, pp. 423–428, Springer, https://doi.org/10.1007/978-3-319-51382-9_46.
36. Sağlam S., Güzelçimen F., Bingöl D., Özkan U.Y., Ekici B. (2022), An evaluation on acoustic behavior of untreated tree trunks depending on bark physical properties and fiber cell structure, *Journal of Natural Fibers*, **19**(4): 1507–1521, https://doi.org/10.1080/15440478.2020.1779901.

37. Samuel B.T., Barburski M., Witczak E., Puszkarz A.K. (2023), Enhancement of low and medium frequency sound absorption using fabrics and air gaps, *Textile Research Journal*, **93**(21–22): 5112–5123, https://doi.org/10.1177/00405175231186176.

38. Santoni A., Bonfiglio P., Magnani A., Marescotti C., Pompoli F., Fausti P. (2021), A hybrid approach for modelling the acoustic properties of recycled fibre mixtures for automotive applications, *Applied Acoustics*, **182**: 108272, https://doi.org/10.1016/j.apacoust.2021.108272.

39. Selver E. (2019), Acoustic properties of hybrid glass/flax and glass/jute composites consisting of different stacking sequences, *Tekstil ve Mühendis*, **26**(113): 42–51.

40. Sezgin H., Duru S.C., Candan C. (2022), Evaluation of acoustic performance of hemp and cotton fiber reinforced thermoplastic composites, *Uludağ Üniversitesi Mühendislik Fakültesi Dergisi*, **27**(1): 151–162, https://doi.org/10.17482/uumfd.925486.

41. Sezgin H., Kucukali-Ozturk M., Berkalp O.B., Yalcin-Enis I. (2021), Design of composite insulation panels containing 100% recycled cotton fibers and polyethylene/polypropylene packaging wastes, *Journal of Cleaner Production*, **304**: 127132, https://doi.org/10.1016/j.jclepro.2021.127132.

42. Takagi H. (2019), Review of functional properties of natural fiber reinforced polymer composites: thermal insulation, biodegradation and vibration damping properties, *Advanced Composite Materials*, **28**(5): 525–543, https://doi.org/10.1080/09243046.2019.1617093.

43. Textile Exchange (2024), *Materials Market Report*, https://textileexchange.org/knowledge-center/reports/materials-market-report-2024/ (access: 9.02.2026).

44. Unterweger C., Brüggemann O., Fürst C. (2014), Synthetic fibers and thermoplastic short-fiber-reinforced polymers: properties and characterization, *Polymer Composites*, **35**(2): 227–236, https://doi.org/10.1002/pc.22654.

45. Yalcin-Enis I., Kucukali-Ozturk M., Sezgin H. (2019), Risks and management of textile waste, [in:] *Nanoscience and Biotechnology for Environmental Applications*, pp. 29–35, https://doi.org/10.1007/978-3-319-97922-9_2.

46. Yan L., Chouw N., Jayaraman K. (2014), Flax fibre and its composites – a review, *Composites Part B: Engineering*, **56**: 296–317, https://doi.org/10.1016/j.compositesb.2013.08.014.

Archives of Acoustics Vol. 51, No. 3, pp. 443–458 (2026), https://doi.org/10.24423/archacoust.2026.4353

Research Paper

Modeling and Investigation of a Multi-Resonant Metamaterial Sound Insulator

Zacharie Laly[(1)]*, Raymond Panneton[(1)], Noureddine Atalla[(1)], Sebastian Ghinet[(2)]

[(1)] *Centre for Research in Acoustics-Signal-Human of the Université de Sherbrooke*
Sherbrooke, Canada

[(2)] *National Research Council Canada, Aerospace*
Ottawa, Canada

*Corresponding Author: Zacharie.Laly@usherbrooke.ca

Received October 8, 2025; revised April 11, 2026; accepted April 15, 2026;
available online April 24, 2026; version of record July 16, 2026; published issue September 23, 2026.

A design of a multi-resonant acoustic metamaterial sound insulator is proposed to attenuate multi-tonal and broadband noise. The unit cell of the metamaterial consists of a porous material with four, nine, and sixteen embedded Helmholtz resonators and its sound attenuation performance is studied using the finite element method (FEM). A theoretical approach based on the transfer matrix method is proposed and the theoretical results of the sound absorption and the transmission loss (TL) agree well with FEM results. Compared to the porous layer with one embedded resonator that presents one TL and one sound absorption resonant peak, the metamaterial made of four, nine, and sixteen embedded resonators exhibits, respectively, four, nine, and sixteen sound absorption and TL resonant peaks. The sound absorption at various incidence angles and the diffuse field sound absorption of the proposed metamaterial are studied. At each resonant frequency, the acoustic pressure reaches its maximum in the cavity of the corresponding resonator while in the associated neck, the acoustic velocity peaks alongside significant acoustic power dissipation density. As the incidence angle increases, the acoustic pressure in the cavity at the resonant frequencies decreases as well as the velocity in the necks and the power dissipation density becomes negligible, resulting in a decrease of the sound absorption. As the number of resonators within the porous layer increases, the sound absorption and the TL frequency band enlarge. Compared to the conventional porous layer, the sound absorption, and the TL are significantly improved at the resonant frequencies of the resonators.

Keywords: porous material, Helmholtz resonators, sound absorption, transmission loss, acoustic metamaterial.



Copyright © 2026 The Author(s).
This work is licensed under the Creative Commons Attribution 4.0 International CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Porous materials are widely used for noise reduction in many applications such as construction, transport, and manufacturing. Their microstructure featuring high porosity, a complex and interconnected pore network, and a large internal surface area allows the sound waves to enter the medium that dissipates their energy through viscous and thermal losses (Allard, Atalla, 2009). Porous materials are inefficient at low frequencies while they provide efficient sound dissipation at high frequencies. This limitation can be overcome by the use of inner inclusions within the porous layer. Boutin and Becot (2015) reported theoretical and experimental analysis on the acoustics of gas saturated rigid porous media with inner resonators. They observed that the resonators modify the effective bulk modulus of the medium which induces strong velocity dispersion and high attenuation around the resonance frequencies. They tested experimentally several prototypes of such media and compared the measurement with the theory. The sound attenuation performance of a cellular porous material with resonant inclusions was investigated by Doutres *et al.* (2015). Using theoretical, numerical and experimental approaches,

they demonstrated the transmission loss (TL) improvement of the material at the resonance frequency of the embedded Helmholtz resonator (HR). They illustrate that the impact of the resonator on the material's sound absorption performance can be positive or negative. Lagarrigue *et al.* (2013) studied the sound attenuation performance of porous material with arranged resonant inclusions (slotted cylinders) that are periodically embedded into the porous matrix and they observed the sound absorption improvement. Zhu *et al.* (2019) proposed a sound absorber composed of a porous layer embedded with periodic resonators made of two 180° twisted split rings. They demonstrated that a single resonator embedded in the porous material can achieve perfect absorption at its resonance frequency, which can be tuned by adjusting the resonator's inner radius. When four twisted split rings were incorporated into the porous layer, the sound absorption coefficient was improved. Abbad *et al.* (2019) studied a front membrane-cavity Helmholtz resonator embedded within a porous material and showed the transmission loss enhancement at the Helmholtz resonance. However, they found that the presence of the resonator can reduce the overall sound absorption performance, depending on the membrane's resonance frequency and they recommended the transmission configuration for application of the proposed design. Ghinet *et al.* (2020) studied the single and double wall configurations of glass wool layers with embedded periodic Helmholtz resonators for aircraft cabin applications. The TL of the studied design was experimentally evaluated under diffuse field excitation. Kone *et al.* (2021; 2022) investigated glass wool with an embedded resonator having a complex neck shape for aircraft cabin applications and they illustrated multiple sound absorption peaks. Magliacano *et al.* (2024) used the finite element method (FEM) and the transfer matrix method (TMM) to investigate the sound attenuation characteristics of porous material with an embedded periodic pattern such as hollow cylindrical inclusions, half-cut hollow cylindrical inclusions and Helmholtz resonators. The FEM results were compared with experiments. Helmholtz resonators with porous material lining were modeled by Zhang and Xin (2024) using the double porosity method. They demonstrated that the porous lining, as an additional damping helps to satisfy the impedance matching condition leading to the sound absorption improvement compared to the classic HR. Laly *et al.* (2022; 2023; 2024a) proposed transfer matrix methods to model acoustic metamaterial consisting of porous materials with embedded periodic Helmholtz resonators. The results of the TMM were validated by comparison with the FEM in single and double wall configurations. They demonstrated the impacts of the parameters of the periodic resonator on the TL and studied the influences of the elasticity of the resonator's walls on the sound attenuation performance (Laly *et al.*, 2024b). A porous material with a conventional embedded Helmholtz resonator provides only one resonant peak with a narrow frequency band.

In this study, a multi-resonant acoustic metamaterial based on a porous material with multiple embedded Helmholtz resonators is presented for multi-tonal and broadband noise reduction. The theoretical model using the transfer matrix method shows good agreement with the finite element method. The transmission loss and the diffuse field sound absorption of the metamaterial are significantly improved at the resonant frequencies of the resonators with the frequency band that increases as the number of resonators within the unit cell increases. The sound attenuation mechanism of the metamaterial is studied. The paper is organized as follows. In Sec. 2, the design of the proposed metamaterial is presented and Sec. 3 shows the theoretical modelling. The comparison of the theory with the finite element method is shown in Sec. 4 and the analysis of the sound attenuation performance is performed in Sec. 5 and finally the conclusion is presented in Sec. 6.

2. Finite element modeling of multi-resonant acoustic metamaterial

Low-frequency noise poses a challenge in many engineering applications due to its long wavelength, which makes conventional porous materials ineffective. In this paper, an acoustic metamaterial design consisting of multiple resonators periodically embedded within a porous material is proposed to effectively control multi-tonal low-frequency noise. Figure 1 presents the unit cell of the proposed metamaterial that consists of a porous material with embedded multiple Helmholtz resonators. In Fig. 1b, four different resonators are integrated within the porous material and Fig. 1c illustrates the porous layer with nine different embedded resonators. In Fig. 1d, sixteen different Helmholtz resonators are embedded within the porous layer and in Fig. 1a, the porous layer with one embedded resonator is shown for comparison. The cavities of the resonators have cylindrical shapes as well as

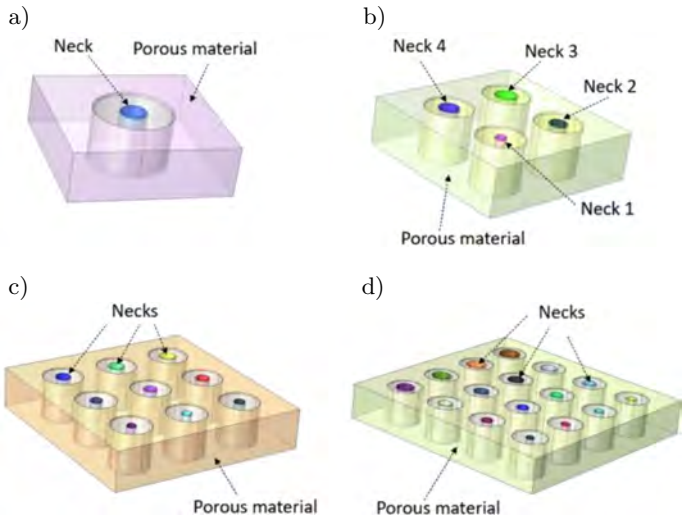


FIG. 1. Acoustic metamaterial based on porous layer with embedded multiple resonators: a) one periodic resonator, b) four periodic resonators, c) nine periodic resonators, d) sixteen periodic resonators.

the necks that are extended into the cavities. In Fig. 1, all the resonators have identical cavities with a diameter of 40 mm and in Fig. 1b, Fig. 1c, and Fig. 1d, the necks of the resonators have different geometric parameters illustrated by different colors while the cavities are separated from each other by an equal distance. The distance between the cavities axes of two adjacent resonators is 60 mm. In Fig. 1c, the nine cavities are identical and the geometric parameters of the nine necks are different, illustrated by different colors. Similarly, the sixteen cavities in Fig. 1d are identical with sixteen necks having different geometric parameters. The walls of the necks and the cavities in Fig. 1 are considered rigid and the thickness of the porous material is equal to the length of each resonator. The resonators are arranged in parallel within the porous material.

The sound attenuation performances of the metamaterial in Fig. 1 are studied numerically using COMSOL Multiphysics and the simulations are conducted with one periodic unit cell (PUC). Figure 2 shows the numerical models of the PUC for the transmission loss calculations. The PUC is made of an incident fluid, a transmission fluid and the porous layer with multiple embedded resonators. The incident and transmission fluids have identical length.

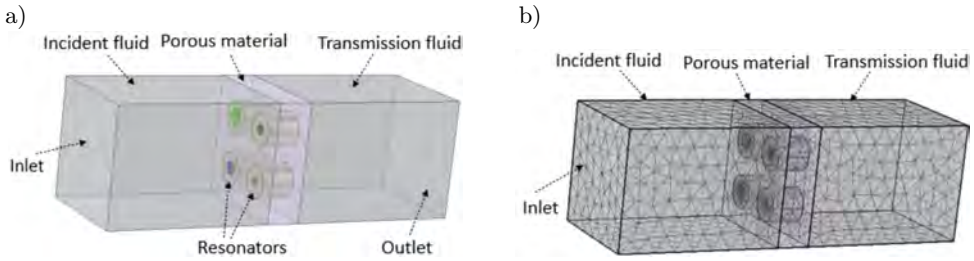


FIG. 2. PUC of the metamaterial: a) geometry with four HR, b) mesh with four HR.

A normal incidence plane wave with pressure amplitude of 1 Pa is imposed on the inlet plane in Fig. 2 and plane wave radiation condition is applied on the inlet and outlet planes while periodic boundary conditions are applied on all parallel planes in Fig. 2. The transmission loss is given by

$$\mathrm{TL} = 10 \log_{10}\left(\frac{W_i}{W_t}\right), \tag{1}$$

where W_t and W_i are the transmitted and incident sound power at the outlet and inlet planes, respectively.

Figure 3 presents the finite element models of the PUC for studying the diffuse field sound absorption performance of the metamaterial. The PUC in Fig. 3 consists of a perfectly matched layer (PML); a background

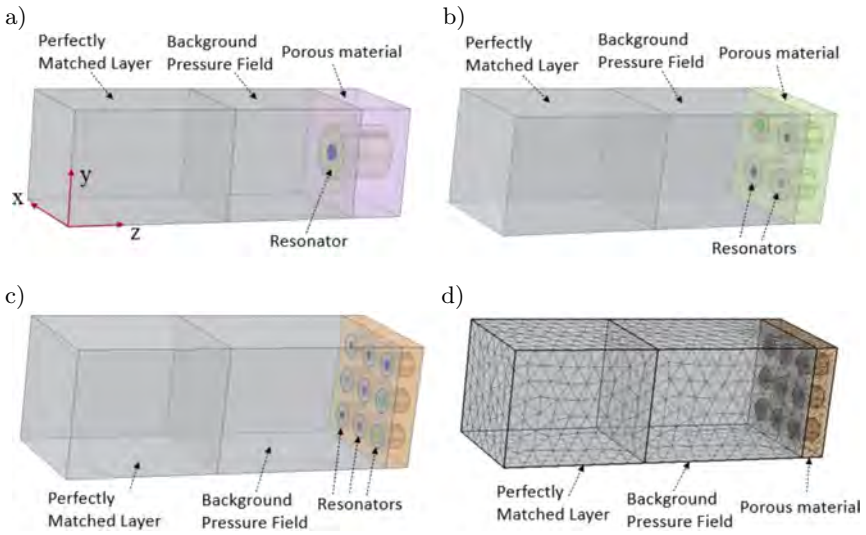


FIG. 3. PUC of the metamaterial for sound absorption estimation:
a) PUC with 1 HR, b) PUC with 4 HR, c) PUC with 9 HR, d) mesh with 9 HR.

acoustic pressure field and the porous material with embedded multiple resonators. Figure 3 shows, respectively, the PUC with one, four, nine resonators, and the mesh with nine resonators. The background pressure field provides an incident plane wave in the z-direction with an amplitude of 1 Pa, an incident angle θ, and an azimuthal angle ψ. The PML creates a virtual anechoic incident domain to eliminate the reflection of acoustic waves from the metamaterials.

The porous material as well as the air inside the neck of each resonator in Fig. 2 and Fig. 3 are modeled using the equivalent fluid model of Johnson–Champoux–Allard with the equivalent density $\rho_{eq}(\omega)$ and bulk modulus $K_{eq}(\omega)$ given by ALLARD and ATALLA (2009):

$$\rho_{eq}(\omega) = \alpha_\infty \rho_0 \left(1 + \frac{\sigma\phi}{j\omega\alpha_\infty\rho_0}\sqrt{1 + \frac{4j\rho_0\omega\eta\alpha_\infty^2}{\phi^2\sigma^2\Lambda^2}}\right), \tag{2}$$

$$K_{eq}(\omega) = \gamma P_0 \left[\gamma - \frac{\gamma - 1}{1 + \frac{8\eta}{j\omega\rho_0 \mathrm{Pr}\Lambda'^2}\sqrt{1 + \frac{j\omega\rho_0\mathrm{Pr}\Lambda'^2}{16\eta}}}\right]^{-1}, \tag{3}$$

where ϕ is the porosity, σ is the static airflow resistivity, α_∞ is the tortuosity, ρ_0 is the density, Λ is the viscous characteristic length, Λ' is the thermal characteristic length, η is the dynamic viscosity, Pr is the Prandtl number, γ is the specific heat ratio, and P_0 is the atmospheric static pressure. The tortuosity and the porosity of the air inside each neck are set to 1. The static airflow resistivity of the i-th neck is given by $\sigma_i = \frac{8\eta}{\phi_i R_i^2}$ with ϕ_i the porosity of the air inside the i-th neck and R_i the radius of the i-th neck whose viscous characteristic length is equal to R_i. The rest of the domains in Fig. 2 and Fig. 3 are characterized using the pressure acoustics module of COMSOL.

The oblique incident acoustic plane wave reaching the surface of the porous layer with embedded resonators is given by

$$P_i(\theta, \psi) = P_A e^{-j(k_x x + k_y y + k_z z)}, \tag{4}$$

where $P_A = 1$ Pa and j is the imaginary unit ($j^2 = -1$). The acoustic wave numbers k_x, k_y, and k_z in Eq. (4) are given by

$$\begin{cases} k_x = k_0 \sin\theta\cos\psi, \\ k_y = k_0 \sin\theta\sin\psi, \\ k_z = k_0 \cos\theta, \end{cases} \tag{5}$$

with $k_0 = \omega/c_0$, where ω is the angular frequency and c_0 is the speed of sound in air. Floquet periodic boundary conditions are imposed on all opposing parallel boundaries of the geometry in Fig. 3. The $\mathbf{k}$ wave vector for Floquet periodicity is given by $\mathbf{k} = [k_x\ k_y\ 0]$. The surface average of the total acoustic pressure P_t at the inlet surface of the metamaterial in Fig. 3 is determined numerically and the complex reflection coefficient is then obtained by

$$R(\theta,\psi) = \frac{P_t(\theta,\psi) - P_i(\theta,\psi)}{P_i(\theta,\psi)}. \tag{6}$$

The normalized surface acoustic impedance $Z_s(\theta,\psi)$ and the sound absorption coefficient $\alpha(\theta,\psi)$ of the metamaterial for each incidence angle θ are expressed by

$$Z_s(\theta,\psi) = \frac{1 + R(\theta,\psi)}{(1 - R(\theta,\psi))\cos(\theta,\psi)}, \qquad \alpha(\theta,\psi) = 1 - |R(\theta,\psi)|^2. \tag{7}$$

The diffuse field sound absorption coefficient for different incident waves with an incidence angle θ varying between 0 and $\theta_{\max}$, where $\theta_{\max}$ is the maximum incidence angle is then determined by (Laly *et al.*, 2025):

$$\alpha_D = \frac{\int_0^{\theta_{\max}} \alpha(\theta)\sin(2\theta)\,\mathrm{d}\theta}{\int_0^{\theta_{\max}} \sin(2\theta)\,\mathrm{d}\theta}. \tag{8}$$

A discretized form of this equation is used in the numerical calculations with $\theta_{\max}$ set to $\pi/2$ and the azimuthal angle ψ set to zero.

The thickness of the porous layer is set to 40 mm which is equal to the length of the cavity of each resonator. The static airflow resistivity of the porous layer is 26 000 $\mathrm{N\cdot s\cdot m^{-4}}$ with open porosity of 99 %, a tortuosity of 1.02 and the viscous and thermal characteristic lengths are respectively 150 µm and 300 µm. With one resonator within the porous material as shown in Fig. 3a, the radius of the neck is set to 4 mm with a length of 20 mm and the lateral dimensions of the PUC are set to 80 mm × 80 mm. For the metamaterial with four resonators embedded within the porous layer, the radii of the four necks are 4.5 mm, 5 mm, 5.5 mm, and 6 mm, respectively, with an identical length of 20 mm. In Fig. 3c where nine resonators are embedded within the porous layer, the radii of the nine resonators are, respectively, 4.5 mm, 5 mm, 5.5 mm, 6 mm, 6.5 mm, 7 mm, 7.5 mm, 8 mm, and 8.5 mm with the same length of 20 mm and the lateral dimensions of the PUC are 200 mm × 200 mm. In Fig. 3d, the length of the sixteen different necks of the resonators is equal to 20 mm and the radii are varied between 3 mm and 10.5 mm with a step of 0.5 mm.

3. Theoretical modeling

In this section, a theoretical approach is presented to model the porous material with embedded multiple resonators. Figure 4 shows the porous material with four embedded resonators having different necks illustrated by different colors. The periodic unit cell is divided into four regions denoted by layer 1 to layer 4 where the

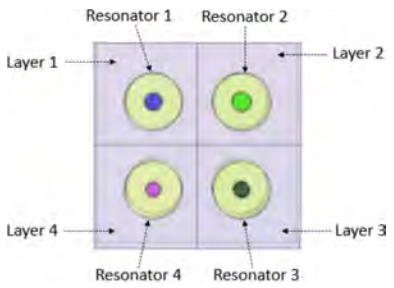


Fig. 4. Porous material with four embedded resonators.

lateral dimensions of each sub-porous layer i ($i = 1,2,3,4$) are $70\,\text{mm} \times 70\,\text{mm}$ and that of the unit cell are $140\,\text{mm} \times 140\,\text{mm}$. The layer i is made of the sub-porous layer i with the embedded resonator i.

The following assumptions are made:

– normal incidence plane wave propagation,
– there is no pressure diffusion between adjacent parallel elements,
– the wavelength is much longer than the PUC,
– each element namely the porous layer and each resonator can be modeled by a 2×2 transfer matrix.

The acoustic impedance of the Helmholtz resonator alone without the porous layer can be expressed as (Doutres *et al.*, 2015; Laly *et al.*, 2023):

$$Z_H^i = \frac{1}{s_e^i}\left(Z_{\text{neck}}^i - j s_v^i Z_0 \cot\left(kL_c^i\right)\right), \tag{9}$$

where Z_0 is the characteristic impedance of air, Z_{neck}^i is the impedance of the i-th neck, L_c^i is the i-th resonator cavity thickness, s_e^i and s_ν^i represent, respectively, the ratio of the i-th neck aperture surface over the cross section area of the i-th external elementary cell and an internal elementary cell. The acoustic impedance of the i-th neck is given by (Doutres *et al.* 2015; Laly *et al.*, 2024a; 2024b):

$$Z_{\text{neck}}^i = R_s\left(4 + \frac{2H_i}{R_i}\right) + j\omega\rho_0\left(\delta_i + \varepsilon_i + H_i\right), \tag{10}$$

where R_i and H_i are the radius and the length of the i-th neck, δ_i and ε_i are the correction lengths, and R_s is a surface resistance that is given by $R_s = 0.5\sqrt{2\eta\rho_0\omega}$. Since each resonator neck is circular, the approximation of the two corrections lengths for $s_e^i < 0.16$ is

$$\delta_i = \varepsilon_i = 0.48\sqrt{S_n^i}\left(1 - 1.14\sqrt{s_e^i}\right), \tag{11}$$

where S_n^i is the neck surface of the i-th resonator. The transfer matrix T_H^i of the i-th resonator can be written as

$$T_H^i = \begin{bmatrix} 1 & 0 \\ 1/Z_H^i & 1 \end{bmatrix}\begin{bmatrix} 1 & j\omega M_s^i \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} T_{H11}^i & T_{H12}^i \\ T_{H21}^i & T_{H22}^i \end{bmatrix}, \tag{12}$$

where M_s^i is the mass per unit area of the i-th resonator ($M_s^i = \rho_0\left(\delta_i + \varepsilon_i + H_i\right)$). From Eq. (12), the admittance matrix Y_H^i of the i-th resonator can be obtained (Verdière *et al.*, 2013):

$$Y_H^i = \frac{1}{T_{H12}^i}\begin{bmatrix} T_{H22}^i & T_{H21}^i T_{H12}^i - T_{H22}^i T_{H11}^i \\ 1 & -T_{H11}^i \end{bmatrix}. \tag{13}$$

The transfer matrix T_p^i of each layer i of the porous material in Fig. 4 is determined using the equivalent density $\rho_{eq}(\omega)$ and the bulk modulus $K_{eq}(\omega)$ given by Eq. (2) and Eq. (3). It is expressed by

$$T_P^i = \begin{bmatrix} \cos\left(k_{eq}^i L_i\right) & j Z_{eq}^i \sin\left(k_{eq}^i L_i\right) \\ j\dfrac{\sin\left(k_{eq}^i L_i\right)}{Z_{eq}^i} & \cos\left(k_{eq}^i L_i\right) \end{bmatrix} = \begin{bmatrix} T_{P11}^i & T_{P12}^i \\ T_{P21}^i & T_{P22}^i \end{bmatrix}, \tag{14}$$

where L_i is the thickness of the i-th layer, Z_{eq}^i and k_{eq}^i are the characteristic impedance and the wave number of the equivalent fluid of the i-th layer given by

$$Z_{eq}^i = \sqrt{\rho_{eq} K_{eq}}/\phi, \qquad k_{eq}^i = \omega\sqrt{\rho_{eq}/K_{eq}}. \tag{15}$$

From Eq. (15), the admittance matrix Y_p^i of each layer of the porous material in Fig. 4 is obtained. It is given by

$$Y_P^i = \frac{1}{T_{P12}^i}\begin{bmatrix} T_{P22}^i & T_{P21}^i T_{P12}^i - T_{P22}^i T_{P11}^i \\ 1 & -T_{P11}^i \end{bmatrix}. \tag{16}$$

Each i-th layer of porous material with the embedded i-th resonator in Fig. 4 are assembled in parallel, so their equivalent transfer matrix M^i is determined using the parallel transfer matrix method (Verdière *et al.*, 2013):

$$M^i = \frac{-1}{\sum r_n^i Y_{n21}^i}\begin{bmatrix} \sum r_n^i Y_{n22}^i & -1 \\ \sum r_n^i Y_{n22}^i \sum r_n^i Y_{n11}^i - \sum r_n^i Y_{n12}^i \sum r_n^i Y_{n21}^i & -\sum r_n^i Y_{n11}^i \end{bmatrix}, \tag{17}$$

where Y_n^i represents the admittance matrix components of Eq. (13) and Eq. (16) with $n = H$ for the resonator and $n = P$ for the porous substrate and r_n^i is the surface ratio of each element given by

$$r_H^i = S_i / S_P^i, \qquad r_P^i = 1 - r_H^i, \tag{18}$$

where S_P^i is the surface of the i-th unit cell layer and S_i is the surface of the i-th resonator. The transfer matrices of layer 1, layer 2, layer 3, and layer 4 in Fig. 4 are then, respectively, denoted by M^1, M^2, M^3 and M^4. Thus, the admittance matrix of each i-th layer in Fig. 4 is obtained from Eq. (17) and is expressed by

$$Y^i = \frac{1}{M_{12}^i}\begin{bmatrix} M_{22}^i & M_{21}^i M_{12}^i - M_{22}^i M_{11}^i \\ 1 & -M_{11}^i \end{bmatrix}. \tag{19}$$

In Fig. 4, layer 1 and layer 2 are assembled in parallel, thus the equivalent transfer matrix TM^a of layer 1 and layer 2 is determined by

$$TM^a = \frac{-1}{\sum r_m Y_{21}^m}\begin{bmatrix} \sum r_m Y_{22}^m & -1 \\ \sum r_m Y_{22}^m \sum r_m Y_{11}^m - \sum r_m Y_{12}^m \sum r_m Y_{21}^m & -\sum r_m Y_{11}^m \end{bmatrix}. \tag{20}$$

In Eq. (20), $m = 1, 2$, Y^m is obtained from Eq. (19), $r_1 = S_1^L / S_{12}^L$, $r_2 = 1 - r_1$ with S_1^L the surface of layer 1 and S_{12}^L the surface of both layer 1 and layer 2.

Similarly, layer 3 and layer 4 in Fig. 4 are assembled in parallel, so their equivalent transfer matrix TM^b is given by

$$TM^b = \frac{-1}{\sum r_u Y_{21}^u}\begin{bmatrix} \sum r_u Y_{22}^u & -1 \\ \sum r_u Y_{22}^u \sum r_u Y_{11}^u - \sum r_u Y_{12}^u \sum r_u Y_{21}^u & -\sum r_u Y_{11}^u \end{bmatrix}, \tag{21}$$

where $u = 3, 4$, Y^u is given by Eq. (19), $r_3 = S_3^L / S_{34}^L$, $r_4 = 1 - r_3$ with S_3^L the surface of layer 3 and S_{34}^L the surface of both layer 3 and layer 4. If the parallel assembly of layer 1 and layer 2 is denoted by material A and the parallel assembly of layer 3 and layer 4 by material B, then materials A and B are arranged in parallel and their admittance matrices can be expressed as

$$YM^q = \frac{1}{TM_{12}^q}\begin{bmatrix} TM_{22}^q & TM_{21}^q TM_{12}^q - TM_{22}^q TM_{11}^q \\ 1 & -TM_{11}^q \end{bmatrix}. \tag{22}$$

In Eq. (22), $q = a, b$ and TM^q is obtained from Eq. (20) and Eq. (21). The transfer matrix of the porous material with four embedded Helmholtz resonators, in Fig. 4, is finally determined by

$$TG = \frac{-1}{\sum r_q YM_{21}^q}\begin{bmatrix} \sum r_q YM_{22}^q & -1 \\ \sum r_q YM_{22}^q \sum r_q YM_{11}^q - \sum r_q YM_{12}^q \sum r_q YM_{21}^q & -\sum r_q YM_{11}^q \end{bmatrix}. \tag{23}$$

The transfer matrix in Eq. (23) can be expressed as $TG = \begin{bmatrix} TG_{11} & TG_{12} \\ TG_{21} & TG_{22} \end{bmatrix}$. The theoretical sound absorption coefficient α_{theo} of the metamaterial in Fig. 4 can be calculated using the components of the matrix TG:

$$\alpha_{\text{theo}} = 1 - \left| \frac{TG_{11} - Z_0\, TG_{21}}{TG_{11} + Z_0\, TG_{21}} \right|^2 . \tag{24}$$

The theoretical transmission loss TL_{theo} of the metamaterial consisting of the porous layer with four embedded resonators is obtained from Eq. (23) and is given by

$$\text{TL}_{\text{theo}} = -20 \log_{10} \left| \frac{2}{TG_{11} + TG_{12}/Z_0 + Z_0\, TG_{21} + TG_{22}} \right|. \tag{25}$$

4. Comparison of the finite element method and the theory

In this section, the sound absorption coefficient and the TL predicted theoretically using Eq. (24) and Eq. (25) are compared with the results of the finite element method. Figure 5 presents the comparison of the sound absorption coefficient of the metamaterial of Fig. 1a made of porous material with one embedded periodic resonator and the comparison of the corresponding TL is shown in Fig. 6. The absorption coefficient of the material with one embedded resonator presents a peak of 0.9 at 364 Hz while without the resonator the absorption coefficient value is 0.34 at 364 Hz. The theoretical calculation agrees well with the FEM. In Fig. 6, the FEM result of the TL with the resonator shows a peak of 18 dB at 364 Hz and the theoretical TL peak at this frequency is 17 dB. Without the resonator, the TL of the porous material alone is lower than 8 dB over the entire frequency range.

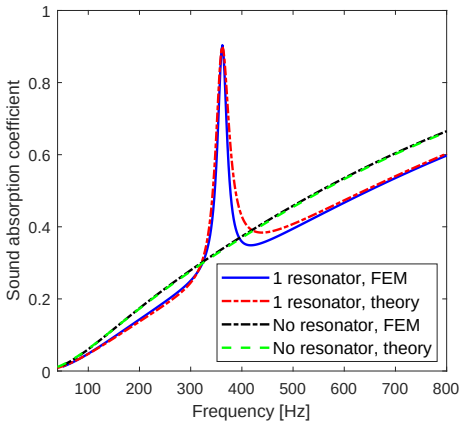


Fig. 5. Comparison of the sound absorption coefficient of the metamaterial with one embedded resonator.

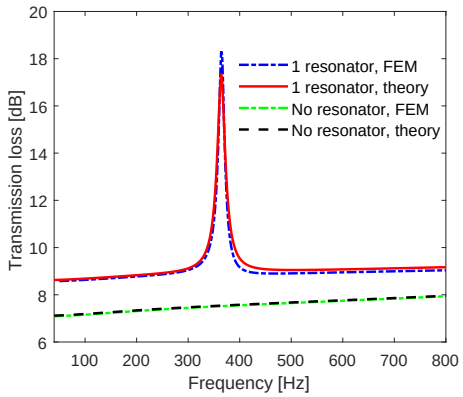


Fig. 6. Comparison of the transmission loss of the metamaterial with one embedded resonator.

Figure 7 presents the comparison of the theoretical absorption coefficient and the FEM result for the metamaterial of Fig. 1b consisting of the porous layer with four embedded different resonators and the comparison of the corresponding TL is shown in Fig. 8. The absorption coefficient as well as the TL exhibit four resonance peaks at 408 Hz, 448 Hz, 490 Hz, and 530 Hz. The FEM peak values of the sound absorption are 0.98, 0.99, 0.98, and 0.92, respectively, while the theoretical values are 0.94, 0.99, 0.99, and 0.97. The peaks of the TL predicted by the FEM are 13 dB, 14 dB, 15.2 dB, and 16 dB, respectively, and for the theory, the peaks of the TL are 12.7 dB, 13.5 dB, 14 dB, and 15 dB. The results of the theory and the FEM in Fig. 7 and Fig. 8 are in good agreement. With four parallel resonators embedded within the porous layer, the sound absorption and the TL present four different resonant peaks while only one resonant peak is observed in Fig. 5 and Fig. 6 for one resonator incorporated within the porous material.

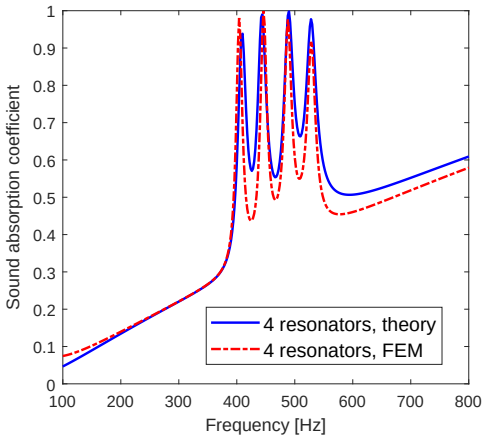


FIG. 7. Comparison of the sound absorption coefficient of the metamaterial with four embedded resonators.

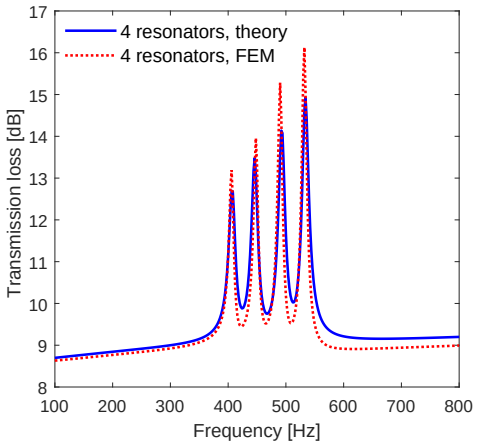


FIG. 8. Comparison of the transmission loss of the metamaterial with four embedded resonators.

For a plane wave impinging on the metamaterial surface with an incidence angle θ, the theoretical sound absorption coefficient and the transmission loss can be determined by (VERDIÈRE *et al.*, 2013):

$$\alpha_\theta = 1 - \left| \frac{T_{11}\cos\theta - Z_0 T_{21}}{T_{11}\cos\theta + Z_0 T_{21}} \right|^2, \tag{26}$$

$$\mathrm{TL}_\theta = 20\log_{10}\left(\frac{1}{2}\left|\left(T_{11} + \frac{T_{12}\cos\theta}{Z_0} + \frac{Z_0 T_{21}}{\cos\theta} + T_{22}\right)\right|\right). \tag{27}$$

In Eq. (26) and Eq. (27), T_{11}, T_{12}, T_{21}, and T_{22} are the components of the global matrix of the metamaterial made of the porous layer with embedded resonators that can be obtained using Eq. (17) and Eq. (23). Figure 9 presents the comparison of the sound absorption coefficient of the porous layer with one and four

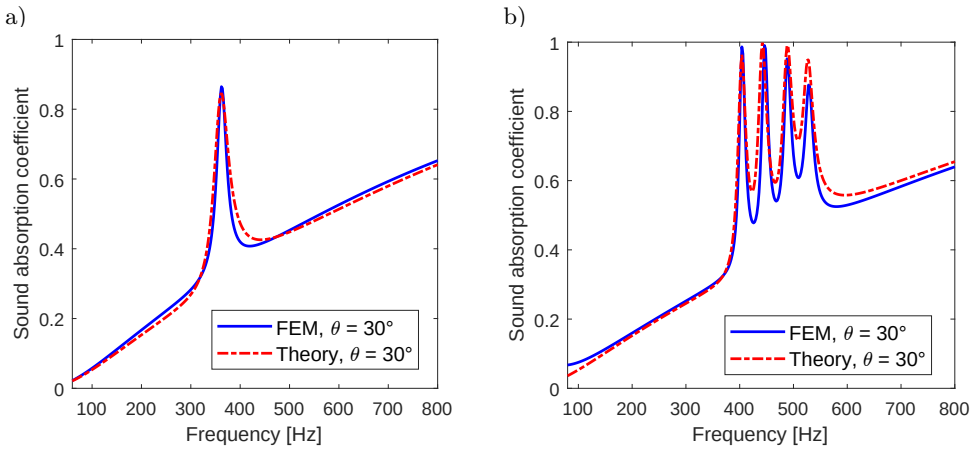


FIG. 9. Sound absorption coefficient for oblique incidence: a) one embedded resonator, b) four embedded resonators.

embedded resonators for an incidence angle of 30°. It can be observed that the oblique incidence sound absorption predicted theoretically agrees with the FEM.

Figure 10a and Fig. 10b show the porous material with four embedded resonators and the distance L separating the axes of two adjacent resonators is varied to analyze the interaction effect between the resonators. Figure 10c, Fig. 10d, and Fig. 10e illustrate the porous material with one, four, and nine perforations in place of the resonators. The diameter of each perforation is identical to the diameter of the cavity of each resonator, which is 40 mm. The absorption coefficient of the porous material with embedded resonators and with perforations is studied in the following section. In Fig. 10a and Fig. 10b, the lateral dimensions are 200 mm × 200 mm and the distance between the axes of two adjacent resonators is set to 45 mm in Fig. 10a, while it is 110 mm in Fig. 10b. The normal incidence sound absorption coefficient of these two configurations is illustrated in Fig. 11, where configuration A represents the case of L equal to 45 mm. With configuration A, the values of the first and fourth absorption peaks are 0.86 and 0.97, while they are 0.83 and 0.98, respectively, for configuration B. A slight difference is also observed for frequencies above 550 Hz that should be caused by the interaction effect between the resonators, which is not considered in the analytical model. The minor deviation observed between the theoretical model and the FEM results in Fig. 7 and Fig. 8 may also result from coupling effects between the resonators.

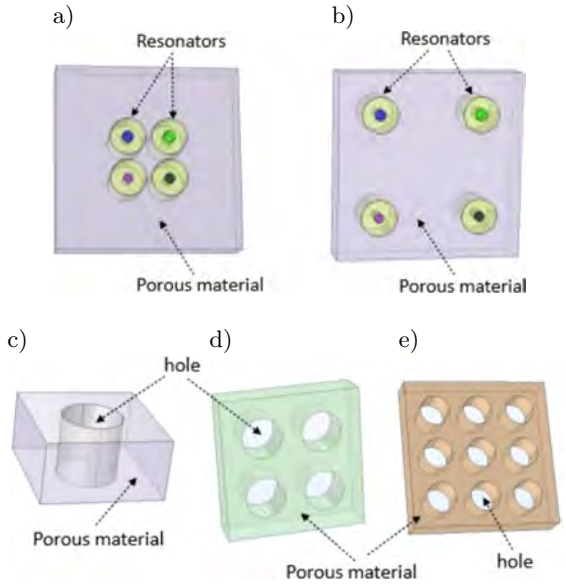


FIG. 10. Geometrical models of: a) porous layer with four resonators, L = 45 mm, b) porous layer with four resonators, L = 110 mm, c) porous layer with one perforation, d) porous layer with four perforations, e) porous layer with nine perforations.

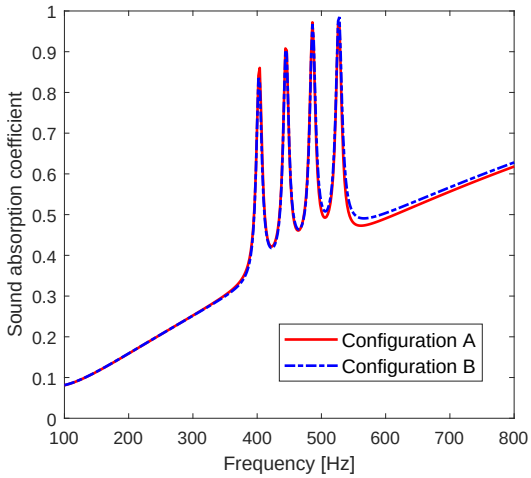


FIG. 11. Sound absorption coefficient.

5. Analysis of the sound attenuation performance

In this section, the sound attenuation performance of the proposed metamaterial is studied numerically. The diffuse field sound absorption coefficient is obtained using Eq. (8) where $\theta_{\max}$ is set to $\pi/2$.

Figure 12 shows the sound absorption coefficient of the porous layer with one embedded periodic Helmholtz resonator (Fig. 1a) at different oblique incidence angles θ. The sound absorption coefficients of the porous layer without resonators and with one and four perforations in the place of resonators as illustrated in Fig. 10c and Fig. 10d are also presented. The normal incidence sound absorption coefficient in Fig. 12 presents a resonant peak of 0.9 at 362 Hz. For $\theta = 60°$, the absorption coefficient shows a peak of 0.77 at 372 Hz while the diffuse field sound absorption coefficient presents a peak of 0.76 at 364 Hz. The normal incidence sound absorption coefficient of the porous material alone is lower than 0.5 for frequencies below 500 Hz. With one or four perforations within the porous layer, the sound absorption is lower than that of porous layer alone.

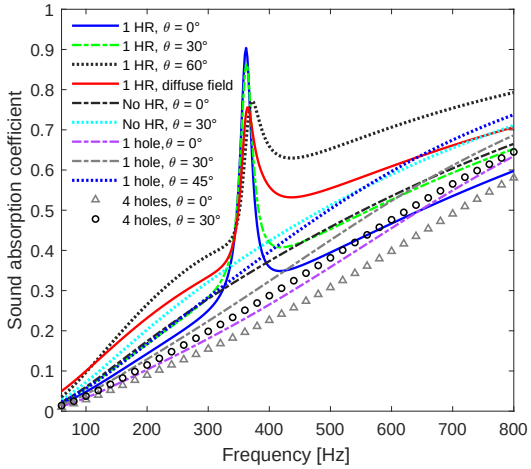


FIG. 12. Sound absorption coefficient of the metamaterial.

Figure 13 shows the oblique incidence and the diffuse field sound absorption coefficient of the porous material with four different incorporated resonators as illustrated in Fig. 1b. As θ increases up to 40°, the sound absorption coefficient with 4 HR shows four identical resonant frequencies of 408 Hz, 448 Hz, 490 Hz, and 530 Hz. The sound absorption peaks at these four resonant frequencies are 0.98, 0.99, 0.976, and 0.92, respectively, for $\theta = 0°$ while for $\theta = 40°$, they are 0.985, 0.99, 0.94, and 0.84. The diffuse field sound absorption with 4 HRs presents four resonant peaks of 0.95, 0.94, 0.87, and 0.76, respectively, at the resonant frequencies, while without HR, the values of the absorption coefficient at the resonant frequencies are 0.5, 0.53, 0.57, and 0.6, respectively. Above 70°, the

absorption coefficient is lower as the incidence angle increases. Indeed, for an incidence angle of 88°, the absorption coefficient without HR is lower than 0.2 and presents four peaks of 0.28, 0.3, 0.32, and 0.4 with 4 HRs.

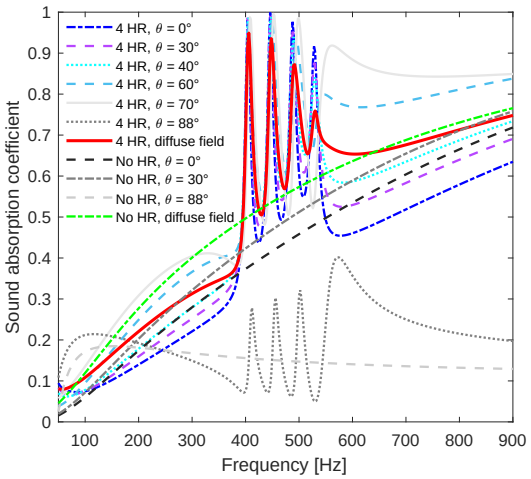


FIG. 13. Diffuse field sound absorption of the metamaterial with four embedded resonators.

The sound absorption coefficient at various incidence angles as well as the diffuse field absorption coefficient of the metamaterial in Fig. 1c made of porous material with nine embedded resonators are shown in Fig. 14 and compared to the sound absorption without the resonators. The corresponding normal incidence TL is presented in Fig. 15 as well as the TL of the porous layer alone and that of the porous layer with nine perforations. When θ increases up to 50°, the absorption coefficient in Fig. 14 with 9 HRs presents nine resonant frequencies around 408 Hz, 448 Hz, 490 Hz, 530 Hz, 572 Hz, 612 Hz, 654 Hz, 694 Hz, and 734 Hz. The absorption peak values for normal incidence in Fig. 14 are 0.82, 0.9, 0.94, 0.98, 0.98, 0.99, 0.984, 0.98, and 0.95 while the diffuse field absorption peak values are 0.87, 0.9, 0.93, 0.94, 0.95, 0.96, 0.91, 0.9, and 0.87, respectively. Without the embedded resonators, the diffuse field absorption coefficient increases from 0.47 at 370 Hz to 0.73 at 800 Hz. The absorption coefficient is significantly improved at the nine resonant frequencies of the nine resonators. The TL of the metamaterial with 9 HR in Fig. 15 exhibits nine resonance peaks of 12.9 dB, 14.4 dB, 14 dB, 15.7 dB, 18 dB, 15.7 dB, 16 dB 16.5 dB, and 13 dB at the resonance frequencies and outside these frequencies the TL is about 9 dB while without the resonators, the TL is between 7 and 8 dB. The TL of the porous layer with nine perforations is lower than 5 dB over the entire frequency range. The diffuse field sound absorption and the TL in Fig. 14 and Fig. 15 are significantly enhanced at the resonance frequencies of the resonators. Compared to Fig. 13, it can be observed that the frequency band of the absorption coefficient increases as the number of resonators within the unit cell increases.

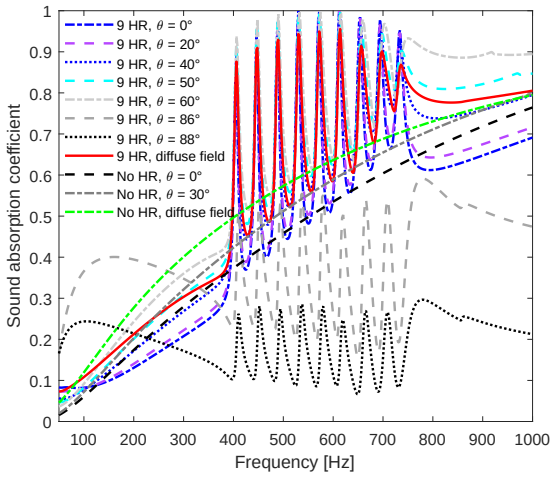


FIG. 14. Diffuse field sound absorption of the metamaterial with nine embedded resonators.

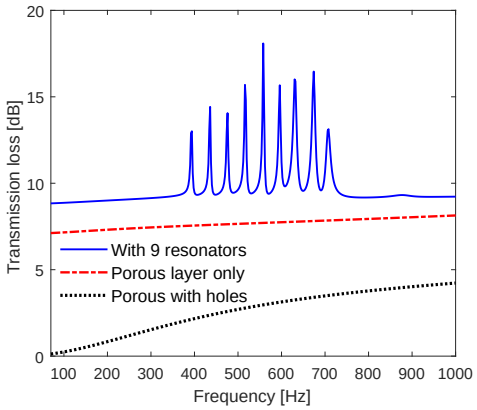


FIG. 15. Transmission loss of the metamaterial with nine embedded resonators.

Figure 16 shows the absorption coefficient at different incidence angles and the diffuse field sound absorption of the porous material with sixteen different embedded resonators as illustrated in Fig. 1d. With sixteen embedded resonators, the absorption coefficient for each incidence angle shows sixteen resonance peaks in Fig. 16 with the peak values that are between 0.54 and 0.95 for the diffuse field sound absorption. The frequency band of the sound absorption in Fig. 16 is larger than the one in Fig. 14 for the porous layer with nine embedded resonators. The absorption coefficient decreases mainly when the incidence angle is higher than 75°.

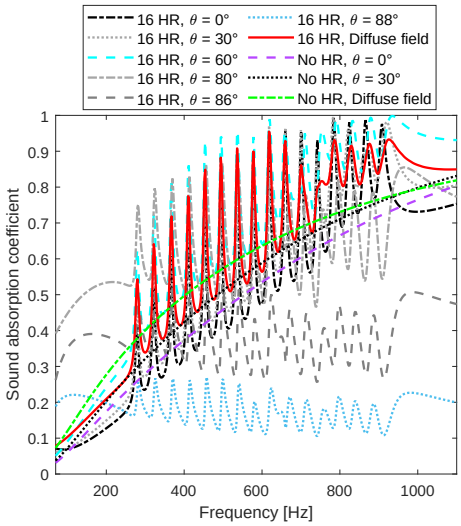


FIG. 16. Diffuse field sound absorption of the metamaterial with sixteen embedded resonators.

The sound attenuation mechanism of the metamaterial is investigated in the following. One considers the material design of Fig. 1b where the PUC is made of four parallel resonators that are incorporated into the porous layer. The sound absorption coefficient in Fig. 13 showed four resonant peaks at 408 Hz, 448 Hz, 490 Hz, and 530 Hz, and when the incidence angle is greater than 70°, the absorption coefficient decreases. Figure 17 illustrates the total root mean square (RMS) acoustic pressure through the metamaterial at 300 Hz, 408 Hz, 448 Hz, 490 Hz, 530 Hz, and 700 Hz for incidence angles of 0° and 75°. The total acoustic velocity at the four resonant frequencies are shown in Fig. 18, while Fig. 19 presents the acoustic power dissipation density.

For a normal incidence, it is observed in Fig. 17a that the RMS acoustic pressure at each resonance frequency reaches its maximum value in the cavity of the corresponding resonator and is 30 Pa. Outside these resonance frequencies the acoustic pressure is very low in the cavities. The acoustic velocity in Fig. 18 reaches a maximum value 0.4 m/s in the neck of each resonator at the corresponding resonance frequency and the acoustic power dissipation density is significant in the neck at the resonance frequency. As the frequency increases above the highest resonance frequency associated with HRs, the power dissipation density in the porous layer reaches its

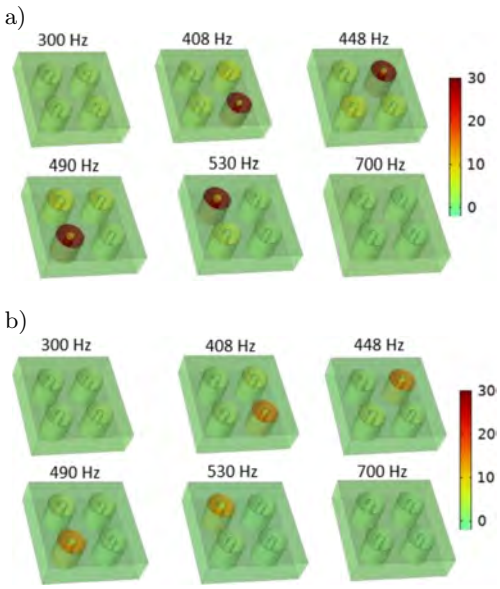


FIG. 17. Total RMS acoustic pressure [Pa]: a) $\theta = 0°$, b) $\theta = 75°$.

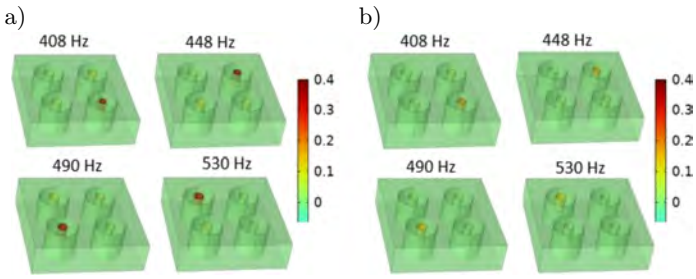


FIG. 18. Total acoustic velocity [m/s]: a) $\theta = 0°$, b) $\theta = 75°$.

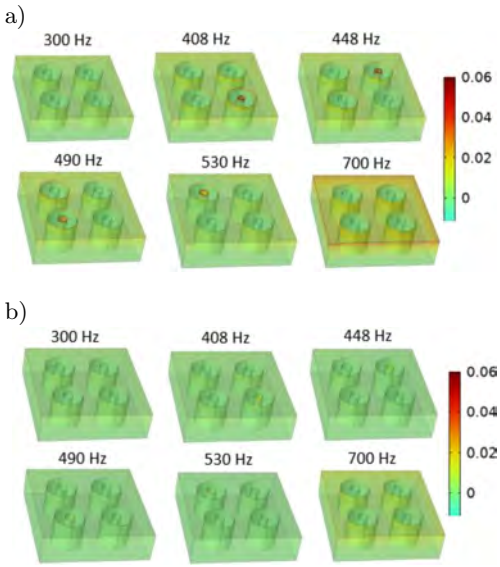


FIG. 19. Acoustic power dissipation density [W/m^3]: a) $\theta = 0°$, b) $\theta = 75°$.

highest value, and sound absorption is mainly due to the porous material matrix. The four parallel resonators act independently resulting in four absorption peaks. For an incidence angle of 75°, the acoustic pressure in the cavities at the resonance frequencies decreases, the acoustic velocity in the necks (Fig. 18b) becomes low and the power dissipation density is negligible in Fig. 19b and consequently the absorption coefficient becomes low as observed in Fig. 13.

6. Conclusion

A design of acoustic metamaterial made of porous material with multiple embedded Helmholtz resonators was proposed and studied for multi-tonal and broadband noise reduction. A theoretical approach based on the transfer matrix method was presented and the theoretical results of the sound absorption and the transmission loss showed good agreement with the FEM. Compared to the porous layer with one embedded Helmholtz resonator that presents one resonant peak, the metamaterial consisting of the porous material with four, nine, and sixteen embedded Helmholtz resonators exhibits respectively four, nine, and sixteen sound absorption and TL resonant peaks. The impacts of the incidence angles were demonstrated and the diffuse field sound absorption was significantly improved at the resonant frequencies of the resonators. The TL and the sound absorption frequency band increase as the number of resonators within the porous layer increases. The proposed metamaterial can help in multiple engineering applications for multi-tonal and broadband noise attenuation.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that there are no known competing financial interests or personal relationships that could have influenced the work described in this paper.

Authors' contributions

Zacharie Laly contributed to conceptualization, investigation, methodology, data analysis and interpretation, validation, and writing. Raymond Panneton contributed to supervision, project administration, resources, data analysis and interpretation, review and editing. Noureddine Atalla contributed to supervision, project administration, resources, data analysis and interpretation, review and editing. Sebastian Ghinet contributed to investigation, data analysis and interpretation, review and editing. All authors reviewed and approved the final manuscript.

References

1. Abbad A., Atalla N., Ouisse M., Doutres O. (2019), Numerical and experimental investigations on the acoustic performances of membraned Helmholtz resonators embedded in a porous matrix, *Journal of Sound and Vibration*, **459**: 114873, https://doi.org/10.1016/J.JSV.2019.114873.
2. Allard J.F., Atalla N. (2009), *Propagation of Sound in Porous Media: Modelling Sound Absorbing Materials*, John Wiley & Sons, https://doi.org/10.1002/9780470747339.
3. Boutin C., Becot F.X. (2015), Theory and experiments on poro-acoustics with inner resonators, *Wave Motion*, **54**: 76–99, https://doi.org/10.1016/J.WAVEMOTI.2014.11.013.
4. Doutres O., Atalla, N., Osman H. (2015), Transfer matrix modeling and experimental validation of cellular porous material with resonant inclusions, *The Journal of the Acoustical Society of America*, **137**(6): 3502–3513, https://doi.org/10.1121/1.4921027.
5. Ghinet S., Bouche P., Padois T., Pires L. (2020), Experimental validation of acoustic metamaterials noise attenuation performance for aircraft cabin applications, [in:] *Proceedings of 2020 International Congress on Noise Control Engineering*.
6. Kone T.C., Ghinet S., Panneton R., Grewa A. (2021), Optimization of metamaterials with complex neck shapes for aircraft cabin noise improvement, [in:] *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, https://doi.org/10.3397/IN-2021-2567.

7. Kone T.C., Ghinet S., Panneton R., Laly Z., Mechefske C., Grewal A. (2022), Control and broadening of multiple noise frequencies using an assembly of sub-metamaterials connected by membranes for aircraft noise mitigation, [in:] *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, https://doi.org/10.3397/IN_2022_0663.

8. Lagarrigue C., Groby J.P., Tournat V., Dazel O., Umnova O. (2013), Absorption of sound by porous layers with embedded periodic arrays of resonant inclusions, *The Journal of the Acoustical Society of America*, **134**(6): 4670–4680, https://doi.org/10.1121/1.4824843.

9. Laly Z., Mechefske C., Ghinet S., Kone C.T. (2024a), Finite element modeling of acoustic metamaterial based on periodic Helmholtz resonator with a membrane in the cavity, *Noise Control Engineering Journal*, **72**(6): 514–542, https://doi.org/10.3397/1/377239.

10. Laly Z., Mechefske C., Ghinet S., Kone C.T., Atalla N. (2023), Design and analysis of periodic acoustic metamaterial sound insulator using finite element method, *Noise Control Engineering Journal*, **71**(5): 344–364, https://doi.org/10.3397/1/377128.

11. Laly Z., Mechefske C., Ghinet S., Kone C.T., Panneton R., Atalla N. (2025), Diffuse field sound absorption performance of porous material with embedded periodic Helmholtz resonators, [in:] *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, https://doi.org/10.3397/in_2025_1074286.

12. Laly Z., Mechefske C., Ghinet S., Kone C.T., Atalla N., Ashrqafi B. (2024b), The effects of Helmholtz resonator wall elasticity on the sound attenuation performance, *Noise Control Engineering Journal*, **72**(6): 543–563, https://doi.org/10.3397/1/377240.

13. Laly Z., Panneton R., Atalla N. (2022), Characterization and development of periodic acoustic metamaterials using a transfer matrix approach, *Applied Acoustics*, **185**: 108381, https://doi.org/10.1016/J.APACOUST.2021.108381.

14. Magliacano D., Catapane G., Petrone G., Verdière K., Robin O. (2024), Sound transmission properties of a porous meta-material with periodically embedded Helmholtz resonators, *Mechanics of Advanced Materials and Structures*, **31**(25): 6748–6756, https://doi.org/10.1080/15376494.2023.2237699.

15. Verdière K., Panneton R., Elkoun S., Dupont T., Leclaire P. (2013), Transfer matrix method applied to the parallel assembly of sound absorbing materials, *The Journal of the Acoustical Society of America*, **134**(6): 4648–4658, https://doi.org/10.1121/1.4824839.

16. Zhang W., Xin F. (2024), Broadband low-frequency sound absorption via Helmholtz resonators with porous material lining, *Journal of Sound and Vibration*, **578**: 118330, https://doi.org/10.1016/J.JSV.2024.118330.

17. Zhu X.F., Lau S.-K., Lu Z., Jeon W. (2019), Broadband low-frequency sound absorption by periodic metamaterial resonators embedded in a porous layer, *Journal of Sound and Vibration*, **461**: 114922, https://doi.org/10.1016/J.JSV.2019.114922.

Archives of Acoustics Vol. 51, No. 3, pp. 459–475 (2026), https://doi.org/10.24423/archacoust.2026.4359

Research Paper

Broadband Vibration-Isolating Metamaterial for MRI-Induced Vibrations

Swapnil Arawade*, Janusz Piechowicz[2]

Department of Mechanics and Vibroacoustics, AGH University of Krakow
Kraków, Poland

*Corresponding Author: sarawade@agh.edu.pl

Received October 22, 2025; revised January 30, 2026 accepted April 24, 2026; available online May 4, 2026; version of record September 10, 2026; published issue September 23, 2026.

Magnetic resonance imaging (MRI) is an important diagnostic tool in the medical field. Vibrations originating from the gradient coils of an MRI system lead to loud operational noise. This study explores a numerical investigation of elastic metamaterial designed for the attenuation of MRI-induced vibrations that could radiate as sound at the air-structure interface. Elastic metamaterials with embedded resonators can significantly attenuate wave propagation by opening a local resonance bandgap. In this study, the wave propagation properties of two elastic metamaterial models are analysed. The resulting dispersion diagrams are studied using a parametric analysis to investigate the influence of different geometric features on bandgap generation. The metamaterial models are analysed using finite element analysis (FEA) to investigate their wave attenuation performance. A frequency analysis of the multicellular model shows a high transmission loss corresponding to the bandgaps in the calculated dispersion diagrams. The results reveal that the optimised metamaterial model shows an overlap of 72.3 % with the vibration frequency range of the MRI scanner, while the frequency analysis of the multicellular arrangement reports a transmission loss of up to 115.7 dB. These results demonstrate the potential of such elastic metamaterials for targeted vibro-acoustic control and provide a framework for designing structures for wave attenuation in sensitive acoustic environments.

Keywords: elastic metamaterials, local-resonance bandgap, vibro-acoustic coupling, magnetic resonance imaging (MRI), MRI noise, wave attenuation.



Copyright © 2026 The Author(s).
This work is licensed under the Creative Commons Attribution 4.0 International CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The magnetic resonance imaging (MRI) is a very important tool in medical diagnosis without the use of harmful ionising radiation (Bruno *et al.*, 2019; Reda *et al.*, 2021; Landini *et al.*, 2018). For generating images, an MRI system utilises a static magnetic field, a dynamic magnetic field gradient, and radiofrequency pulses. The interaction of the static magnetic field and the gradient field leads to the generation of a Lorentz force, setting the gradient coil into vibration. This vibration propagates to the scanner surface, and its vibro-acoustic coupling with the surrounding air generates the typical MRI noise (Zhang *et al.*, 2025; Leo, D'Orazio, 2025; Hedeen, Edelstein, 1997; Mansfield *et al.*, 1998; Moelker *et al.*, 2003). Literature has reported sound pressure levels reaching up to 133 dB, which may induce anxiety, fear, and a potential temporary loss of hearing (Moelker, Pattynama, 2003; Hutter *et al.*, 2018; Arawade, Piechowicz, 2024a; 2024b). Furthermore, this vibration also propagates through the structure and reaches the scanning zone, reducing the quality of images with blurring and other unknown artifacts (Gallichan *et al.*, 2010; Fuhrer *et al.*, 2019; Frollo *et al.*, 2007). Generally, MRI noise and vibrations are present in the frequency range of 300 Hz to 3000 Hz, with major peaks observed between 1000 Hz and 2000 Hz. Many approaches have been explored to reduce the generation and propagation of the MRI vibration and noise, such as modifying the input electric current pulses to the gradient coil, gradient coil modification, and active and passive noise control (McJury, 2022; Kanal *et al.*, 1992; Al-Meer, 2022; Li *et al.*, 2011; Lee *et al.*, 2017; Lasota, Meller, 2020).

In addition to these system-level approaches, significant industrial efforts have been directed towards vibration isolation and structural decoupling of MRI equipment from its surroundings. AMC MECANOCAUCHO offers their Vibrabsorber + Sylomer® mounts, which effectively minimise vibration propagation[1]. Kinetics Noise Control offers precision isolation systems, including elastomeric isolators, to mitigate airborne and structure-borne noise[2]. Unisorb Installation Technologies (2025) presents engineered isolation materials, such as IB-500™ blocks for minimising vibrations[3]. International Design Engineering (IDE) delivers both passive and active vibration isolators to enhance MRI imaging performance[4]. Additionally, acoustic panels and floating ceiling and wall treatments are also commonly employed in the MRI rooms to absorb airborne noise and enhance the room acoustics[5,6]. Despite these developments, most existing solutions focus on isolating the MRI scanner or treating the room and do not directly suppress vibro-acoustic couplings at the panel. This highlights the need for strategies that address vibration attenuation at the source.

Metamaterials are artificial materials that possess specially tailored properties that natural materials do not have (Kumar *et al.*, 2022; Wei *et al.*, 2025; Wu *et al.*, 2025; Zhang, Zhao, 2025). Metamaterials are explored for various counterintuitive properties, such as frequency bandgaps (Miao *et al.*, 2025; Liu *et al.*, 2025), cloaking (Ahmad *et al.*, 2025; Fu *et al.*, 2025), negative refraction (Ruks *et al.*, 2025; Zhou *et al.*, 2025), negative mass density (Hao *et al.*, 2025; Wang *et al.*, 2025), and a negative Poisson's ratio (Li, 2025; Shunshun, Zhao, 2025). Mechanical metamaterials are a class of metamaterials having properties designed to regulate the propagation of elastic waves. Literature shows that their properties are highly influenced by structural parameters instead of solely the composition of the constituent materials (Chua *et al.*, 2025; He *et al.*, 2024). Elastic metamaterials are composed of periodic composites of two or more materials with dissimilar mechanical properties, leading to the generation of elastic bandgaps in wave dispersion. Elastic waves cannot propagate within the frequency range of a bandgap and therefore this property of elastic metamaterials makes them important for potential applications in vibration isolation and damping.

The methods to obtain a bandgap over a specific frequency range are mainly classified as the Bragg scattering effect and the local resonance mechanism (Kaina *et al.*, 2013). Phononic crystals-based metamaterials having scatterers embedded in a host material utilise the principle of Bragg scattering for wave propagation control. In this method, wave reflection and interference within the periodic structure lead to the formation of a frequency-dependant bandgap, restricting wave propagation at specific frequencies (Chuang *et al.*, 2020; Liu, Zhu, 2019). Phononic crystals show a bandgap when the structural periodicity is on the order of the wavelength, where the bandgap characteristics typically depend upon the lattice arrangement, scatterer size, and the impedance contrast between the constituent materials in the unit cell (Qi *et al.*, 2018; Xiang *et al.*, 2020). These cellular structures show a high strength-to-mass ratio, which is favourable in different applications requiring robust and broad bandgaps (Wen *et al.*, 2020; Wickeler, Naguib, 2020). Phononic crystals are primarily suitable for high-frequency applications because of the requirement of destructive interference of waves between the periodic cells. In the case of low- to mid-frequency range vibration control, using Bragg scattering phononic crystals becomes impractical due to the requirement of a larger structural size (Zhao *et al.*, 2024). The local resonance metamaterial based on a periodic structure was first introduced by Liu *et al.* (2000), achieving subwavelength bandgaps using local resonance in the structure. Locally resonant metamaterials work on the principle of negative indexes, where the frequency-dependant properties become negative in the tuned frequency range. In this negative property region, the local resonators oscillate with maximum potential, which provides restoring forces opposing the incoming excitation, and thus elastic/acoustic waves cannot propagate through the structure (Lee, Wright, 2016; Lee *et al.*, 2022). The resonating units have been reported as spring-mass systems, Helmholtz resonators, mass-on-membrane resonators, and other configurations (Hu *et al.*, 2019; Li *et al.*, 2022; Peiró-Torres *et al.*, 2019; Naify *et al.*, 2010). Through the modification of design parameters such as mass, stiffness, damping, and

[1] https://www.mecanocaucho.com/en/news/magnetic-resonance-en/ (access: 26.10.2025).
[2] https://kineticsnoise.com/precision/isolation (access: 26.10.2025).
[3] https://unisorb.com/industries/medical-equipment/ (access: 26.10.2025).
[4] https://www.ideworld.com/en/active_vibration_isolation.html (access: 26.10.2025).
[5] https://www.acousticalsurfaces.com/blog/soundproofing/soundproofing-a-noisy-mri-scanner/ (access: 26.10.2025).
[6] https://faradayshielding.com.au/acoustic-wall-panels-2/ (access: 26.10.2025).

geometry, local resonance metamaterials can be tuned to attenuate specific targeted frequencies (Fan *et al.*, 2021; Hu *et al.*, 2021). Local resonance metamaterials show smaller lattice cells with low- to mid-frequency bandgaps, highlighting their potential in multiple applications of structural monitoring, vibration isolation, and noise mitigation (Ma, Sheng, 2016; Pires *et al.*, 2022). Generally, they show narrow bandgaps, and research is actively underway to optimise locally resonant metamaterials to produce wide bandgaps.

Despite extensive research in the field of metamaterials, very few studies have explored their potential to control the vibration and noise generated by MRI systems. This work explores the inverse design of a locally resonant elastic metamaterial specifically targeting the vibration frequencies generated by the gradient coils of an MRI scanner that propagate and couple to the acoustic field at the air-structure interface. The local resonance mechanism is selected because it enables the formation of bandgaps within the target low- to mid-frequency range using subwavelength unit cells, resulting in a compact design suitable for the confined geometry of the MRI system. Dispersion relation analysis is employed to characterise the intrinsic wave propagation behaviour of the periodic cell and identify the subwavelength-resonance-induced bandgaps that govern the attenuation behaviour. The study investigates the influence of constituent materials and geometric parameters on bandgap generation and vibration attenuation performance by using the finite element method (FEM). By designing the bandgap to correspond with the MRI system vibration spectra, this investigation presents a systematic methodology for designing metamaterials for targeted vibro-acoustic control in sensitive medical environments.

2. Model and methods

This section mainly presents the geometric structure of the locally resonant elastic metamaterial and introduces the method for evaluation of dispersion of its characteristic curves and transmission loss spectrum of the elastic metamaterial using the solid mechanics module of the numerical simulation software COMSOL, followed by a material compatibility analysis.

2.1. Design of the elastic metamaterial

The study focuses on a 2D elastic metamaterial having periodic arrangement in the $X-Y$ plane and a uniform structure in the out-of-plane direction. The inclusion material is considered to be an isotropic hard solid periodically embedded in a soft matrix material. The unit cell forms a square lattice structure with a lattice constant a, representing the distance between two adjacent unit cells. The diameter of the inclusion is denoted by d_1. The study includes analysis of two models: the first model consists of a hard solid inclusion in a single soft matrix, while the second model consists of a hard inclusion wrapped in an additional layer of soft elastic material and embedded together in the soft matrix. The thickness of this soft wrapping layer is denoted by w. The proposed geometric models and their corresponding unit cells are shown in Fig. 1.

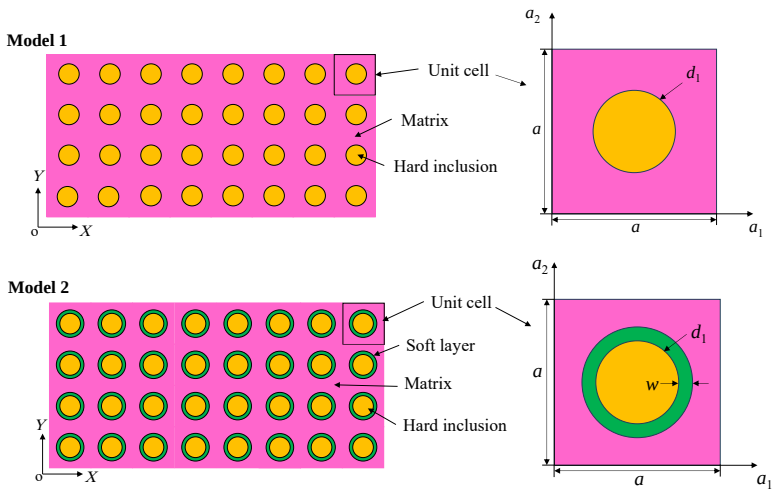


Fig. 1. Structural schematic of the elastic metamaterial presenting configuration of metamaterials models and their unit cell geometries.

The geometry of the metamaterial layer is defined by the initial geometric parameters of the unit cell presented in Table 1. In this study, in addition to a non-magnetic nature, chemical inertness and corrosion resistance are critical attributes that are required for the constituent materials to ensure suitability for application in an MRI environment. Therefore, the proposed metamaterial uses tungsten as the solid inclusion, a silicone (polysiloxane) polymer as the host elastic matrix, and soft rubber as the wrapping layer. The mechanical material properties considered for the numerical model are presented in Table 2.

TABLE 1. Geometric parameters in the model.

Parameter	Value [mm]
Lattice constant of unit cell – a	5
Diameter of hard inclusion – d_1	3
Thickness of the soft wrapping layer – w	0.2

TABLE 2. Material properties in the model.

Material	Density [kg/m^3]	Modulus of elasticity [GPa]	Poisson's ratio
Tungsten	18 950	360	0.28
Silicone	1080	4.5E-03	0.46
Rubber	1300	1.175E-04	0.47

In an MRI system, the gradient coil system is usually cylindrical with a radius of approximately $R \approx 300\,\text{mm}$. In this study, the maximum size of the metamaterial unit cell is constrained to below $a = 8\,\text{mm}$. This yields a curvature parameter $a/R = 0.026$. This small ratio indicates a weakly curved configuration, for which the structure can be locally approximated as planar at the scale of the unit cell. Consequently, the plane wave hypothesis employed in the dispersion analysis remains valid for the considered frequency range. In addition, the silicone host has a low elastic modulus, which allows it to conform to the cylindrical geometry of the MRI bore with negligible curvature-induced stress at the unit cell scale. Considering these conditions, the mechanical excitation of the metamaterial layer base can be approximated as locally uniform, and curvature effects are not expected to significantly influence the local resonant bandgap characteristics.

2.2. FEA

To characterise the intrinsic wave propagation behaviour of the periodical metamaterial presented in models 1 and 2, a dispersion relation analysis is performed using an eigenfrequency formulation. This approach enables the identification of subwavelength bandgaps associated with the locally resonant unit cell, within which wave propagation is restricted. By solving the Bloch–Floquet eigenvalue problem for a single unit cell, the dispersion relations are obtained independently of finite size and boundary effects, providing the fundamental basis for prediction of vibration attenuation. The theoretical background and conceived method for the calculation of the dispersion diagram are explained in this section (SUN *et al.*, 2022; HUANG *et al.*, 2021; LI *et al.*, 2023; YANG *et al.*, 2023; TANG *et al.*, 2024). The direct basic vectors in orthogonal Cartesian coordinates are $a_1 = a(1,0)$ and $a_2 = a(0,1)$. In the reciprocal space, these vectors are defined as follows:

$$b_1 = \frac{2\pi}{a}(1,0); \qquad b_2 = \frac{2\pi}{a}(0,1). \tag{1}$$

The finite element analysis (FEA) enables the evaluation of the dispersion diagram by solving the extreme values of the multi-order eigenfrequencies in the first Brillouin zone. The square lattice of unit cells exhibits $^1/_8$ symmetry, and the corresponding irreducible Brillouin zone (IBZ) with key points of highest symmetry Γ, X, and M are shown in Fig. 2.

Wave propagation in the linear elastic, periodic lattice media is expressed by the following governing equation:

$$-\rho\omega^2\mathbf{u} = (\lambda+\mu)\nabla(\nabla\cdot\mathbf{u}) + \mu\nabla^2\mathbf{u}, \tag{2}$$

where $\mathbf{u}$ is the displacement vector, ω is the angular frequency, ρ is the density, ∇ is the Hamilton operator, and λ and μ are the Lamé constants of the constituent materials.

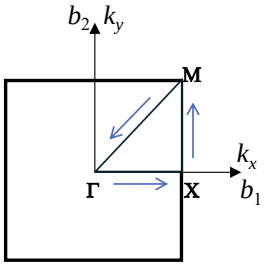


FIG. 2. Illustration of IBZ Γ-X-M in the square lattice unit cell.

According to Bloch–Floquet theory, an infinite structure with a lattice constant can be substituted by a unit cell for the basic calculation. The Bloch–Floquet theorem defines the displacement field of periodic media as

$$\mathbf{u}(\mathbf{r},t) = \mathbf{u}(\mathbf{r})e^{i\omega t}, \tag{3}$$

where t is time and $\mathbf{r} = (x, y)$ is the position vector.

The displacement vector $\mathbf{u}(\mathbf{r})$ can be expressed as

$$\mathbf{u}(\mathbf{r}) = U(\mathbf{r})e^{-i\mathbf{k}\cdot\mathbf{r}}, \tag{4}$$

where $U(\mathbf{r})$ is the eigenvalue amplitude and $\mathbf{k} = (k_x, k_y)$ is the Bloch wave vector. The displacement $\mathbf{u}(\mathbf{r})$ is also periodic, consistent with the lattice vector $\mathbf{R}$ (Bravais lattice vector), which is expressed as

$$\mathbf{u}(\mathbf{r}+\mathbf{R}) = U(\mathbf{r}). \tag{5}$$

Thus, the displacement u satisfies the following relation:

$$\mathbf{u}(\mathbf{r}+\mathbf{R}) = e^{i\mathbf{k}\cdot\mathbf{r}}\mathbf{u}(\mathbf{r}). \tag{6}$$

This relation is applied as a boundary condition to the outer edges of the unit cell in the x- and y-directions. Combining the governing equation with these boundary conditions yields a standard eigenvalue problem:

$$(\mathbf{K} - \omega^2\mathbf{M})\mathbf{q} = \mathbf{F}, \tag{7}$$

where $\mathbf{K}$ is the stiffness matrix, $\mathbf{M}$ is the mass matrix, and $\mathbf{q}$ and $\mathbf{F}$ are the generalised displacement and force vectors, respectively. The formulated eigenvalue problem is solved numerically by using an eigenfrequency analysis in COMSOL Multiphysics. The dispersion relation is calculated by sweeping the Bloch wave vector $\mathbf{k}$ along the boundaries of the IBZ. The FEM of the metamaterial cell for evaluation of dispersion relations was discretised by using a free triangular mesh. The mesh density was defined based on the smallest geometric feature of the model. Particularly, for model 2 with the thin wrapping layer of soft rubber around tungsten, a minimum of two finite elements were ensured across the thickness of the layer. A mesh refinement study was performed to ensure numerical accuracy, confirming that the results are insensitive to a further increase in mesh density.

To maintain the applicability of the elastic metamaterial in an MRI system for attenuating the vibration produced by the gradient coil, the size of the metamaterial was limited to 40 mm in the Y-direction. A parametric analysis was performed to investigate the influence of the geometric features in model 1 and model 2. The lattice constant a was varied from 5 mm to 8 mm, the diameter d_1 of the hard inclusion was varied from 3 mm to 7.5 mm, and the thickness w of the rubber layer was varied from 0.1 mm to 0.7 mm. Furthermore, the bandgap coverage (BGC) percentage was calculated to investigate the suitability of the metamaterial for attenuating MRI vibrations, according to the following expression:

$$\%\ \mathrm{BGC} = \left(\frac{\sum_{j=1}^{M} \max\left(0,\ \left(\min\left(f_{d2}, f_{g2}^{(j)}\right) - \max\left(f_{d1}, f_{g1}^{(j)}\right)\right)\right)}{f_{d2} - f_{d1}} \right) \times 100, \tag{8}$$

where f_{d1} and f_{d2} are the lower and upper limits of the desired bandgap, respectively, M is the total number of bandgaps in the desired frequency range, and $f_{g1}^{(j)}$ and $f_{g2}^{(j)}$ are the lower and upper limits of the j-th obtained bandgap, respectively.

A frequency-domain analysis was performed to evaluate the transmission loss (TL) of a finite number of elastic metamaterial cells, validating the band structure obtained from the eigenfrequency analysis. In the 2D $X-Y$ model, the TL is evaluated along the radial Y-direction, corresponding to the direction normal to the MRI panel surface. The FEA model used to calculate TL is shown in Fig. 3. Both models were meshed with free triangular elements, with the mesh setup and refinement checks consistent with those performed in the dispersion analysis. A harmonic displacement excitation with a frequency ranging from 0 Hz to 3000 Hz was applied as the input to the first cell of the elastic metamaterial, and the output was measured at the end of the last cell. The TL is calculated as

$$\mathrm{TL} = 20\log_{10}\left(\frac{P_{\mathrm{in}}}{P_{\mathrm{out}}}\right), \tag{9}$$

where P_{out} is the output displacement and P_{in} is the input displacement. An artificial domain known as a perfectly matched layer (PML) was considered on the input and output sides of the model. The TL was calculated for a number of iterations by varying the geometric parameters of the model 1 and model 2 elastic metamaterials.

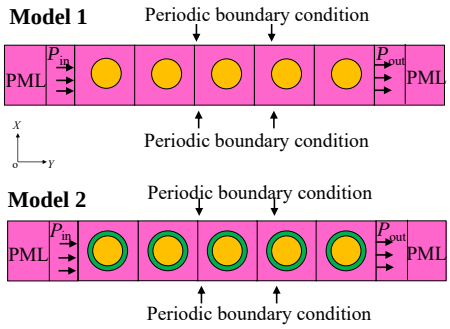


Fig. 3. FEA setup for evaluation of transmission loss in the Y-direction through a multicellular arrangement of metamaterial cells.

2.3. MRI compatibility analysis of selected materials

While Subsec. 2.1 and Subsec. 2.2 presented the mechanical necessity of the selected materials for achieving specific wave propagation characteristics, it is also crucial to verify their electromagnetic compatibility for integration into an MRI scanner. This sections analyses the electromagnetic characteristics of the constituent materials to ensure they do not disturb the imaging process. Properties such as magnetic susceptibility and electrical conductivity are particularly important to determine whether the structure might perturb the static magnetic field, influence the gradient field distribution, or induce currents during MRI operation. In the presented metamaterial models, the primary components are tungsten rods acting as inertial resonators, silicone serving as the host matrix, and a soft rubber wrapping layer functioning as additional layer around the tungsten rods specifically in the second model.

2.3.1. Influence of magnetic susceptibility on field homogeneity

Tungsten is a paramagnetic material with a volume magnetic susceptibility of approximately $\chi_v \approx 76\times10^{-6}$, calculated from the reported mass susceptibility of $\chi_{\mathrm{mass}} = 4\times10^{-9}$, m^3/kg, (Enghag, 2008) and an industrial-grade density of 18 950 $\mathrm{kg/m}^3$. The silicone has volume diamagnetic susceptibility of $\chi_v = -8.03\times10^{-6}$ for RTV 615 silicone, which is comparable to that of water and biological tissues (Wapler *et al.*, 2014). Soft rubber is a diamagnetic material and shows a volume susceptibility on the order of $\chi_v \approx -1\times10^{-5}$ (Callister, Rethwisch, 2020). As a result, the overall magnetic contrast introduced by the elastomeric components is minimal.

As discussed by Schenck, the localised susceptibility-mismatch-induced magnetic field distortion ΔB decays with distance s according to an inverse-cubic relation: $\Delta B \propto 1/s^3$, where s is the scalar distance from the

susceptibility inclusion (Schenck, 1996). Given the relatively weak susceptibility contrast of tungsten compared to ferromagnetic materials and the peripheral placement of the metamaterial relative to the scanning volume, any field distortion decays substantially before reaching the isocentre, resulting in a negligible influence on the global field homogeneity and the linearity of the gradient within the scanning volume. For context, the volume magnetic susceptibility of tungsten is approximately half that of titanium alloys commonly used in MRI-conditional implants (Kajima *et al.*, 2020; Chiba *et al.*, 2019).

2.3.2. Influence of electrical conductivity and eddy currents

In an MRI system, the dynamic magnetic fields from the gradient coils are known to induce eddy currents responsible for heating, electromotive forces, and torques on conductors and imaging artifacts. Tungsten is an electrically conductive material and could, in principle, support induced eddy currents when exposed to time-varying gradient magnetic fields. The magnitude of the eddy currents and the heating effects in a typical MRI environment strongly depends on the geometry, location, and orientation of the conductor (Buchli *et al.*, 1988; Stroud *et al.*, 2019; Lee *et al.*, 2024).

The tungsten rods are oriented parallel to the main magnetic field, which minimises magnetically induced displacement and torque. Conversely, time-varying transverse gradient magnetic fields are perpendicular to the rod axis and may induce eddy currents within the conductor. However, these currents are expected to remain localised and limited in magnitude due to the design features of the metamaterial. Each tungsten rod is encapsulated by the dielectric silicone in model 1, while model 2 has an additional layer of dielectric rubber wrapped around the tungsten rods. This configuration prevents electrical continuity between adjacent tungsten inclusions, consequently minimising the possibility of large-scale conduction loops and resulting eddy-current heating. Additionally, the relatively small diameter of the tungsten rods minimises the effective loop area exposed to changing magnetic flux, further reducing the eddy-current amplitudes.

Considering the radiofrequency (RF) pulses during an MRI scan, the RF field couples with the conductor in the MRI scan bore and induces heat at the tips of the conductor. This heating effect is maximised when the length of the conductor corresponds to half of the RF excitation wavelength in the MRI medium, a phenomenon known as the antenna effect (Mattei *et al.*, 2008; Winter *et al.*, 2021). Taking advantage of the 2D design of the metamaterial, the length of the tungsten rods can be explicitly controlled to make them electrically short compared to the RF wavelength at clinical field strength, thereby minimising their ability to resonate with the RF field. Furthermore, the rods are embedded in dielectric material, ensuring galvanic isolation from the scanner and the subject. The isolation limits the RF-induced eddy-current pathways and prevents conduction loops.

Based on the above considerations, the selected materials support the assumption that the proposed structure introduces negligible perturbation to the MRI magnetic field and minimal eddy-current generation, while producing effective vibration control through the local resonance mechanism.

3. Results and discussion

This section presents the band structures of elastic metamaterial model 1 and model 2 calculated using the method mentioned in the previous section. The mechanisms of bandgap formation and the parametric influences on these characteristic are thoroughly investigated. Furthermore, the TL is evaluated through a frequency-domain analysis as an indicator of reduced vibro-acoustic coupling.

3.1. Bandgap characteristics and vibration modes

Initially, the first eight eigenfrequencies were considered for the calculation of the dispersion curves. The dispersion relations obtained from the eigenfrequency analysis of model 1 in its initial configuration ($a = 5\,\mathrm{mm}$, $d_1 = 3\,\mathrm{mm}$) are shown in Fig. 4. The dispersion relations of model 1 notably show a complete bandgap highlighted in the grey region, where the upper limit of the eigenfrequency solutions was observed to be above 11 000 Hz. As per requirements of the application in an MRI scanner, the bandgap should appear within the frequency range

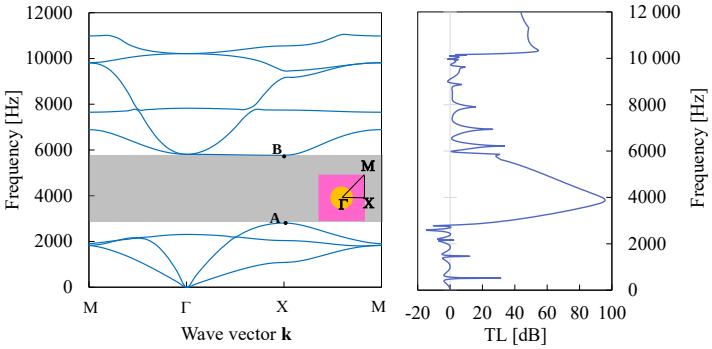


FIG. 4. Dispersion diagram of elastic metamaterial model 1 presenting the identified bandgap shaded in the grey region, with points A and B representing the lower and upper limits of the bandgap, accompanied by the TL spectrum.

of 300 Hz to 3000 Hz. To assess the validation of the dispersion curve characteristics of the initial configuration of model 1, the TL characteristics were evaluated in the frequency range from 0 Hz to 12 000 Hz, as depicted in Fig. 4.

It is evident that the dispersion curves of model 1 exhibit a single ultrawide complete bandgap spanning from 2804.4 Hz to 5765.7 Hz between the 3rd and 4th eigenfrequency bands. The attenuation peak exceeds 76 dB within the bandgap, with an average TL exceeding 58 dB within this forbidden band. The vibration attenuation region in the TL spectrum roughly corresponds to the attenuation region of the bandgap observed in the dispersion curve, providing a strong validation for the accuracy of the calculated dispersion diagram. The overall BGC concerning the MRI vibration range of model 1 was found to be 7.24 % for the given configuration. The vibration r f the bandgap in model 1, are sh vibration, where red indicates th

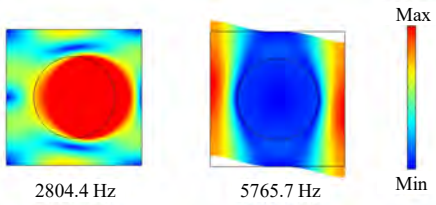


FIG. 5. Vibration modes of bandgap in model 1 at points A and B shown in the dispersion diagram.

Model 2 of the elastic metamaterial was analysed using a similar configuration of unit cell of model 1 (a = 5 mm; d_1 = 3 mm) and a thickness of soft wrapping layer of w = 0.2 mm. To maintain consistency in the methodology, the analysis was limited to first eight eigenfrequencies, and the dispersion relation was calculated and plotted

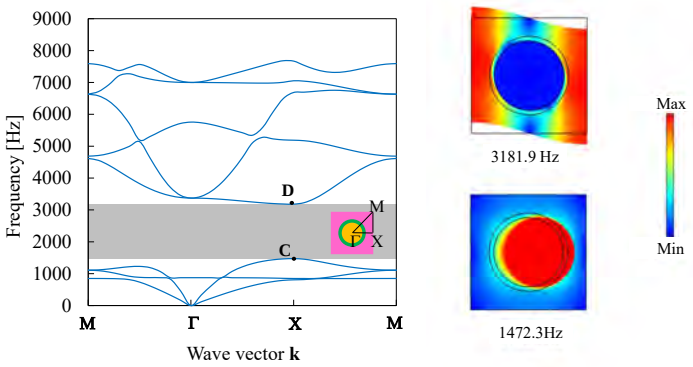


FIG. 6. Dispersion diagram of elastic metamaterial model 2 depicting the bandgap shaded in the grey region, along the bandgap vibration modes at points C and D.

as shown in Fig. 6. Model 2 also shows a single wide complete bandgap starting at 1472.3 Hz and closing at 3181.9 Hz. The presence of the soft layer around the hard inclusion shifted the bandgap to lower frequencies. The vibration mode shapes corresponding to the opening and closing of the bandgap at points C and D are presented in Fig. 6. The BGC calculated for the analysed configuration of model 2 was found to be 56.57 %.

3.2. Parametric analysis of metamaterial models

After identifying bandgaps closer to the desired frequency range, the influence of the geometric parameters on both elastic metamaterial models was explored. The diameter d_1 of the hard inclusion was varied from 3 mm to 0.5 mm less than the lattice constant a with an incremental step of 0.5 mm. During the subsequent analysis, the lattice constant parameter a was varied from 5 mm to 8 mm, with an incremental step of 1.0 mm. The variation in the obtained bandgap characteristics of model 1 are shown in the Fig. 7.

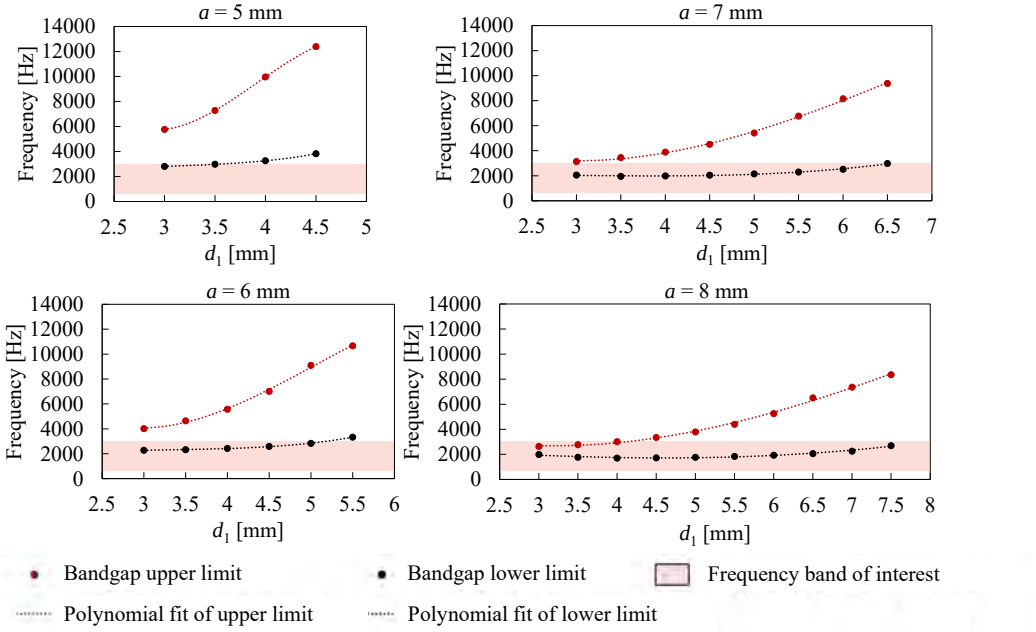


Fig. 7. Bandgap characteristics of metamaterial model 1 for different diameters of inclusion in unit cell with lattice constants, depicting the overlap with the frequency range of interest.

The results from the parametric analysis highlight that the bandgap characteristics are significantly influenced by the inclusion diameter and the lattice constant of the metamaterial. As observed in Fig. 7, the upper limit of the bandgap showed a significant increase with an increase in the diameter of the hard inclusion, resulting in a widening of the bandgap. The lower limit of the bandgap also presented a gradual increase corresponding to the increase in the diameter of the hard inclusion. Model 1 with the configuration a = 5 mm, d_1 = 4.5 mm presented the widest bandgap spanning from 3820.7 Hz to 12 389.6 Hz. Increasing the diameter in the other configurations having lattice constants of a = 6 mm, 7 mm, and 8 mm revealed a similar trend in the upper and lower limits of the bandgap frequencies. Importantly, an increase in the lattice constant a shifts the bandgaps downward in the dispersion diagrams. The minimum lower limit of the bandgap was observed to be 1708.6 Hz, with a corresponding upper limit of 3017.7 Hz for the configuration of a = 8 mm, d_1 = 4 mm in model 1.

Model 2 of the elastic metamaterial was analysed to investigate the influence of geometric parameters on the bandgap characteristics. As in the case of model 1, the lattice constant was similarly varied from 5 mm to 8 mm with a step of 1 mm, and the diameter of the hard inclusion was changed from 3 mm with a step of 0.5 mm, keeping the thickness of the soft wrapping layer at 0.2 mm. The results from the parametric analysis of model 2 are presented in Fig. 8. An increase in the inclusion diameter d_1, shows a significant increase in the upper limit of the bandgap but a small decrement in the lower limit of the bandgap, leading to a widening of the overall bandgap width. The widest bandgap was observed for the configuration of a = 5 mm, d_1 = 4.5 mm, w = 0.2 mm

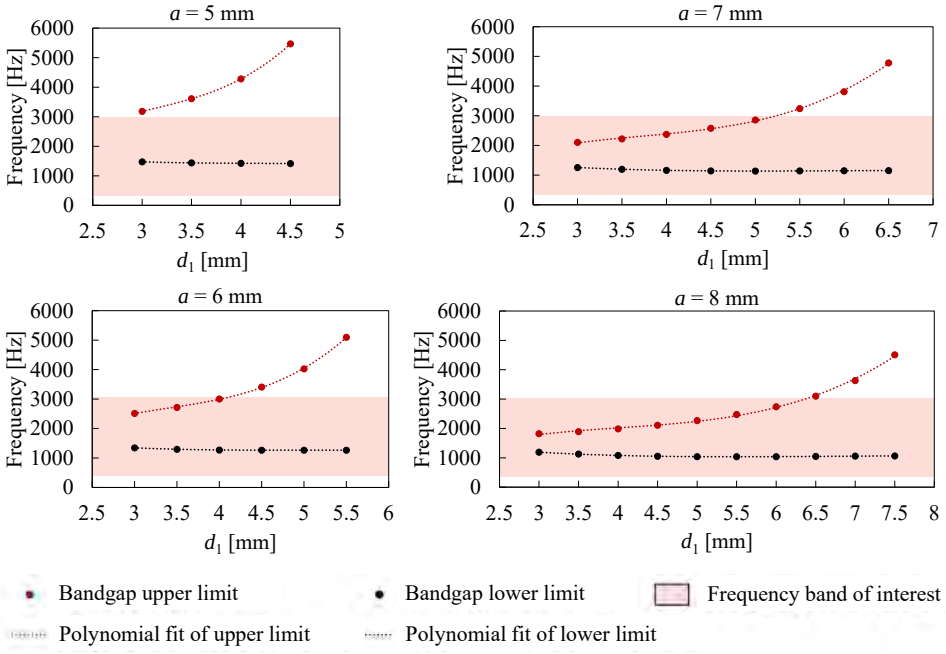


FIG. 8. Bandgap characteristics of metamaterial model 2 with 0.2 mm thick rubber layer for varying diameters of inclusion in unit cell with lattice constants, depicting the overlap of the bandgaps with the frequency range of interest.

in the range of 1413.9 Hz to 5466.7 Hz. Meanwhile, the bandgap with the minimum lower limit of 1034.9 Hz and a corresponding upper limit of 2472.4 Hz was obtained for the configuration of a = 8 mm, d_1 = 5.5 mm, w = 0.2 mm in model 2.

A comprehensive analysis of the parametric studies for model 1 and model 2 was conducted to investigate their performance concerning the attenuation of MRI system vibrations, as shown in the Fig. 9. The performance factor was calculated as the % BGC according to Eq. (8). It was revealed that metamaterial model 1 exhibits a maximum BGC of 47.83 % for the configuration of a = 8 mm, d_1 = 4 mm. In model 2, the maximum BGC of 72.35 % was achieved with the configuration of a = 8 mm, d_1 = 6.5 mm, w = 0.2 mm, which was consequently selected as the optimum model for subsequent investigations. Comparing the corresponding plots in Fig. 7, Fig. 9, and Fig. 10, the presence of rubber layer is essential to shift the bandgap into the lower frequency range.

Model 1

a [mm] \ d1 [mm]	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5
5	7.24	0.87	0.00	0.00	NA	NA	NA	NA	NA	NA
6	26.74	25.13	21.34	15.40	6.36	0.00	NA	NA	NA	NA
7	35.00	38.79	37.73	35.28	31.50	26.18	18.16	0.91	NA	NA
8	23.96	37.26	47.83	47.09	45.41	42.83	39.32	34.57	27.25	11.12

Model 2

a [mm] \ d1 [mm]	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5
5	56.57	57.86	58.39	58.74	NA	NA	NA	NA	NA	NA
6	43.34	52.59	63.94	64.25	64.22	64.22	NA	NA	NA	NA
7	31.42	37.83	44.89	53.24	63.76	68.83	68.56	68.39	NA	NA
8	23.29	28.41	33.55	39.14	45.54	53.24	62.96	72.35	71.96	71.71

100
0
% BGC

FIG. 9. BGC percentage (% BGC) calculated for different configurations of metamaterial model 1 and model 2 with a wrapping rubber thickness of w = 0.2 mm for the desired frequency range of 300 Hz to 3000 Hz.

As the configuration of metamaterial model 2 with an inclusion diameter of d_1 = 6.5 mm and a lattice constant of a = 8 mm led to the maximum % BGC, it was selected for further evaluation to investigate the influence of the rubber wrapping layer thickness on the bandgap characteristics. The results from the analysis are depicted

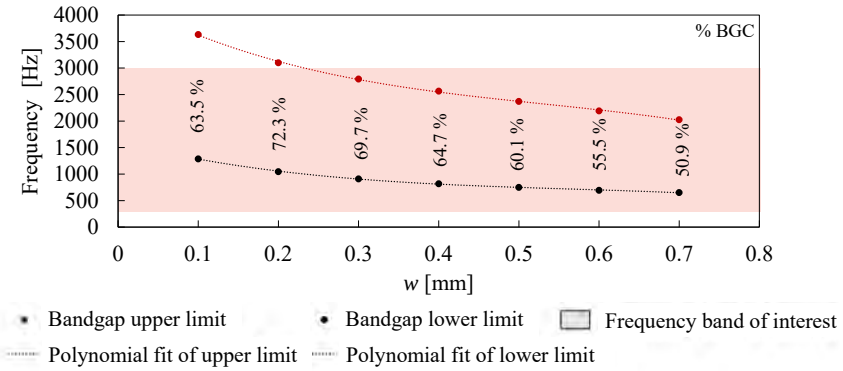


FIG. 10. Bandgap characteristics obtained by varying the wrapping layer thickness w in model 2 with a resonator diameter of $d_1 = 6.5$ mm and a lattice constant of $a = 8$ mm.

in Fig. 10. The thickness of the wrapping layer was varied from 0.1 mm to the 0.7 mm with the incremental step size of 0.1 mm. As the layer thickness increased, the bandgap shifted to lower frequencies, though this compromised the bandgap width. The maximum BGC of 72.3 % was observed for the configuration with a 0.2 mm wrapping layer thickness. Furthermore, an increase in thickness resulted in a gradual reduction of the BGC from the maximum of 72.3 % to a minimum of 50.9 %. The lowest frequency associated with the opening of the bandgap was found to be 650.6 Hz for model 2 with the configuration of $a = 8$ mm, $d_1 = 6.5$ mm, $w = 0.7$ mm.

The polynomial fit equations of the upper and lower limits of the bandgap, derived from the complete parametric analysis of model 1 and model 2, are presented in Appendix.

3.3. Frequency domain analysis of the optimum model

The unit cell of model 2, exhibiting the maximum BGC of 72.3 % identified in the previous section, was considered for analysing the influence of the number of unit cells on the elastic wave propagation phenomenon. The TL curves were plotted for arrangement of three, four, and five cells, as shown in Fig. 11. A harmonic input displacement of 0.5 mm was applied to the left edge of the first cell, and the response displacement on the opposite side of the last cell was analysed to calculate the TL in the model. The results showed that the position of bandgap characteristics concerning the frequency range remained identical, but the vibration attenuation performance (TL) increased with an increase in the number of cells. Also, the results of the displacement propagation in the five-unit-cell model for frequencies inside and outside the bandgap is shown in Fig. 11. It can be clearly seen that in the bandgap region, the wave cannot propagate to the other side of the metamaterial.

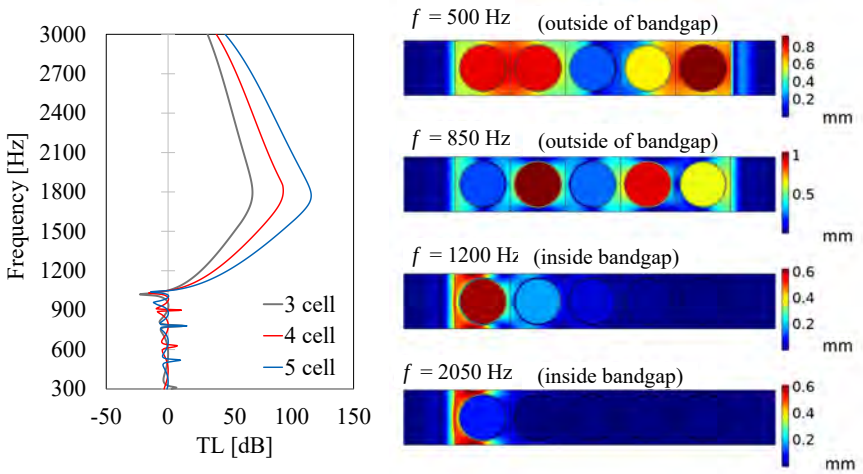


FIG. 11. Transmission loss characteristics calculated for the composite system of elastic metamaterial model 2 with arrangement of three, four, and five unit cells using the configuration ($a = 8$ mm, $d_1 = 6.5$ mm, and $w = 0.2$ mm), along with the displacement wave propagation in metamaterial model 2 with five-unit cells at different input frequencies.

This study demonstrates that a high contrast in the physical properties of the constituent materials can produce local resonance bandgaps in mid-range frequencies, specifically from 300 Hz to 3000 Hz, advancing conventional metamaterial designs that require unit cell sizes of few centimetres to target a similar frequency range (Gao *et al.*, 2023; Yin *et al.*, 2024). Material selection was challenging due to MRI-specific and medical requirements. Model 2, with its optimum configuration, achieved a BGC of 72.3 % and high TL within bandgaps, suggesting potential for reducing vibro-acoustic coupling at the air-structure interface. Although the optimum model covers a high-noise frequency zone, frequencies below the bandgap may still contribute to noise. Hence, it is important to explore more configurations to fully cover the MRI noise spectrum. Currently, the weight of the identified metamaterial is high due to the large fill fraction of the high-density inclusion. Future work will focus on enhancing the attenuation capabilities of the elastic metamaterial while ensuring it to be lightweight and non-magnetic, and fully compatible with the MRI environment.

4. Conclusion

In summary, two models of an elastic metamaterial composed of a soft silicone matrix, a rubber layer, and hard tungsten cylindrical inclusions were analysed for targeted vibration attenuation in the MRI gradient coil frequency range It was found that the both models show bandgaps in the desired frequency range. However, the BCG of model 1 was very low compared to that of model 2. As observed in model 1, the bandgap shifted to lower frequencies with an increase in the inclusion diameter, but it could not achieve a BGC higher than 47.83 %, with the bandgap opening at 1708.6 Hz. Conversely, the optimum configuration of model 2 demonstrated a BGC of up to 72.3 %, with a lower bandgap opening frequency at 1046.6 Hz.

The characteristics of the presented elastic metamaterial are highly dependent on its geometric parameters. The parametric analysis revealed that larger inclusion diameters, larger lattice constants, and an increased thickness of the soft layer shift the bandgap towards lower frequencies, though this narrows the bandgap width. A frequency-domain analysis demonstrated significant TL within the forbidden bands, validating the vibration attenuation performance. These results suggest that further studies should focus on increasing BGC through exploring multi-material configurations for enhanced performance.

Overall, this numerical study highlights the capability of elastic metamaterials to attenuate MRI-induced vibration. Additionally, the high TL achieved in the bandgap suggests that the presented elastic metamaterial may provide a means to reduce vibro-acoustic coupling in MRI systems.

Appendix

Trendline equations of the bandgap limits in model 1 from the parametric analysis.

Configuration	Upper limit of the bandgap	Lower limit of the bandgap
$a = 5$ mm, $d_1 \in [3, 4.5]$	$f_{g2} = -1864.1d_1^3 + 21880d_1^2 - 79995d_1 + 99160$	$f_{g1} = 222.27d_1^3 - 2111.4d_1^2 + 7011.1d_1 - 5227.6$
$a = 6$ mm, $d_1 \in [3, 5.5]$	$f_{g2} = -277.3d_1^3 + 4202.4d_1^2 - 17597d_1 + 26525$	$f_{g1} = 80.29d_1^3 - 818.66d_1^2 + 2903.8d_1 - 1239.3$
$a = 7$ mm, $d_1 \in [3, 6.5]$	$f_{g2} = -53.57d_1^3 + 1181.9d_12 - 5653.8d_1 + 10967$	$f_{g1} = 23.636d_1^3 - 197.09d_1^2 + 462.25d_1 + 1780.1$
$a = 8$ mm, $d_1 \in [3, 7.5]$	$f_{g2} = -20.172d_1^3 + 591.54d_1^2 - 3166.2d_1 + 7421.9$	$f_{g1} = 4.7403d_1^3 + 30.757d_1^2 - 582.69d_1 + 3286.3$

Trendline equations of the bandgap limits of model 2 from the parametric analysis.

Configuration	Upper limit of the bandgap (f_{g2})	Lower limit of the bandgap (f_{g1})
$a = 5$ mm, $w = 0.2$ mm $d_1 \in [3, 4.5]$	$f_{g2} = 361.56d_1^3 - 3308.8d_1^2 + 10883d_1 - 9448.6$	$f_{g1} = -20.277d_1^3 + 253.14d_1^2 - 1070.6d_1 + 2953.3$
$a = 6$ mm, $w = 0.2$ mm $d_1 \in [3, 5.5]$	$f_{g2} = 156.2d_1^3 - 1586.6d_1^2 + 5809.9d_1 - 4857.8$	$f_{g1} = -10.95d_1^3 + 164.3d_1^2 - 818.35d_1 + 2618.9$
$a = 7$ mm, $w = 0.2$ mm $d_1 \in [3, 6.5]$	$f_{g2} = 77.66d_1^3 - 864.36d_1^2 + 3479.1d_1 - 2667.8$	$f_{g1} = -6.4297d_1^3 + 113.04d_1^2 - 648.65d_1 + 2359.7$
$a = 8$ mm, $w = 0.2$ mm $d_1 \in [3, 7.5]$	$f_{g2} = 43d_1^3 - 519.71d_1^2 + 2277.3d_1 - 1526.4$	$f_{g1} = -4.1224d_1^3 + 82.899d_1^2 - 536.63d_1 + 2164.7$
$a = 8$ mm, $d_1 = 6.5$ mm $w \in [0.1, 0.7]$	$f_{g2} = -7708.7w^3 + 12277w^2 - 8101.6w + 4317$	$f_{g1} = -3417.4w^3 + 5814.9w^2 - 3762.6w + 1604.1$

Fundings

This work was supported and financed by Department of Mechanics and Vibroacoustics, Faculty of Mechanical Engineering and Robotics, AGH University of Krakow, in Kraków, Poland, from the finance source: 10000.501.00.130000.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' contributions

Swapnil Arawade conceptualized the study, wrote the original draft of the manuscript, conducted the formal analysis, and interpreted the results. Janusz Piechowicz supervised the study, contributed to the interpretation of the results, reviewed and edited the manuscript, and acquired the funding. All authors reviewed and approved the final manuscript.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Ahmad M.J. *et al.* (2025), Non-conformal thermal cloak metamaterial by continuous metal fiber embedded 3D printing, *International Journal of Heat and Mass Transfer*, **242**: 126796, https://doi.org/10.1016/j.ijheatmasstransfer.2025.126796.
2. Al-Meer M.H. (2022), MRI acoustic noise cancellation using CNN, *Journal of Engineering Research*, https://doi.org/10.36909/jer.17661.
3. Arawade S., Piechowicz J. (2024a), Acoustic characterization of the 1.5 Tesla MRI facility in mobile imaging trailer, *Vibrations in Physical Systems*, **34**(2): 20242019, http://doi.org/10.21008/j.0860-6897.2024.2.19.
4. Arawade S., Piechowicz J. (2024b), Spectral analysis of MRI sound signal, *International Journal of Electronics and Telecommunications*, **70**(4): 839–847, http://doi.org/10.24425/ijet.2024.152068.
5. Bruno F. *et al.* (2019), Advanced magnetic resonance imaging (MRI) of soft tissue tumors: techniques and applications, *La Radiologia Medica*, **124**: 243–252, https://doi.org/10.1007/s11547-019-01035-7.
6. Buchli R., Boesiger P., Meier D. (1988), Heating effects of metallic implants by MRI examinations, *Magnetic Resonance in Medicine*, **7**(3): 255–261, https://doi.org/10.1002/mrm.1910070302.
7. Callister W.D. Jr, Rethwisch D.G. (2020), *Callister's Materials Science and Engineering*, John Wiley & Sons.
8. Chiba Y. *et al.* (2019), Quantification of metal-induced susceptibility artifacts associated with ultrahigh-field magnetic resonance imaging of spinal implants, *JOR Spine*, **2**(3): e1064, https://doi.org/10.1002/jsp2.1064.
9. Chua J.W., Poh D.K.W., Ding S., Pei H., Li X. (2025), Unveiling the structure-property relationships of multilayered Helmholtz resonance-based acoustic metamaterials, *Smart Materials in Manufacturing*, **3**: 100073, https://doi.org/10.1016/j.smmf.2025.100073.
10. Chuang K.C., Yuan Z.W., Guo Y.Q., Lv X.F. (2020), Extracting torsional band gaps and transient waves in phononic crystal beams: method and validation, *Journal of Sound and Vibration*, **467**: 115004, https://doi.org/10.1016/j.jsv.2019.115004.
11. Enghag P. (2008), *Encyclopedia of the Elements: Technical Data – History – Processing – Applications*, John Wiley & Sons.

12. Fan L., He Y., Chen X.A., Zhao X. (2021), A frequency response function-based optimization for metamaterial beams considering both location and mass distributions of local resonators, *Journal of Applied Physics*, **130**(11): 115101, https://doi.org/10.1063/5.0059025.

13. Frollo I., Andris P., Pribil J., Juras V. (2007), Indirect susceptibility mapping of thin-layer samples using nuclear magnetic resonance imaging, *IEEE Transactions on Magnetics*, **43**(8): 3363–3367, https://doi.org/10.1109/TBME.2019.2923693.

14. Fu Y. *et al.* (2025), A multi-polarization multi-band reconfigurable cloaking-illusion metasurface, *Advanced Optical Materials*, **13**(4): 2402381, https://doi.org/10.1002/adom.202402381.

15. Fuhrer E., Jouda M., Klein C.O., Wilhelm M., Korvink J.G. (2019), Gradient-induced mechanical vibration of neural interfaces during MRI, *IEEE Transactions on Biomedical Engineering*, **67**(3): 915–923, https://doi.org/10.1109/TBME.2019.2923693.

16. Gallichan D., Scholz J., Bartsch A., Behrens T.E., Robson M.D., Miller K.L. (2010), Addressing a systematic vibration artifact in diffusion-weighted MRI, *Human Brain Mapping*, **31**(2): 193–202, https://doi.org/10.1002/hbm.20856.

17. Gao W., Zhang Q., Sun J., Guo K. (2023), A novel 3D-printed magnesium alloy phononic crystal with broadband bandgap, *Journal of Applied Physics*, **133**(8): 085103, https://doi.org/10.1063/5.0135770.

18. Hao L. *et al.* (2025), Tunable tri-band negative mass density acoustic metamaterials with nested three-layer hollow tubes, *Physica B: Condensed Matter*, **707**: 417194, https://doi.org/10.1016/j.physb.2025.417194.

19. He Y. *et al.* (2024), Design and mechanical properties analysis of hexagonal perforated honeycomb metamaterial, *International Journal of Mechanical Sciences*, **270**: 109091, https://doi.org/10.1016/j.ijmecsci.2024.109091.

20. Hedeen R.A., Edelstein W.A. (1997), Characterization and prediction of gradient acoustic noise in MR imagers, *Magnetic Resonance in Medicine*, **37**(1): 7–10, https://doi.org/10.1002/mrm.1910370103.

21. Hu G., Austin A.C.M., Sorokin V., Tang L. (2021), Metamaterial beam with graded local resonators for broadband vibration suppression, *Mechanical Systems and Signal Processing*, **146**: 106982, https://doi.org/10.1016/j.ymssp.2020.106982.

22. Hu G., Tang L., Xu J., Lan C., Das R. (2019), Metamaterial with local resonators coupled by negative stiffness springs for enhanced vibration suppression, *Journal of Applied Mechanics*, **86**(8): 081009, https://doi.org/10.1115/1.4043827.

23. Huang T.T. *et al.* (2021), Based on auxetic foam: a novel type of seismic metamaterial for Lamb waves, *Engineering Structures*, **246**: 112976, https://doi.org/10.1016/j.engstruct.2021.112976.

24. Hutter J. *et al.* (2018), Quiet echo planar imaging for functional and diffusion MRI, *Magnetic Resonance in Medicine*, **79**(3): 1447–1459, https://doi.org/10.1002/mrm.26810.

25. Kaina N., Fink M., Lerosey G. (2013), Composite media mixing Bragg and local resonances for highly attenuating and broad bandgaps, *Scientific Reports*, **3**: 3240, https://doi.org/10.1038/srep03240.

26. Kajima Y., Takaichi A., Tsutsumi Y., Hanawa T., Wakabayashi N., Kawasaki A. (2020), Influence of magnetic susceptibility and volume on MRI artifacts produced by low magnetic susceptibility Zr-14Nb alloy and dental alloys, *Dental Materials Journal*, **39**(2): 256–261, https://doi.org/10.4012/dmj.2018-426.

27. Kanal E., Shellock F.G., SMRI Safety Committee (1992), Policies, guidelines, and recommendations for MR imaging safety and patient management, *Journal of Magnetic Resonance Imaging*, **2**(2): 247–248, https://doi.org/10.1002/jmri.1880020222.

28. Kumar R., Kumar M., Chohan J.S., Kumar S. (2022), Overview on metamaterial: history, types and applications, *Materials Today: Proceedings*, **56**: 3016–3024, https://doi.org/10.1016/j.matpr.2021.11.423.

29. Landini L., Positano V., Santarelli M. [Eds] (2018), *Advanced Image Processing in Magnetic Resonance Imaging*, CRC Press, Boca Raton, https://doi.org/10.1201/9781420028669.

30. Lasota A., Meller M. (2020), Iterative learning approach to active noise control of highly autocorrelated signals with applications to machinery noise, *IET Signal Processing*, **14**(8): 560–568, https://doi.org/10.1049/iet-spr.2020.0064.

31. Lee G., Lee D., Park J., Jang Y., Kim M., Rho J. (2022), Piezoelectric energy harvesting using mechanical metamaterials and phononic crystals, *Communications Physics*, **5**(1): 94, https://doi.org/10.1038/s42005-022-00869-4.

32. Lee N., Park Y., Lee G.W. (2017), Frequency-domain active noise control for magnetic resonance imaging acoustic noise, *Applied Acoustics*, **118**: 30–38, https://doi.org/10.1016/j.apacoust.2016.11.003.

33. Lee S.H., Wright O.B. (2016), Origin of negative density and modulus in acoustic metamaterials, *Physical Review B*, **93**(2): 024302, https://doi.org/10.1103/PhysRevB.93.024302.

34. Lee S.K., Tarasek M., Li K., Spence D.K. (2024), Eddy current heating of metallic shells and plates: application to gradient-induced conductor heating in MRI, *IEEE Transactions on Magnetics*, **60**(10): 1–20, https://doi.org/10.1109/TMAG.2024.3432915.

35. Leo P., D'Orazio A. (2025), Lorentz force and vibrations in transverse gradient coils in MRI, *International Journal of Mechanical Sciences*, **288**: 110011, https://doi.org/10.1016/j.ijmecsci.2025.110011.

36. Li G., Chen Y., Chen W., Liu J., He H. (2022), Local resonance – Helmholtz lattices with simultaneous solid-borne elastic waves and air-borne sound waves attenuation performance, *Applied Acoustics*, **186**: 108450, https://doi.org/10.1016/j.apacoust.2021.108450.

37. Li M., Rudd B., Lim T.C., Lee J.H. (2011), In situ active control of noise in a 4 T MRI scanner, *Journal of Magnetic Resonance Imaging*, **34**(3): 662–669, https://doi.org/10.1002/jmri.22694.

38. Li Y. (2025), A review on porous structure negative Poisson's ratio metamaterials, *Journal of Engineering Research and Reports*, **27**(1): 189–204, https://doi.org/10.9734/jerr/2025/v27i11378.

39. Li Y., Yan S., Peng Y. (2023), Broadband vibration attenuation characteristic of 2D phononic crystals with cross-like pores, *Thin-Walled Structures*, **183**: 110418, https://doi.org/10.1016/j.tws.2022.110418.

40. Liu M., Zhu W.D. (2019), Modeling and analysis of in-plane and out-of-plane elastic wave propagation in a phononic-crystal circular beam, *Journal of Sound and Vibration*, **462**: 114886, https://doi.org/10.1016/j.jsv.2019.114886.

41. Liu X., Chen S., Wang B., Tan X., Cao B., Yu L. (2025), A mechanical metamaterial with real-time tunable bandgap based on pneumatic actuation, *International Journal of Mechanical Sciences*, **289**: 110045, https://doi.org/10.1016/j.ijmecsci.2025.110045.

42. Liu Z., Zhang X., Mao Y., Zhu Y.Y., Yang Z., Chan C.T., Sheng P. (2000), Locally resonant sonic materials, *Science*, **289**(5485): 1734–1736, https://doi.org/10.1126/science.289.5485.1734.

43. Ma G., Sheng P. (2016), Acoustic metamaterials: from local resonances to broad horizons, *Science Advances*, **2**(2): e1501595, https://doi.org/10.1126/sciadv.1501595.

44. Mansfield P., Glover P.M., Beaumont J. (1998), Sound generation in gradient coil structures for MRI, *Magnetic Resonance in Medicine*, **39**(4): 539–550, https://doi.org/10.1002/mrm.1910390406.

45. Mattei E. *et al.* (2008), Complexity of MRI induced heating on metallic leads: experimental measurements of 374 configurations, *Biomedical Engineering OnLine*, **7**(1): 11, https://doi.org/10.1186/1475-925X-7-11.

46. McJury M.J. (2022), Acoustic noise and magnetic resonance imaging: a narrative/descriptive review, *Journal of Magnetic Resonance Imaging*, **55**(2): 337–346, https://doi.org/10.1002/jmri.27525.

47. Miao Z. *et al.* (2025), Design of multi-bandgap metamaterial plate based on composite cylindrical resonators, *Materials & Design*, **250**: 113570, https://doi.org/10.1016/j.matdes.2024.113570.

48. Moelker A., Pattynama P.M. (2003), Acoustic noise concerns in functional magnetic resonance imaging, *Human Brain Mapping*, **20**(3): 123–141, https://doi.org/10.1002/hbm.10134.

49. Moelker A., Wielopolski P.A., Pattynama P.M. (2003), Relationship between magnetic field strength and magnetic-resonance-related acoustic noise levels, *Magnetic Resonance Materials in Physics, Biology and Medicine*, **16**: 52–55, https://doi.org/10.1007/s10334-003-0005-9.

50. Naify C.J., Chang C.M., McKnight G., Nutt S. (2010), Transmission loss and dynamic response of membrane-type locally resonant acoustic metamaterials, *Journal of Applied Physics*, **108**(11): 114905, https://doi.org/10.1063/1.3514082.

51. Peiró-Torres M.P., Castiñeira-Ibáñez S., Redondo J., Sánchez-Pérez J.V. (2019), Interferences in locally resonant sonic metamaterials formed from Helmholtz resonators, *Applied Physics Letters*, **114**(17): 171901, https://doi.org/10.1063/1.5092375.

52. Pires F.A., Sangiuliano L., Denayer H., Deckers E., Desmet W., Claeys C. (2022), The use of locally resonant metamaterials to reduce flow-induced noise and vibration, *Journal of Sound and Vibration*, **535**: 117106, https://doi.org/10.1016/j.jsv.2022.117106.

53. Qi X.Q., Li T.J., Zhang J.L., Zhang Z., Tang Y.Q. (2018), Band gap structures for 2D phononic crystals with composite scatterer, *Applied Physics A*, **124**(5): 364, https://doi.org/10.1007/s00339-018-1789-9.

54. Reda R., Zanza A., Mazzoni A., Cicconetti A., Testarelli L., Di Nardo D. (2021), An update of the possible applications of magnetic resonance imaging (MRI) in dentistry: a literature review, *Journal of Imaging*, **7**(5): 75, https://doi.org/10.3390/jimaging7050075.

55. Ruks L., Ballantine K.E., Ruostekoski J. (2025), Negative refraction of light in an atomic medium, *Nature Communications*, **16**(1): 1433, https://doi.org/10.1038/s41467-025-56250-w.

56. Schenck J.F. (1996), The role of magnetic susceptibility in magnetic resonance imaging: MRI magnetic compatibility of the first and second kinds, *Medical Physics*, **23**(6): 815–850, https://doi.org/10.1118/1.597854.

57. Shunshun R., Zhao G. (2025), Mechanical characterization of a metamaterial with negative Poisson's ratio under compressive loading: experimental along with FEM, *Mechanics of Advanced Materials and Structures*, **33**(1): 2477229, https://doi.org/10.1080/15376494.2025.2477229.

58. Stroud J., Stupic K., Walsh T., Celinski Z., Hankiewicz J.H. (2019), Local heating of metallic objects from switching magnetic gradients in MRI, [in:] *Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Application*, **10954**: 150–161, https://doi.org/10.1117/12.2512903.

59. Sun P. *et al.* (2022), Mechanics and extreme low-frequency band gaps of auxetic hexachiral acoustic metamaterial with internal resonant unit, *Applied Acoustics*, **200**: 109046, https://doi.org/10.1016/j.apacoust.2022.109046.

60. Tang R., Lu T., Zheng W. (2024), Low-frequency ultrawide band gap study of symmetric conical scatterer phononic crystal, *Journal of Vibration Engineering & Technologies*, **12**(3): 3667–3676, https://doi.org/10.1007/s42417-023-01077-2.

61. Wang Z. *et al.* (2025), A broadband zero-frequency seismic metamaterial based on negative effective mass density, *Physics Letters A*, **531**: 130151, https://doi.org/10.1016/j.physleta.2024.130151.

62. Wapler M.C., Leupold J., Dragonu I., von Elverfeld D., Zaitsev M., Wallrabe U. (2014), Magnetic properties of materials for MR engineering, micro-MR and beyond, *Journal of magnetic resonance*, **242**: 233–242, https://doi.org/10.1016/j.jmr.2014.02.005.

63. Wei Y., Pan F., Lin X., Zhang L., Xiang J., Chen Y. (2025), On-demand reprogrammable mechanical metamaterial driven by structure performance relations, *Advanced Materials*, **37**(9): 2410865, https://doi.org/10.1002/adma.202410865.

64. Wen S., Xiong Y., Hao S., Li F., Zhang C. (2020), Enhanced band-gap properties of an acoustic metamaterial beam with periodically variable cross-sections, *International Journal of Mechanical Sciences,* **166**: 105229, https://doi.org/10.1016/j.ijmecsci.2019.105229.

65. Wickeler A.L., Naguib H.E. (2020), Novel origami-inspired metamaterials: design, mechanical testing and finite element modelling, *Materials and Design*, **186**: 108242, https://doi.org/10.1016/j.matdes.2019.108242.

66. Winter L., Seifert F., Zilberti L., Murbach M., Ittermann B. (2021), MRI-related heating of implants and devices: a review, *Journal of Magnetic Resonance Imaging*, **53**(6): 1646–1665, https://doi.org/10.1002/jmri.27194.

67. Wu X., Yan H., Zhou Y., Zhang P., Lu Q., Shi H. (2025), Review of additive manufactured metallic metamaterials: design, fabrication, property and application, *Optics & Laser Technology*, **182**: 112066, https://doi.org/10.1016/j.optlastec.2024.112066.

68. Xiang H., Ma X., Xiang J. (2020), Band gaps and transmission characteristics analysis on a two-dimensional multiple-scatter phononic crystal structure, *Materials*, **13**(9): 2106, https://doi.org/10.3390/ma13092106.

69. Yang H.Y. *et al.* (2023), Low frequency elastic waves and vibration control mechanism of innovative phononic crystal thin plates, *Physica B: Condensed Matter*, **667**: 415189, https://doi.org/10.1016/j.physb.2023.415189.

70. Yin W., Zhu J., Tong Z., Wang L., Li D., Wang L. (2024), 3D printed three-dimensional elastic metamaterial with surface resonant units for low-frequency vibration isolation, *Virtual and Physical Prototyping*, **19**(1): e2382159, https://doi.org/10.1080/17452759.2024.2382159.

71. Zhang W., Zhao Y. (2025), Sound absorption characteristics of the metamaterial with stochastic parameters, *International Journal of Mechanical Sciences*, **287**: 109929, https://doi.org/10.1016/j.ijmecsci.2025.109929.

72. Zhang Y., Kong X., He W., Xu Z. (2025), Eddy current characterization using efficient finite element analysis for z-gradient coils in a low-field MRI scanner, *IEEE Transactions on Instrumentation and Measurement*, **74**: 1–9, https://doi.org/10.1109/TIM.2025.3544743.

73. Zhao H., Zhang E., Lu G. (2024), Study on the equivalent stiffness of a local resonance metamaterial concrete unit cell, *Buildings*, **14**(4): 1035, https://doi.org/10.3390/buildings14041035.

74. Zhou Y., Sang T., Wang Y. (2025), Negative refraction guided by a glide-reflection symmetric crystal interface, *Materials*, **18**(6): 1210, https://doi.org/10.3390/ma18061210.

Archives of Acoustics Vol. 51, No. 3, pp. 477–493 (2026), https://doi.org/10.24423/archacoust.2026.4443

Research Paper

Examination of Low-Frequency Sound Decay in Two Connected Enclosed Spaces Having Different Sound Absorption on Boundary Surfaces

Mirosław Meissner

Institute of Fundamental Technological Research, Polish Academy of Sciences
Warsaw, Poland

e-mail: mmeissn@ippt.pan.pl

Received February 22, 2026; revised April 20, 2026; accepted April 27, 2026; available online April 30, 2026; version of record August 17, 2026; published issue September 23, 2026.

In this study, a sound decay in two connected enclosed spaces was investigated. The research concerned the low-frequency range, because in this situation the sound decay is predominantly non-exponential. Numerical simulations were performed for a system of two connected rectangular subrooms. The larger subroom was highly reverberant, while the smaller subroom had walls that provided moderate sound damping. Simulation results showed that after switching off the harmonic source, a non-exponential behavior of sound decay is usually observed. This is because many eigenmodes are involved in the sound decay, hence this process strongly depends on both the source frequency and its position. Reverberant properties of the system were also determined using energy decay curves calculated by the reverse-time integration of the squared room impulse response. Simulations showed that the effect of the significant difference in sound absorption in subrooms is a wide range of modal reverberation time values, i.e., it is large for eigenmodes localized in a highly reverberant room and small for eigenmodes localized in a damped room. The analyzed room system is therefore characterized by an interesting feature resulting from the localization of eigenmodes. When the source is located in the larger subroom, slowly decaying modes localized in this subroom dominate the sound decay process. However, when the source is located in the smaller subroom, the sound decay in the larger subroom is much faster since non-localized modes with a shorter modal reverberation time are excited within this subroom.

Keywords: coupled spaces, sound decay, sound absorption, room impulse response, decay times, localization of eigenmodes, modal reverberation time.



Copyright © 2026 The Author(s).
This work is licensed under the Creative Commons Attribution 4.0 International CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The phenomenon of sound decay inside enclosed spaces is of fundamental importance in room acoustics, as predicted decay times are the main quantities utilized to assess the room acoustic quality. Accurate determination of decay times is essential when evaluating the acoustics of concert halls (Ermann, Johnson, 2005; Bradley, Wang, 2007; Luizard *et al.*, 2015), irregularly shaped spaces such as churches (Martellotta, 2016; Girón *et al.*, 2017; Sü Gül, 2021), and architectural structures in the form of connected enclosed spaces, also referred to as coupled spaces (Jasa, Xiang, 2009; Luizard, Katz, 2014; Das, Abel, 2021). A special feature of coupled spaces is the possibility of achieving non-exponential sound decays (Bradley, Wang, 2009; Xiang *et al.*, 2011). This attribute is exploit in concert halls to obtain the double slope sound decay, characterized by a fast initial and a slow late decay, which listeners perceive as a beneficial balance between the reverberance and the sound clarity (Bradley, Wang, 2010).

A number of theoretical approaches and numerical implementations have been developed to investigate acoustical properties of coupled spaces. They utilize statistical-acoustic models, the diffusion-equation model, geometric

methods, the improved Fourier series method and the modal analysis also known as the modal expansion method. Statistical-acoustic models (Summers *et al.*, 2004; Summers, 2012) are based on the diffuse sound field assumption, according to which sound energy is homogeneous in acoustic field and moves in all directions with the same probability. These hypotheses are roughly correct for sound frequencies above the Schroeder frequency (Schroeder, 1996). The diffusion-equation model (Xiang *et al.*, 2009; Luizard *et al.*, 2014; Sü Gül *et al.*, 2019) is an extension of the statistical theory to spatially changing reverberant sound fields and uses the analogy of the sound energy density with a density of 'sound particles' travelling along straight lines. Geometrical methods are applicable at high sound frequencies, and most of the existing geometrical room acoustic modeling techniques are based on ray-tracing and beam tracking algorithms (Nijs *et al.*, 2002; Summers *et al.*, 2005). However, these methods do not allow all relevant propagation paths involving edge diffraction to be found. Therefore, the calculation results obtained by these methods may be inaccurate in the case of coupled spaces and irregularly shaped enclosures. The improved Fourier series method is used in studies of structural vibrations and vibro-acoustic couplings, and has been successfully applied to the study of acoustics of two coupled spaces with a coupling opening of variable size (Shi *et al.*, 2018).

Wave phenomena in the sound field, such as diffraction and wave interferences, are fully described by modal analysis, but this method is more difficult to apply to enclosed spaces with irregular shapes. In this method, the response of an enclosure to sound excitation can be described by the modal expansion of the sound pressure and the decay coefficient associated with each acoustic mode (Meissner, 2010; Dance, Van Buuren, 2013). Modal analysis has proven to be particularly useful for systems of coupled spaces, as it enables the identification of phenomena such as the localization of eigenmodes (Meissner, 2009), the formation of acoustic energy vortices in the active sound intensity vector field (Meissner, 2012), and the occurrence of large dips in the spatial distribution of the sound pressure level under steady-state conditions (Meissner, Wiśniewski, 2019).

The objective of this study is to examine the sound decay process in the system of two connected enclosed spaces with different sound absorption on boundary surfaces. The research focuses on the low-frequency range, since in this case a non-exponential nature of sound decay is usually observed. The theoretical model uses the modal expansion method, in which the reverberant sound pressure is decomposed into a series of eigenmodes with specific eigenfrequencies. The study takes into account two types of system responses. In the first case, a steady-state sound field is produced by a harmonic point source, and the sound decay process begins immediately after the source is switched off. In the second case, the system response is initiated by a point source emitting the Dirac delta time impulse, and decay times are determined from a slope of energy decay curve calculated by the reverse-time integration of the squared room impulse response. Numerical simulations were performed for a system of two connected rectangular rooms, where the larger room was highly reverberant and walls of the smaller room provided moderate sound damping.

2. Theoretical analysis

2.1. Problem definition

The object of theoretical analysis is a system of two enclosed spaces Ω_1 and Ω_2 connected by a transparent opening (Fig. 1). The spaces are filled with air, hence the product ρc represents the impedance of air, where ρ is the

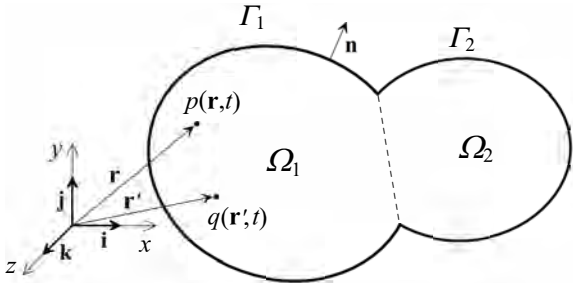


Fig. 1. Analyzed system consisting of two enclosed spaces Ω_1 and Ω_2 connected by a transparent opening denoted by dashed line.

air density and c is the sound speed. The sound pressure $p(\mathbf{r},t)$ in the system is generated by the volume sound source $q(\mathbf{r}',t)$, where $\mathbf{r}=(x,y,z)$ is a position of a field point and $\mathbf{r}'=(x',y',z')$ is a position of a source point.

In general case, the time dependence of source function $q(\mathbf{r}',t)$ can be arbitrary, therefore the presented theory allows to determine the interior sound field in both the steady-state and reverberant conditions. A boundary surface of the system is marked by Γ and it is divided in two parts Γ_1 and Γ_2 representing boundary surfaces of the spaces Ω_1 and Ω_2. It is assumed that the surfaces Γ_1 and Γ_2 are characterized by different sound absorption, therefore the sound pressure on these surfaces satisfies the following boundary conditions:

$$\mathbf{r}\in\Gamma_i: \quad \nabla p\cdot\mathbf{n}=-\frac{1}{cZ_i}\frac{\partial p}{\partial t}, \qquad i=1,2, \tag{1}$$

where ∇ is the gradient operator, the dot is a scalar product, $\mathbf{n}$ is the outward normal vector, and $Z_i=|Z_i|e^{\mathrm{j}\phi_i}$ is the impedance on the surface Γ_i normalized by the impedance of air. Due to random directions of incident sound waves, damping properties of the surface Γ_i are determined by the random-incident absorption coefficient α_i given by (Paris, 1928):

$$\alpha_i=\frac{8\cos(\phi_i)}{|Z_i|^2}\left\{|Z_i|+\frac{\cos(2\phi_i)}{\sin(\phi_i)}\arctan\left[\frac{|Z_i|\sin(\phi_i)}{1+|Z_i|\cos(\phi_i)}\right]-\ln\left[1+2|Z_i|\cos(\phi_i)+|Z_i|^2\right]\cos(\phi_i)\right\}. \tag{2}$$

Since it was assumed that the system is characterized by different sound absorption on the boundary surfaces Γ_1 and Γ_2, from theoretical and practical points of view, the case when the absorption coefficients α_1 and α_2 differ significantly deserves careful analysis. The research focuses on the low-frequency range, since in this case the sound decay strongly depends on the excitation frequency. At low frequencies, typical absorptive materials provide a small sound damping, so the system under consideration can be treated as lightly damped.

2.2. Low-frequency sound decay

At low frequencies, a physically correct description and interpretation of acoustics in enclosed spaces requires the use of wave-based methods. This is particularly necessary when simulating sound decay in enclosures with irregular shapes and systems of coupled rooms, because in such cases the theoretical model should take sound diffraction into account. In room acoustics, the most convenient way to describe a low-frequency sound field is to use the room impulse response (RIR). This is because a knowledge of the RIR function $h(\mathbf{r}',\mathbf{r},t)$, which describes the pressure response at the receiving point $\mathbf{r}=(x,y,z)$ to the time impulse generated at the source point $\mathbf{r}'=(x',y',z')$, makes it possible to determine the response of a room to any sound excitation. When the spatial and temporal properties of the source function $q(\mathbf{r}',t)$ are known, the pressure response to this excitation can be determined using the convolution integral in the time domain (Damelin, Miller, 2012):

$$p(\mathbf{r},t)=\int_V q(\mathbf{r}',t)*h(\mathbf{r}',\mathbf{r},t)\,\mathrm{d}^3\mathbf{r}'=\int_{-\infty}^{t}\int_V q(\mathbf{r}',\tau)h(\mathbf{r}',\mathbf{r},t-\tau)\mathrm{d}^3\mathbf{r}'\,\mathrm{d}\tau, \tag{3}$$

where V is the volume of a room system, the asterisk denotes a convolution operation, and $\mathrm{d}^3\mathbf{r}'=\mathrm{d}x'\,\mathrm{d}y'\,\mathrm{d}z'$ is the volume element. A method for finding the RIR function was presented by Meissner and Wiśniewski (2020) and the result obtained is as follows:

$$h(\mathbf{r}',\mathbf{r},t)=c^2\sum_{m=1}^{\infty}\frac{e^{-\xi_m t}\sin(\psi_m t)\Phi_m(\mathbf{r}')\Phi_m(\mathbf{r})}{\psi_m}, \tag{4}$$

where m is the mode number, Φ_m are the eigenfunctions, which fulfill the orthonormal property in the volume V, the damping coefficients ξ_m are computed from the equation:

$$\xi_m=r_m+\mathrm{j}\varphi_m=\frac{c}{2}\left[\int_{\Gamma_1}\frac{\Phi_m^2(\mathbf{r}_s)\,\mathrm{d}\Gamma}{Z_1}+\int_{\Gamma_2}\frac{\Phi_m^2(\mathbf{r}_s)\,\mathrm{d}\Gamma}{Z_2}\right], \tag{5}$$

where $\mathbf{r}_s$ is the coordinate of the position on boundary surfaces Γ_1 and Γ_2, $\mathrm{j} = (-1)^{1/2}$ is the imaginary unit, ψ_m are the complex frequencies determined from the formula:

$$\psi_m = \Omega_m + \mathrm{j}\vartheta_m = \sqrt{\frac{a_m + \sqrt{a_m^2 + b_m^2}}{2}} + \mathrm{j}\sqrt{\frac{-a_m + \sqrt{a_m^2 + b_m^2}}{2}}, \tag{6}$$

where $a_m = \omega_m^2 - r_m^2 + \varphi_m^2$, $b_m = -2r_m\varphi_m$, and ω_m are the eigenfrequencies, which are related to the corresponding eigenfunctions Φ_m through the equations:

$$\Delta\Phi_m + \left(\frac{\omega_m}{c}\right)^2 \Phi_m = 0, \tag{7}$$

where $\Delta = \nabla \cdot \nabla$ is the Laplace operator.

Two methods are mainly used to describe the process of sound decay inside room systems. In the first method, the sound decay process is analyzed in the case of switching off the harmonic point source that created the steady-state sound field inside a room system. Therefore, in this case, the source function in Eq. (3) assumes the form:

$$q(\mathbf{r}', \tau) = Q\delta(\mathbf{r}' - \mathbf{r}_0)e^{\mathrm{j}\omega\tau}, \tag{8}$$

where $Q = (8\pi\rho cW)^{1/2}$, W is the source power (Kinsler *et al.*, 2000), ω is the source angular frequency, and $\mathbf{r}_0 = (x_0, y_0, z_0)$ determines the source position. From a theoretical point of view, the steady-state sound field is achieved when the time interval between the present time τ and the moment of activation of the source is infinitely long. Therefore, in Eq. (3), it should be assumed that the source was turned on at the moment $\tau = -\infty$. Furthermore, to simplify the analysis, it is assumed that the source is stopped at the moment $\tau = 0$, hence after substituting Eq. (8) into Eq. (3) and performing integration we obtain the formula:

$$p(\mathbf{r}, t) = c^2 Q \sum_{m=1}^{\infty} \frac{e^{-\xi_m t}\Phi_m(\mathbf{r}_0)\Phi_m(\mathbf{r})}{\omega_m^2 - \omega^2 + 2\mathrm{j}\omega\xi_m}\left[\cos(\psi_m t) + \frac{(\xi_m + \mathrm{j}\omega)\sin(\psi_m t)}{\psi_m}\right] \tag{9}$$

describing sound decay after switching off a harmonic point source. A quantity often used to evaluate this decay is the modal reverberation time T_m, defined as the time required for the pressure level to decay by 60 dB in the m-th mode, so

$$T_m = \frac{3\ln(10)}{r_m}, \tag{10}$$

where the damping coefficient r_m, which is the real part of ξ_m, can be calculated from Eq. (5). Using Eq. (9), the spatial distribution of the pressure amplitude can be determined for different times t, which allows us to predict changes in this distribution as time increases. Furthermore, at a given position $\mathbf{r}$ of the field, Eq. (9) allows one to calculate the changes in pressure amplitude over time in order to illustrate the nature of sound decay and, if possible, to estimate decay times.

The most commonly used method for analyzing sound decay in enclosed spaces is the method based on the reverse-time integration of the squared RIR (Schroeder, 1996). This method produces a smoothed decay profile, commonly known as the energy decay curve (EDC), which is used to estimate decay times. Assuming that a time impulse is emitted by a source located at the point $\mathbf{r}_0$ and transforming the right-hand side of Eq. (4), we obtain:

$$h(\mathbf{r}_0, \mathbf{r}, t) = \sum_{m=1}^{\infty} [\alpha_m(t) + \mathrm{j}\beta_m(t)]\Phi_m(\mathbf{r}_0)\Phi_m(\mathbf{r}), \tag{11}$$

where $\alpha_m(t)$ and $\beta_m(t)$ can be determined from the equations:

$$\begin{aligned}\alpha_m(t) = {} & \frac{c^2 e^{-r_m t}}{\Omega_m^2 + \vartheta_m^2}\Big\{[\Omega_m \sin(\varphi_m t) + \vartheta_m \cos(\varphi_m t)]\cos(\Omega_m t)\sinh(\vartheta_m t) \\ & + [\Omega_m \cos(\varphi_m t) - \vartheta_m \sin(\varphi_m t)]\sin(\Omega_m t)\cosh(\vartheta_m t)\Big\},\end{aligned} \tag{12}$$

$$\beta_m(t) = \frac{c^2 e^{-r_m t}}{\Omega_m^2 + \vartheta_m^2}\Big\{[\Omega_m \cos(\varphi_m t) - \vartheta_m \sin(\varphi_m t)]\cos(\Omega_m t)\sinh(\vartheta_m t) - [\Omega_m \sin(\varphi_m t) + \vartheta_m \cos(\varphi_m t)]\sin(\Omega_m t)\cosh(\vartheta_m t)\Big\}. \tag{13}$$

As can be seen from Eq. (11), the RIR function takes complex values, so the quantity corresponding to the squared RIR is the square of the modulus of $h(\mathbf{r}_0, \mathbf{r}, t)$:

$$|h(\mathbf{r}_0, \mathbf{r}, t)|^2 = \left[\sum_{m=1}^{\infty} \alpha_m(t)\Phi_m(\mathbf{r}_0)\Phi_m(\mathbf{r})\right]^2 + \left[\sum_{m=1}^{\infty} \beta_m(t)\Phi_m(\mathbf{r}_0)\Phi_m(\mathbf{r})\right]^2, \tag{14}$$

thus, EDC can be calculated using the reverse-time integration:

$$\mathrm{EDC}(t) = \int_t^{\infty} |h(\mathbf{r}_0, \mathbf{r}, \tau)|^2 \,\mathrm{d}\tau. \tag{15}$$

A decibel version of EDC, here denoted by L_{EDC}, can be determined from the formula:

$$L_{\mathrm{EDC}}(t) = 10\log\left[\frac{\mathrm{EDC}(t)}{\mathrm{EDC}(0)}\right], \tag{16}$$

and based on the slope of L_{EDC} at different stages of the sound decay process, the corresponding decay times can be predicted.

3. Numerical study

3.1. System of coupled spaces under consideration

A process of sound decay was numerically tested for the system of two connected rectangular subrooms of the same height. The choice of the research object was dictated by practical considerations, as buildings and public facilities often contain spaces in the form of two-room coupled systems. This system together with the associated coordinate system is depicted in Fig. 2. Therefore, the wall areas of subrooms A and B are as follows:

$$S_1 = 2(d_1 + h)l_1 + (d_1 + 2d_2)h, \tag{17}$$

$$S_2 = 2(d_3 + h)l_2 + d_3 h. \tag{18}$$

In the numerical study, the following dimensions of the room system were assumed: $l_1 = 15\,\mathrm{m}$, $l_2 = 5\,\mathrm{m}$, $d_1 = 10\,\mathrm{m}$, $d_2 = 2\,\mathrm{m}$, $d_3 = 6\,\mathrm{m}$, and $h = 4\,\mathrm{m}$.

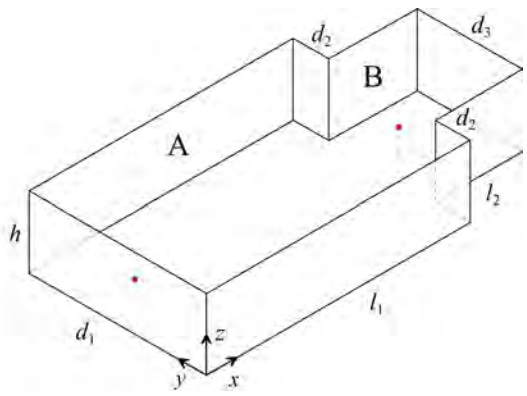


Fig. 2. Room system under investigation containing two connected rectangular rooms marked with letters A and B. Red dots indicate locations of the sound source.

As shown by the theoretical analysis, the RIR function (Eq. (4)) and the reverberant sound field (Eq. (9)) are dependent on the sound source position. Therefore, in computer simulations it was assumed that the source is placed either at the point $\mathbf{r}_0 = (2\,\mathrm{m}, 4\,\mathrm{m}, 1.6\,\mathrm{m})$ or at the point $\mathbf{r}_0 = (17\,\mathrm{m}, 4\,\mathrm{m}, 1.6\,\mathrm{m})$ located inside subrooms A and B, respectively (Fig. 2). The reverberant sound field distribution and decay times were determined on a horizontal plane located 1.2 m above the floor.

Since a characteristic feature of the analyzed room system is a different sound absorption on boundary surfaces, it was assumed that subroom A has almost acoustically hard walls, whose absorption properties can be described by the real-valued surface impedance (Thydal *et al.*, 2021). Therefore, the surface impedance Z_1 was assumed to have the value of 250, and using Eq. (2), it can be calculated that this value corresponds the absorption coefficient α_1 of 0.031. On the other hand, it was assumed that subroom B is moderately damped, because the surface of its walls has the impedance $Z_2 = 10e^{\mathrm{j}\pi/6}$, for which the absorption coefficient α_2 is 0.44. The absorption coefficient characterizing a sound damping in the whole room system is defined as (Vorländer, 2020):

$$\overline{\alpha} = \frac{\alpha_1 S_1 + \alpha_2 S_2}{S_1 + S_2}, \tag{19}$$

so, using Eq. (17) to Eq. (19), it can be found that $\overline{\alpha} = 0.12$, which is a value low enough to treat the considered room system as lightly damped.

According to the wave theory, in the low-frequency range enclosed spaces may be treated as acoustic resonators with characteristic modes determined by the eigenfunctions Φ_m and eigenfrequencies ω_m. Since the analyzed room system has a complex geometry, it is not possible to determine the functions Φ_m analytically, so numerical methods must be used. However, the fact, that the system under consideration is lightly damped allows the functions Φ_m to be approximated by the eigenfunctions determined for the rigid walls. Thus, in this case Φ_m can be described using the functions in the form (Kuttruff, 2017):

$$\Phi_m(\mathbf{r}) = \Psi_\kappa(x, y)\Theta_\mu \cos\left(\frac{\pi\mu z}{h}\right), \tag{20}$$

where the modal numbers κ, $\mu = 0, 1, 2, \ldots$ are not simultaneously equal to zero, the eigenfunctions Ψ_κ satisfy the orthonormal property over the surface S of a horizontal cross section of the system, $\Theta_\mu = 1/\sqrt{h}$ for $\mu = 0$ and $\Theta_\mu = \sqrt{2/h}$ for $\mu > 0$. Finally, the formula for calculating the frequency of the m-th mode is given by

$$f_m = \sqrt{f_\kappa^2 + \left(\frac{\mu c}{2h}\right)^2}, \tag{21}$$

where the eigenfrequencies f_κ are associated with the functions Ψ_κ, and since $\Psi_0 = 1/\sqrt{S}$, the eigenfrequency f_κ for $\kappa = 0$ is equal to zero.

Using the finite element method, the eigenfunctions Ψ_κ were calculated for the first 5450 m-indexed eigenmodes with the frequencies f_m up to 400 Hz. Examples of contour maps of the eigenfunctions Ψ_κ with corresponding modal frequencies f_κ are shown in Fig. 3. These results indicate that for some modes the acoustic energy may be concentrated within one of the subrooms (Fig. 3a, Fig. 3c, Fig. 3f, Fig. 3h). This phenomenon is called the localization of eigenmodes and occurs in enclosed spaces of irregular shape. It should be emphasized that among modes calculated for the analyzed room system, there are also modes which are totally delocalized. This is because they represent modes of a cuboidal space with height h and the width d_1, like the system under study, and the length being the sum of l_1 and l_2 (Fig. 2). The contour map of the eigenfunction Ψ_κ for such a mode is shown in Fig. 3d.

3.2. Analysis of sound decay after turning off harmonic source

As shown in Subsec. 2.2, the theoretical model allows for determining the spatial distribution of the pressure amplitude after the harmonic source is turned off. In this case, as indicated by Eq. (9), the sound pressure $p(\mathbf{r}, t)$ is the sum of decaying modal oscillations. From this it follows that the amplitude of this pressure depends on

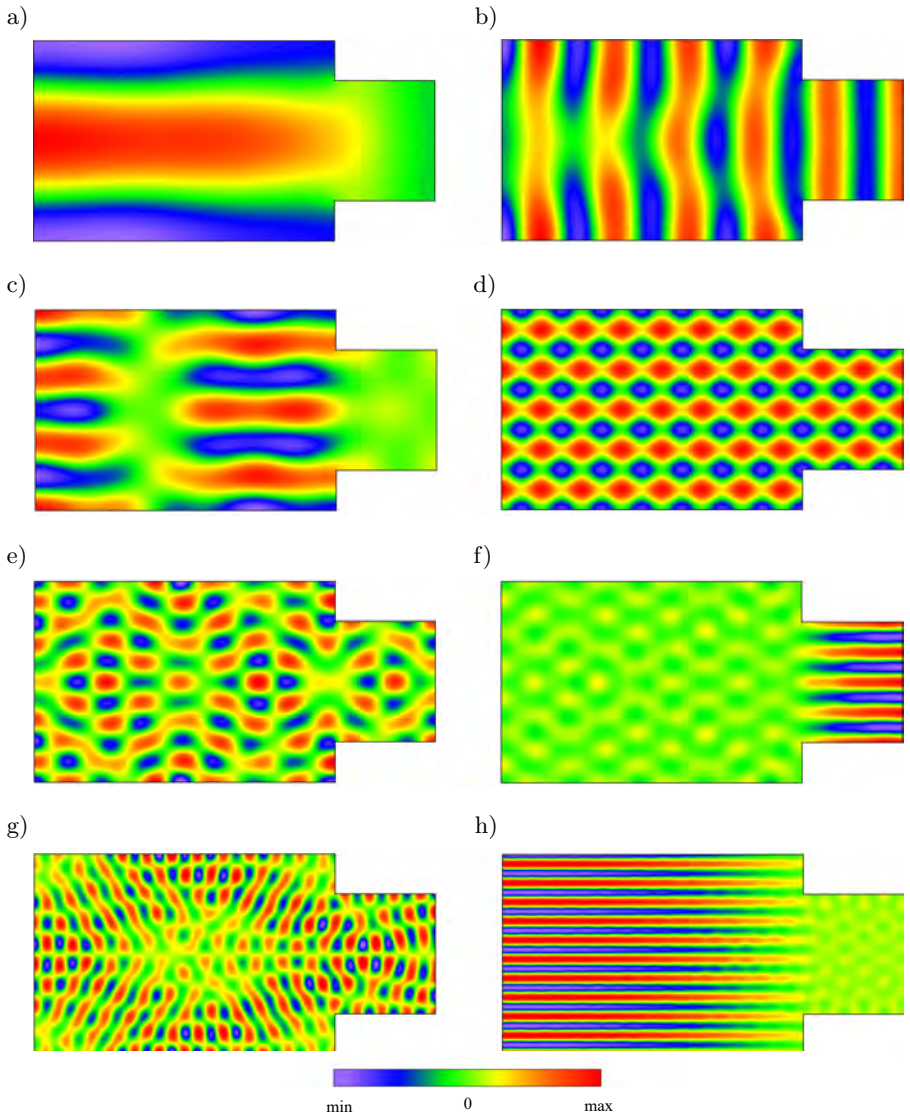


FIG. 3. Contour maps of eigenfunctions Ψ_κ for the modal number κ: a) 9, b) 50, c) 60, d) 154, e) 200, f) 271, g) 500, h) 652. Frequencies f_κ for these functions are (in Hz): a) 34.5, b) 93.4, c) 104, d) 171.5, e) 195.1, f) 229.2, g) 313.9, h) 360.2.

time, therefore, to determine this amplitude, the discrete Hilbert transform H_d was used, in which the pressure signal was represented by $p_r(\mathbf{r}, n)$, $n = 0, 1, ..., N-1$, i.e., the digitally sampled real part of $p(\mathbf{r}, t)$. Thus, the formula for calculating the pressure amplitude is as follows:

$$P(\mathbf{r}, n) = \sqrt{p_r^2(\mathbf{r}, n) + H_d^2\{p_r(\mathbf{r}, n)\}}, \tag{22}$$

where the discrete Hilbert transform of $p_r(\mathbf{r}, n)$ is determined by (KAK, 1970):

$$H_d\{p_r(\mathbf{r}, n)\} = \begin{cases} \dfrac{2}{\pi} \displaystyle\sum_{m=\text{odd}}^{N-1} \dfrac{p_r(\mathbf{r}, n)}{n-m}, & n \text{ even}, \\ \dfrac{2}{\pi} \displaystyle\sum_{m=\text{even}}^{N-1} \dfrac{p_r(\mathbf{r}, n)}{n-m}, & n \text{ odd}. \end{cases} \tag{23}$$

In the numerical study, the distributions of pressure amplitude P on the observation plane were computed for selected source frequencies and different times t from the moment of switching off the source ($t = 0$). Furthermore, in order to illustrate the character of sound decay, the changes in sound pressure p_r and pressure level L_w over

time at the receiving point $\mathbf{r} = (8\,\mathrm{m}, 7\,\mathrm{m}, 1.2\,\mathrm{m})$ located on the observation plane were calculated. The formula used to determine the pressure level L_w was as follows:

$$L_w = 20\log\left(\frac{P}{P_{\max}}\right), \tag{24}$$

where $P_{\max}$ is a maximum of P recorded for the time t greater than or equal to zero. Colored contour maps in Fig. 4 illustrate distributions of P on the observation plane in the case when the harmonic source is located in subroom A. They were obtained for frequencies of 104 Hz and 257 Hz and times t of 0 s, 2 s, and 4 s. These data show that after turning off the source, it takes as long as 4 s for the pressure amplitude level to drop by approximately 36 dB for the first frequency and by 47 dB for the second frequency. This nature of the sound decay is obviously due to the fact that subroom A has walls that are almost acoustically hard.

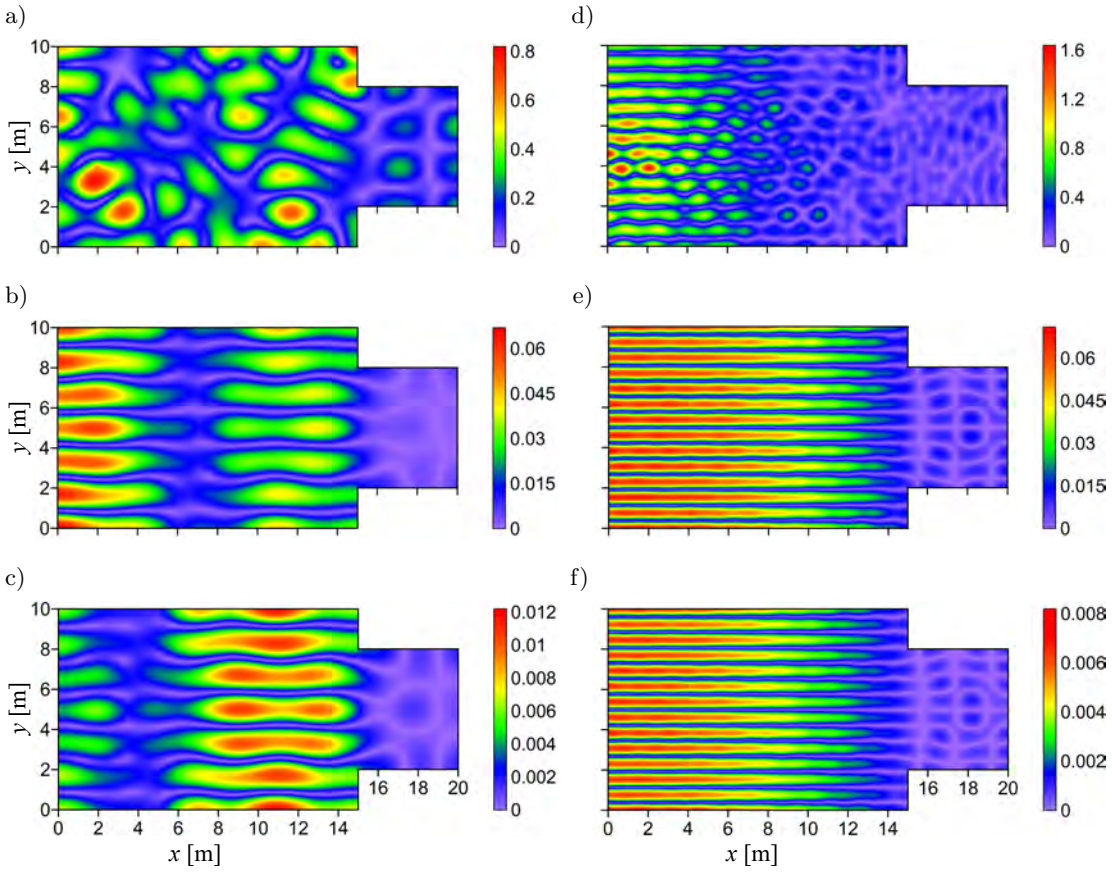


FIG. 4. Changes in the pressure amplitude P (in pascals) on observation plane for excitation frequency: a)–c) 104 Hz, d)–f) 257 Hz, and the time t: a), d) 0 s, b), e) 2 s, c), f) 4 s. Sound source located in subroom A.

As may be seen from Fig. 4, the distributions of P on the observation plane may change as time t increases (Fig. 4a to Fig. 4c), or after a certain moment, a further increase in t does not affect these distributions (Fig. 4d to Fig. 4f). The first situation occurs when strongly excited eigenmodes have eigenfrequencies very close to the source frequency, while the second one occurs when, after some time, the sound decay becomes almost perfectly exponential. The explanation of these finding can be found in Fig. 5, which depicts variations in the sound pressure p_r and the pressure level L_w at the receiving point for the frequencies considered. As can be seen in Fig. 5a and Fig. 5b, for an excitation frequency of 104 Hz, the acoustic signal recorded at the receiving point has a form of decaying beating oscillations, and these beats are so large that the pressure level L_w reaches its maximum value after 0.6 s from the moment the source is turned off. A detailed analysis of the numerical data allowed us to conclude that two eigenmodes with the modal number m of 130 and 134, having frequencies of 103.03 Hz and 103.99 Hz, are responsible for the strong beating effect. The shape of the eigenfunction Ψ_κ for the

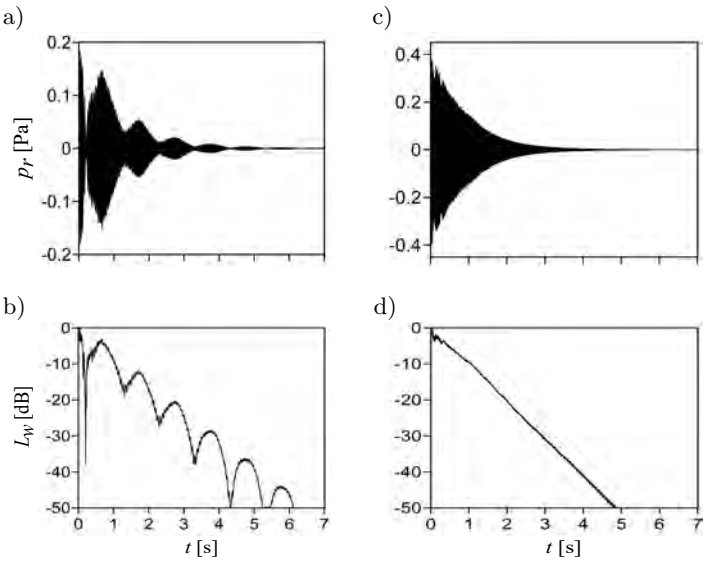


FIG. 5. Variations in of the sound pressure p_r and the pressure level L_w at receiving point for excitation frequency: a), b) 104 Hz, c), d) 257 Hz. Sound source located in subroom A.

134th eigenmode is shown in Fig. 3c. Since these eigenmodes are localized in subroom A, the modal reverberation times determined for these modes from Eq. (10) have large values, i.e., 6.19 s for the 130th eigenmode and 5.14 s for 134th eigenmode.

In the case of the excitation frequency of 257 Hz, the nature of the sound decay is different, because after a certain time has elapsed since the sound source was turned off, the acoustic response of the room system is dominated by one eigenmode. As a consequence, beyond this time, the observed temporal changes in the pressure level L_w for this eigenmode are almost perfectly linear (Fig. 5d). As it results from Fig. 4e and Fig. 4f, this eigenmode is localized in subroom A and based on the analysis of numerical data it was recognized as the 1544th eigenmode. The modal reverberation time for this mode is 5.87 s and this result agrees very well with a decay time of 5.85 s predicted from changes in the pressure level L_w for time above 2 s.

In the next part of the numerical study, the influence of changing the sound source position on the sound decay process was investigated. The calculation results presented in Fig. 6 show changes in the pressure amplitude P on the observation plane simulated for the same excitation frequencies as before and the times t of 0 s, 1 s, and 2 s in the case when the sound source is located in subroom B. From the comparison of Fig. 4 and Fig. 6 it follows that there is a qualitative similarity of pressure distributions some time after the sound source is turned off, caused by the influence of slowly decaying modes localized in subroom A on the sound decay. However, this comparison also indicates significant differences in the sound decay process, because the data in Fig. 6 show that for a frequency of 104 Hz the sound level drops by as much as 47 dB just 2 s after the source is turned off, and by 43 dB for a frequency of 257 Hz. This means that moving the source from subroom A to the subroom B results in a faster initial decay and a slower late decay. This conclusion is confirmed by the calculation results shown in Fig. 7, illustrating variations in the sound pressure p_r and the pressure level L_w at the receiving point for the frequencies considered in the case when the sound source is located in subroom B.

The analysis clearly shows that in the considered system of coupled spaces, the sound decay process after turning off harmonic source is extremely complex. This results not only from the fact that this process is strongly dependent on frequency, but, as shown in Eq. (9) and the obtained calculation results, its nature is also significantly influenced by the location of the sound source. Another phenomenon that has a major impact on sound decay is the localization of eigenmodes, and when the eigenmode localized in a subroom with very low sound damping dominates this process, a long-term exponential sound decay can be observed. Apart from this special case, the decay of the sound behaves very irregularly due to the excitation of eigenmodes with very similar modal frequencies, resulting in a beating effect, or the excitation of eigenmodes characterized by different modal decay times.

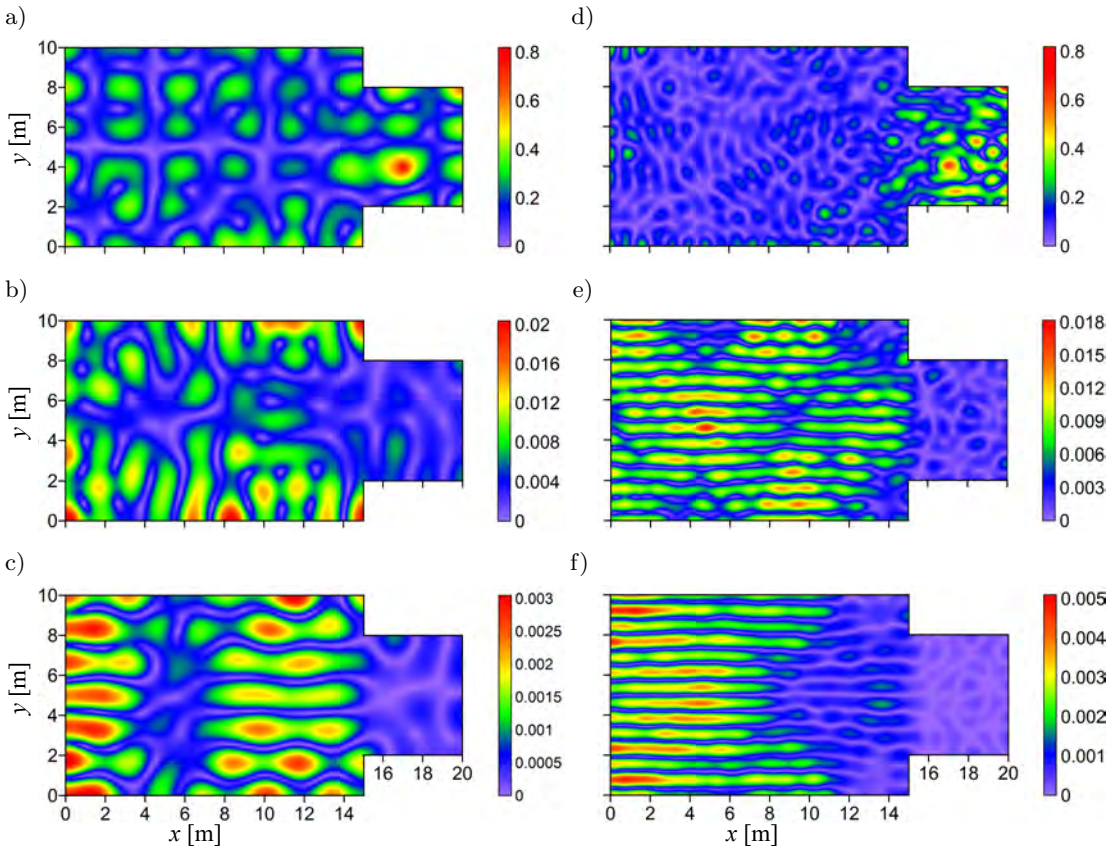


FIG. 6. Changes in the pressure amplitude P (in pascals) on observation plane for excitation frequency: a)–c) 104 Hz, d)–f) 257 Hz, and the time t: a), d) 0 s, b), e) 1 s, c), f) 2 s. Sound source located in subroom B.

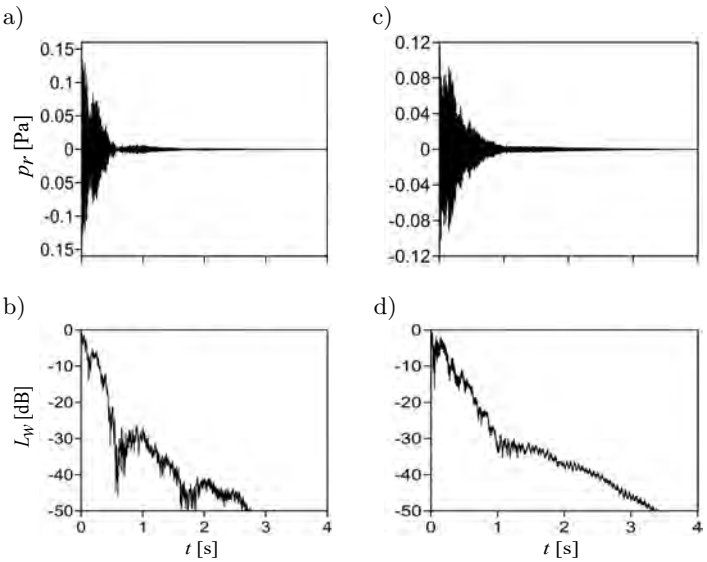


FIG. 7. Variations in the sound pressure p_r and the pressure level L_w at receiving point for excitation frequency: a), b) 104 Hz, c), d) 257 Hz. Sound source located in subroom B.

3.3. Results of decay time prediction from variations of L_{EDC}

As stated in Subsec. 2.2, the method based on the reverse-time integration of the squared RIR is a procedure often employed to analyze the sound decay in enclosed spaces. Using Eq. (16), L_{EDC}, the decibel version of EDC, was determined and the application of linear regression to relevant parts of L_{EDC} enabled the prediction of the early decay time (EDT), the decay times T_{10}, T_{20}, T_{30}, and the late decay time (LDT). EDT is the decay time

evaluated from a decrease in L_{EDC} from 0 dB to −10 dB, multiplied by a factor of 6. T_{10} is defined as the decay time estimated from a drop of L_{EDC} from −5 dB to −15 dB, multiplied by a factor of 6. T_{20} is defined as the decay time estimated from a drop of L_{EDC} from −5 dB to −25 dB, multiplied by a factor of 3. Further, T_{30} is the decay time predicted from a decrease in L_{EDC} from −5 to −35, multiplied by a factor of 2. Finally, LDT is the decay time with the limits of −25 dB and −35 dB in L_{EDC}, multiplied by a factor of 6.

Figure 8 shows shapes of L_{EDC} determined at the previously assumed receiving point for the sound source located in subroom A (blue line) and subroom B (red line). The first observation is that for the source located in subroom A, changes in L_{EDC} are almost linear when L_{EDC} drops more than 15 dB, and the L_{EDC} curve has a clearly concave shape when the source is located in subroom B. The second observation is more important. Namely, the decay curves in Fig. 8 prove that moving the source from subroom A to subroom B causes significant reduction of decay times. This conclusion is supported by the data in Table 1, which presents the results of predicting the decay times from the decay curves in Fig. 8.

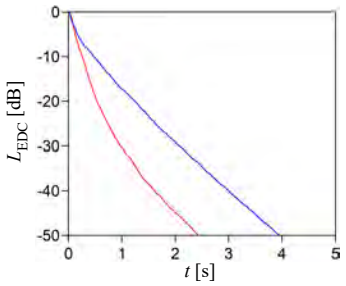


FIG. 8. Shapes of L_{EDC} at receiving point for source located in subroom A (blue line) and subroom B (red line).

TABLE 1. Decay times EDT, T_{10}, T_{20}, T_{30}, and LDT determined from decay curves in Fig. 8.

Source location	Decay times [s]				
	EDT	T_{10}	T_{20}	T_{30}	LDT
Subroom A	2.97	4.11	4.53	4.87	5.59
Subroom B	1.52	1.71	1.93	2.42	3.37

The data in Table 1 indicate that locating the source in subroom B causes an approximately twofold reduction of EDT and T_{30}, and also a more than twofold decrease in the decay times T_{10} and T_{20}. It also follows from Fig. 8 that the concave nature of the decay curve is much greater for the source located in subroom B, and furthermore in this case there is no typical double-sloped energy decay because this decay is multi-sloped. The different shapes of the decay curves for the source located in subrooms A and B result from the influence of two factors on the sound decay process. The first of them is a different sound absorption on boundary surfaces because subroom A is highly reverberant, while walls of subroom B provide moderate sound damping. The second is the phenomenon of localization of eigenmodes, which in the considered system of coupled spaces manifests itself by the concentration of eigenmode energy in subroom A or subroom B (Fig. 3a, Fig. 3c, Fig. 3f, Fig. 3h).

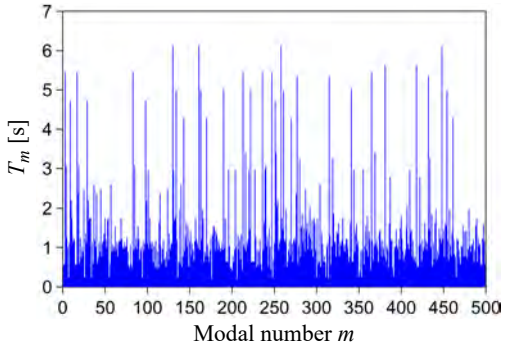


FIG. 9. Modal reverberation time T_m for first 500 eigenmodes.

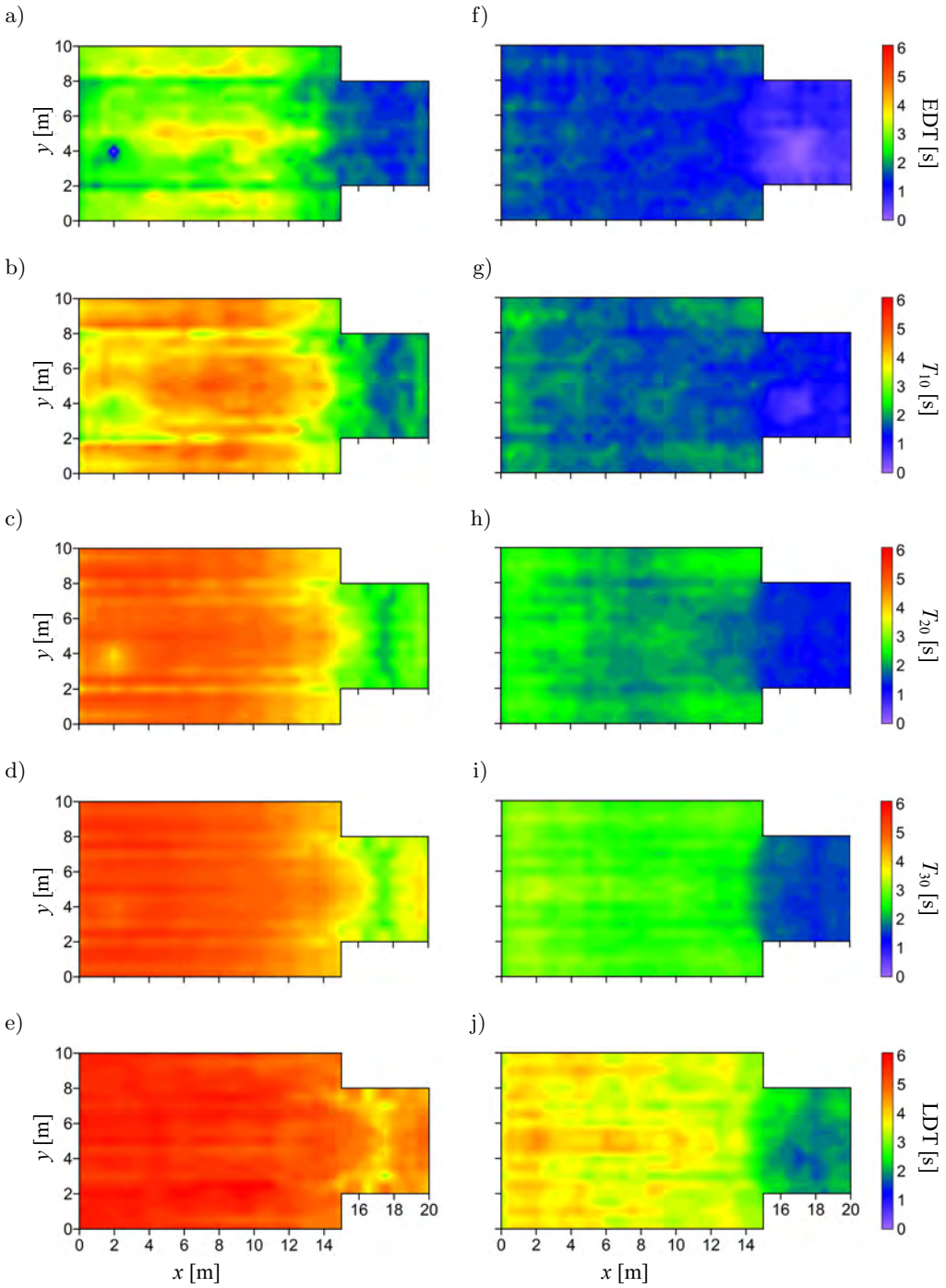


FIG. 10. Distribution of decay times EDT, T_{10}, T_{20}, T_{30}, and LDT on observation plane for source located in: a)–e) subroom A and f)–j) subroom B.

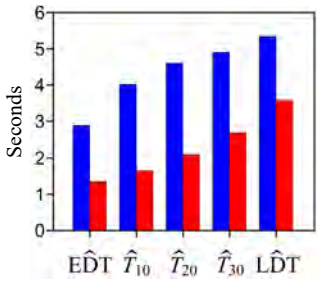


FIG. 11. Mean decay times $\widehat{\text{EDT}}$, $\widehat{T}_{10}$, $\widehat{T}_{20}$, $\widehat{T}_{30}$, $\widehat{\text{LDT}}$ on observation plane in subroom A for source located in subrooms A (blue bars) and subroom B (red bars).

The localization of eigenmodes is a consequence of the irregular distribution of the eigenfunctions $\Phi_m(\mathbf{r})$ in (x, y) plane, hence, according to Eq. (5) and Eq. (10), the parameter that unambiguously characterizes this phenomenon is the modal reverberation time T_m. In numerical study, 5450 eigenmodes were used, therefore, to clearly show the changes in T_m, in Fig. 9 the modal number m was limited to 500. The calculation results presented in Fig. 9 are particularly interesting because they show how a different sound absorption in subrooms can affect the sound decay process in the considered two-room coupled system. It turns out that the significant difference in sound absorption in subrooms results in a wide range of modal decay time values. Indeed, the data in Fig. 9 show that for modes strongly localized in subroom A, the modal reverberation time T_m is even above 6 s, while for modes strongly localized in subroom B the value of T_m is even below 0.3 s.

Based on the foregoing analysis, the results in Fig. 8 can be explained as follows. If the sound source is located in subroom A, the response of the room system to this excitation is dominated by slowly decaying modes that are localized in this subroom. Therefore, decay times increase from about 3 s in the early decay stage to as much as 5.59 s in the late decay stage, as may be seen from the data in Table 1. In contrast, when the source is located in subroom B, non-localized modes with shorter modal reverberation time are excited in subroom A in the early stage of sound decay, so the EDT is 1.57 s, which is approximately half that for the source located in subroom A. The influence of modes localized in subroom A is, however, visible in the late decay stage, because the LDT reaches a value of 3.37 s.

An important conclusion arising from the analysis is that, in the considered room system, the nature of sound decay at the receiving point located in subroom A can be significantly modified by changing the position of the source. To check whether this also holds true for other points, the distributions of all considered decay times in the observation plane were calculated. The obtained results in the form of colored contour maps are shown in Fig. 10. They confirm and simultaneously generalize the previous findings, as it follows from Fig. 10 that placing the source in subroom B causes a significant shortening of all decay times in the observation plane.

It can be seen from Fig. 10 that the decay times predicted in subroom A vary moderately. This fact allows to characterize the sound decay process in subroom A by the mean values of the decay times on the observation plane. Since calculations were carried out in a finite number of observation points, the mean decay times were determined using an arithmetic averaging. The results of calculating these quantities, labelled $\widehat{\mathrm{EDT}}$, $\widehat{T}_{10}$, $\widehat{T}_{20}$, $\widehat{T}_{30}$, and $\widehat{\mathrm{LDT}}$, are collated in Fig. 11 and illustrated by vertical bars. Similar to the data in Table 1, this figure shows that placing the source in subroom B leads to an approximately twofold reduction of $\widehat{\mathrm{EDT}}$ and $\widehat{T}_{30}$,

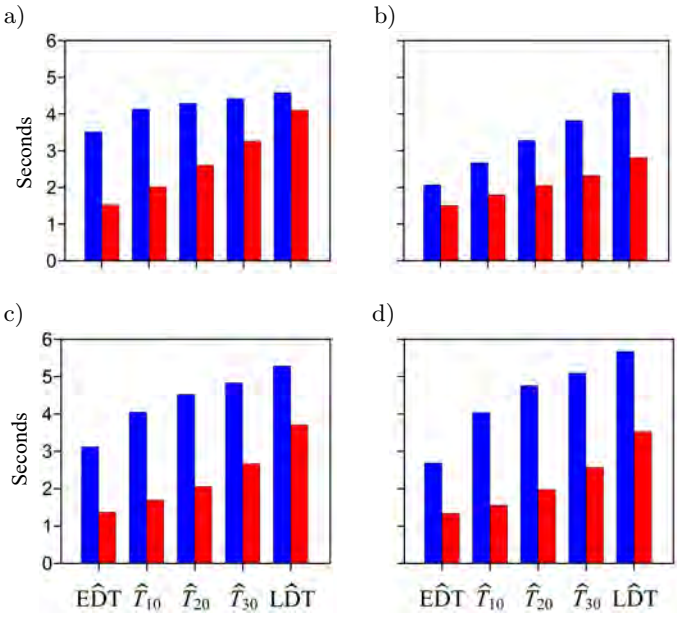


FIG. 12. Mean decay times $\widehat{\mathrm{EDT}}$, $\widehat{T}_{10}$, $\widehat{T}_{20}$, $\widehat{T}_{30}$, $\widehat{\mathrm{LDT}}$ on observation plane in subroom A for octave bands: a) 31.5 Hz, b) 63 Hz, c) 125 Hz, d) 250 Hz. Results marked with blue and red bars correspond to source located in subrooms A and B, respectively.

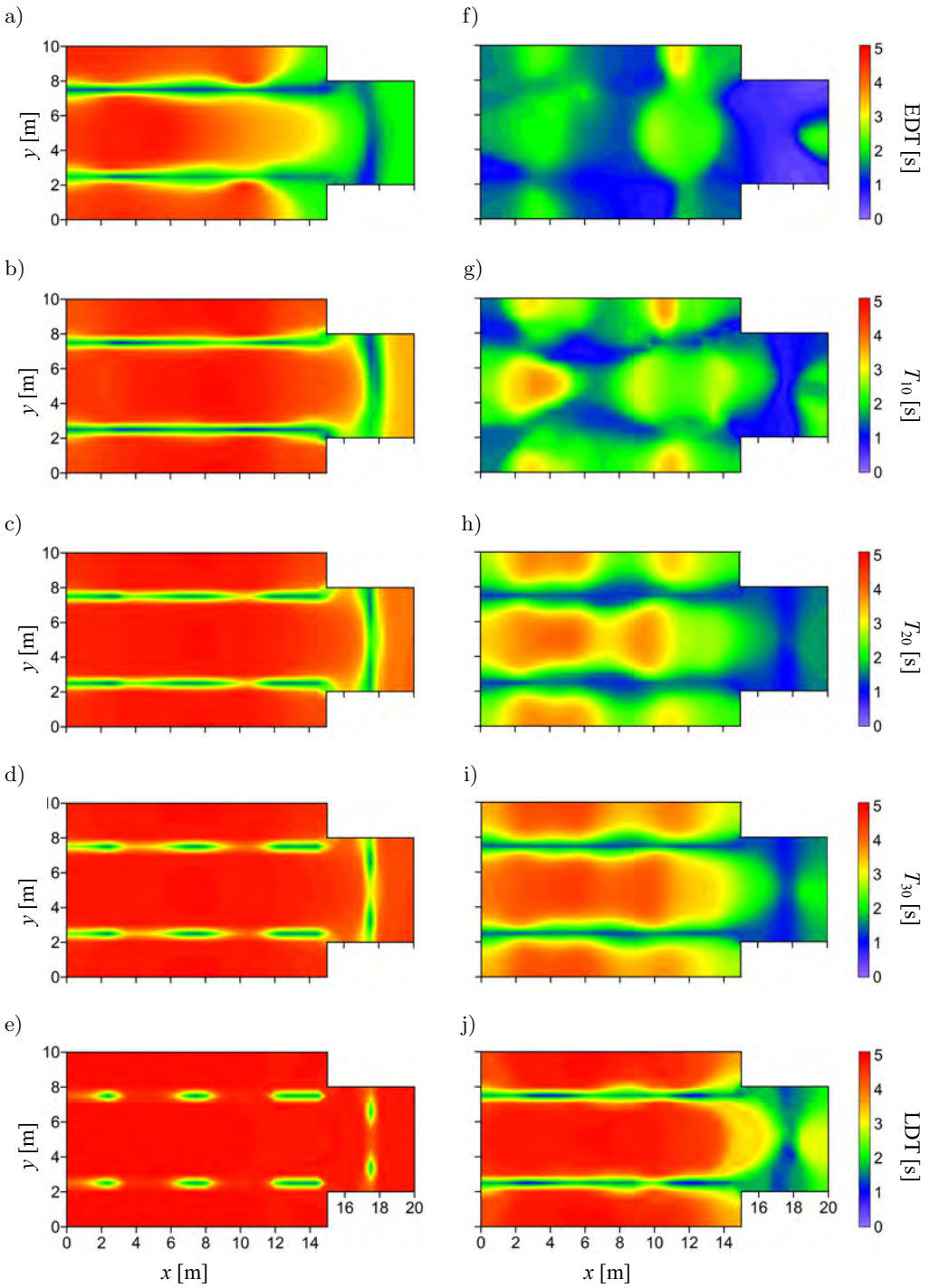


FIG. 13. Distribution of decay times EDT, T_{10}, T_{20}, T_{30}, LDT on observation plane for octave band 31.5 Hz and source located in: a)–e) subroom A and f)–j) subroom B.

as well as a more than twofold decrease in $\widehat{T}_{10}$ and $\widehat{T}_{20}$. However, such significant reduction of $\widehat{\text{LDT}}$ is not observed, because slowly decaying modes localized in subroom A start to dominate in the late decay stage.

The calculation results in Fig. 10 were obtained when eigenmodes with the frequencies f_m up to 400 Hz were used to determine the RIR function (Eq. (11)). However, since in a low-frequency range the character of sound decay depends on frequency, it is equally important to predict the decay times in frequency bands such as octaves. In the developed theoretical model, a bandpass filtering of the RIR function can be realized by summing such components of the series in Eq. (11) for which $f_l < f_m < f_u$, where f_l and f_u are the lower and upper limits of octave bands. Then, using Eq. (16), L_{EDC} for the particular octave band can be determined

and by applying the linear regression to appropriate parts of L_{EDC}, the decay times at observation points can be calculated, and then, from their values, the mean decay times on the observation plane lying in subroom A can be predicted.

Since the frequency range is limited to 400 Hz, the mean decay times were computed for octave bands with the center frequency of 31.5 Hz, 63 Hz, 125 Hz, and 250 Hz. The calculation results are presented in Fig. 12. They prove that for octave bands with the center frequency of 125 Hz and 250 Hz the mean decay times $\widehat{\mathrm{EDT}}$, $\widehat{T}_{10}$, $\widehat{T}_{20}$, $\widehat{T}_{30}$ are approximately twice as small when the source is located in subroom B. These results are very similar to those obtained for the frequency band up to 400 Hz (Fig. 11), and this is because in these octaves there is a large number of eigenmodes, 473 and 3312, respectively. Obviously, such a rule is not visible for octave bands with a center frequency of 31.5 Hz and 63 Hz, as the number of eigenmodes in these octaves is very small, i.e., 11 and 69, respectively. This is particularly true for the first octave, which contains a strongly localized eigenmode with the frequency of 34.5 Hz and the modal reverberation time of 4.9 s. The shape of the eigenfunction Ψ_κ for this eigenmode is shown in Fig. 3a. As can be seen in Fig. 13, this eigenmode dominates the sound decay process when the source is located in subroom A. Moreover, when the source is placed in subroom B, this mode strongly influences the sound decay process in the late decay stage.

4. Summary and conclusions

Sound decay is a fundamental and easily noticeable acoustic feature of enclosed spaces, hence one of the basic goals of room acoustics is to assess sound decay times based on the reverberant sound field in enclosed spaces. In this study, the sound decay in the system of two connected enclosed spaces with a different sound absorption on boundary surfaces was analyzed. The theoretical model was based on the modal expansion of sound pressure, since at low frequencies acoustic modes excited in enclosed spaces can have a significant influence on the sound decay process. Two types of reverberant responses of the system were examined. In the first case, the system was driven by a harmonic point source, and the sound decay process began immediately after the source was turned off. In the second case, the reverberant response of the system was initiated by the Dirac delta time impulse emitted by a point source.

Using the presented theoretical model, numerical simulations of the sound decay process was performed for a system of two connected rectangular subrooms of the same height. The larger subroom marked A was highly reverberant, while the smaller subroom marked B had walls that provided moderate sound damping. The eigenfunctions and eigenfrequencies of this system were determined using the finite element method. The obtained calculation results showed that after switching off the harmonic source, the sound decay curves usually have an irregular shape. This is because the sound decay process, which has been shown to strongly depend on both the source frequency and the source location, is generally non-exponential. Only when this process is dominated by the eigenmode localized in subroom A, an exponential sound decay was observed.

Another method used to evaluate the reverberant properties of the analyzed two-room coupled system was based on the reverse-time integration of RIR. By applying this method, smooth EDCs were obtained and based of their slope, the decay times EDT, T_{10}, T_{20}, T_{30}, and LDT were determined. Calculation results demonstrated that a different sound absorption in subrooms significantly impacts the sound decay process because it results in a wide range of modal decay time values. This is because these values are large for eigenmodes strongly localized in subroom A and very small for eigenmodes strongly localized in subroom B. Thus, the system under consideration has an interesting feature that is the result of the phenomenon of localization of eigenmodes. Namely, if the source is located in subroom A, the sound decay process in subroom A is dominated by slowly decaying modes that are strongly localized in this subroom. However, when the source is located in subroom B, non-localized modes with shorter modal reverberation time are excited in subroom A in the early stage of sound decay. Therefore, as calculation results showed, in this case the decay times in subroom A are approximately half that for the source located in subroom A. Obviously, this regularity does not occur in the late stage of sound decay, because in this stage the modes localized in subroom A start to influence the sound decay.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The author declares that there are no known competing financial interests or personal relationships that could have influenced the work described in this paper.

Author's contribution

The author conceptualized the study, performed the analysis, wrote the original draft, and approved the final manuscript.

References

1. Bradley D.T., Wang L.M. (2007), Comparison of measured and computer-modeled objective parameters for an existing coupled volume concert hall, *Building Acoustics*, **14**(2): 79–90, https://doi.org/10.1260/135101007781448028.
2. Bradley D.T., Wang L.M. (2009), Quantifying the double slope effect in coupled volume room systems, *Building Acoustics*, **16**(2): 105–123, https://doi.org/10.1260/135101009788913275.
3. Bradley D.T., Wang L.M. (2010), Optimum absorption and aperture parameters for realistic coupled volume spaces determined from computational analysis and subjective testing results, *Journal of the Acoustical Society of America*, **127**(1): 223–232, https://doi.org/10.1121/1.3268604.
4. Damelin S.B., Miller W. (2012), *The Mathematics of Signal Processing*, Cambridge University Press, New York.
5. Dance S.M., Van Buuren G. (2013), Effects of damping on the low-frequency acoustics of listening rooms based on an analytical model, *Journal of Sound and Vibration*, **332**(25): 6891–6904, https://doi.org/10.1016/j.jsv.2013.07.011.
6. Das O., Abel J.S. (2021), Grouped feedback delay networks for modeling of coupled spaces, *Journal of the Audio Engineering Society*, **69**(7/8): 486–496, https://doi.org/10.17743/jaes.2021.0026.
7. Ermann M., Johnson M. (2005), Exposure and materiality of the secondary room and its impact on the impulse response of coupled-volume concert halls, *Journal of Sound and Vibration*, **284**(3–5): 915–931, https://doi.org/10.1016/j.jsv.2004.07.030.
8. Girón S., Álvarez-Morales L., Zamarreño T. (2017), Church acoustics: a state-of-the-art review after several decades of research, *Journal of Sound and Vibration*, **411**: 378–408, https://doi.org/10.1016/j.jsv.2017.09.015.
9. Jasa T., Xiang N. (2009), Efficient estimation of decay parameters in acoustically coupled-spaces using slice sampling, *Journal of the Acoustical Society of America*, **126**(3): 1269–1279, https://doi.org/10.1121/1.3158934.
10. Kak S.C. (1970), The discrete Hilbert transform, *Proceedings of the IEEE*, **58**(4): 585–586, https://doi.org/10.1109/PROC.1970.7696.
11. Kinsler L.E., Frey A.R., Coppens A.B., Sander J.V. (2000), *Fundamentals of Acoustics*, 4th ed., John Wiley & Sons, New York.
12. Kuttruff H. (2017), *Room Acoustics*, 6th ed., CRC Press, New York.
13. Luizard P., Katz B.F.G. (2014), Investigation of the effective aperture area of sliding and hinged doors between coupled spaces, *Journal of the Acoustical Society of America*, **136**(2): 1269–1279, http://doi.org/10.1121/1.4890202.
14. Luizard P., Polack J-D., Katz B.F.G. (2014), Sound energy decay in coupled spaces using a parametric analytical solution of a diffusion equation, *Journal of the Acoustical Society of America*, **135**(5): 2765–2776, http://doi.org/10.1121/1.4870706.
15. Luizard P., Katz B.F.G., Guastavino C. (2015), Perceived suitability of reverberation in large coupled volume concert halls, *Psychomusicology: Music, Mind, and Brain*, **25**(3): 317–325, https://doi.org/10.1037/pmu0000109.
16. Martellotta F. (2016), Understanding the acoustics of Papal Basilicas in Rome by means of a coupled-volumes approach, *Journal of Sound and Vibration*, **382**: 413–427, http://doi.org/10.1016/j.jsv.2016.07.007.

17. Meissner M. (2009), Spectral characteristics and localization of modes in acoustically coupled enclosures, *Acta Acustica united with Acustica*, **95**(2): 300–305, https://doi.org/10.3813/AAA.918152.

18. Meissner M. (2010), Simulation of acoustical properties of coupled rooms using numerical technique based on modal expansion, *Acta Physica Polonica A*, **118**(1): 123–127, https://doi.org/10.12693/APhysPolA.118.123.

19. Meissner M. (2012), Acoustic energy density distribution and sound intensity vector field inside coupled spaces, *Journal of the Acoustical Society of America*, **132**(1): 228–238, https://doi.org/10.1121/1.4726030.

20. Meissner M., Wiśniewski K. (2019), Influence of room modes on low-frequency transients: theoretical modeling and numerical predictions, *Journal of Sound and Vibration*, **448**: 19–33, https://doi.org/10.1016/j.jsv.2019.02.012.

21. Meissner M., Wiśniewski K. (2020), Investigation of damping effects on low-frequency steady-state acoustical behaviour of coupled spaces, *Royal Society Open Science*, **7**(8): 200514, https://doi.org/10.1098/rsos.200514.

22. Nijs L., Jansens G., Vermeir G., van der Voorden M. (2002), Absorbing surfaces in ray-tracing programs for coupled spaces, *Applied Acoustics*, **63**(6): 611–626, https://doi.org/10.1016/S0003-682X(01)00063-9.

23. Paris E.T. (1928), On the coefficient of sound-absorption measured by the reverberation method, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **5**(29): 489–497, https://doi.org/10.1080/14786440308565092.

24. Schroeder M.R. (1965), New method of measuring reverberation time, *Journal of the Acoustical Society of America*, **37**(3): 409–412, https://doi.org/10.1121/1.1909343.

25. Schroeder M.R. (1996), The "Schroeder frequency" revisited, *Journal of the Acoustical Society of America*, **99**(5): 3240–3241, https://doi.org/10.1121/1.414868.

26. Shi S., Jin G., Xiao B., Liu Z. (2018), Acoustic modeling and eigenanalysis of coupled rooms with a transparent coupling aperture of variable size, *Journal of Sound and Vibration*, **419**: 352–366, https://doi.org/10.1016/j.jsv.2018.01.024.

27. Summers J.E., Torres R.R., Shimizu Y. (2004), Statistical-acoustics models of energy decay in systems of coupled rooms and their relation to geometrical acoustics, *Journal of the Acoustical Society of America*, **116**(2): 958–969, https://doi.org/10.1121/1.1763974.

28. Summers J.E., Torres R.R., Shimizu Y., Dalenbäk B-I.L. (2005), Adapting a randomized beam-axis-tracing algorithm to modeling of coupled rooms via late-part ray tracing, *Journal of the Acoustical Society of America*, **118**(3): 1491–1502, https://doi.org/10.1121/1.2000772.

29. Summers J.E. (2012), Accounting for delay of energy transfer between coupled rooms in statistical-acoustics models of reverberant-energy decay, *Journal of the Acoustical Society of America*, **132**(2): 129–134, https://doi.org/10.1121/1.4734591.

30. Sü Gül Z., Odabaş E., Xiang N., Çalışcan M. (2019), Diffusion equation modeling for sound energy flow analysis in multi domain structures, *Journal of the Acoustical Society of America*, **145**(4): 2703–2717, https://doi.org/10.1121/1.5095877.

31. Sü Gül Z. (2021), Exploration of room acoustics coupling in Hagia Sophia of Istanbul for its different states, *Journal of the Acoustical Society of America*, **149**(1): 320–339, https://doi.org/10.1121/10.0002971.

32. Thydal T., Pind F., Jeong C.-H., Engsig-Karup A.P. (2021), Experimental validation and uncertainty quantification in wave-based computational room acoustics, *Applied Acoustics*, **178**: 107939, https://doi.org/10.1016/j.apacoust.2021.107939.

33. Vorländer M. (2020), *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, 2nd ed., Springer, Berlin.

34. Xiang N., Jing Y., Bockman A.C. (2009), Investigation of acoustically coupled enclosures using a diffusion-equation model, *Journal of the Acoustical Society of America*, **126**(3): 1187–1198, https://doi.org/10.1121/1.3168507.

35. Xiang N., Goggans P., Jasa T., Robinson P. (2011), Bayesian characterization of multiple-slope sound energy decays in coupled-volume systems, *Journal of the Acoustical Society of America*, **129**(2): 741–752, https://doi.org/10.1121/1.3518773.

Archives of Acoustics Vol. 51, No. 3, pp. 495–506 (2026), https://doi.org/10.24423/archacoust.2026.4431

Research Paper

Low Noise Optimization of Automotive Air Conditioning Based on Response Surface Methodology

Cheng Qian[(1)], Dajian Wang[(2)], Hao Wang[(2)], Fushen Tang[(1)], Min Lin[(1)], Jinjin Xu[(1)*]

[(1)] *School of Energy and Power Engineering, Jiangsu University of Science and Technology*
Zhenjiang, People's Republic of China

[(2)] *Xiezhong International Thermal Management Systems (Jiangsu) Co., LTD.*
Nanjing, People's Republic of China

*Corresponding Author: xujinjin@just.edu.cn

Received February 4, 2026; revised May 7, 2026; accepted May 11, 2026; version of record July 28, 2026; published issue September 23, 2026.

Air conditioning is an important contributor to in-vehicle noise. In order to achieve low noise optimization, investigations were performed in a typical air conditioning unit. Computational fluid dynamics and computational aeroacoustics were combined to simulate the noise characteristics. Steady and transient simulations were conducted to identify noise sources, revealing that the dipole noise is mainly distributed on the fan blades and the volute tongue, and the fan contributes more significantly to the far-field noise. Based on the response surface methodology (RSM), three key structural parameters of the fan-volute tongue clearance, volute tongue radius, and blade outlet angle were selected as design variables, while the far-field *A*-weighted sound pressure level (SPL) was set as the response index. The final optimization parameter are: volute tongue clearance of 13.36 mm, volute tongue radius of 11.83 mm, and blade outlet angle of 110°. The far-field SPL is reduced from 61.8 dB to 55.5 dB, with a relative reduction of 10.2 %. This research validates the effectiveness and reliability of the proposed method, providing an efficient and systematic technical reference for the noise reduction design of automotive air conditioning systems.

Keywords: vehicle air conditioning, aerodynamic noise, response surface methodology, noise reduction optimization.



Copyright © 2026 The Author(s).
This work is licensed under the Creative Commons Attribution 4.0 International CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Aerodynamic noise caused by vehicle air conditioning is identified as an important contributor to the interior noise. It could reach the interior cabin without any sound isolation and can impact customer comfort (Lee *et al.*, 2011; Singh, Mohanty, 2018; Ravichandran *et al.*, 2023). A study by the World Health Organization (WHO) points out that long-term exposure to traffic noise above 55 dB may give rise to a variety of health risks, such as cardiovascular system diseases, reduced sleep quality, and impaired mental health (Liu, Li, 2022). Therefore, how to scientifically and effectively solve the aerodynamic noise caused by vehicle air conditioning has become an important problem in the automotive industry.

To reduce internal aerodynamic noise, some studies have adjusted and optimized the structure of the air conditioning system, such as the air supply ducts and the geometric shape of the air conditioning box (Aissaoui *et al.*, 2015; Zhang *et al.*, 2021; Wu *et al.*, 2023). More studies focused on conducting targeted aerodynamic optimization on the crucial noise source, that is the centrifugal fan. A study designed five different fan models and developed a design-analysis method for testing Sirocco fan noise, whose performance and noise characteristics

were predicted within a few percent relative error (Lee, Kil, 2010). The trailing edge structures of blades were optimized. Wide serrations with a wavelength (λs) to the amplitude (2 h) ratio of $\lambda s/h = 0.6$ reduce the overall sound pressure level (SPL) by up to 11 dB (Moreau, Doolan, 2016). Research developed low-noise centrifugal fans by modifying the linear trailing edge lines of the fan blades into inclined S-shaped lines. The study confirmed the noise reductions achieved by using the inclined S-shaped trailing edges by comparing the predicted BPF noise components of the developed fans with those of the existing fan (Heo *et al.*, 2011). These researches mainly focused on a single structure optimization of the fan to achieve low noise levels, while an effective method needs to be developed that would comprehensively consider the various relevant factors affecting the fan noise.

The response surface methodology (RSM) can comprehensively take structural parameters affecting aerodynamic noise into consideration, and then obtain a noise reduction optimization solution. Compared with traditional trial-and-error optimization and single-factor analysis methods, RSM can efficiently characterize the nonlinear coupling relationship between multiple structural design variables and aerodynamic noise responses without extensive computational trials. It can construct a high-precision surrogate model with limited sample points, effectively reduce the workload of numerical simulation and parametric modeling, and significantly improve the overall optimization efficiency. The RSM has been widely applied in the research of aerodynamic noise reduction optimization. A study combined the RSM with the partial differential equation (PDE) method to investigate the aerodynamic design problem of flying wings (Sevant *et al.*, 2000). Another study established an optimization function between optimization variables and multi-objective performance by integrating the RSM and the entropy weight method (Zhang *et al.*, 2017). The RSM is also used for structural noise reduction optimization of fans. Researchers adopted the RSM for the rational optimization of fan structural parameters, achieving the optimal control of aerodynamic noise levels (Yang *et al.*, 2013). Others designed a Box–Behnken test based on the RSM, and established the relationship between fan performance and boundary parameters using a quadratic regression surrogate model. A comprehensive evaluation function for fan performance is proposed in the research, therefore obtaining the optimal layout strategy for the dual-fan system (Guo *et al.*, 2022).

However, the studies mostly take the SPL of the fan itself as the response index. Considering the practical application of fan in automobile air conditioning system, noise characteristics and fan noise reduction optimization need to be explored and discussed in the whole automobile air conditioning box. The far-field A-weighted SPL of the vehicle air conditioning cabinet was taken as the response index in the study and the optimal fan structural parameter was determined for reducing the noise.

2. Methodology

2.1. Aeroacoustic theoretical equations

The basic aeroacoustic equations are the theoretical foundation for solving aerodynamic noise problems. British mathematician James Lighthill reorganized and transformed the Navier–Stokes (N-S) equations. He proposed the Lighthill equation and the eighth-power law. The specific expressions of the Lighthill equation are shown in Eq. (1) and Eq. (2), and the expression of the eighth-power law is given in Eq. (3):

$$\frac{\partial^2 \rho'}{\partial t^2} - c_0^2 \nabla^2 \rho' = \frac{\partial^2 T_{ij}}{\partial x_i \partial x_j}, \tag{1}$$

$$T_{ij} = \rho u_i u_j - \sigma_{ij} + \left[(p - p_0) - c_0^2 (\rho - \rho_0)\right] \delta_{ij}, \tag{2}$$

where ρ' is the disturbance quantity of fluid density, c_0 is the speed of sound, T_{ij} is the Lighthill stress tensor, which represents the turbulent stress inside the fluid, p_0 is the ambient pressure, δ_{ij} is the Kronecker delta function;

$$P \propto \rho_0 U^8 a_0^{-5} L^2, \tag{3}$$

where P is the radiated sound power of turbulent noise, ρ_0 is the fluid density, U is the characteristic velocity of the flow (e.g., turbulent velocity), L is the characteristic length scale.

Newby Curle, a British mathematician, proposed the Curle equation based on the Lighthill theory, further improving the theoretical system of aeroacoustics. The specific expression of the Curle equation is as follows:

$$\rho'(\mathbf{x},t) = \frac{1}{4\pi c_0^2}\int\limits_V \frac{\partial^2 T_{ij}}{\partial y_i \partial y_j}\frac{1}{|\mathbf{x}-\mathbf{y}|}\,\mathrm{d}^3\mathbf{y} + \frac{1}{4\pi c_0^2}\int\limits_S \left(\frac{\partial p}{\partial n}\frac{1}{|\mathbf{x}-\mathbf{y}|} - \frac{\rho_0 v_n}{|\mathbf{x}-\mathbf{y}|^2}\right)\mathrm{d}S(\mathbf{y}), \tag{4}$$

where $\rho'(\mathbf{x},t)$ is the density fluctuation in the sound field, c_0 is the speed of sound, T_{ij} is the Lighthill stress tensor, p is the pressure on the solid boundary, v_n is the normal velocity on the solid boundary, S is the solid boundary surface, V is the fluid volume.

Ffowcs Williams and Hawkings proposed the Ffowcs Williams–Hawkings (FW-H) equation. This equation expands the application scope of the aeroacoustics theory. The mathematical form of the FW-H equation is shown in Eq. (5):

$$\frac{1}{c_0^2}\frac{\partial^2 p}{\partial t^2} - \nabla^2 p = \frac{\partial^2}{\partial x_i \partial x_j}T_{ij}H(f) - \frac{\partial}{\partial x_i}[p_{ij}n_i + \rho u_i(u_n - v_n)]\delta(f) - \frac{\partial}{\partial t}[\rho_0 v_n + \rho(u_n - v_n)]\delta(f), \tag{5}$$

$$H(f) = \begin{cases} 1 & f > 0, \\ 0 & f < 0, \end{cases} \tag{6}$$

$$\delta(f) = \frac{\partial H}{\partial f} = \begin{cases} 1 & f = 0, \\ 0 & f \neq 0, \end{cases} \tag{7}$$

where c_0 is the far-field speed of sound, t is the reception time, p is the sound pressure, p_{ij} is the pressure stress tensor, n_j is the unit normal vector pointing outward to the external region, u_i is the fluid velocity component in the i-th direction, u_n is the normal fluid velocity of the integral surface, v_n is the normal velocity of the integral surface, ρ and ρ_0 are the perturbed density and the initial density, $\delta(f)$ is the Dirac function, which is used to describe the source term on the boundary, $H(f)$ is the Heaviside function, which is used to distinguish between the fluid region and the solid region.

2.2. Physical model

The automotive air conditioning system consists of several key components, including the internal and external circulation air inlets, filter, fan, fan volute, evaporator, heater, and air control valves. By adjusting the air valve, different air conditioning blowing and working modes can be switched according to actual requirements. The study took the full-cool face blowing mode as an example to analyze the aerodynamic noise characteristics and noise propagation mechanisms. A simplified simulation model was developed based on the physical structure of the air conditioning unit. Minor components and non-critical structures were reasonably simplified, and the filter and evaporator were treated as porous media. The simplified air conditioning system model is shown in Fig. 1.

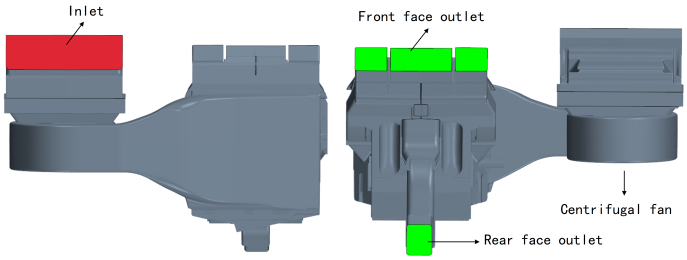


Fig. 1. Simplified model diagram of the air conditioning system.

2.3. Mesh generation and grid-independence verification

Considering that the simulation model has a complex structure, with large curvature changes and rich detailed features in local areas, triangular meshes were used for surface meshing, with the mesh size controlled within the

range of 1 mm to 4 mm. Local densification was applied to the air inlets, air outlets, fan blades, porous media regions, pipeline corners, and areas with significant changes in structural parameters. Finally, the number of surface meshes reached 1.15 million, as shown in Fig. 2.

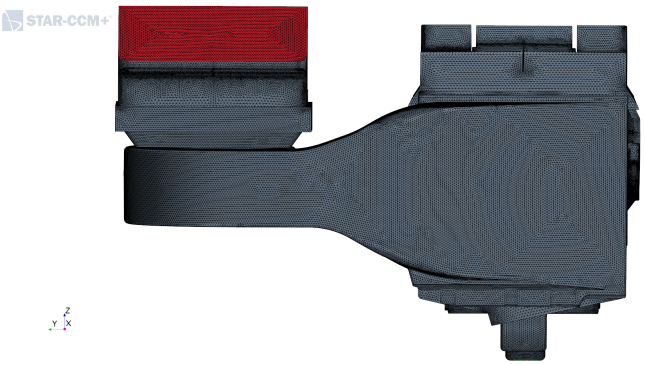


FIG. 2. Surface mesh schematic diagram.

Polyhedral meshes are adopted as the volume mesh type, which exhibit extremely high geometric adaptability and strong flexibility. The volume mesh was generated based on the surface mesh. A prism layer mesh was selected for the boundary layer, with a total thickness of 2 mm divided into 3 layers and a growth rate of 1.2. The maximum volume mesh size was set to 5 mm.

After determining the mesh type and the basic size, mesh independence verification was conducted to ensure the reliability of calculation results. The mass flow rate at the outlet and the far-field SPL were taken as evaluation criteria. Their values were compared at six kinds of meshes with different grid numbers as shown in Fig. 3.

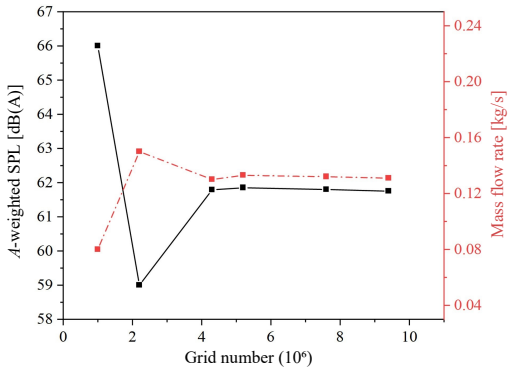


FIG. 3. Grid independence verification.

When the number of mesh elements increases from 4.23 million to 9.8 million, their difference of the flow rate at the outlets is less than 2 %, and the far-field SPL is less than 0.5 dB(A). Therefore, the final numerical calculation model adopts a mesh with 4.23 million elements, and the mesh generation result is shown in Fig. 4.

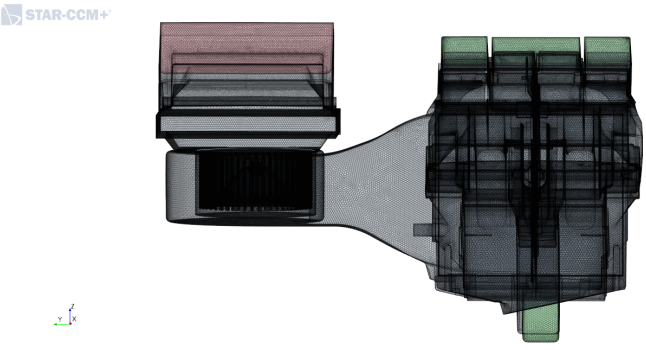


FIG. 4. Volume mesh schematic diagram.

2.4. Boundary conditions and solution

According to the actual operating conditions of the automotive vehicle air conditioning system, the inlet of the computational domain is set as a stagnation inlet, and the outlet is a pressure outlet. A no-slip wall condition is applied to accurately simulate the interaction between the fluid and the wall surfaces, including the development of the boundary layer and flow separation phenomena. The $y+$ wall treatment method is selected in this study. All settings are shown in Table 1.

Table 1. Steady-state calculation solver parameter settings.

Setting	Setting value
Temporal discretization scheme	Steady
Flow	Separated flow
Equation of state	Constant density
Reynolds-averaged turbulence	K-epsilon turbulent flow
K-epsilon turbulence models	Realizable k-ε two-layer model
Wall treatment method	Two-layer full $y+$ wall treatment
Rotating model	Moving reference frames
Acoustic model	Broadband noise source (BNS) model

For the internal flow field of the automotive air conditioning cabinet, the realizable k-ε model was selected as the physical model for the steady-state calculation. Based on the steady-state calculation, it is necessary to select an appropriate acoustic model to analyze the noise characteristics of the air conditioning system. This study adopts the broadband noise model, which is based on the Curle and Proudman equations for the solution. This model can effectively predict the distribution characteristics of broadband noise, providing a theoretical basis for subsequent noise optimization. The porosity of filter and evaporator was set to 0.8 according to the existing studies (Li, 2016; Zhang, 2019).

Transient calculation is performed after convergent steady-state calculation. In this study, the detached eddy simulation (DES) model is selected as the turbulence model for transient simulation. The rotating speed was set at 2266 rpm. Correspondingly, the transient calculation time step size was set at 7.3×10^{-5} s and the total simulation duration was 0.106 s. The FW-H acoustic method is adopted to obtain far-field aerodynamic noise of an air conditioning box under transient flow conditions. Far-field noise monitoring points are set at 1 m directly in front of the air conditioning cabinet air outlet (monitoring point A) and 1 m at a 45° diagonal direction (monitoring point B). The specific layout positions are shown in Fig. 5.

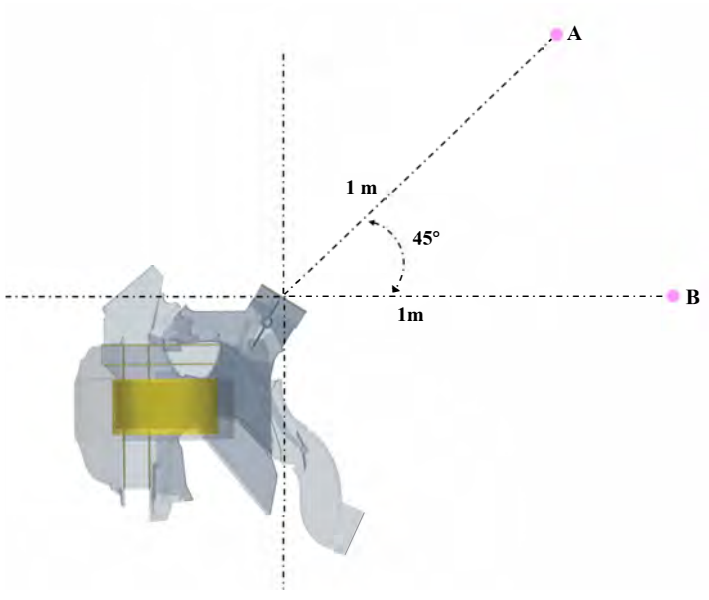


Fig. 5. Schematic diagram of far-field monitoring point locations.

2.5. Response surface methodology

The quadratic polynomial model is the most commonly used model for response surface experiments, and its general form is shown in Eq. (8):

$$Y = \beta_0 + \sum_{k}^{i=1} \beta_i X_i + \sum_{k}^{i=1} \beta_{ii} X_i^2 + \sum_{i<j} \beta_{ij} X_i X_j + \epsilon, \tag{8}$$

where Y is the response variable, X_i is the i-th independent variable (factor), β_0 is the constant term (intercept), β_i is the coefficient of the linear term, which determines the slope and direction of the response surface, β_{ii} is the coefficient of the quadratic term, representing the influence of the quadratic term of the i-th independent variable on the response variable, β_{ij} is the coefficient of the interaction term, which determines the distortion degree of the response surface and reflects the interaction between the independent variables, ϵ is the error term, which is usually assumed to follow a normal distribution, representing the random error unexplained by the model; the smaller the error term is, the stronger the predictive ability of the model is. The variance of the error term σ^2 is used to evaluate the uncertainty of the model, k is the number of independent variables.

For the convenience of calculation and analysis, the quadratic polynomial model can be expressed in matrix form as follows:

$$Y = \mathbf{X}\beta + \epsilon, \tag{9}$$

where Y is the vector of response variables, X is the design matrix containing the linear terms, quadratic terms and interaction terms of the independent variables, β is the vector of model coefficients, ϵ is the vector of error terms. Each row of the design matrix $\mathbf{X}$ corresponds to an experimental point, and each column corresponds to a model (e.g., linear, quadratic or interaction).

The model coefficients are usually estimated by the least squares method:

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y, \tag{10}$$

where $\widehat{\beta}$ is the estimated value of the model coefficients, X^{T} is the transpose of the design matrix, $(X^{\mathrm{T}}X)^{-1}$ is the inverse matrix of the design matrix. The objective of the least squares method is to minimize the sum of squared residuals, as shown in Eq. (11), where $\widehat{Y}_i$ is the predicted value of the model for the i-th experimental point:

$$\mathrm{SS}_{\mathrm{res}} = \sum_{k}^{i=1} (Y_i - \widehat{Y}_i)^2. \tag{11}$$

The coefficient of determination R^2 is used to evaluate the goodness of fit of the model:

$$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}}. \tag{12}$$

The adjusted coefficient of determination R^2_{adj} takes into account the number of model parameters:

$$R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{SS}_{\mathrm{res}}/(n-p)}{\mathrm{SS}_{\mathrm{tot}}/(n-1)}, \tag{13}$$

where R^2 represents the proportion of the variation in the response variable explained by the model, R^2_{adj} takes into account the number of model parameters to prevent overfitting. Higher values of R^2 and R^2_{adj} indicate that the model has a good goodness of fit.

3. Aerodynamic noise characteristics

As shown in Fig. 6, the vehicle air conditioning simulation results indicate that dipole noise is mainly distributed on the surface of the fan blades and the outer area of the volute, with the maximum noise value at the

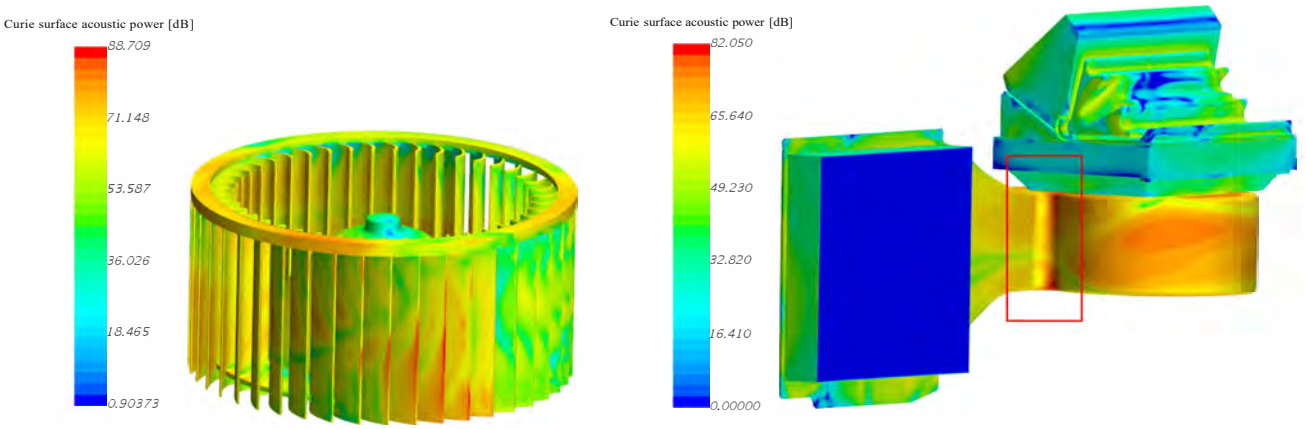


FIG. 6. Distribution of dipole noise on fan blades and volute under the operating condition of 2266 rpm.

blade edges reaching 88.7 dB. In addition, due to the small gap between the blades and the volute tongue, strong pressure fluctuations are likely to occur when air flows through. Consequently, the maximum dipole noise value at the volute consistently appears in the volute tongue area, with the maximum value being 82.0 dB.

To further analyze the contribution of different noise sources in the far-field SPL, the fan blades, the entire air conditioning system, and the air conditioning unit casing were respectively inputted as sound source. The SPL spectra are shown in Fig. 7. When the air conditioning unit casing is set as the noise source, the overall SPL spectrum is relatively low. And the noise intensity is higher in the low-frequency range around 500 Hz, which may be related to casing vibration or resonance of internal airflow. However, when the fan blades or the entire air conditioning system are used as the noise source, obvious peaks appear in the spectrum curve at the fundamental frequency and its harmonic frequencies. Due to the influence of the axial radiation characteristics of rotational noise, the peak amplitude at monitoring point A (directly facing the air outlet) is higher than that at monitoring point B (at a 45° diagonal angle).

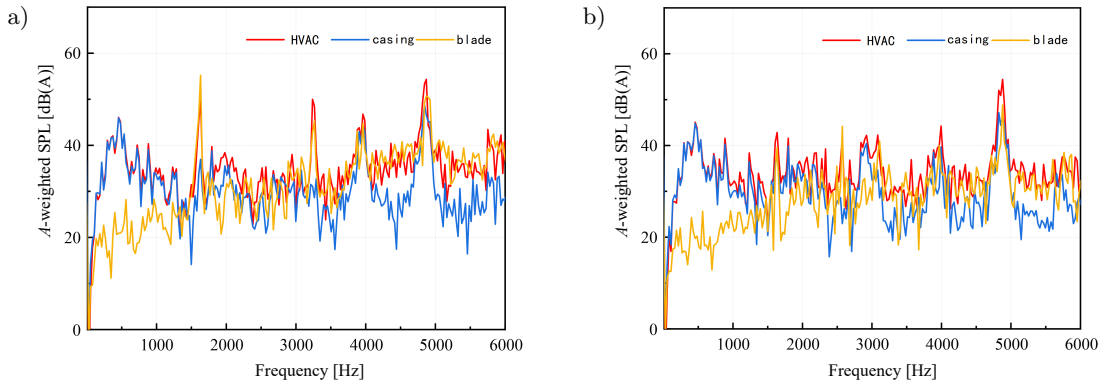


FIG. 7. Far-field SPL frequency spectrum curve: a) monitoring point A, b) monitoring point B.

After calculating the obtained sound pressure data, the total sound pressure levels of far-field noise are obtained, as shown in Table 2. When the entire air conditioning system and the fan are taken as noise sources, the far-

TABLE 2. Comparison of far-field total SPLs from different noise sources.

Noise source	Total SPL [dB(A)]
Entire vehicle air conditioning system	61.8
Casing	58.5
Fan blade	60.2

field SPLs reach 61.8 dB(A) and 60.2 dB(A), respectively. When the casing is regarded as the noise source, the total SPL is 58.5 dB(A). It could be found that the fan contributes more significantly to the far-field noise.

Based on the above analysis, the low noise optimization design would focus on the fan and its volute tongue part in subsequent work.

4. Noise reduction optimization based on RSM

According to the studies of former researchers (Wei *et al.*, 2023; Huang *et al.*, 2025; Bai *et al.*, 2024), the fan volute tongue gap (A), volute tongue radius (B), and blade outlet angle (C) were specifically selected as the variables for the designed experiment. Taking the far-field SPL (Y) of the automotive vehicle air conditioning system as the response, a full-factor experiment with 3 factors and 2 levels was designed. The levels of the experimental factors are shown in Table 3.

Table 3. Full factorial experimental factors and levels.

Factor levels	Volute tongue gap (A) [mm]	Volute tongue radius (B) [mm]	Blade outlet angle (C) [°]
1	8	10	110
2	12	16	150

In this study, the central composite design (CCD) method was selected for the design of the RSM experiment. The specific experimental design and the corresponding result are shown in Table 4.

Table 4. Response surface experimental design table and results.

Run order	Point type	Volute tongue gap (A) [mm]	Volute tongue radius (B) [mm]	Blade outlet angle (C) [°]	Far-field SPL (Y) [dB(A)]
1	1	8	10	110	58.7
2	1	12	10	110	57.0
3	1	8	16	110	59.4
4	1	12	16	110	58.1
5	1	8	10	150	61.1
6	1	12	10	150	60.8
7	1	8	16	150	62.1
8	1	10	16	150	63.0
9	−1	10	13	130	57.8
10	−1	10	13	130	58.3
11	−1	10	13	130	58.0
12	−1	10	13	130	58.3
13	−1	10	13	130	58.1
14	−1	10	13	130	58.4
15	0	6.63	13	130	58.7
16	0	13.36	13	130	57.8
17	0	10	7.9546	130	59.9
18	0	10	18.0454	130	62.3
19	0	10	13	96.36	57.9
20	0	10	13	163.6	64.2

Analysis of variance (ANOVA) is performed on the designed experimental results, as shown in Table 5. It can be seen from the ANOVA table that the model is significant ($P < 0.05$) and there is no significant LOF ($P > 0.05$), except for the quadratic term AA (volute tongue gap × volute tongue gap). This indicates that the model includes all significant terms, while the term AA is insignificant. Next, it is necessary to evaluate the model fitting performance by examining the coefficient of determination (R^2) and the adjusted coefficient of determination (R^2_{adj}). The results are shown in Table 6.

Table 5. Analysis of variance (ANOVA).

Source	Degree of freedom (DOF)	Adjusted sum of squares	Adjusted mean squares	F-value	P-value
Model	9	78.2561	8.6951	137.40	0.001
Linear	3	50.0515	16.6838	263.64	0.001
Volute tongue gap (A)	1	1.0632	1.0632	16.80	0.002
Volute tongue radius (B)	1	5.8159	5.8159	91.90	0.008
Blade outlet angle (C)	1	43.1724	43.1724	682.20	0.015
Quadratic	3	26.0821	8.6940	137.38	0.005
A^2	1	0.0006	0.0006	0.01	0.922
B^2	1	14.1351	14.1351	223.36	0.00
C^2	1	13.9841	13.9841	220.97	0.000
Two-factor interaction (TFI)	3	2.1226	0.7075	11.18	0.002
AB	1	0.3200	0.3200	5.06	0.048
AC	1	1.5138	1.5138	23.92	0.001
BC	1	0.2888	0.2888	4.56	0.058
Error	10	0.6328	0.0633	–	–
Lack of fit (LOF)	5	0.3887	0.0777	1.59	0.311
Pure error (PE)	5	0.2441	0.0488	–	–
Total	19	78.8890	–	–	–

Table 6. Model summary.

R^2	R^2_{adj}
0.992	0.9848

It can be seen from Table 6 that the coefficient of determination (R^2) for the overall fitting performance and the correlation factor (R^2_{adj}) are both higher than 0.95, indicating that the model exhibits excellent fitting and predictive performance.

Since the quadratic term AA (volute tongue gap × volute tongue gap) in the model is insignificant, and no abnormalities were observed in the model residuals, only the AA term was removed for model refinement and refitting. As shown in Table 7, the adjusted coefficient of determination (R^2_{adj}) increased by 0.13 %, respectively, compared with the original model. It indicates that the model's fitting performance has been improved after eliminating the insignificant term, and it exhibits enhanced predictive capability. After removing the insignificant term, the regression equation for the response value Y (far-field SPL) is obtained as follows:

$$Y = 129.22 - 1.987A - 3.39B - 0.7017C + 0.11011B^2 + 0.002464C^2 + 0.0333AB + 0.01087AC + 0.00317BC. \quad (14)$$

Table 7. Comparative model summary: original vs. reduced model.

	Original model	Processed model	Change
R^2	0.992	0.992	–
R^2_{adj}	0.9848	0.9861	↑

Based on the fitted regression equation, response surface plots were generated to analyze the effects of factors A (volute tongue gap), B (volute tongue radius), and C (blade outlet angle) on the response Y (far-field SPL). The results are shown in Fig. 8.

It can be observed from Fig. 8a that when the blade outlet angle is kept constant at 130°, the far-field SPL decreases with the increase of the volute tongue gap. As the volute tongue radius increases, the far-field SPL first decreases and then increases, indicating that there exists a minimum value within the selected factor level range. This indicates that the effect of the volute tongue radius in the far-field SPL is not a linear relationship. It can be seen from Fig. 8b that when the volute tongue radius is maintained at 13 mm, the far-field SPL first decreases and then increases as the blade outlet angle increases. It can also be observed from Fig. 8c that the response

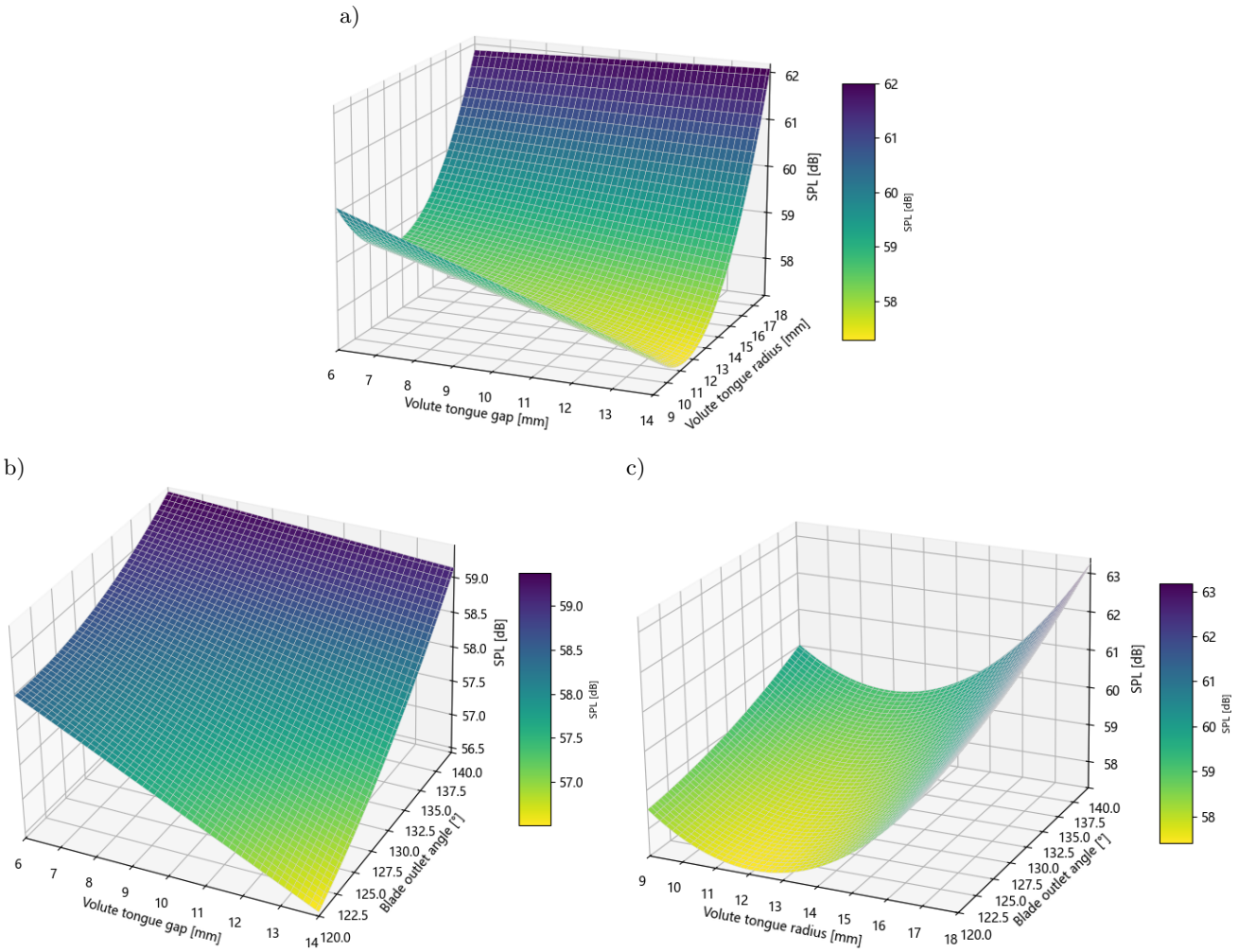


FIG. 8. Far-field SPL response surface plot: a) A vs. B, b) A vs. C, c) B vs. C.

surface plot exhibits a curved surface, confirming that the minimum value of the far-field SPL exists within the selected factor level range.

After the response surface test, the response optimizer function in Minitab software can be used to calculate the predicted values and confidence intervals of the model under optimal conditions. The response optimizer is employed to predict the optimal conditions for the model, and the factor combination corresponding to the minimum value of the response Y (far-field SPL) is determined, as detailed in Table 8.

TABLE 8. Response optimization solution.

Volute tongue gap (A) [mm]	Volute tongue radius (B) [mm]	Blade outlet angle (C) [°]
13.36	11.83	110

Simulation verification was performed on the factor combination of the optimal solution obtained by Minitab software, and a comparison was conducted with the original model data. The results are shown in Table 9. The total SPL at the far-field monitoring point was reduced by 6.3 dB(A), indicating a significant optimization effect.

TABLE 9. Comparative analysis: optimized vs. original model.

Scheme	Far-field SPL [dB(A)]
Original structure	61.8
Optimized structure	55.5

5. Conclusion

A reliable vehicle air conditioning system simulation model was established, in which non-critical structures were simplified. Filters/evaporators were treated as porous media (porosity = 0.8). Triangular surface meshes, polyhedral volume meshes with local densification and 3-layer prism boundary meshes were adopted. Grid-independence verification ensured reliability. Steady-state simulations used realizable k-ε and BNS models; transient simulations adopted DES and FW-H methods (time step = 7.3×10^{-5} s) to capture flow and acoustic characteristics.

Dipole noise mainly distributed on the fan blades (max 88.7 dB) and volute tongue (max 82.0 dB) due to pressure fluctuations. Different sound sources contribute differently to far-field noise. When the air conditioning unit casing is inputted as the noise source, the far-field SPL is 58.5 dB(A) and noise intensity is higher in the low-frequency range around 500 Hz. When the fan blades and the entire system is used as the noise source, the far-field SPL reaches 60.2 dB(A) and 61.8 dB(A), respectively, and obvious peaks appear in the spectrum curve at the fundamental frequency and its harmonic frequencies.

With volute tongue gap (A), radius (B), and blade outlet angle (C) as variables, and SPL as the response, a CCD experiment was conducted for RSM. An insignificant term was removed, and ANOVA confirmed the quadratic model's significance ($P < 0.05$) and good fitting ($R^2 = 0.992$, $R^2_{\text{adj}} = 0.9861$). The model exhibits excellent fitting and predictive performance. The regression equation for the response value was obtained, and the optimal parameter combination ($A = 13.36$ mm, $B = 11.83$ mm, $C = 110°$) reduced SPL from 61.8 dB(A) to 55.5 dB(A).

Fundings

This work was supported by Xiezhong International Thermal Management Systems (Jiangsu) Co., LTD. (contract no. 2024321103000540).

Conflict of interests

The authors declare that there are no known competing financial interests or personal relationships that could have influenced the work described in this paper.

Authors' contributions

Cheng Qian performed simulation, and wrote the manuscript, Dajian Wang and Hao Wang established the model and determined main parameters. Fushen Tang conducted data collection and data analysis. Min Lin performed simulation and made the plots. Jinjin Xu checked and optimized the manuscript and guided the whole study process. All authors reviewed and approved the final manuscript.

References

1. Aissaoui A. *et al.* (2015), Flow-induced noise optimization of SUV HVAC system using a lattice Boltzmann method, *SAE International Journal of Passenger Cars – Mechanical Systems*, **8**(3): 1053–1062, https://doi.org/10.4271/2015-01-2323.
2. Bai M., Liu Z., Ling Y., Tan H. (2024), Effect of impeller structure on aerodynamic performance and noise reduction of a small multi-blade centrifugal fan, *Science and Technology for the Built Environment*, **30**(6): 563–578, https://doi.org/10.1080/23744731.2024.2357526.
3. Guo R., Mi T., Li L., Luo R. (2022), Research on aerodynamic performance and noise reduction of high-voltage fans on fuel cell vehicles, *Applied Acoustics*, **186**: 108454, https://doi.org/10.1016/j.apacoust.2021.108454.
4. Heo S., Cheong C., Kim T.-H. (2011), Development of low-noise centrifugal fans for a refrigerator using inclined S-shaped trailing edge, *International Journal of Refrigeration*, **34**(8): 2076–2091, https://doi.org/10.1016/j.ijrefrig.2011.07.003.

5. Huang Y., Zhang H., Liu Z. (2025), Realizable k-ε model-based gradual-radius volute tongue on aerodynamic performance and noise of multi-wing centrifugal fan, *Applied Sciences*, **15**(17): 9471, https://doi.org/10.3390/app15179471.
6. Lee C., Kil H.-G. (2010), An aero-acoustic performance prediction method of Sirocco fan, [in:] *Proceedings of the ASME 2010 3rd Joint US-European Fluids Engineering Summer Meeting collocated with 8th International Conference on Nanochannels, Microchannels, and Minichannels. ASME 2010 3rd Joint US-European Fluids Engineering Summer Meeting*, pp. 813–817, https://doi.org/10.1115/FEDSM-ICNMM2010-31040.
7. Lee D., Perot F., Ih K.-D., Freed D. (2011), Prediction of flow-induced noise of automotive HVAC systems, *SAE 2011 World Congress & Exhibition*, https://doi.org/10.4271/2011-01-0493.
8. Li S. (2016), *Flow noise analysis of Class A automotive HVAC systems* [in Chinese], MA Thesis, *Jilin University*.
9. Liu S., Li S. (2022), A brief discussion on the hazards and sources of automobile noise [in Chinese], *Times Automobile*, (4): 29–31.
10. Moreau D.J., Doolan C.J. (2016), The generation of tonal noise from sawtooth trailing-edge serrations at low Reynolds numbers, *The Aeronautical Journal*, **120**(1228): 971–983, https://doi.org/10.1017/aer.2016.39.
11. Ravichandran A., Logdesser A., Peller N., Manhart M. (2023), Acoustic resonances in an automotive HVAC outlet, *Journal of Sound and Vibration*, **569**: 118051, https://doi.org/10.1016/j.jsv.2023.118051.
12. Sevant N.E., Bloor M.I.G., Wilson M.J. (2000), Aerodynamic design of a flying wing using response surface methodology, *Journal of Aircraft*, **37**(4): 562–571, https://doi.org/10.2514/2.2665.
13. Singh S., Mohanty A.R. (2018), HVAC noise control using natural materials to improve vehicle interior sound quality, *Applied Acoustics*, **140**: 100–109, https://doi.org/10.1016/j.apacoust.2018.05.013.
14. Wei Y., Li L., Lun Y., Wang Z., Yang H., Zhang W. (2023), Effect of volute-tongue clearance on the aerodynamic performance and noise of multi-wing centrifugal fan for air conditioning, *Journal of Applied Fluid Mechanics*, **16**(7): 1483–1498, https://doi.org/10.47176/jafm.16.07.1514.
15. Wu Z., Zhou S., Li Y., Jin W., Luo Y. (2023), Air duct optimization design based on local turbulence loss analysis and IMOCS algorithm, *Machines*, **11**(2): 129, https://doi.org/10.3390/machines11020129.
16. Yang Z., Gu Z., Wang Y., Yan J., Yang X. (2013), Prediction and optimization of aerodynamic noise in an automotive air conditioning centrifugal fan, *Journal of Central South University*, **20**(5): 1245–1253, https://doi.org/10.1007/s11771-013-1608-6.
17. Zhang C., Li A., Li J., Hou Y., Chen X. (2021), Radiation noise control of a 90° rectangular elbow in ventilation and air conditioning systems, *Journal of Building Engineering*, **37**: 102157, https://doi.org/10.1016/j.jobe.2021.102157.
18. Zhang F. (2019), Analysis of aerodynamic noise transfer characteristics and noise reduction optimization of automotive air conditioning blowers [in Chinese], MA Thesis, *Southwest Jiaotong University*.
19. Zhang L., Wu K., Liu Y. (2017), Investigation on multi-objective performance optimization algorithm application of fan based on response surface method and entropy method, *Journal of Thermal Science*, **26**(6): 533–539, https://doi.org/10.1007/s11630-017-0970-3.

Archives of Acoustics Vol. 51, No. 3, pp. 507–514 (2026), https://doi.org/10.24423/archacoust.2026.4326

Research Paper

The Impact of the VR Interface on the Accuracy and Ergonomics of Spatial Hearing Tests

Grzegorz Rusinek, Agnieszka Paula Pietrzak*

Institute of Radioelectronics and Multimedia Technology, Warsaw University of Technology
Warsaw, Poland

*Corresponding Author: agnieszka.pietrzak@pw.edu.pl

Received September 8, 2025; revised March 6, 2026; accepted May 5, 2026; version of record July 20, 2026; published issue September 23, 2026.

In recent years, virtual reality (VR) has been an attractive alternative to traditional methods of conducting hearing tests, offering both a natural way to indicate sound direction and complete digital data recording. The aim of this study was to analyze how the use of VR goggles and motion controllers affects the accuracy of localizing acoustic stimuli and the subjective comfort of participants, compared to a WEB-based application operated with a mouse. The study was performed for 16 azimuthal samples (0° elevation) and 8 elevation samples at azimuth 90°. Mean absolute angular errors (MAAE) were analyzed, and the paired-samples t-tests were performed. The results show that using the VR interface can reduce horizontal MAAE by 7° compared to the WEB-based interface. In the elevation tasks, no statistically significant differences were observed. Respondents rated VR the highest in terms of intuitiveness, comfort, speed, and perceived precision. These findings confirm the potential of VR to be used in localization tests, especially in the horizontal plane, and suggest directions for further research on HRTF personalization and interface optimization for elevation tasks.

Keywords: virtual reality (VR), spatial sound localization, user experience.



Copyright © 2026 The Author(s).
This work is licensed under the Creative Commons Attribution 4.0 International CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Spatial audio quality listening tests often involve sound source localization. Participants hear short stimuli and then indicate the direction from which the sound originates. Although this simple concept stayed mostly unchanged throughout the years of research, researchers have tried multiple methods of gathering results from the participants.

For the localization task, experiments require an external response that maps participants' judgments onto measurable 3D coordinates. Many traditional methods are intuitive for listeners but require a supervisor to observe and transcribe responses, limiting scalability. Such methods may include: voice description of direction (Mason *et al.*, 2001), pointing with head/nose (Haber *et al.*, 1993; Bronkhorst, 1995; Fujii *et al.*, 2007), pointing with the torso/chest (Haber *et al.*, 1993; Fujii *et al.*, 2007), indicating direction with hand or with an auxiliary instrument, such as a baton (Haber *et al.*, 1993; Oldfield, Parker, 1984). These methods are comfortable from the perspective of the test participants, since their only task is to point in the direction of sound, which is an intuitive task to do. However, it is problematic in terms of collecting results, since in most cases it requires another person (a supervisor) actively observing the test participant, reading the indicated direction and recording the results.

Desktop and WEB-based tests often replace physical pointing with a mouse or gamepad that rotates a virtual sphere, providing broad reach but at the cost of indirect, cursor-driven mapping. Although convenient for large-

scale data collection, such interfaces typically inflate completion times and, for elevation, report angular errors well above laboratory baselines.

Commodity head-mounted displays (HMDs) with six-degree-of-freedom (6-DoF) wands that track hand position with sub-millimeter precision allow natural reach-and-point actions. For the participant, pointing with the virtual reality (VR) controller is as intuitive as pointing with their hand. The controller can read its own rotation and position in space, which are then processed by a computer program, making it possible to determine the position of the sound source indicated by the test subject with sufficient precision. Ahrens *et al.* (2019) demonstrated that simply visualizing the controller tip as a laser beam inside the scene significantly reduced the elevation error compared with an HMD-only baseline. Valzolgher *et al.* (2023) examined the learning effects, showing that embodied actions (pointing or reaching) accelerated adaptation to spectral perturbations, yet they did not benchmark absolute localization performance against a browser-based task.

Previous research has contrasted different response modes within VR (laser versus reaching, natural versus symbolic actions) or compared VR to loudspeaker rigs without a corresponding graphical-user-interface (GUI) control. No study has systematically benchmarked a consumer VR pointing controller against a standard web/gamepad interface under matched acoustic rendering, using identical target sets and collecting both objective and subjective metrics. Such results can help the evidence-based decisions when choosing between low-cost browser tests and immersive setups for large-scale spatial-hearing assessments.

The aim of this study was to compare the VR pointing interface with a conventional mouse-driven WEB-based interface. Thirty-three non-expert listeners performed horizontal and vertical localization tasks using the same generic head-related transfer function (HRTF) stimuli, presented through the same headphones, with the difference only in the response modality. The main goal was to assess whether the use of the VR interface would result in the lower mean absolute angular errors (MAAE) than the WEB interface (in the horizontal or vertical plane) and whether participants would rate the VR interface higher in terms of intuitiveness, comfort, speed, and perceived precision.

2. Methodology

In this study, a listening test was designed and conducted to examine the effect of using the VR interface on the accuracy of virtual sound source localization, compared to a WEB-based application interface operated by a mouse. Additionally, a questionnaire survey was conducted to examine the subjective comfort of the participants when using the VR interface compared to the WEB-based one.

Thirty-three voluntary participants (10 female, 23 male) aged 18–26 years (M = 21.4, SD = 2.1), with no self-reported hearing problems completed the test. All of the participants were non-experts. Handedness was not assessed, but participants were asked to operate each interface with their dominant hand.

All participants took a virtual sound source localization test, using the VR interface and the WEB interface. The interface order was randomized. A break ranging from 5 min to 10 min was provided between blocks. The listening test consisted of twenty-four samples for each interface. Sixteen samples were presented in the horizontal plane (0° elevation) at fixed azimuth angles of 45°, 90°, 105°, 165°, 225°, 270°, 315°, 345°, each presented twice in random order. Eight samples were presented in the vertical plane (azimuth of 90°) at fixed elevation angles of −60°, −45°, 30°, 90°, each presented twice in random order. Among the twenty-four samples, half were white noise samples (eight samples with variable azimuth and four with variable elevation), and the other half were samples of female voice with analogous division into samples with variable azimuth and elevation. Two stimulus types (white noise and speech) were included to reduce task predictability. Each time, the samples were played in random order, but the samples with variable azimuth were played first, followed by those with variable elevation. For the test in both forms (using VR and using the WEB application), the same audio samples were used. The hardware used for the VR test were the Meta Quest 2 VR goggles with their dedicated spatial motion controllers with Audio-Technica ATH-M50X closed-back headphones. The results were compared with the same test conducted using a WEB application on a PC computer with the same Audio-Technica ATH-M50X closed-back headphones. Stimuli were presented at the sound level adjusted by the user, with inter-channel level matched

within ±0.5 dB. Rendering used third-order ambisonics (fs = 48 kHz) and the IEM Binaural Decoder (default HRTF with headphone compensation for ATH-M50x).

In the case of the WEB application, the participants were asked to mark the sound sources on a 3D simulation displayed on a monitor, with an interactive image of a listening dummy and a sphere-shaped cage surrounding it, on which possible locations of the sound source were marked (Fig. 1). The response grid had 15° resolution in both azimuth and elevation. The participant had to click on one of the possible locations to give an answer. When conducting a test using VR, the test subject was placed in a virtual scene that consisted of a vast plain, with a virtual test stand placed in the middle. The stand contained a large sphere-shaped cage, with the participant placed in the middle of it (Fig. 2). The sphere-shaped cage helped with better orientation in space and corresponded to the cage used in the WEB application. Participants responded by pointing at one of the predefined locations on the grid (ray-casting) and confirming their choice by pressing a button on the virtual stand. Because the binaural scene was head-locked and did not update with head orientation, head tracking was not used in either condition, and participants were instructed to keep their heads still while listening.

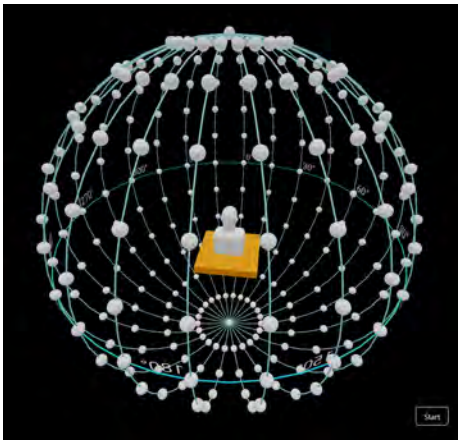


FIG. 1. View of the WEB sound-localization interface.

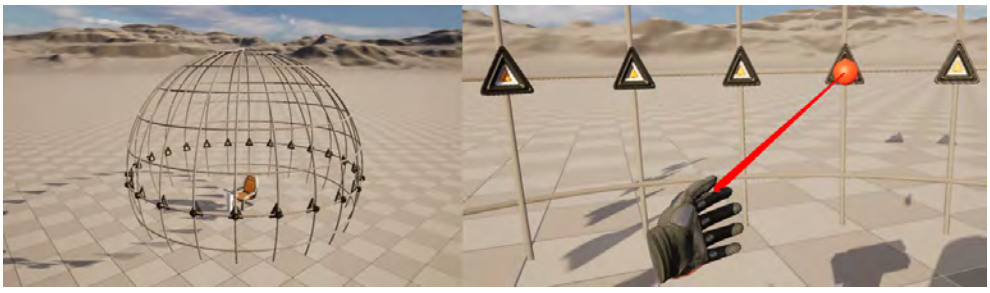


FIG. 2. View of the VR sound-localization interface.

The primary outcome was the MAAE [°], computed per participant separately for VR and WEB in the horizontal (H) and vertical (V) planes. The results were compared for VR and WEB within subjects using two-tailed paired t-tests for H, V, and the per-participant mean of $H + V$. Because each spatial position was presented only twice, the number of observations per angle was considered too small to provide reliable statistical comparisons at the level of individual spatial positions. Therefore, the statistical analysis focused on plane-level performance.

When conducting the test, special emphasis was also placed on comparing the performance of participants in the task with both approaches (does the use of VR with its innovative input method improve the ability of localizing sound over the conventional way of performing the test?) and also gathering feedback from participants regarding the methods of conducting the test – both in terms of their perceived performance in the task and their personal method of preference based on comfort, ease of use and other factors.

3. Results

Localization accuracy was assessed for 33 listeners who completed the task with both interfaces: immersive VR (head-mounted display + handheld controller) and conventional WEB (desktop display + mouse). Separate analyses were carried out for the horizontal (azimuth) and vertical (elevation) planes, followed by an overall ($H + V$) comparison.

All participants' data were analysed for the identification of potentially 'unreliable' listeners (MAAE > mean + 2SD), and although partial results of two participants exceeded the considered threshold, a closer examination revealed no significant impact of this data on the overall conclusions. Furthermore, these participants did not report any issues during the tests. Therefore, especially because one of the goals was to assess the intuitiveness and ease of use of the two interfaces, it was ultimately decided to retain the data of all participants ($N = 33$) for further analysis.

3.1. Localization in horizontal plane

The distribution of horizontal-plane localization errors (MAAE) obtained with the two test interfaces is presented in Fig. 3. Results for the test performed with the use of the VR interface show a noticeably lower median ($M = 39.9°$) and a narrower distribution of errors than for the WEB interface, for which the median $M = 46.9°$ and the upper whisker reaches almost 90°. It suggests that the listeners not only made smaller azimuth errors with the immersive headset and hand-held controller, but they also did so more consistently, whereas the mouse-based WEB interface produced a broader spread that includes several exceptionally large errors.

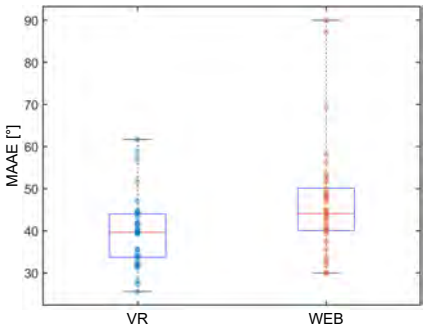


Fig. 3. Boxplots of MAAE [°] in the horizontal plane across all participants for the VR and WEB interfaces, with individual participant data points overlaid.

In the horizontal plane participants localized sounds by 7.0° more accurately in VR than on the WEB interface (Table 1). The paired-sample t-test (Table 2) revealed a statistically significant difference, ($t(32) = -2.65$, $p = 0.012$), which confirms a meaningful VR advantage in the azimuth task.

Table 1. MAAE and SD for the VR and WEB interfaces in the horizontal, vertical, and combined planes.

	MAAE [°]		SD [°]		Δ MAAE [°]
Plane	VR	WEB	VR	WEB	VR-WEB
Horizontal	39.87	46.90	8.59	13.68	−7.03
Vertical	45.44	40.97	10.62	13.66	4.47
$H + V$	42.66	43.93	5.59	11.15	−1.28

Table 2. Paired comparisons of MAAE (VR-WEB) across planes, mean difference (Δ), 95 % CI, paired t-test and Wilcoxon signed-rank statistics.

	Δ MAAE [°]	CI95_L	CI95_H	t-paired	p-value	Cohen's dz	W	p-value
Plane	VR-WEB							
Horizontal	−7.03	−12.44	−1.62	−2.65	0.012	−0.46	150.5	0.020
Vertical	4.47	−0.63	9.58	1.78	0.084	0.31	379	0.078
$H + V$	−1.28	−5.04	2.48	−0.69	0.494	−0.12	257	0.675

3.2. Localization in vertical plane

Figure 4 compares vertical-plane localization errors for the two considered interfaces. In this elevation task the pattern reverses: the WEB interface exhibits a slightly lower median MAAE (M = 41.0°) than the VR set-up (M = 45.4°), indicating that, on average, listeners judged elevation slightly more accurately with the mouse-based GUI. At the same time the WEB distribution is wider, signifying greater inter-individual variability. By contrast, the VR errors have a narrower inter-quartile range, suggesting more consistent, although less accurate, performance for the elevation angles.

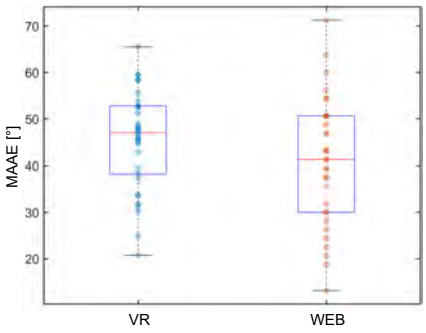


Fig. 4. Boxplots of MAAE [°] in the vertical plane across all participants for the VR and WEB interfaces, with individual participant data points overlaid.

In the vertical plane, the MAAE was by 4.5° lower for the WEB than the VR interface (Table 1). However, neither the parametric (p = 0.084) nor the non-parametric (p = 0.078) test (Table 2) reached the 0.05 significance threshold, suggesting that any vertical-plane advantage of the WEB interface is rather reflecting inter-individual variability rather than a systematic effect.

3.3. Overall localization comparison

When horizontal and vertical errors were averaged for each participant to obtain a single overall-MAAE score, the two interfaces converged. The VR interface resulted in the MAAE of 42.7° (SD = 5.6°), whereas the mouse-based WEB interface produced 43.9° (SD = 11.2°). Thus, the overall difference was −1.3°, far smaller than the typical within-listener spread. What could be more informative is the difference in the standard deviation and inter-quartile range: the distribution of errors for the VR interface is noticeably narrower. However, the overall analysis suggests that the significant horizontal advantage of VR (≈7°) is mostly offset by its moderate vertical shortfall (≈5°), leaving overall accuracy virtually unchanged yet noticeably more consistent in the immersive condition. This balance explains why statistical test failed to reach significance, and it underscores that the overall benefit of VR may lie in reducing inter-individual variability rather than in shifting the grand mean.

3.4. Perceived usability and participant experience

Immediately after completing the localization test with both interfaces, participants rated four aspects of perceived usability on 5-point Likert scales (1 = very poor, 5 = very good): intuitiveness, comfort, accuracy (perceived, meaning the confidence in one's indications), and speed. For each interface an overall usability score was also computed as the mean of these four items. Participants were also asked to choose their preference between VR and WEB interface, and leave an open comment if necessary.

The ratings show that participants favored VR over WEB across all aspects (Table 3). VR received higher ratings for intuitiveness (4.58 vs 3.94), comfort (4.45 vs 3.88), perceived accuracy (4.06 vs 3.48), and operating speed (4.09 vs 3.58). Paired t-tests (N = 33) indicated significantly higher ratings for VR on all aspects.

Based on per-participant averages, 73 % preferred VR, 18 % preferred the WEB interface, and 9 % reported no preference. In open comments participants most often praised the 'natural' aiming with the VR controller and cited 'difficulty achieving precise clicks' with the WEB cursor. Occasional remarks referred to the slight page

Table 3. Mean Likert-scale ratings (1–5) for usability aspects comparing VR and WEB; Δ = VR-WEB, with paired t-tests indicating significantly higher scores for VR on all items, including overall usability.

Aspect	VR	WEB	Δ (VR-WEB)	t-paired	p-value
Intuitiveness	4.58	3.94	0.64	3.37	0.002
Comfort	4.45	3.88	0.58	3.03	0.005
Accuracy (perceived)	4.06	3.48	0.58	2.70	0.011
Speed	4.09	3.58	0.52	2.52	0.017
Overall usability score	4.30	3.72	0.576	4.62	<.001

latency or headset pressure. These findings support VR as the preferred interaction modality for spatial hearing experiments, particularly when naturalistic pointing and efficient supervision are a priority.

4. Discussion

Two interaction modalities used in spatial sound localization experiments were compared: a mouse-driven WEB interface and an immersive VR interface using a handheld controller. While both conditions relied on identical binaural stimuli and headphones, the response method differed. The results demonstrate that the VR interface improved localization accuracy in the horizontal plane, whereas no statistically significant advantage was observed in the vertical plane.

The improvement in horizontal localization accuracy when using the VR interface may be explained primarily by the more natural response modality. In the immersive condition, participants indicated the perceived sound direction by pointing with a tracked controller, which closely resembles natural reaching and pointing actions. Previous studies have shown that embodied interaction can facilitate spatial perception tasks and reduce cognitive load associated with translating auditory perception into a response (Valzolgher *et al.*, 2023). In contrast, the The WEB interface required participants to translate a perceived direction into cursor movement on a 2D display, which introduces an additional visuomotor mapping step. Such indirect interaction may increase response uncertainty and contribute to larger angular errors.

This advantage of the VR interface was observed only for the horizontal plane. In the elevation task, the WEB interface yielded slightly lower mean errors, although the difference was not statistically significant. One possible explanation is that vertical localization relies strongly on spectral cues encoded in the head-related transfer function (HRTF). Because the present experiment used a generic HRTF, listeners may have experienced front-back or elevation ambiguities that cannot be easily compensated by improved interaction methods. Previous research has shown that elevation localization remains challenging when using non-individualized HRTFs, even in immersive environments. Another relevant factor is the absence of dynamic auditory cues. In the present implementation the binaural scene was head-locked and did not update with head movements. Head motion is known to provide additional information that helps disambiguate elevation and front-back confusions. Studies in spatial hearing have shown that even small head movements can significantly improve localization performance by providing dynamic changes in interaural cues and spectral filtering. Because head tracking was intentionally disabled in this experiment to maintain identical acoustic rendering across conditions, participants could not benefit from these dynamic cues. This limitation may have particularly affected elevation judgments.

Despite the limited improvement in objective accuracy, the subjective evaluations clearly favored the VR interface. Participants rated VR higher in terms of intuitiveness, comfort, perceived accuracy, and operating speed, and the majority expressed a preference for the immersive condition. These findings suggest that response method that resembles natural pointing may reduce task difficulty and increase participant engagement, which can be beneficial for large-scale listening studies. The results also suggest that the VR interfaces may produce a narrower distribution of errors among participants, indicating more consistent performance.

Several limitations of this study should be acknowledged. First, the use of generic HRTFs may have limited elevation accuracy and introduced perceptual ambiguities. Second, the binaural rendering was head-locked, preventing the use of dynamic auditory cues that are known to improve spatial perception. Third, the participant

sample consisted mainly of young non-expert listeners, which may limit the generalizability of the results. Finally, the number of tested spatial positions was relatively small and constrained to discrete grid points. Future work should investigate localization performance in VR environments that include head-tracked dynamic binaural rendering and individualized HRTFs where possible. Such improvements could allow participants to utilize natural head movements and more accurate spectral cues, potentially leading to improved elevation localization. Additionally, larger and more diverse participant groups would help determine whether the advantages of the VR interfaces generalize across populations and experimental contexts.

5. Conclusions

This study compared an immersive virtual-reality interface with a conventional mouse-driven WEB application for spatial sound-localization tasks. The VR condition reduced the horizontal mean-absolute error by 7° (≈15 %). No statistically significant advantage was found in the vertical plane, suggesting that elevation cues still pose a challenge when rendered with a generic HRTF. Subjective ratings favored VR on intuitiveness, comfort, speed and perceived accuracy, with 73 % of participants preferring the immersive setup.

While the findings confirm that natural pointing with motion controllers can improve the ergonomics of psychoacoustic testing, the overall localization accuracy remained limited and the vertical performance did not improve. Future work should examine the use of individualised HRTFs where possible and include larger, more diverse samples to validate the observed horizontal-plane benefit. Because in this study the VR implementation was head-locked, which is also a limitation, future works could enable world-stabilized, head-tracked micromovements to quantify the expected gain in elevation accuracy and test whether VR's advantage grows when dynamic cues are available. An optimised VR protocol could provide researchers with a scalable, supervisor-independent platform for high-quality spatial-hearing experiments.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' contributions

Agnieszka Paula Pietrzak conceptualized the study, supervised the project, and wrote the original draft. Grzegorz Rusinek developed the software (WEB/VR implementation). Both authors designed the methodology and the experiment, curated the data, performed the formal analysis, and prepared the figures and tables, reviewed and edited the manuscript. All authors have accepted the responsibility for the entire content of this manuscript and approved its submission.

References

1. Ahrens A., Lund K.D., Marschall M., Dau T. (2019), Sound source localization with varying amount of visual information in virtual reality, *PLOS ONE*, **14**(3): e0214603, https://doi.org/10.1371/journal.pone.0214603.
2. Bronkhorst A.W. (1995), Localization of real and virtual sound sources, *The Journal of the Acoustical Society of America*, **98**(5): 2542–2553, https://doi.org/10.1121/1.413219.
3. Fujii A., Ohsugi Y., Yamamoto Y., Nakamura T., Sugiura T., Tauchi M. (2007), Comparison of different pointing methods for sound localizability measurement in the vision impaired subjects, *Journal of Physiological Anthropology*, **26**(3): 381–385, https://doi.org/10.2114/jpa2.26.381.

4. Haber L., Haber R.N., Penningroth S., Novak K., Radgowski H. (1993), Comparison of nine methods of indicating the direction to objects: data from blind adults, *Perception*, **22**(1): 35–47, https://doi.org/10.1068/p220035.
5. Mason R., Ford N., Rumsey F., de Bruyn B. (2001), Verbal and nonverbal elicitation techniques in the subjective assessment of spatial sound reproduction, *Journal of the Audio Engineering Society*, **49**(5): 366–384.
6. Oldfield S.R., Parker S.P.A. (1984), Acuity of sound localisation: a topography of auditory space. I. Normal hearing conditions, *Perception*, **13**(5): 581–600, https://doi.org/10.1068/p130581.
7. Valzolgher C., Capra S., Sum K., Pavani F., Picinali L. (2023), Training spatial hearing skills in virtual reality through a sound-reaching task, [in:] *Forum Acusticum 2023* (Torino, Italy), https://doi.org/10.61782/fa.2023.0341.

Archives of Acoustics Vol. 51, No. 3, pp. 515–532 (2026), https://doi.org/10.24423/archacoust.2026.4269

Research Paper

Adversarial Audio Inpainting with Selective State Space Model, Efficient Attention, and Large-Scale Pre-Trained Model

Junkang Yang[(1)], Hongqing Liu[(2)*], Liming Shi[(3)], Lu Gan[(4)], Hiromitsu Nishizaki[(1)], Chee Siang Leow[(1)]

[(1)] *Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences, University of Yamanashi* Kofu, Japan

[(2)] *Chongqing Key Lab of Mobile Communications Technology, Chongqing University of Posts and Telecommunications* Chongqing, China

[(3)] *School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications* Chongqing, China

[(4)] *College of Engineering, Design and Physical Science, Brunel University* London, United Kingdom

*Corresponding Author: hongqingliu@cqupt.edu.cn

Received June 25, 2025; revised March 14, 2026; accepted April 12, 2026; available online April 15, 2026; version of record July 7, 2026; published issue September 23, 2026.

Recent deep-learning based speech enhancement algorithms have many applications in the areas of noise reduction, de-reverberation, bandwidth extension, echo cancellation, to name a few. Packet loss is also one of the main causes of voice quality degradation in Voice over Internet Protocol (VoIP) calls. Currently, generative adversarial networks (GANs) have shown a strong ability in image generation, and many of those models also work well in speech tasks. In this work, we propose a lightweight model based on GAN to handle the task of audio packet loss concealment. Specifically, we use a U-shaped network operating in the time-frequency domain as a generator, which is trained by the Mel-GAN discriminator with multi-loss. In addition, to enhance the model's performance under unfavorable channels, we introduce noise and bandwidth loss in the training data. The experiments show that our method outperforms the baseline in both objective and subjective metrics under an ideal channel with no other distortions, and it still largely maintains its performance in the presence of noise and bandwidth loss.

Keywords: adversarial learning, audio inpainting, selective state space model, attention mechanism.



Copyright © 2026 The Author(s).
This work is licensed under the Creative Commons Attribution 4.0 International CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Speech signals are usually broken down into frames for transmission in the channel, and these frames can be acknowledged as data packets during transmission. These packets are often transmitted in a communication system in a disordered fashion and are reassembled into an ordered sequence at the receiving end. When there is interference in the channel or the communication link experiences interruptions, it will produce missing packets or high packet jitter, causing the degradation of speech quality. In applications such as mobile digital communications, videoconferencing systems, and Voice over Internet Protocol (VoIP) calls, packet loss can dramatically affect the quality of speech to the extent that people cannot understand semantic information from intermittent speech. At the same time, due to the variety of distortion types in the channel and the high real-time nature of these scenarios, it is a challenging task to consider both robustness and lightweight when repairingral networks.

Packet loss concealment (PLC) is a technology to reconstruct the lost part of speech using existing information (Mohamed, Schuller, 2020). Some techniques refer to a similar task with the terms audio inpainting (Miotello *et al.*, 2024), waveform interpolation (Lagrange *et al.*, 2005), or extrapolation (Maher, 1994; Aironi *et al.*, 2023). Early PLC methods were mainly based on statistical modeling, such as hidden Markov models (HMMs) (Rodbro *et al.*, 2006) or different coding approaches (Janicki, Księżak, 2008). As learning-based methods continue to improve, the performance of these methods gradually becomes obsolete. Using deep learning techniques, Lee and Chang (2015) first proposed a DNN-based model to conceal the degradation caused by lost packets, generating the estimated log-power spectra and phases of missing frames and fixing the waveform by putting these features into a decoder. In the subsequent works, recurrent network structures and convolutional modules have been widely used for this task (Mohamed, Schuller, 2020; Lotfidereshgi, Gournay, 2018; Lin *et al.*, 2021; Davy *et al.*, 2023). These methods can better capture all sorts of dependencies between samples and improve the overall performance. In recent years, many generative models such as GAN and diffusion models have achieved good performance in the field of speech enhancement (Fu *et al.*, 2021; Yen *et al.*, 2023). In the PLC task, the GAN based model can allow the generator to achieve better performance, compared to autoregressive algorithms, with less parameters through adversarial training (Ebner, Eltelt, 2020). For example, Zhao (2023) uses a GAN network to repair the magnitude together with a phase reconstruction algorithm, implementing the speech inpainting for audio editing software. Diffusion models are often applied to audio joint task including PLC or inpainting. As in the study by Moliner *et al.* (2023), in which issues related to bandwidth extension, de-clipping, and inpainting problems with a general diffusion model are solved and convincing results are obtained, unfortunately, it seems not suitable for speech and real-time communication systems. In Interspeech 2022, Microsoft organized the 1st Audio Deep Packet Loss Concealment Challenge (Diener *et al.*, 2022) and it received many effective models (Li *et al.*, 2022; Liu *et al.*, 2022; Westhausen, Mayer, 2022; Valin *et al.*, 2022; Guan *et al.*, 2022), including GAN, RNN, and the deep learning methods mixed with traditional ways. This challenge complements this research field, and it is a great reference for later works.

Currently, the difficulties in the PLC task still exist. When expecting a stronger capability in loss packet reconstruction, more parameters are required in most instances. However, the deployment of the model on the resource-limited devices is so sensitive to this aspect. Furthermore, signals are facing various types of loss in real-world channels. When other distortions exist together with packet loss, most of the systems will have a large degradation in performance, and most of the related work does not evaluate the performance of the model in real-world scenarios. Besides, after front-end speech enhancement algorithms come various downstream tasks such as automatic speech recognition (ASR), speech emotion recognition (SER), and multi-model large language models. Improving effectiveness in front-end processing for downstream tasks is also an issue that needs attention.

To handle these problems, this work proposes the GAN-based model to conceal the packet loss. Our main contributions are as follows:

1. By introducing a state space model and efficient local attention in the generator, our method achieves a better performance than the baselines while remaining lightweight.
2. By using ASR and speech quality evaluation pretrained models to generate the loss function at the training stage, we improve the model's performance in downstream ASR tasks and subjective evaluation experiments.
3. We constructed training data containing bandwidth loss and noise, which is closer to the real-world situation. Experiments show that our model trained on these data presents a higher robustness in the real-world data.

2. Proposed method

Figure 1 illustrates the GAN proposed in this paper, which is primarily composed of a generator and a discriminator. The generator is responsible for transforming input spectrograms or waveforms into enhanced outputs, while the discriminator evaluates the quality of the generator's outputs and generates multi-loss to optimize the entire network.

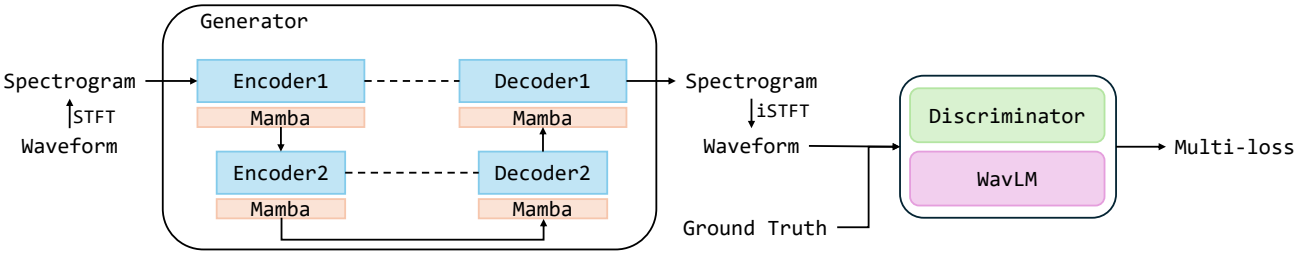


FIG. 1. Structure of proposed generative adversarial network.

The generator is the core component of the network, accepting spectrograms as input, where it is first converted into spectrograms via the short-time Fourier transform (STFT). The spectrograms are then fed into two separate encoder-decoder paths within the generator. Each encoder and decoder incorporates a Mamba block for feature modeling and fusion. At the end, the output spectrogram is converted back into a waveform via inverse short-time Fourier transform (iSTFT). It is worth mentioning that there is a skip connection between each encoder and its corresponding decoder.

The discriminator is also a crucial component of this network, whose primary function is to distinguish between fake waveforms and real waveforms. It takes the output waveforms from the generator and real waveforms as input, analyzes and evaluates the input data through a multi-layer neural network structure, and outputs a probability value or feature representation indicating the authenticity of the input data. The discriminator provides the generator with an adversarial loss function, encouraging the generator to continuously optimize the quality of the generated waveforms, making them closer to the distribution and characteristics of real waveforms.

Furthermore, WavLM (CHEN *et al.*, 2022) is a transformer (VASWANI *et al.*, 2017) based on a pre-trained audio language model which serves as an auxiliary discriminative component in this system, working alongside the discriminator as part of a multi-loss calculation framework. WavLM conducts in-depth semantic and feature-level analysis of input waveforms, leveraging its pre-trained knowledge and feature extraction capabilities acquired from large-scale audio data to evaluate the differences and similarities between generated and real waveforms from multiple perspectives. This provides richer supervisory signals during model training, helping the generator produce waveforms that better align with the semantic and acoustic characteristics of real audio. By enhancing both the quality and naturalness of the generated audio, WavLM enables the entire model to holistically consider various factors such as waveform authenticity and semantic features in audio generation tasks.

The entire workflow is as follows. The input waveform is converted into a spectrogram via STFT, and the spectrogram is fed into the generator. After feature processing by the encoder, the decoders reconstruct the spectrogram. Then the output spectrogram is converted back into a waveform via iSTFT. The generated waveform and the ground-truth waveform are then served as inputs into the discriminator and the pre-trained model to compute the adversarial loss and audio quality loss, respectively. The total loss is used to update the parameters of the generator through backpropagation and gradient descent, progressively improving the generator's output quality.

Next we introduce each component of the network in detail.

2.1. GENERATOR

The generator mainly contains encoders, decoders and state space model blocks (Mamba).

2.1.1. ENCODER AND DECODER

The encoder–decoder in Fig. 2 is designed under 3 practical constraints of PLC. First, PLC requires restoring locally time-frequency patterns such as harmonics. These features are typically disrupted by burst losses. So we use lightweight convolutions as the main building modules to efficiently capture local structures with low latency. Besides, accurate inpainting also needs contextual cues beyond a short receptive field, especially when consecutive packets are missing; therefore, we need to enhance cross-frequency interactions and long-context

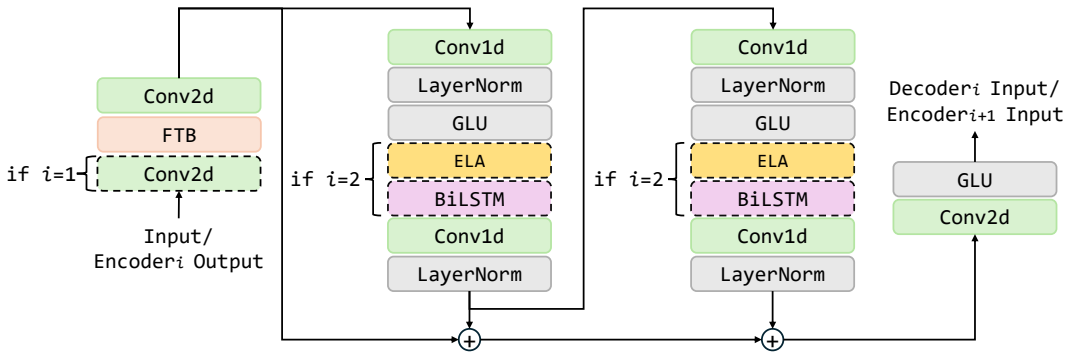


FIG. 2. Structure of encoder.

modeling at the bottleneck, rather than stacking heavy global attention everywhere. Additionally, the model is intended for real-time or resource-limited deployment, so each additional module must provide clear gains per computational cost.

Based on these considerations, we adopt a U-Net style encoder–decoder with skip connections to preserve fine-grained spectral details and to reduce over-smoothing during reconstruction. As illustrated in Fig. 2, the encoder first applies 2D convolution to extract local time-frequency features, followed by a frequency transformer block (FTB) (DAI *et al.*, 2024) to strengthen frequency-wise dependency modeling, since packet loss often breaks spectral continuity across frequency bins. When operating in the first encoder layer ($i = 1$), an additional 2D convolution and GLU activation are used to selectively filter informative channels and suppress artifacts introduced by lossy segments. Subsequently, 1D convolution and layer normalization are employed to capture short-term temporal correlations and stabilize training. In the second encoder layer ($i = 2$), we introduce an efficient local attention (ELA) method (XU, WAN, 2024) together with a bidirectional long short-term memory (BiLSTM) near the bottleneck, where the temporal resolution is already reduced; this design allows the model to capture salient regions and longer-range dependencies with minimal computational overhead. The processed features are then combined with earlier-stage representations through skip connections, mitigating gradient vanishing and facilitating multi-level feature fusion.

ELA is an efficient local attention mechanism for deep convolutional neural networks. It improves upon coordinate attention by using 1D convolution and group normalization to enhance feature representation, avoiding the generalization issues of batch normalization in coordinate attention and the negative effects of channel dimension reduction on attention generation. Specifically, as Fig. 3 depicts, ELA first employs strip pooling to

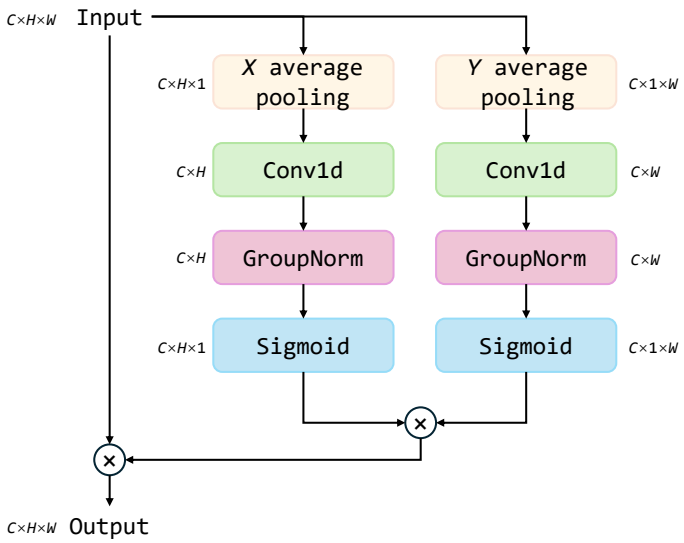


FIG. 3. Efficient local attention.

extract 1D feature vectors in the horizontal and vertical directions from the input feature map. Given the input tensor $X \in \mathbb{R}^{C\times H\times W}$, the operations are:

$$X_h(c,w) = \frac{1}{H}\sum_{i=1}^{H} X(c,i,w), \tag{1}$$

$$X_w(c,h) = \frac{1}{W}\sum_{j=1}^{W} X(c,h,j), \tag{2}$$

where $X_h \in \mathbb{R}^{C\times W}$ and $X_w \in \mathbb{R}^{C\times H}$ are 2D feature maps (matrices) obtained by average pooling along the height and width dimensions, respectively. Here, $c \in \{1,...,C\}$ denotes the channel index, $h \in \{1,...,H\}$ and $w \in \{1,...,W\}$ denote the spatial indices along height and width, respectively.

After pooling, a 1D convolution is applied along the remaining spatial dimension to capture local contextual interactions. Specifically, the convolution operates along the width dimension for X_h and along the height dimension for X_w, while being applied independently for each channel. After that, group normalization and Sigmoid activation functions are used to process the feature maps, generating positional attention predictions for both directions, given by

$$Y_h = \sigma(\mathrm{GN}(\mathrm{Conv}(X_h))), \tag{3}$$

$$Y_w = \sigma(\mathrm{GN}(\mathrm{Conv}(X_w))), \tag{4}$$

where $\sigma(\cdot)$, $\mathrm{GN}(\cdot)$, $\mathrm{Conv}(\cdot)$ represent the Sigmoid fuction, group normalization, and 1D convolution, respectively. Finally, the predictions from both directions are combined via a product operation to obtain the final positional attention map. That is,

$$Y = X \cdot Y_h \cdot Y_w, \tag{5}$$

where $\cdot$ denotes elementwise multiplication. The attention maps Y_h and Y_w are broadcast along the height and width dimensions, respectively, before being applied to the input feature map X.

ELA can accurately locate regions of interest without reducing the channel dimension, thus avoiding the loss of feature information caused by the channel dimension reduction. Additionally, ELA has a lightweight structure with low computational complexity. For example, in the medical image detection task (Sun *et al.*, 2025; Hao *et al.*, 2024), the model using ELA as the core component effectively improved the accuracy while having fewer parameters and computational complexity than the baselines. ELA also shows good generalization ability, stably enhancing the performance of both large and small networks. For instance, it significantly improves the accuracy of ResNet series networks (He *et al.*, 2016) with minimal impact on model parameters (Xu, Wan, 2024). These conclusions can also be verified in our experiments.

2.1.2. State space model block

Packet loss concealment requires a model to infer missing speech segments from the surrounding context, particularly under burst loss conditions where consecutive frames are unavailable. While convolutional layers are effective at modeling local structures, their receptive field grows slowly with depth, and recurrent or attention-based modules often incur high computational costs. To address this challenge, we introduce a selective state space model (SSM) based on the Mamba architecture (Gu, Dao, 2023) as a core temporal modeling component in the generator. Its detailed description is shown in Fig. 4.

The key motivation for using Mamba lies in its ability to model long-range temporal dependencies with linear computational complexity. Unlike Transformer-based attention, whose cost scales quadratically with the sequence length, the state space formulation allows efficient propagation of contextual information across long time spans, which is particularly important for reconstructing speech segments affected by continuous packet losses. Moreover, the selective mechanism in Mamba enables the model to dynamically control information flow, allowing it to emphasize informative regions while attenuating the less relevant context.

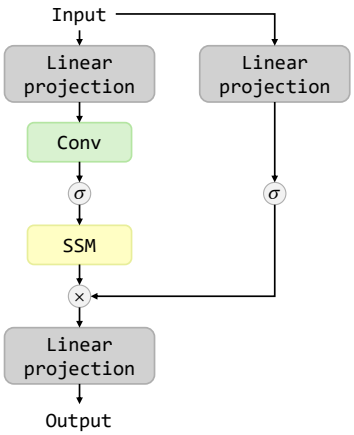


FIG. 4. Mamba block.

In our architecture, the Mamba block is placed within the encoder–decoder pipeline to complement convolutional feature extraction. While convolutional layers and ELA focus on local time-frequency structures, the Mamba block aggregates long-range temporal information across frames, thereby improving the continuity and naturalness of reconstructed speech. This design allows the proposed model to effectively handle both short and long packet losses while maintaining a favorable trade-off between performance and computational efficiency.

2.2. DISCRIMINATOR

Inspired by YANG *et al.* (2024), we employed a multi-scale discriminator architecture which is shown in Fig. 5 as the core component of the adversarial training framework to optimize the generator's performance. This discriminator consists of three independent sub-discriminators $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$, each designed with inspiration from MelGAN (KUMAR *et al.*, 2019). Specifically, each sub-discriminator comprises seven convolutional layers, with the first four layers featuring downsampling capabilities to progressively compress the spatiotemporal resolution of the input signal and capture multi-scale features. The input to the discriminator undergoes differentiated processing: $\mathcal{D}_1$ receives the raw waveform, $\mathcal{D}_2$ processes a 2 times downsampled waveform, and $\mathcal{D}_3$ analyzes a 4 times downsampled waveform. This design enables the generator to maintain consistency across different temporal scales, thereby enhancing the naturalness of the output speech and its spectral coherence. Furthermore,

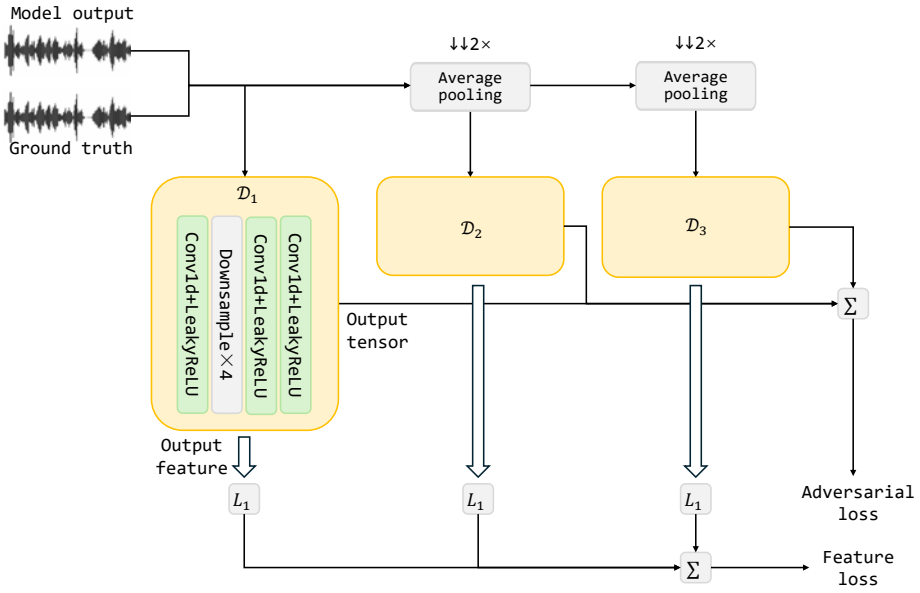


FIG. 5. Structure of discriminator.

the multi-level feature output by the discriminator are used to compute the adversarial loss and feature loss. The former optimizes the generator by minimizing the difference between generated and real samples, while the latter constrains the generator to align its intermediate-layer feature distribution with that of real data.

2.3. WavLM

The architecture of WavLM (Chen *et al.*, 2022) is shown in Fig. 6, and it is composed of a CNN encoder and a transformer encoder, with the latter incorporating gated relative position bias (Chi *et al.*, 2022) to better capture the sequence order of input speech. The CNN encoder is tasked with extracting features from the input waveform. It consists of 7 temporal convolutions, each followed by layer normalization and a GELU activation layer. The convolutional kernels have 512 channels, with strides of (5, 2, 2, 2, 2, 2, 2) and widths of (10, 3, 3, 3, 3, 2, 2), ensuring that each output represents approximately 25 ms of audio with a stride of 20 ms. The features x extracted by the CNN are masked and then fed into the transformer encoder as input. At the bottom of the transformer encoder is a convolution-based relative position embedding layer (Raffel *et al.*, 2020) with a kernel size of 128 and 16 groups. To enhance model performance, WavLM introduces gated relative position bias, which is encoded based on the offset between the key and the query in self-attention. Specifically, for the input hidden states $\{\mathbf{h}_i\}_{i=1}^{T}$, each $\mathbf{h}_i$ is linearly projected into a query, key, and value $(\mathbf{q}_i, \mathbf{k}_i, \mathbf{v}_i)$. Here, $\mathbf{h}_i \in \mathbb{R}^D$ denotes the hidden representation at a time step i output by the WavLM transformer encoder, where D is the hidden dimension of the pre-trained model. In our implementation, we use WavLM base, for which $D = 768$. The self-attention outputs are calculated by incorporating attention logits with gated relative position bias r_{i-j}. The calculations are:

$$g_i^{(\text{update})}, g_i^{(\text{reset})} = \sigma(\mathbf{q}_i \cdot \mathbf{u}), \sigma(\mathbf{q}_i \cdot \mathbf{w}), \tag{6}$$

$$\widetilde{r}_{i-j} = \mathbf{w} \cdot g_i^{(\text{reset})} \cdot d_{i-j}, \tag{7}$$

$$r_{i-j} = d_{i-j} + g_i^{(\text{update})} \cdot d_{i-j} + (1 - g_i^{(\text{update})}) \cdot \widetilde{r}_{i-j}, \tag{8}$$

where $\mathbf{u}$ and $\mathbf{w}$ are the learnable parameter tensors with the same dimensions as the query representation $\mathbf{q}_i$, the dot product in Eq. (6) denotes an inner product that produces scalar gating values, and $\sigma(\cdot)$ represents sigmoid function. The gating mechanism allows the relative position bias to be dynamically adjusted based on the current speech content. Additionally, d_{i-j} is calculated using bucketed relative position embeddings, with its range varying according to the offset. Consequently, WavLM learns rich speech representation capabilities through its pre-training framework of masked speech denoising and prediction, enabling it to capture multi-level information in speech signals, such as content, speaker characteristics, and environmental noise.

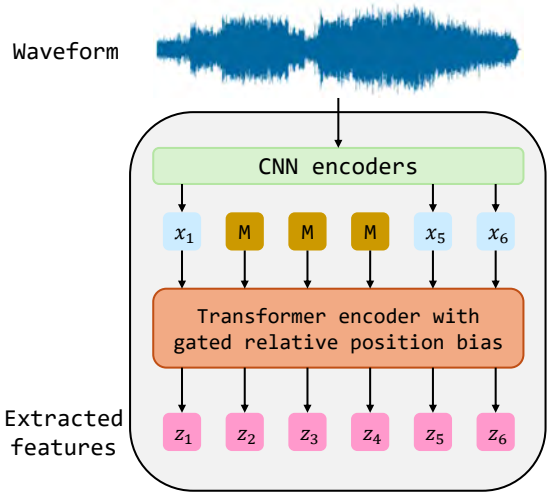


Fig. 6. Structure of WavLM.

In our task, using the deep features extracted by WavLM to compute the loss function can guide the model to not only focus on waveform-level matching but also optimize higher-level semantic and perceptual characteristics of speech. The WavLM-based loss can measure the distribution difference between enhanced speech and clean speech in the latent space, thereby avoiding the over-smoothing issues that may arise from low-level metrics, resulting in more natural enhanced speech with better-preserved details.

2.4. Loss

The loss function of our model adopts an adversarial training approach, combining multiscale STFT loss, generator adversarial loss, feature loss, and pre-trained model loss. The total loss function can be expressed as

$$\mathcal{L} = \mathcal{L}_{\text{MSTFT}} + \mathcal{L}_{\text{adv}}^{\mathcal{G}} + \lambda_f \mathcal{L}_f + \mathcal{L}_{\text{pre}}, \tag{9}$$

where $\lambda_f = 100$, which is used to balance the weight of feature loss.

The multi-scale STFT loss L_{MSTFT} is

$$\mathcal{L}_{\text{MSTFT}} = \mathbb{E}_{(x,y)\sim p_{\text{data}}}\left[\sum_{m=1}^{3}\left(\frac{\|s(y,\theta_m) - s(\mathcal{G}(x),\theta_m)\|_F}{\|s(y,\theta_m)\|_F} + \frac{1}{N}\left\|\log\frac{s(y,\theta_m)}{s(\mathcal{G}(x),\theta_m)}\right\|\right)\right], \tag{10}$$

where $s(\mathcal{G}(x),\theta_m)$ represents the STFT magnitude spectrum of the output signal $\mathcal{G}(x)$ under the parameter θ_m, $\|\cdot\|_F$ and $\|\cdot\|$ denote the Frobenius norm and ℓ_1 norm, respectively, and N is the number of elements in the magnitude spectrum. The multi-scale parameters θ_m cover different FFT sizes (512, 1024, 2048) and hop length (50, 120, 240). The logarithm in Eq. (10) is elementwise.

The generator adversarial loss $L_{adv}^{\mathcal{G}}$ and feature loss L_f are:

$$\mathcal{L}_{\text{adv}}^{\mathcal{G}} = \mathbb{E}_{x\sim p_x}\left[\frac{1}{K}\sum_k \max(0, 1 - \mathcal{D}_k(\mathcal{G}(x)))\right], \tag{11}$$

$$\mathcal{L}_f = \mathbb{E}_{(x,y)\sim p_{\text{data}}}\left[\frac{1}{KL}\sum_{k,l}\|\mathcal{D}_k^l(y) - \mathcal{D}_k^l(\mathcal{G}(x))\|_1\right], \tag{12}$$

where p_x is the distribution of lossy speech, p_{data} denotes the empirical joint distribution of lossy speech and clean speech from the training dataset, K is the number of discriminators ($K = 3$), L is the number of layers in the discriminator, and $\mathcal{D}_k^l$ denotes the feature output of the l-th layer in the k-th discriminator. This loss design achieves noise-robust bandwidth extension by jointly optimizing time-frequency domain reconstruction accuracy (STFT loss) and adversarial training (adversarial loss and feature loss).

The pre-trained model loss $\mathcal{L}_{\text{pre}}$ is the KL divergence between the output features of prediction x and ground truth y from WavLM.

2.5. Datasets

We used clean speech data and noise data from 2023 DNS challenge (Dubey *et al.*, 2024) and VCTK (Yamagishi *et al.*, 2019), and WHAM! (Wichern *et al.*, 2019) in the training phase of the model. We first added noise to the clean speech data with a probability of 0.5, making the SNR of speech in the range of −10 dB to 0 dB, and then randomly masked it to simulate the packet loss effect, with a mask window length of $t_{\text{win}} \in [20, 120]$ [ms]. We synthesized a total of 300 000 samples, each of which was normalized to a length of 5 s and resampled to 16 kHz, with a total duration of 416.67 h. During training, we used 90 % of the data as the training set and 10 % as the validation set.

In the testing phase, we used the blind test set from the 2022 PLC challenge (Diener *et al.*, 2022) and the last 5 speakers in the VCTK dataset (Yamagishi *et al.*, 2019) for testing. For the VCTK test data, we set the duration of lost packets to a multiple of 20 ms, and the maximum packet loss time length to 120 ms,

equivalent to 6 consecutive packets. The distribution of lost parts when the packet loss rate is in the range of 10 % to 40 % is completely based on the settings in (Aironi *et al.*, 2023). In addition, we used real speech collected from the wireless communication system of marine ships to evaluate the performance of the model in real scenarios. According to the coarse estimation, in this dataset, the packet loss rates approximately range from 0 % to 40 % with burst durations ranging from 0 ms to 100 ms. We made this part of the data publicly available[1].

2.6. Detailed configurations

During model training, we employ the Adam optimizer (with β_1 = 0.8 and β_2 = 0.999) to jointly optimize the generator and discriminator under a fixed learning rate of 1×10^{-3}. The training is conducted on 2 NVIDIA RTX3090 GPUs for 100 epochs, and the checkpoint achieving the best performance on the validation set is selected for testing. For the STFT configuration in our network, the FFT size is set to 512 with a hop length of 128. In addition, we used the WavLM base pre-trained model provided by the official repository[2] to generate pre-trained model loss.

2.7. Evaluation metrics

In the evaluations of the model, we selected a set of multidimensional metrics to comprehensively measure speech quality, comprehensibility, and recognition performance. Specifically, we conducted tests on the following metrics in the experiments.

DNSMOS (Reddy *et al.*, 2022) is an objective speech quality evaluation index based on neural networks, which simulates human subjective auditory perception using deep networks. PLCMOS (Diener *et al.*, 2023), like DNSMOS, is based on neural networks. The difference is that it focuses on evaluating the quality of packet loss concealment in communication scenarios, modeling through perceptual linear prediction coefficients and auditory masking effects. Its calculation process considers the frequency band energy distribution and psychoacoustic characteristics. The range of values for these two metrics is 1 to 5, and they are both non-intrusive, which means that no reference signal is required during the evaluation.

Additionally, we employed the PESQ (Rix *et al.*, 2001), a widely used objective metric in the field of speech enhancement. PESQ is an objective speech quality assessment method based on the human auditory model, primarily used to measure the perceptual quality of speech signals after processes such as encoding or decoding, transmission, or noise reduction. Its output scores typically range from −0.5 to 4.5, with higher scores indicating better quality. We also used STOI (Taal *et al.*, 2011), which focuses on the objective evaluation of speech intelligibility. By analyzing the time-frequency characteristics of speech signals, STOI predicts how easily a listener can understand the speech, with scores ranging from 0 to 1, and a score closer to 1 indicates higher intelligibility. It is worth mentioning that both metrics require reference signals for testing.

On the other hand, we randomly selected 25 native English-speaking males and 25 native English-speaking females online and asked them to listen to speech signals processed by different models to obtain mean opinion score (MOS). This subjective evaluation follows the MOS listening test in accordance with the general guidelines of ITU-T Recommendation P.800. MOS is a subjective evaluation metric where human listeners rate speech quality based on auditory perception, typically on a scale from 1 (poor) to 5 (excellent), directly reflecting subjective human perception. The average of all individual scores was taken as the final metric.

In order to measure the impact of different models on downstream tasks, we use speech processed by these models for ASR and use the word accuracy (WAcc) as an indicator. We use the same ASR model for all methods.

2.8. Baseline models

We tested the proposed model and all the models submitted to the PLC Challenge (Diener *et al.*, 2022) on the blind test set provided by the challenge, and all training and testing conditions were kept consistent.

[1] https://drive.google.com/file/d/182xMlO768kH3ntOO8C˙vNIOqYZ0IR4cu/view?usp=sharing
[2] https://github.com/microsoft/unilm/tree/master/wavlm

In addition, we compared 3 traditional algorithms: Opus (Valin *et al.*, 2016), WebRTC (Blum *et al.*, 2021), EVS (Lecomte *et al.*, 2015), and several deep learning-based models such as PLAAE (Pascual *et al.*, 2021), TFGAN (Wang *et al.*, 2021), bin2bin (Aironi *et al.*, 2023) based on generative adversarial networks, a model based on recurrent neural networks (Lotfidereshgi, Gournay, 2018), and the MFM pre-training method and the model in (Yang *et al.*, 2023) on VCTK and our collected speech datasets. We adopted the settings claimed in the papers of these models and retrained and reimplemented them when necessary. We also referred to the experimental results in (Aironi *et al.*, 2023).

3. Experiment results

Table 1 presents a comparison of the performance of different models based on the 2022 PLC challenge blind test data, focusing on key metrics such as PLCMOS, DNSMOS, CMOS, and WAcc, where CMOS is defined by

$$\text{CMOS} = \text{MOS} - \text{MOS}_{\text{clean}}. \tag{13}$$

Table 1. Performance comparison based on 2022 PLC challenge blind test data.

Team ID*	PLCMOS	DNSMOS	CMOS	WAcc	#Parameters
Clean (reference)	4.51	3.89	0	0.98	–
#12	4.28	3.80	–0.55	0.85	2.36
#1	3.74	3.79	–0.64	0.84	5.9
#9	3.83	3.68	–0.81	0.80	–
#14	3.98	3.69	–0.84	0.79	4.97
#11	3.28	3.51	–1.10	0.75	–
#10	3.48	3.74	–1.04	0.74	–
#6	2.90	3.48	–1.31	0.71	3.77
Proposed model	**4.34**	**3.85**	**–0.42**	**0.89**	**1.71**

*Since the methods and names of most models in this challenge are unknown, team numbers are used to represent the models.

In PLCMOS, which evaluates the quality of inpainting, the score of clean speech is 4.51. The proposed model achieves 4.34, the closest to the reference among all teams, significantly outperforming others like #12 at 4.28 and #1 at 3.74. This highlights the model's ability to maintain signal quality close to the ideal reference. For DNSMOS, measuring overall audio quality, the reference's score is 3.89. The proposed model scores 3.85, again leading other teams such as #9 at 3.68 and #14 at 3.69. This indicates the model's effectiveness in preserving audio quality. As for CMOS, which indicates the subjective feeling of human beings' distortion, the reference is 0. The proposed model's score of –0.42 is much closer to zero compared to teams like #6 at −1.31 and #11 at −1.10, showing a better experience brought to the listeners. Finally, in WAcc, the ASR performance metric, the reference is 0.98, and the proposed model attains 0.89, surpassing other teams like #12 at 0.85 and #1 at 0.84. This shows the proposed model demonstrates superior performance across all key metrics, effectively balancing signal quality, audio fidelity, and minimal distortion, making it the best solution among these models. With a parameter number of 1.71M, our model achieves a PLCMOS score of 4.34, a DNSMOS score close to 3.85, and an impressive WAcc of 0.89, outperforming all other models across these three dimensions. Besides, model #1, with a parameter number of 5.90M, which is nearly three times that of our model, only reaches a WAcc of 0.85, while its PLCMOS and DNSMOS scores show no significant advantage. Model #12, with 2.36M parameters, which is still 0.65M more than ours, achieves a WAcc of approximately 0.85, and its PLCMOS and DNSMOS performance is also lower than our model. Model #14, with 4.97M parameters, also fails to match our model's levels in PLCMOS, DNSMOS, and WAcc. Similarly, model #6, with 3.77M parameters, lags behind our model in all three key metrics.

Table 2 compares the average performance of different methods on the VCTK dataset, including traditional methods and learning-based approaches, with performance evaluated using PESQ and STOI metrics. Among the traditional methods, EVS achieves the best performance with a PESQ of 1.89 and an STOI of 0.78. Among the

TABLE 2. Average performance comparison based on VCTK data.

Type	Method	PESQ	STOI
Traditional	Opus	1.77	0.77
	WebRTC	1.70	0.70
	EVS	1.89	0.78
Learning-based	PLAAE	2.04	0.84
	TFGAN	1.97	0.81
	bin2bin	2.72	0.88
	RNN	2.23	0.83
	MFM	2.52	0.85
Learning-based	Proposed model	**3.07***	**0.91***

*Statistically significant improvement over all baseline methods (paired t-test, $p < 0.05$).

learning-based baseline models, bin2bin demonstrates the strongest performance in both metrics. The proposed model achieves a PESQ of 3.07 and an STOI of 0.91. Compared to EVS, the proposed model improves PESQ by 1.18 and STOI by 0.13, and compared to bin2bin, it increases PESQ by 0.35 and STOI by 0.03. This indicates that the proposed model surpasses all baseline models in both key metrics, fully demonstrating its significant advantages in enhancing speech quality and intelligibility.

Figure 7 shows the spectrograms output by the bin2bin model and the proposed model. There are a large number of obvious blank stripes in the lossy speech spectrogram, which represent the packet loss caused by compression or transmission. Although the bin2bin model improves the spectral structure of the lossy speech, there are some sparse noises in the red box that do not conform to the speech characteristics. This shows that the model has a certain effect in hiding packet loss, but fails to improve the overall quality of speech. At the same time, there is an obvious discontinuity in the red oval area below the red box, indicating that the model is still insufficient in recovering low-frequency local details. On the contrary, the speech output by the proposed model

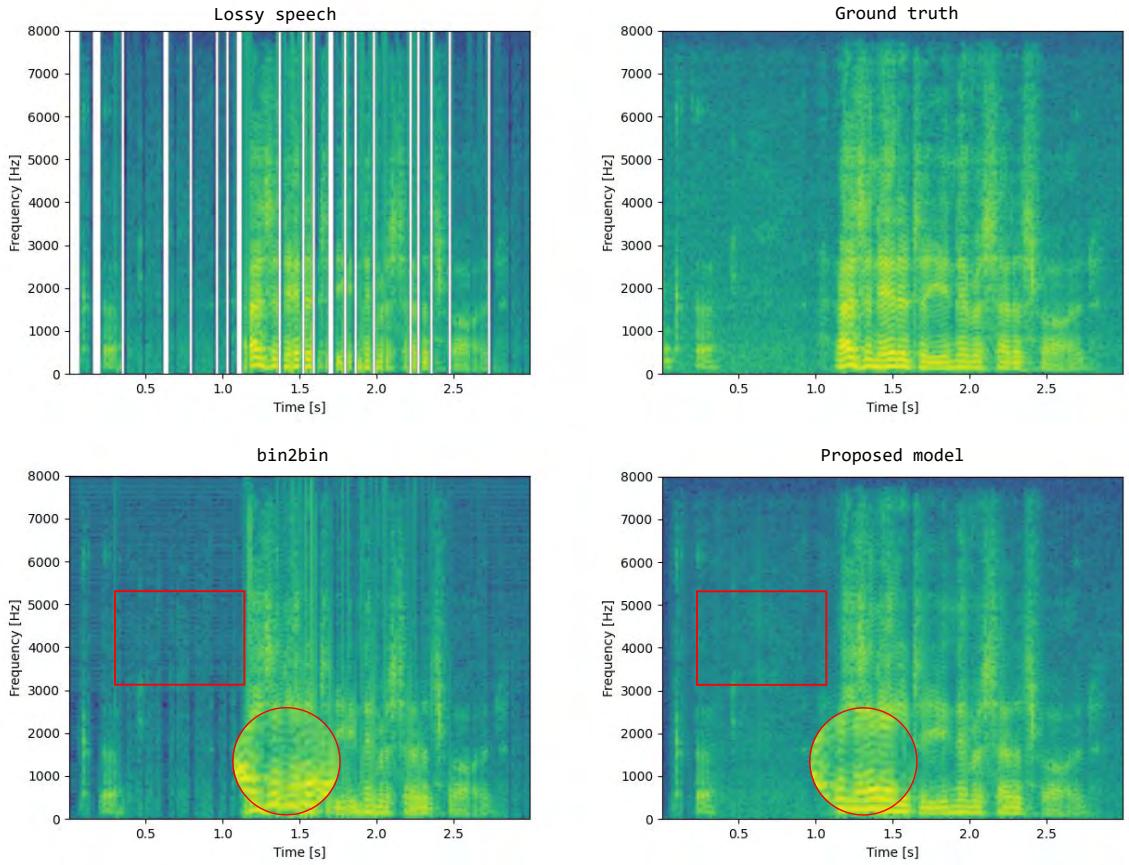


FIG. 7. Visualization of the output spectrogram.

does not have any non-speech features in the area within the red box, where it is replaced by a more continuous and smooth speech spectrum structure. In the high-frequency area above 3000 Hz, the energy distribution of the spectrogram is more uniform, and the details are richer, indicating that the proposed model can effectively recover the high-frequency details of the speech signal. In addition, the low-frequency part of the red circular area has a smooth transition, indicating that the proposed model generates natural and highly intelligible speech features. In general, the proposed model can effectively remove periodic noise and random noise caused by the nonlinear mapping of neural networks while restoring key details of speech. This is due to the network's ability to accurately model the time-frequency structure of speech signals, which enables high-fidelity restoration of the original signal.

Table 3 presents the overall quality scores of the output speech under different packet loss rates and maximum burst loss durations. This score is the average of DNSMOS, PLCMOS, PESQ, STOI, and WAcc after they are normalized to $[0, 1]$. The test data is VCTK. This normalized average is used only as an auxiliary indicator to illustrate overall trends across different loss conditions and is not intended to replace the interpretation of individual metrics, which are reported and discussed separately in other cases. The data in the table reflects the dual impact of the packet loss rate and the burst loss duration on speech quality, with an overall trend showing that as the packet loss rate increases and the burst loss duration extends, the output speech quality gradually declines.

Table 3. Overall quality of the output speech under specific loss rate and burst loss.

Maximum burst loss [ms]	Loss rate [%]			
	10	20	30	40
20	0.89	0.85	0.82	0.78
40	0.86	0.81	0.80	0.77
60	0.82	0.80	0.76	0.73
80	0.79	0.75	0.73	0.70
100	0.76	0.72	0.69	0.67
120	0.74	0.70	0.66	0.62

As the packet loss rate increases, the restoration quality gradually degrades across all burst loss settings, indicating that a higher loss frequency poses a greater challenge even when the burst duration is limited. For a fixed packet loss rate, increasing the maximum burst loss duration leads to a more pronounced degradation in quality, reflecting the difficulty of reconstructing longer consecutive missing segments. Notably, the degradation caused by the burst loss duration is more severe at higher loss rates, suggesting a compounding effect when the frequent and prolonged packet loss occurs simultaneously. These trends indicate that the proposed model is robust to the moderate packet loss and short bursts, while its performance degrades gracefully as loss conditions become more extreme, which is consistent with practical communication scenarios.

Figure 8 further illustrates the phenomenon of speech quality variation with the burst loss duration, as shown in Table 3, through spectrogram visualization. When the burst loss duration is 20 ms and 40 ms, the continuity and details of the spectrogram remain relatively clear, with high-frequency components well preserved, which aligns with the higher quality scores under the corresponding conditions in Table 3. At this point, the main characteristics of the speech signal are retained, and despite a certain degree of packet loss, the network's recovery mechanism still maintains good speech quality. As the burst loss duration increases to 100 ms and 120 ms, the spectrogram begins to show distinct vertical noise streaks and blurred spectral structures, particularly in the high-frequency range (above 4000 Hz). This reflects that prolonged continuous packet loss causes more severe damage to the speech signal, leading to a decline in the quality scores of the output speech in Table 3. Moreover, comparing spectrograms under different packet loss rates reveals that, under the same burst loss duration, a higher packet loss rate results in more noise and missing segments in the spectrogram, along with greater loss of spectral details. For example, at a 20 ms burst loss duration, the spectrogram continuity at a 10 % packet loss rate is significantly better than at a 40 % packet loss rate, which is consistent with the decreasing trend in the same row of data, in Table 3 as the loss rate increases.

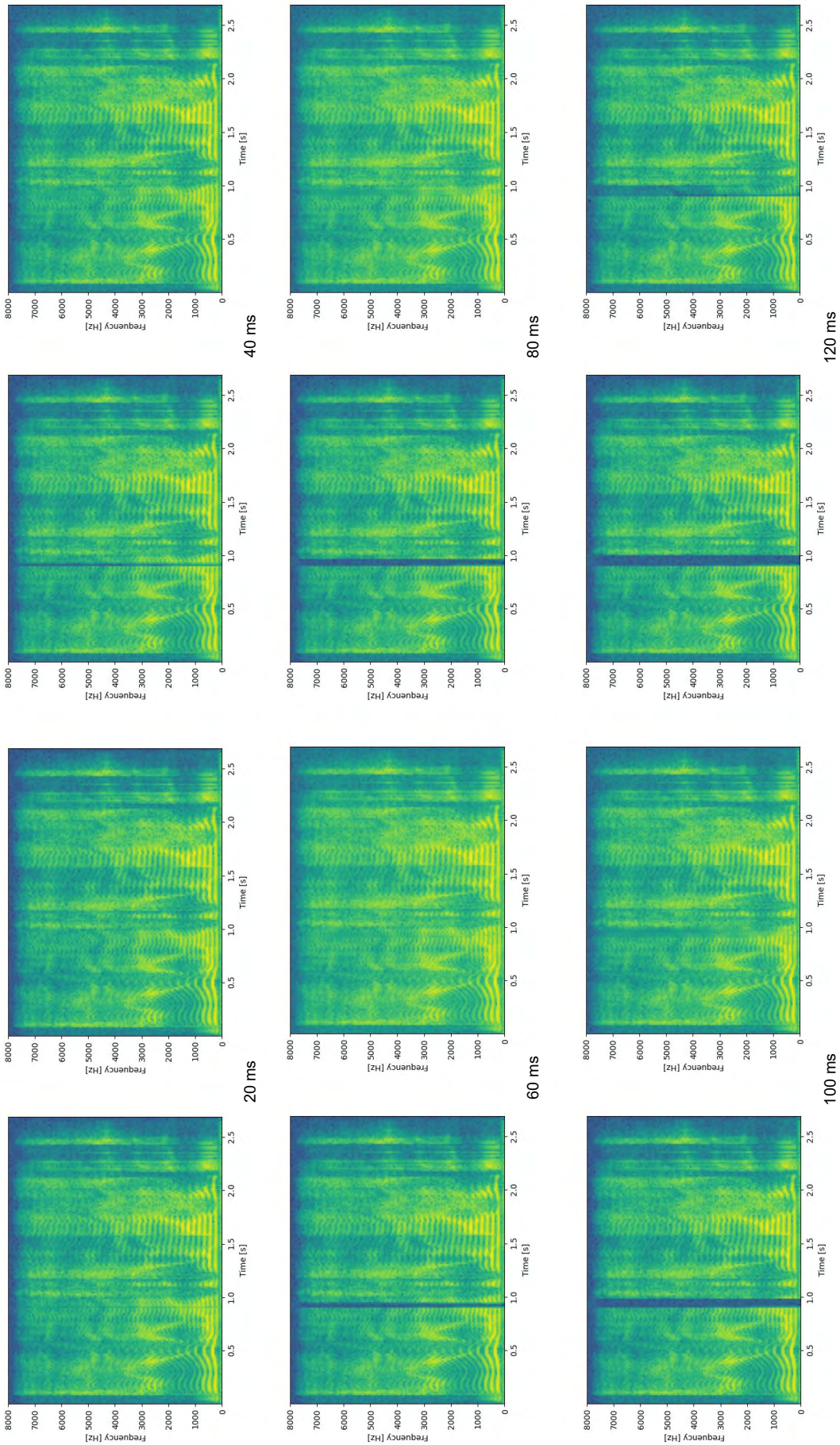


FIG. 8. Visualization of the output spectrogram for different burst loss.

Figure 8 and Table 3 jointly reveal the dual-impact mechanism of the burst loss duration and packet loss rate on speech quality. Short burst losses primarily affect the details of the speech signal, while prolonged burst losses disrupt the continuity and integrity of the speech signal. An increase in the packet loss rate exacerbates this damaging effect, leading to more spectral information loss. Together, these dual influences determine the performance of packet loss concealment systems under different conditions.

Table 4 reports the performance of different packet loss concealment methods on the collected real-world data, reflecting the quality of the restored speech signals. Compared with the degraded input, all PLC methods provide noticeable improvements, indicating that packet loss concealment is effective under practical communication conditions. Traditional codecs such as Opus, WebRTC, and EVS achieve modest gains, with PLCMOS values in the range of 3.12–3.21 and MOS around 3.11–3.24, which reflects the limited capability of their built-in PLC mechanisms in handling combined packet loss and noise. In contrast, learning-based approaches consistently outperform codec-based methods; for example, bin2bin improves PLCMOS to 3.68 and MOS to 3.62, while MFM further increases DNSMOS to 3.52. The proposed model achieves the best overall performance. These results indicate that jointly modeling packet loss concealment and noise suppression is particularly beneficial for real-world scenarios, enabling the proposed model to deliver more natural and intelligible restored speech than both traditional codecs and existing learning-based methods.

Table 4. Test results based on our collected real-world data.

Method	PLCMOS	DNSMOS	MOS
Input	3.02	3.24	2.98
Opus	3.16	3.26	3.20
WebRTC	3.12	3.23	3.11
EVS	3.21	3.29	3.24
PLAAE	3.28	3.40	3.47
TFGAN	3.23	3.32	3.25
bin2bin	3.68	3.51	3.62
RNN	3.36	3.38	3.42
MFM	3.56	3.52	3.55
Proposed model	**3.88***	**3.67***	**3.75***

*Statistically significant improvement over all baseline methods (paired t-test, $p < 0.05$).

Table 5 shows the results of an ablation study designed to evaluate the impact of different components and training strategies on model performance. The item column lists various experimental configurations, including the original setup and several variants where specific components or training methods are removed. The data show that when adversarial training is absent, all three metrics decrease compared to the original setup, indicating that adversarial training positively contributes to improving various aspects of speech quality. For the case without $\mathcal{L}_{\text{MSTFT}}$, the PLCMOS and DNSMOS decline, with PLCMOS dropping significantly from 3.88 in the original setup to 3.62, suggesting that $\mathcal{L}_{\text{MSTFT}}$ plays a crucial role in inpainting. In the absence of $\mathcal{L}_{\text{pre}}$, PLCMOS and DNSMOS show minimal changes, while MOS decreases from 3.75 in the original setup to 3.65. This implies that $\mathcal{L}_{\text{pre}}$ has little effect on the objective metrics of speech quality but significantly impacts the overall MOS, which may encompass broader considerations such as subjective evaluations or downstream task accuracy. In other

Table 5. Results of the ablation study.

Item	PLCMOS	DNSMOS	MOS
Without adversarial training	3.77	3.60	3.63
Without $\mathcal{L}_{\text{MSTFT}}$	3.62	3.55	3.59
Without $\mathcal{L}_{\text{pre}}$	3.86	3.66	3.65
Without mamba block	3.76	3.61	3.66
Without ELA	3.82	3.65	3.69
Original setting	**3.88**	**3.67**	**3.75**

words, $\mathcal{L}_{\text{pre}}$ may be important for certain subjective or downstream task performance but does not markedly improve objective speech quality metrics. When the Mamba block is removed, both PLCMOS and DNSMOS decrease compared to the original setup, with DNSMOS dropping from 3.67 to 3.61, demonstrating that it contributes positively to the modeling of long sequences. When it comes to ELA, PLCMOS decreases from 3.88 in the original setup to 3.82, and DNSMOS decreases from 3.67 to 3.65. This may suggest that ELA has a slightly negative effect on PLCMOS but a positive effect on DNSMOS. Overall, the original setup performs best when considering all factors. This indicates that the components and training strategies in the original setup work synergistically to achieve optimal performance across multiple dimensions.

4. Conclusion

In this work, we proposed an efficient audio inpainting model based on a GAN framework with a selective state space model and adversarial training, achieving robust packet loss concealment while maintaining lightweight. By integrating Mamba blocks for efficient long-sequence modeling and ELA attention for local feature refinement, our model outperforms existing methods in both objective metrics and subjective evaluations, particularly in noisy and bandwidth-limited scenarios. The inclusion of WavLM-based perceptual loss further enhances speech quality and intelligibility for the downstream ASR task. Experiments on real-world and benchmark datasets demonstrate our model's superior performance and generalization ability, making it a practical solution for real-time speech communication systems. Our future work will explore further optimization for edge devices and extension to broader speech enhancement applications.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' contributions

Junkang Yang conceptualized the study and wrote the original draft. Hongqing Liu supervised the project. Lu Gan and Liming Shi performed the analysis of experiments. Hiromitsu Nishizaki and Chee Siang Leow contributed to data interpretation. All authors reviewed and approved the final manuscript.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Aironi C., Cornell S., Serafini L., Squartini S. (2023), A time-frequency generative adversarial based method for audio packet loss concealment, [in:] *2023 31st European Signal Processing Conference (EUSIPCO)*, pp. 121–125, https://doi.org/10.23919/EUSIPCO58844.2023.10290027.
2. Blum N., Lachapelle S., Alvestrand H. (2021), WebRTC: real-time communication for the open web platform, *Communications of the ACM*, **64**(8): 50–54, https://doi.org/10.1145/3453182.
3. Chen S. *et al.* (2022), WavLM: large-scale self-supervised pre-training for full stack speech processing, *IEEE Journal of Selected Topics in Signal Processing*, **16**(6): 1505–1518, https://doi.org/10.1109/JSTSP.2022.3188113.

4. Chi Z. *et al.* (2022), XLM-E: cross-lingual language model pre-training via ELECTRA, [in:] *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6170–6182, https://doi.org/10.18653/v1/2022.acl-long.427.
5. Dai T., Wang J., Guo H., Li J., Wang J., Zhu Z. (2024), FreqFormer: frequency-aware transformer for lightweight image super-resolution, [in:] *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 731–739, https://doi.org/10.24963/ijcai.2024/81.
6. Davy S., Belton N., Tobin J., Zuber O.B., Dong L., Xuewen Y. (2023), A causal convolutional approach for packet loss concealment in low powered devices, [in:] *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, https://doi.org/10.1109/ICASSP49357.2023.10096505.
7. Diener L., Purin M., Sootla S., Saabas A., Aichner R., Cutler R. (2023), PLCMOS – a data-driven non-intrusive metric for the evaluation of packet loss concealment algorithms, [in:] *Proceedings Interspeech 2023*, pp. 2533–2537, https://doi.org/10.21437/Interspeech.2023-1532.
8. Diener L., Sootla S., Branets S., Saabas A., Aichner R., Cutler R. (2022), INTERSPEECH 2022 audio deep packet loss concealment challenge, [in:] *Proceedings Interspeech 2022*, pp. 580–584, https://doi.org/10.21437/Interspeech.2022-10829.
9. Dubey H. *et al.* (2024), ICASSP 2023 deep noise suppression challenge, [in:] *IEEE Open Journal of Signal Processing*, **5**: 725–737, https://doi.org/10.1109/OJSP.2024.3378602.
10. Ebner P.P., Eltelt A. (2020), Audio inpainting with generative adversarial network, *arXiv*, https://doi.org/10.48550/arXiv.2003.07704.
11. Fu S.-W. *et al.* (2021), MetricGAN+: an improved version of MetricGAN for speech enhancement, *Proceedings Interspeech 2021*, pp. 201–205, https://doi.org/10.21437/Interspeech.2021-599.
12. Gu A., Dao T. (2023), Mamba: linear-time sequence modeling with selective state spaces, *arXiv*, https://doi.org/10.48550/arXiv.2312.00752.
13. Guan Y., Yu G., Li A., Zheng C., Wang J. (2022), TMGAN-PLC: audio packet loss concealment using temporal memory generative adversarial network, [in:] *Proceedings Interspeech 2022*, pp. 565–569, https://doi.org/10.21437/Interspeech.2022-644.
14. Hao S., Li X., Peng W., Fan Z., Ji Z., Ganchev I. (2024), YOLO-CXR: a novel detection network for locating multiple small lesions in chest X-ray images, *IEEE Access*, **12**: 156003–156019, https://doi.org/10.1109/ACCESS.2024.3482102.
15. He K., Zhang X., Ren S., Sun J. (2016), Deep residual learning for image recognition, [in:] *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
16. Janicki A., Księżak B. (2008), Packet loss concealment algorithm for VoIP transmission in unreliable networks, [in:] *New Trends in Multimedia and Network Information Systems*, **181**: 23–33, https://doi.org/10.3233/978-1-58603-904-2-23.
17. Kumar K. *et al.* (2019), MelGAN: generative adversarial networks for conditional waveform synthesis, [in:] *Advances in Neural Information Processing Systems*.
18. Lagrange M., Marchand S., Rault J.-B. (2005), Long interpolation of audio signals using linear prediction in sinusoidal modeling, *Journal of the Audio Engineering Society*, **53**(10): 891–905.
19. Lecomte J. *et al.* (2015), Packet-loss concealment technology advances in EVS, [in:] *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5708–5712, https://doi.org/10.1109/ICASSP.2015.7179065.
20. Lee B.-K., Chang J.-H. (2015), Packet loss concealment based on deep neural networks for digital speech transmission, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24**(2): 378–387, https://doi.org/10.1109/TASLP.2015.2509780.
21. Li N., Zheng X., Zhang C., Guo L., Yu B. (2022), End-to-end multi-loss training for low delay packet loss concealment, [in:] *Proceedings Interspeech 2022*, pp. 585–589, https://doi.org/10.21437/Interspeech.2022-11439.
22. Lin J., Wang Y., Kalgaonkar K., Keren G., Zhang D., Fuegen C. (2021), A time-domain convolutional recurrent network for packet loss concealment, [in:] *ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7148–7152, https://doi.org/10.1109/ICASSP39728.2021.9413595.

23. Liu B., Song Q., Yang M., Yuan W., Wang T. (2022), PLCNet: real-time packet loss concealment with semi-supervised generative adversarial network, [in:] *Proceedings Interspeech 2022*, pp. 575–579, https://doi.org/10.21437/Interspeech.2022-10428.
24. Lotfidereshgi R., Gournay P. (2018), Speech prediction using an adaptive recurrent neural network with application to packet loss concealment, [in:] *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5394–5398, https://doi.org/10.1109/ICASSP.2018.8462185.
25. Maher R.C. (1994), A method for extrapolation of missing digital audio data, *Journal of the Audio Engineering Society*, **42**(5): 350–357.
26. Miotello F., Pezzoli M., Comanducci L., Antonacci F., Sarti A. (2024), Deep prior-based audio inpainting using multi-resolution harmonic convolutional neural networks, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **32**: 113–123, https://doi.org/10.1109/TASLP.2023.3324556.
27. Mohamed M.M., Schuller B.W. (2020), Concealnet: an end-to-end neural network for packet loss concealment in deep speech emotion recognition, *arXiv*, https://doi.org/10.48550/arXiv.2005.07777.
28. Moliner E., Lehtinen J., Välimäki V. (2023), Solving audio inverse problems with a diffusion model, [in:] *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, https://doi.org/10.1109/ICASSP49357.2023.10095637.
29. Pascual S., Serrà J., Pons J. (2021), Adversarial auto-encoding for packet loss concealment, [in:] *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 71–75, https://doi.org/10.1109/WASPAA52581.2021.9632730.
30. Raffel C. *et al.* (2020), Exploring the limits of transfer learning with a unified text-to-text transformer, *Journal of Machine Learning Research*, **21**(140): 1–67.
31. Reddy C.K.A., Gopal V., Cutler R. (2022), Dnsmos P.835: a non-intrusive perceptual objective speech quality metric to evaluate noise suppressors, [in:] *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 886–890, https://doi.org/10.1109/ICASSP43922.2022.9746108.
32. Rix A.W., Beerends J.G., Hollier M.P., Hekstra A.P. (2001), Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs, [in:] *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, **2**: 749–752, https://doi.org/10.1109/ICASSP.2001.941023.
33. Rodbro C.A., Murthi M.N., Andersen S.V., Jensen S.H. (2006), Hidden Markov model-based packet loss concealment for voice over IP, *IEEE Transactions on Audio, Speech, and Language Processing*, **14**(5): 1609–1623, https://doi.org/10.1109/TSA.2005.858561.
34. Sun Y. *et al.* (2025), Deliod a lightweight detection model for intestinal organoids based on deep learning, *Scientific Reports*, **15**(1): 5040, https://doi.org/10.1038/s41598-025-89409-y.
35. Taal C.H., Hendriks R.C., Heusdens R., Jensen J. (2011), An algorithm for intelligibility prediction of time–frequency weighted noisy speech, *IEEE Transactions on Audio, Speech, and Language Processing*, **19**(7): 2125–2136, https://doi.org/10.1109/TASL.2011.2114881.
36. Valin J.-M. *et al.* (2022), Real-time packet loss concealment with mixed generative and predictive model, [in:] *Proceedings Interspeech 2022*, pp. 570–574, https://doi.org/10.21437/Interspeech.2022-903.
37. Valin J.-M., Maxwell G., Terriberry T.B., Vos K. (2016), High-quality, low-delay music coding in the opus codec, *arXiv*, https://doi.org/10.48550/arXiv.1602.04845.
38. Vaswani A. *et al.* (2017), Attention is all you need, [in:] *Advances in Neural Information Processing Systems*.
39. Wang J., Guan Y., Zheng C., Peng R., Li X. (2021), A temporal-spectral generative adversarial network based end-to-end packet loss concealment for wideband speech transmission, *The Journal of the Acoustical Society of America*, **150**(4): 2577–2588, https://doi.org/10.1121/10.0006528.
40. Westhausen N.L., Meyer B.T. (2022), tPLCnet: real-time deep packet loss concealment in the time domain using a short temporal context, [in:] *Proceedings Interspeech 2022*, pp. 2903–2907, https://doi.org/10.21437/Interspeech.2022-10157.
41. Wichern G. *et al.* (2019), WHAM!: extending speech separation to noisy environments, [in:] *Proceedings Interspeech*, https://doi.org/10.21437/Interspeech.2019-2821.

42. Xu W., Wan Y. (2024), ELA: efficient local attention for deep convolutional neural networks, *arXiv*, https://doi.org/10.48550/arXiv.2403.01123.

43. Yamagishi J., Veaux C., MacDonald K. (2019), CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92), University of Edinburgh. The Centre for Speech Technology Research (CSTR), https://doi.org/10.7488/ds/2645.

44. Yang D.-H., Kim D., Chang J.-H. (2023), Masked frequency modeling for improving packet loss concealment in speech transmission systems, [in:] *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5, https://doi.org/10.1109/WASPAA58266.2023.10248056.

45. Yang J., Liu H., Gan L., Jing X. (2024), Spectral network based on lattice convolution and adversarial training for noise-robust speech super-resolution, *The Journal of the Acoustical Society of America*, **156**(5): 3143–3157, https://doi.org/10.1121/10.0034364.

46. Yen H., Germain F.G., Wichern G., Le Roux J. (2023), Cold diffusion for speech enhancement, [in:] *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, https://doi.org/10.1109/ICASSP49357.2023.10096064.

47. Zhao H. (2023), A GAN speech inpainting model for audio editing software, [in:] *Proceedings Interspeech 2023*, pp. 5127–5131, https://doi.org/10.21437/Interspeech.2023-904.

Archives of Acoustics Vol. 51, No. 3, pp. 533–548 (2026), https://doi.org/10.24423/archacoust.2026.4251

Research Paper

Audio Strips Network (ASNet) and Amalgamated Audio Features (A2F): A Synergistic Approach for Audio Source Separation

Shunmugalingam Parvathi Sakthidevi*, Chandrasekar Divya

Centre for Information Technology and Engineering, Manonmaniam Sundaranar University
Tirunelveli, India

*Corresponding Author: spsakthidevi2000@gmail.com

Received May 29, 2025; revised December 10, 2025; accepted April 25, 2026; available online April 29, 2026; version of record September 7, 2026; published issue September 23, 2026.

Audio source separation refers to the procedure of decomposing a mixed audio signal into its constituent components. This technique enables numerous applications, including creative music production, educational tools, karaoke, transcription, and music analysis. Despite the recent success of deep learning-based source separation techniques, these techniques often do not perform very accurately and do not provide high-quality separation of sources when many contain complex combinations in their mixtures. Source separation techniques generally rely on temporal or spectral features for analysis, which does not fully capture the complex dynamics of audio signals. To address these limitations, this paper proposes the amalgamated audio features (A2F), a hybrid representation combining temporal and spectral features. Then, the audio strip network (ASNet) is proposed, a novel framework designed to achieve clean and precise separation of individual audio sources with enhanced performance. ASNet utilizes A2F to separate sources more effectively. The model is trained and evaluated on the MUSDB, DSD100, and MUSDB18-HQ datasets, a benchmark for music source separation, and its standard measures, such as the signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR), are used to examine performance. ASNet achieves enhanced separation performance with SDR values of 12.63 for drums, 11.42 for vocal, 12.01 for bass, and 11.14 for other, and SIR values of 9.57 for drums, 9.61 for vocal, 9.66 for bass, and 9.67 for other. This advancement benefits musicians through high-quality remixing and creativity while aiding researchers in improving deep learning and hybrid audio processing models.

Keywords: feature extraction, amalgamated audio features (A2F), source extraction, audio strip network (ASNet).



Copyright © 2026 The Author(s).
This work is licensed under the Creative Commons Attribution 4.0 International CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Audio signal processing (Zölzer, 2022) has expanded greatly in recent years, particularly in the domain of audio source separation: the process of extracting different individual constituent sources from a combined auditory mixture. Source separation is crucial in many areas, including music production (Huber *et al.*, 2023), speech transcription (O'Connell, Kowal, 2009), audio enhancement (Chuang *et al.*, 2022) and restoration, hearing aids (Sanders *et al.*, 2021), and forensic audio analysis (Renukadevi *et al.*, 2024). Even though some advancements have been made, creating realistic source separation remains difficult due to the fact that audio signals are often inherently complex and often exhibit overlapping frequencies, transient characteristics, dense harmonics, environmental noise, and reverberation, all of which can impair model performance.

Advances in the field of deep learning (DL) (Issa, Al-Irhaym, 2021), have transformed the source separation landscape by introducing powerful, data-driven approaches capable of learning complex patterns directly from raw or transformed audio inputs. Various techniques have achieved state-of-the-art performance in source sepa-

ration tasks. For example, convolutional neural network (CNN) such as U-Net (Deng *et al.*, 2024), Conv-TasNet (Défossez *et al.*, 2021), SCNet (Tong *et al.*, 2024), DTTNet (Chen *et al.*, 2024), and DenseNet (Takahashi, Mitsufuji, 2020) have shown great success in modeling spectral and spatial correlations. At the same time, recurrent neural networks (RNNs) (Sun *et al.*, 2021; Luo, Yu, 2023), incorporating gated linear unit (GLU) (Tong *et al.*, 2024; Défossez *et al.*, 2019), and bidirectional long short-term memory (Bi-LSTM) networks (Chen *et al.*, 2024; Deng *et al.*, 2024; Défossez *et al.*, 2021), have demonstrated efficacy in capturing long-range temporal dependencies within audio streams. While CNNs effectively capture spatial and spectral characteristics, RNNs are capable of modeling the temporal dependencies of audio segments while assigning increased computational weight to relevant acoustic regions. Nevertheless, both CNNs and RNNs face limitations, particularly when audio signals are highly overlapping or contaminated with noise.

In addition, there are many existing approaches that are computationally intensive, requiring great computational resources for training and inference, which limit their use in resource-constrained scenarios. Noise and reverberation further complicate these separation difficulties, undermining the efficacy of many current approaches. To address these challenges, a novel methodology is proposed that combines the strengths of DL with advanced feature extraction techniques. This methodology introduces hybrid features that integrate temporal and spectral cues, enabling a more accurate representation of the complex dynamics present in musical mixtures. Built upon this feature set, the proposed hybrid DL model achieves higher SDR and SIR scores across benchmark datasets, demonstrating superior separation quality and robustness compared to existing models. Its relevance is further reflected in practical applications such as music production, remixing, and audio processing research, where high-quality, efficient source separation is essential.

2. Background analysis

This section reviews existing DL methodologies and feature extraction techniques, while addressing the challenges associated with these approaches. The gaps in current methods are identified to establish the foundation for the proposed framework.

2.1. Analysis of feature extraction

Mel frequency cepstral coefficients (MFCCs) (Hamza *et al.*, 2022) mimic human hearing by utilizing the fast Fourier transform (FFT) (Rajaby, Sayedi, 2022), pre-emphasis, framing, a hamming window, Mel filterbanks, log-energy calculation, and the discrete cosine transform (DCT) to extract speech and audio features. However, they face challenges with computational efficiency and noise robustness. While effective under controlled conditions, MFCCs struggle with real-world noise, edge-device computation, and dynamic speech variations (Rezaul *et al.*, 2024). Similarly, cochlear implant strategies comparing the Crystalis method (128-point FFT) and spectral feature extraction (SFE) (512-point FFT + interpolation), show that while SFE improves spectral resolution it lacks validation in noise robustness, efficiency, and generalizability. Future work should explore DL integration and lightweight SFE variants for clinical scalability, as long-term acclimatization and real-world performance remain unassessed (Zhang *et al.*, 2025). Furthermore, the 1-transistor-1-memristor (1T1R) neuromorphic system mimics biological neurons for audio feature extraction, achieving an accuracy range from 85.38 % to 95.91 % on the spiking Heidelberg digits (SHD) dataset but lacks real-world noise robustness testing. Future work should improve complementary metal-oxide-semiconductor (CMOS) compatibility, scalability beyond memristors, and generalization to diverse acoustic environments (Wu *et al.*, 2024).

2.2. Performance of DL approaches

A denoising autoencoder with U-Net and Bi-LSTM (DAEUBL) is applied to denoise multi-level random telegraph signals (RTS) (Deng *et al.*, 2024) contaminated with mixed noise. Unlike single-noise-focused models, the DAEUBL generalizes across various noise profiles and signal lengths via unified training and adaptive segmentation. However, it lacks real-world benchmark validation and a computational efficiency analysis, limiting

scalability insights and practical deployment trade-offs for complex RTS denoising (Deng *et al.*, 2024). Similarly, the sparse conformer network (SCNet) applies the short-time Fourier transform (STFT) and sparse downsampling (SD) to compress binaural audio spectrograms for source separation. It uses group normalization (GroupNorm), gated linear units (GLUs), and dual-path RNN to model sequence dependencies. Despite its success, research gaps for SCNet include the need for better high-frequency modeling, lower complexity, and improved generalization to unseen data (Tong *et al.*, 2024).

In the dilated dense block (D2Block) architecture, each feature map is processed with varying dilation factors to ensure better input coverage without blind spots. The deep dual-path densely connected network (D3Net) extends this architecture by stacking D2Blocks with a repeated dilation pattern, allowing flexible feature transformation across resolutions and improving computational efficiency with channel reduction mechanisms. However, certain shortcomings persist, including the potential for aliasing artifacts, increased model complexity due to dense connections, and challenges in maintaining effective receptive field coverage across layers (Takahashi, Mitsufuji, 2020). Furthermore, the dual-PATH TFC-TDF U-Net (DTTNet) uses time-distributed fully connected (TDF) layers, time-frequency convolutions (TFC), and an improved dual-path module (IDPM) for music source separation. GroupNorm and Bi-LSTM enhance sequence modeling across time and frequency. While its real-world generalization has been tested on patterns such as 'wah-wah' guitar filters and vocal chops beyond the MUSDB18-HQ benchmark, primary challenges remain. These challenges include better fine-tuning for distorted audio patterns and more robust separation across unseen acoustic environments (Chen *et al.*, 2024).

Finally, DL architectures such as Demucs and the convolutional time domain audio separation network (Conv-TasNet) employ convolutional encoders/decoders, GLU and Bi-LSTM to achieve complete source separation in musical mixtures. Improvements, such as weight rescaling at initialization, the shift trick for time-equivariance, and sinc-based resampling, address performance bottlenecks. Nonetheless, key difficulties involve maintaining phase consistency, managing varying loudness levels across entire tracks, preserving time-shift invariance, and optimizing memory usage when working with these large DL models (Défossez *et al.*, 2021).

3. Methodology

This section reveals the innovative DL framework designed to address key challenges in music source separation, featuring two core components: A2F for effective feature extraction and ASNet for robust source separation.

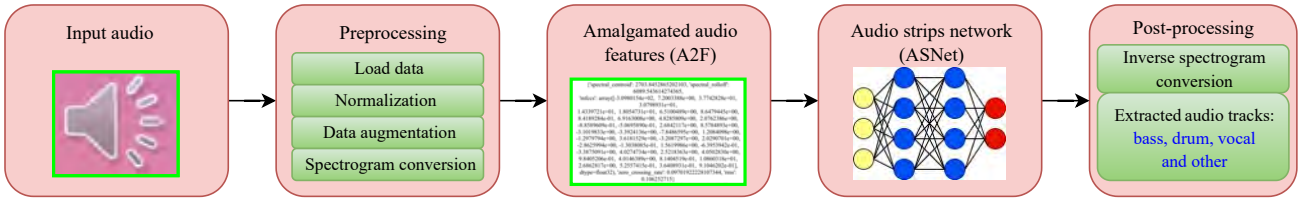


Fig. 1. End-to-end architecture of the proposed framework.

3.1. Dataset

To evaluate and train the proposed model for audio source separation, three benchmark datasets were employed: MUSDB, DSD100, and MUSDB18-HQ. The model was trained independently on each dataset to assess its separation performance and generalization ability across diverse musical compositions, genres, and recording environments. Figure 2 shows accuracy and loss curves for training and validation across the proposed ASNet model evaluated on the MUSDB18-HQ, DSD100, and MUSDB datasets. It includes accuracy vs. epochs and loss

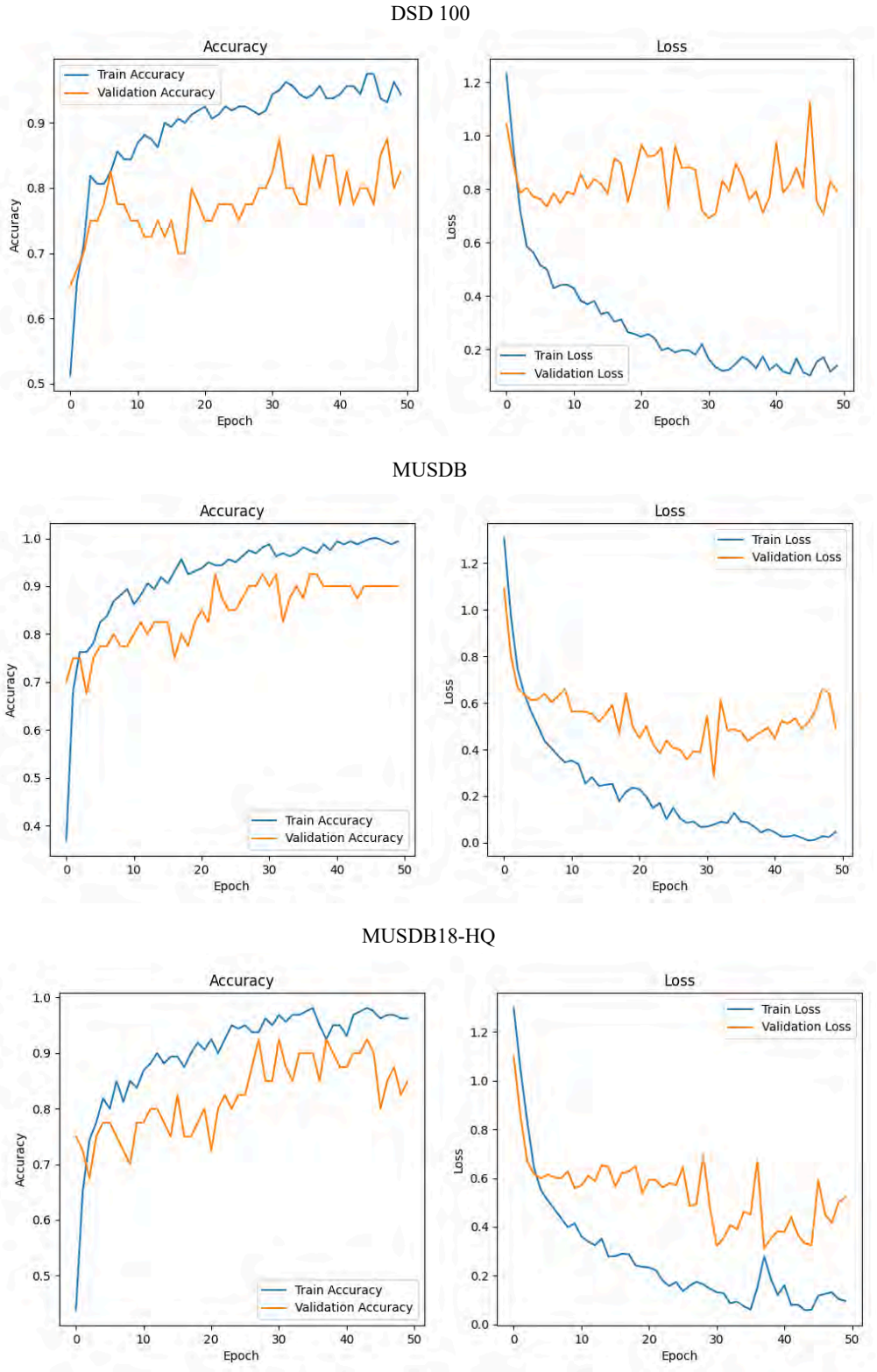


FIG. 2. Accuracy and loss visualization for training and validation across the various datasets.

vs. epochs plots, highlighting the convergence behavior, learning progress, and generalization capability of the ASNet.

The MUSDB, DSD100, and MUSDB18-HQ (RAFII *et al.*, 2019) datasets consist of 166, 100, and 150 full-length music tracks, respectively. The MUSDB and MUSDB18-HQ provide approximately 10 hours of audio data each,

while DSD100 includes tracks ranging from 2 min and 22 s to 7 min and 20 s. These datasets cover a wide range of musical genres and include isolated stems for drums, bass, vocals, and other instruments. All stereophonic audio signals are captured at 44 100 Hz. Each dataset is partitioned into an 80 % split for training and a 20 % part for testing, using A2F audio features as the inputs, while the corresponding source categories (separated signals) function as the ground-truth labels. To maintain reproducibility and consistency, a random state of 42 is used during data splitting. The method is trained on a batch size of 32 over 50 epochs to preserve available memory and limit the computational burden.

For the MUSDB dataset, the training accuracy steadily rises, reaching approximately 0.95, while the validation accuracy also improves and saturates around 0.85, indicating effective learning with a slight generalization gap. The training loss consistently decreases throughout the epochs. However, the validation loss begins to show mild fluctuations after epoch 30. This suggests minor overfitting in the later stages of training. For the DSD100 dataset, the training accuracy steadily improves, reaching nearly 0.98, while the validation accuracy rises and stabilizes around 0.85 after epoch 30. Both training and validation loss decrease during the initial epochs. However, the validation loss becomes highly variable in the later stages. This implies some generalization instability, probably due to the complexity of the dataset. Finally, on the MUSDB18-HQ dataset, a training accuracy is at approximately 0.96, while the validation accuracy increases quickly and has very high confidence values between 0.85 and 0.90, which indicate strong generalization. As optimization continues, the training loss decreases continuously; however, the validation loss varies widely after epochs 25. This indicates possible overfitting, or the fact that the model can be highly sensitive to validation samples.

In summary, all three datasets have learning behavior characterized by increasing accuracy and decreasing loss. The model achieves consistent performance at high levels of validation accuracy for all three datasets. While variability exists, particularly for validation loss, this provides a good opportunity to further improve performance using techniques such as early stopping, regularization, or data augmentation. Among the datasets, MUSDB18-HQ seems to perform best with the highest level of stable validation accuracy; more importantly, the dataset's audio quality appears to have meaningful effects on the model's generalization ability.

3.2. Preprocessing

The audio preprocessing pipeline starts with reading the audio file (such as WAV or MP3), returning a tuple that contains the audio time series (1D if mono or 2D if stereo) and the sample rate, which is fixed at 48 000 samples per second. Then, the time-domain waveform is transformed using the STFT, resulting in the signal transformed to a complex-valued time-frequency representation. If a user provides an MP3 file, it must first be decoded into pulse-code modulation (PCM), as MP3 is a compressed frequency-domain format; this decoding step restores the actual time-domain waveform required for further analysis. The magnitude of this complex STFT output is calculated to obtain a 2D array of values (frequency bins × time frames), representing the signal's energy distribution across frequency over time. To better align with human auditory perception, the magnitude spectrogram is further converted to the decibel scale.

Normalization techniques are applied to ensure numerical stability and consistency across inputs. In the time domain, peak normalization scales waveform amplitudes to the range $[-1, 1]$. In the frequency domain, log-magnitude or decibel scaling is used to manage dynamic range and improve learning efficiency. Data augmentation is employed across both the time and frequency domains to enhance generalization and address the challenge of limited labeled data. In the time domain, augmentations such as time warping and pitch shifting modify the waveform directly. while corresponding spectral transformations are applied to the decibel spectrograms in the frequency domain. These preprocessing steps, spanning both waveform and spectral representations, enable robust feature extraction and improved model generalization. This dual-domain approach is necessary for optimal source separation performance. Figure 3a shows the time-domain waveform of the audio signal, representing amplitude variations over time, while Fig. 3b and Fig. 3c display the magnitude spectrogram and decibel spectrogram, respectively, capturing the frequency content of the signal across time with different scaling.

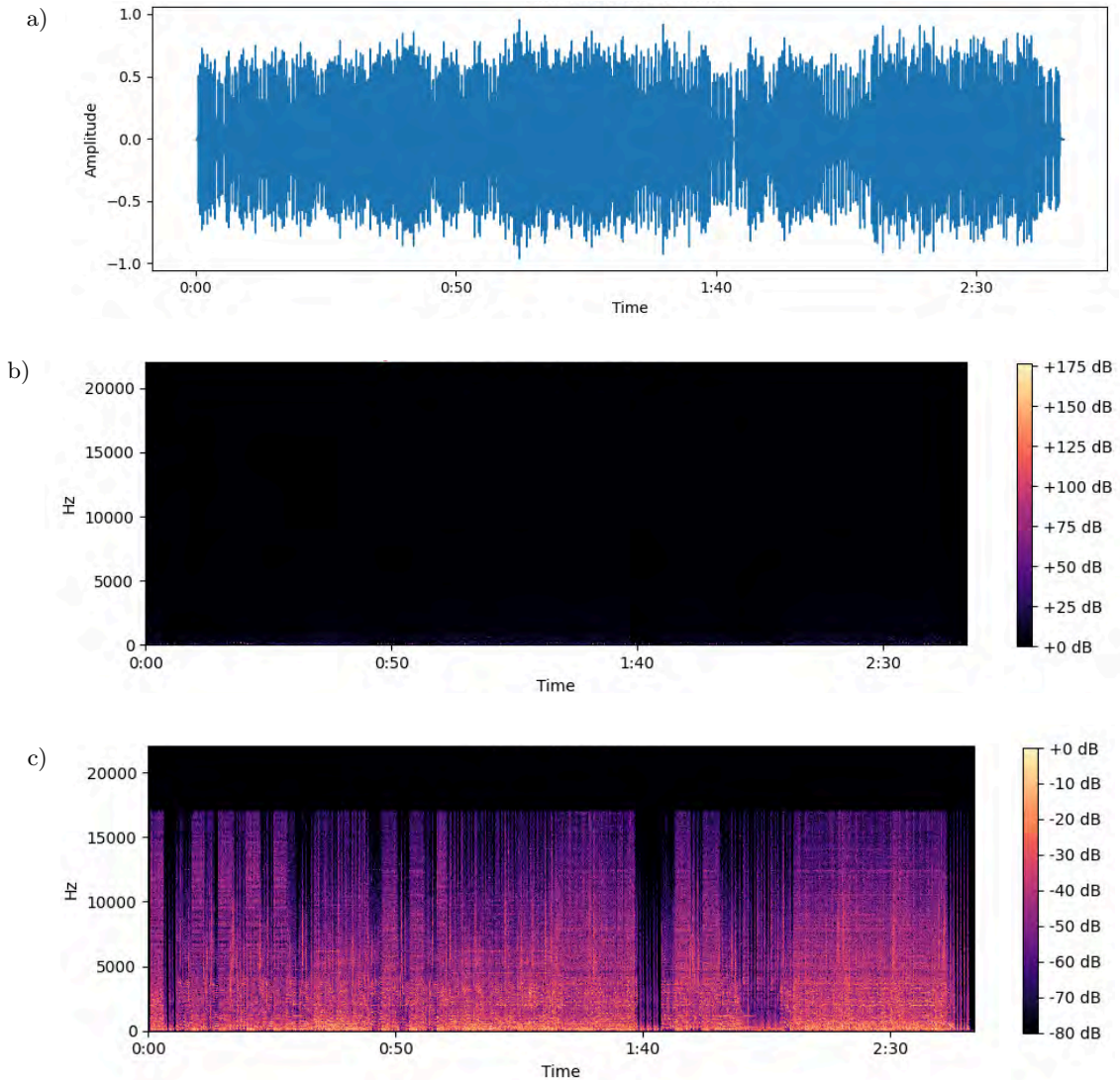


FIG. 3. Visual representations of the audio signal during preprocessing stages:
a) time-domain waveform, b) time-frequency magnitude spectrogram, c) time-frequency decibel spectrogram.

3.3. A2F

A specialized feature extraction pipeline designed for audio source separation tasks is introduced and named A2F. Figure 4 illustrates the proposed A2F framework. This preprocessing system operates on mixed audio inputs to derive discriminative feature representations that enhance the downstream DL model's ability to separate and reconstruct individual sound sources. The A2F framework's hierarchical feature amalgamation process is constructed to capture both local spectro-temporal patterns and global acoustic characteristics, providing the separation network with optimized input features for improved source disentanglement performance.

The input to the framework is a mixed audio signal that contains a blend of multiple sources. This signal is processed in two parallel branches to extract distinct types of features.

Spectral features are derived from the STFT-based spectrogram, which provides information about the frequency distribution, harmonic structure, and timbre characteristics of each source. The spectral centroid indicates the brightness of the signal, helping the model distinguish high-frequency sources such as vocals from low-frequency sources such as bass and drums. The spectral roll off reflects how energy is distributed across the spectrum, enabling the separation of harmonic instruments that exhibit smoother energy decay from percussive instruments whose energy extends sharply into higher frequencies. MFCCs contribute by capturing perceptually

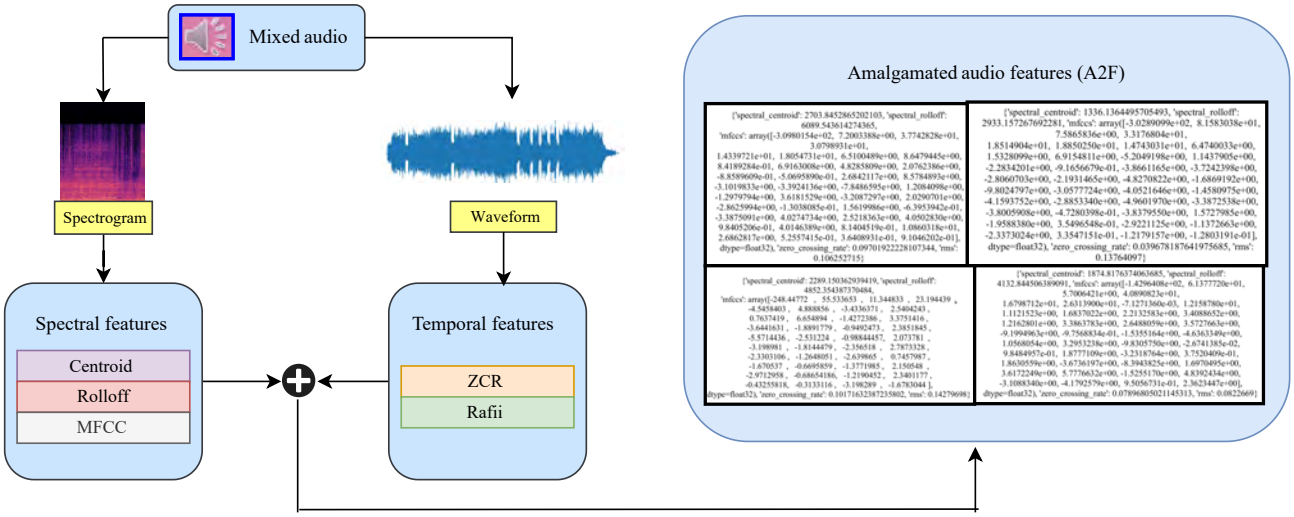


FIG. 4. Design of the proposed A2F framework.

relevant timbral information, allowing the network to identify and cluster tones from different instruments even when their frequency components overlap.

In parallel, temporal features extracted directly from the waveform, which provides insight into the signal's time-domain behavior. The zero-crossing rate (ZCR) helps differentiating between transient, noisy, or percussive sounds, which have high ZCR, and sustained or tonal sounds such as vocals or bass, which show a lower ZCR. Root-mean-square (RMS) energy captures the loudness contour of the signal, helping to segment active and inactive regions and maintain a balanced energy during source reconstruction. Together, these spectral and temporal features offer a holistic representation of the mixed audio, enabling the model to learn clearer boundaries between overlapping sources and ultimately improving the accuracy and robustness of the separation task.

3.3.1. FEATURE FUSION

A rich, multi-domain feature set is created by combining the temporal and spectral features. The resulting A2F representation, shown in the right panel of Fig. 4, contains the numerical values of the extracted spectral and temporal features. This matrix provides a complete set of input features for the source separation model, enabling a comprehensive analysis of the audio signal that specifically makes use of time-domain features to find rhythmic patterns and transients, alongside frequency-domain features to identify timbre and tonal patterns. Thanks to this robust feature representation, the model is better equipped to differentiate between overlapping sources in the mixed signal.

The A2F framework addresses some major limitations in current audio feature extraction techniques by combining noise-aware preprocessing, dual-domain feature extraction, multi-feature fusion, and integration of deep learning. With respect to noise, the proposed model demonstrates improved robustness to noise, better generalization capabilities, less reliance on handcrafted features alone. In the end, this enhances separation performance when used to separate sound events present in realistic audio environments.

3.4. ASNET

The proposed ASNet is composed of a downsampling module, a bottleneck layer, a bidirectional long short-term memory (Bi-LSTM), an upsampling module and a fully connected layer (FCL), as shown in Fig. 5.

The downsampling blocks are designed to gather the most significant musical elements from the incoming spectrogram data. When viewing the first downsampling block (downsample 1) the use of a 2D convolutional (Conv2D) layer expands the input data from 1 to 64 channels. Performing convolution using a 3×3 kernel allows the network to detect both temporal variations in harmony intensities and the harmonic relationship of notes

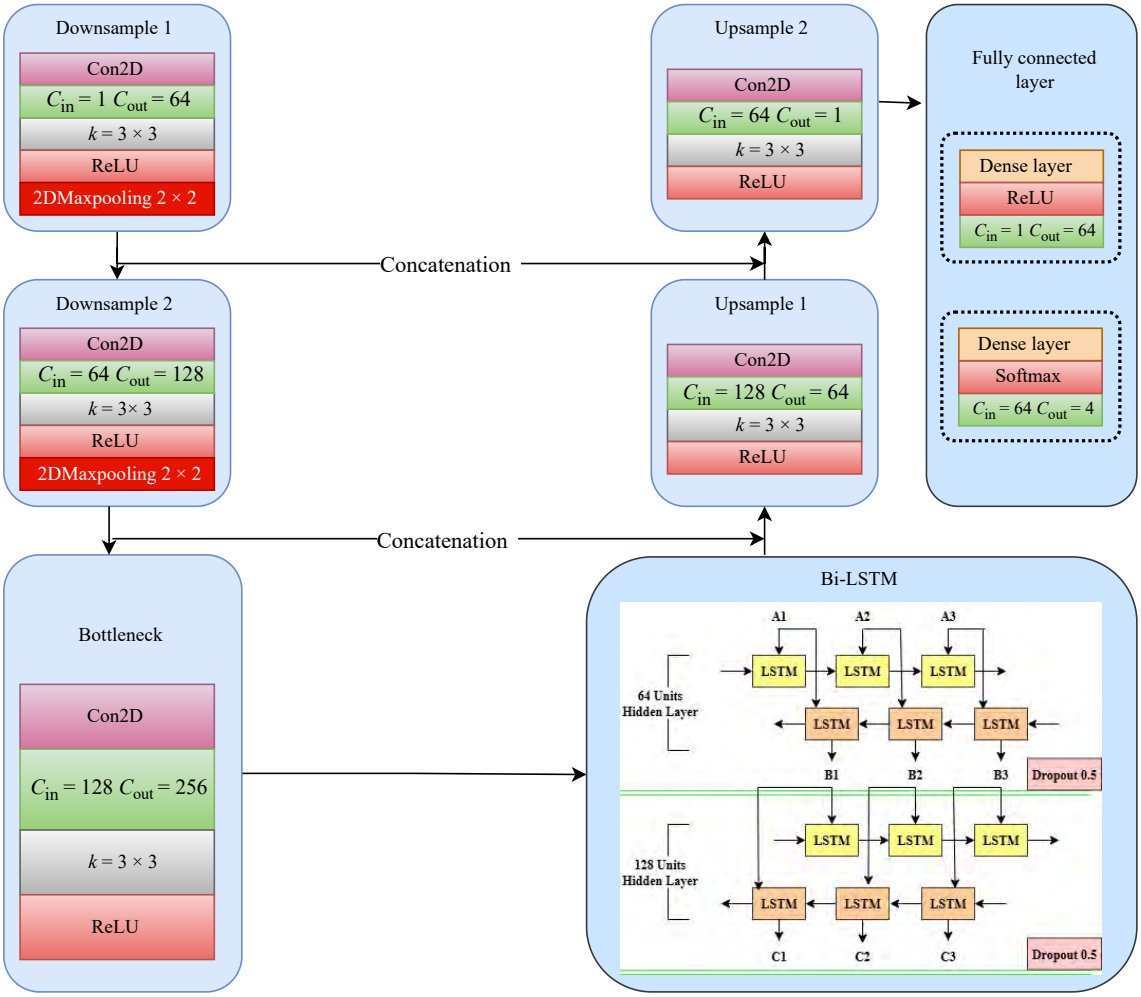


FIG. 5. Proposed ASNet architecture for audio source separation.

(timbral) to one another. Following a rectified linear unit (ReLU) activation, the maximum spectra are retained, creating a new representation with the highest harmonic (tonal) composition and instrument recognition. In the second block, downsample 2, the total number of filters is increased from 64 to 128, creating higher-level features that allow the network to learn additional patterns of timbre associated with certain instruments and notes in the music, as well as additional harmonics. Repeating this process using ReLU activation and max-pooling operations gradually reduces the data, allowing only the most significant musical features to remain, while all extraneous information such as noise is removed. Together, downsample 1 and downsample 2 allow the model to effectively establish and recognize the time-frequency relationships of music required for effective music extraction.

After the compressed features have been created, they are input into a bottleneck layer, which has the scope of expanding the depth of feature maps from 128 to 256. This layer allows the model to find highly abstract patterns in music that cannot be represented by other layers. The bottleneck layer also lets the model capture information related to more complex interactions between timbres, dense overlaps within harmonic structures, and instrument-specific signatures in spectra, leading to more precise separation of sounds. The use of a Conv2D-layer at this point lets the network apply wide-ranging time-frequency dependencies during encoding operations. Furthermore, the increased number of filters gives the network the flexibility to encode finer, smaller details such as vocal-harmonic characteristics, resonances in drums, overtones in bass, and textures from other sounds in the background. Additionally, the ReLU activation functions support the largest and sparsest high-energy musical features, therefore enhancing the detection of key harmonic points and tempo events or vice versa. In conclusion, the bottleneck layer acts as a high-level encoder within the overall structure of an end-to-end

network; it transforms the lower-level features of spectrograms into robust, rich and discriminative representations ready for subsequent use in recurrent multi-task layers, producing improved separation capabilities for each song's music and achieving improved music extraction capabilities.

Temporal modelling occurs at the Bi-LSTM block, which creates two bidirectional layers, containing 64 and 128 units, respectively. These two bidirectional layers within one block allow for capturing short-term transitions, such as note onsets and inflections that occur with voice, as well as longer strands of time, such as sustained harmonic notes, rhythmic patterns, and instrument continuities. This design is aimed at increasing the capability of the model to accurately represent complex, overlapping audio source patterns found in music where many different instruments coexist and interact throughout a time sequence. The use of a Bi-LSTM is critical because musical events rely not only on previous events in the past but also on future events, as this future context plays an important role in separating elements such as chord progressions, vocal phrasing or the sustain of notes. With the inclusion of a 0.5 dropout rate, it is also critical to provide strong regularization to prevent overfitting. The dropout rate will encourage the LSTM to identify effective time dependencies rather than memorizing actual patterns. Ultimately, this layer of the neural network block provides a strong temporal understanding that enables the model to separate, differentiate, and then extract audio sources.

Then, the purpose of the upsampling blocks is to gradually recover the source signal from its compressed representation by progressively increasing the size of the representation while improving the quality of the reconstructed musical information. The first upsampling block reduces the number of channels from 128 to 64 during using a Conv2D layer, so the model still has the basic harmonic and timbral information necessary for the reconstruction of the original audio while eliminating the redundant features created by the encoder. Furthermore, the addition of the ReLU activation layer helps highlight spectral elements such as dominant harmonics and transient signals. In the second upsampling block, reducing the number of channels from 64 to 1 helps the network focus on generating a clean, single-channel spectrogram while preserving the frequency characteristics of the instrument to correctly reconstruct the extracted source. The concatenation operation that occurs at this point allows the model to combine high-level semantic details learned in the deeper layers of the network with low-level spatial information from the early encoder layers. This integration helps the model recover fine frequency details, harmonic continuity, and timbral subtleties that were impacted by the downsampling process. The combination of filters and concatenation operations allows the network to increase the size of the audio representation, increase its quality, and build the target source audio in both time and frequency with sufficient temporal and spectral resolution.

The architecture ends with an FCL that provides the final classification by processing the high-level feature extracted throughout the network. The dense layer with 64 units generates a consolidated learned representation of the temporal and spectral feature sets. It achieves this by utilizing a ReLU activation to discover stronger features (dominant harmonics, transient peaks, and instrument-specific signatures), allowing for more accurate identification of instrument categories. The final softmax layer takes the output from the dense layer and converts it to probabilities, which enables the model to identify what portions of the time-frequency spectrogram belong to each instrument (vocals, drums, bass, and other). The classification-based output from the network is vital to providing targeted instrument assignments for audio extraction, thus facilitating a higher accuracy rate with respect to each separated audio component.

3.5. Post-processing

Post-processing serves as the critical final stage in audio source separation, transforming the model's spectrogram predictions into usable audio signals. As illustrated in Fig. 6, this involves two key conversions: (a) shows the inverse short-time Fourier transform (ISTFT) spectrogram (magnitude), visualizing frequency content over time after ISTFT reconstruction, while (b) displays the reconstructed audio waveform in the time domain, illustrating amplitude variations across time. This stage effectively closes the loop between neural network outputs and practical audio applications, enabling immediate playback or downstream analysis.

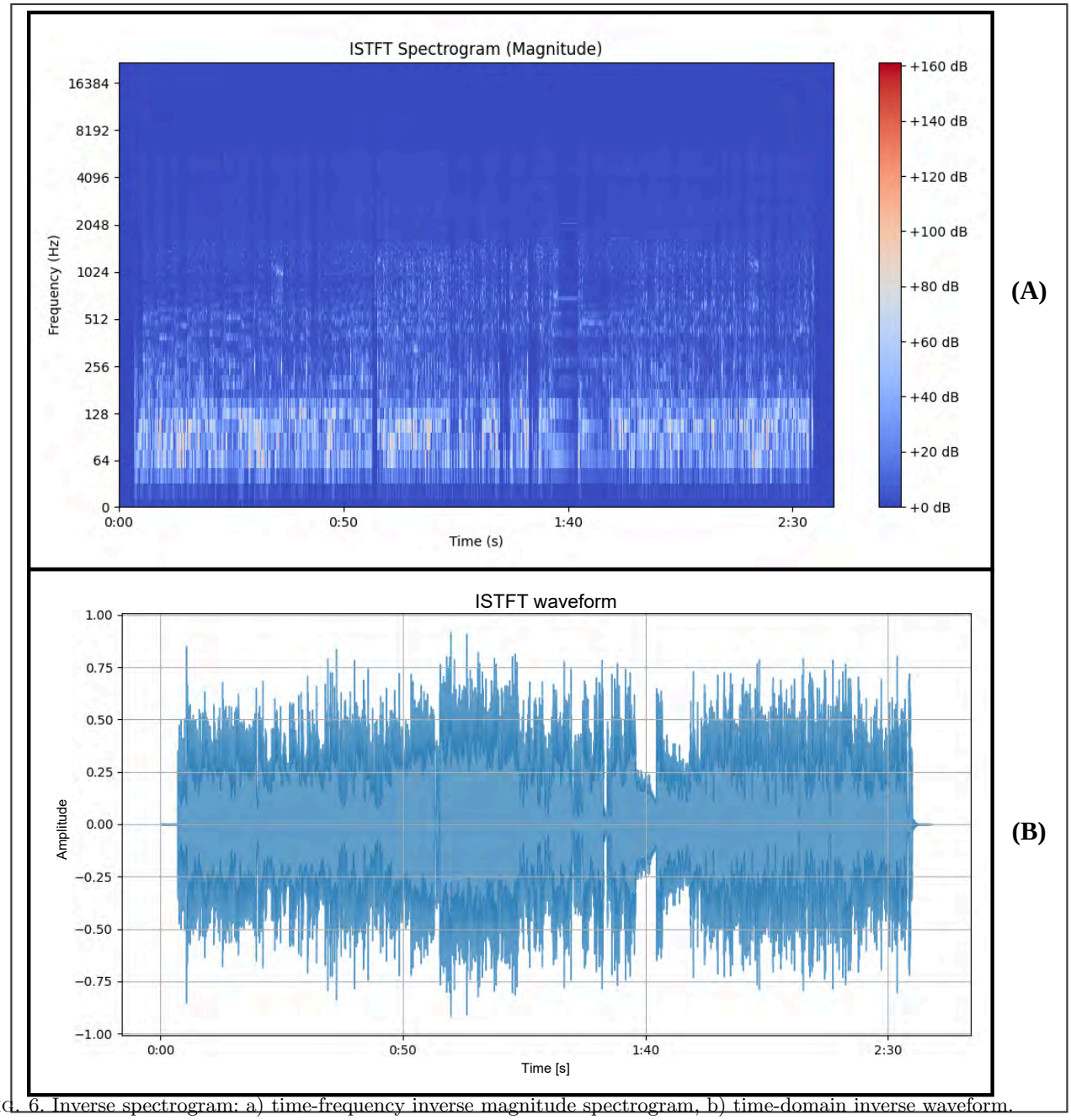


Fig. 6. Inverse spectrogram: a) time-frequency inverse magnitude spectrogram, b) time-domain inverse waveform.

4. Experiment

This section presents the experimental configuration designed to evaluate the proposed ASNet framework. It describes the primary aspects of model validation, the datasets used for training and validating, the evaluation metrics used to evaluate performance, and a comparison with existing methods, to show the improvements achieved by the proposed model. The objective of this experimental study is to demonstrate that the model not only performs well in source separation but also generalizes well across a variety of audio mixtures.

4.1. Post-processing

To objectively evaluate the effectiveness of the proposed audio source separation method, two widely accepted evaluation metrics are employed: signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR). When compared to the original actual sources, these measurements evaluate how well the separated audio sources perform.

SDR describes the fidelity of a separated audio signal by measuring the amount of distortion, artifacts, or errors that remain after separation. An increased SDR score indicates that there are very little distortion or 'artifacts' in the signal, and that the reconstructed audio is a clean and accurate representation of the original source. Conversely, a lower SDR score indicates poorer audio quality and implies that the separated signal contains some distortion, missing components, and/or visible artifacts that diminish clarity.

SIR quantifies how well a model separates the target source signal from any unwanted sounds associated with other instruments or vocal sounds in the original mixture. A higher SIR score indicates that the separated source has been effectively isolated from any competing audio sources and has clearly defined separation boundaries. A lower SIR score indicates that interference remains, with other instruments still audible in the extracted signal; therefore, this indicates that the separation was less efficient. These two metrics evaluate complementary aspects of performance: SDR evaluates the quality of the reconstruction, and SIR evaluates the strength of the separation of the source.

4.2. Performance comparison analysis

This section evaluates and compares the effectiveness of the proposed ASNet models in achieving the desired source separation outcomes. The comparison provides insights into optimal approaches and potential areas for improvement. The metrics SDR and SIR are both measured in decibels.

Table 1 displays the SDR and SIR scores achieved by the proposed ASNet model on three benchmark datasets: MUSDB, DSD100, and MUSDB18-HQ. The evaluation covers four audio source separation tasks: drums (D), vocals (V), bass (B), and other (O). The model demonstrates its best overall performance on the MUSDB18-HQ dataset, achieving the highest SDR and SIR scores among all evaluated datasets. Specifically, the SDR improved by approximately 5.4 % over MUSDB and 12.9 % over DSD100, while the SIR showed an improvement of about 2.7 % over MUSDB and 1.8 % over DSD100. The Conv2D block (256 filters) compresses and enriches features before passing them to the Bi-LSTM block, preserving fine details yielding a higher SDR. Finally, dropout and FCL layers prevent overfitting, boosting model's generalization capacities across datasets.

Table 1. Evaluation of the ASNet model across various benchmark datasets.

Metric	MUSDB				DSD100				MUSDB18-HQ			
	D	V	B	O	D	V	B	O	D	V	B	O
SDR [dB]	11.98	10.73	11.21	10.81	11.19	10.64	10.33	10.68	12.63	11.42	12.01	11.14
SIR [dB]	9.27	9.52	9.51	9.51	9.55	9.46	9.59	9.11	9.57	9.61	9.66	9.67

Descriptive statistics such as the mean, median, and standard deviation (SD) reveal different but complementary dimensions of model behavior when evaluating audio separation, as these metrics provide a snapshot of the model's overall separation capability, a representation of typical separation performance when outliers or other unusual audio samples exist, and an opportunity to assess how consistently well the model separates multiple sample types (i.e., either as a result of the presence of several high-performing samples or as a result of consistently high separation accuracy for many samples). Together, the mean and median provide insight into model strength. While the mean provides an average value of separation success across all samples tested, the median provides a stable assessment of typical audio separation when outliers or other unusual audio samples are present. A third metric, the SD serves as another indicator of separation consistency by providing a measure of variability in the model's performance among multiple recordings. A low SD value implies reliable separation performance, while a high SD value signifies sporadic performance or inconsistent fidelity or separation across samples. Table 2 illustrates these results across the MUSDB, DSD100, and MUSDB18-HQ datasets: the MUSDB dataset has closely aligned mean and median SDR values, and moderate, low variability in the mean values; the DSD100 dataset has a small range of SDR values with no outliers and a much lower SD value; finally, the MUSDB18-HQ dataset has a very high average SDR value of 11.80 dB and a larger SDR SD value (0.570), indicating high level of separation quality alongside consistent performance, as demonstrated by its very low SIR SD

Table 2. Descriptive statistics of SDR and SIR metrics across the MUSDB, DSD100, and MUSDB18-HQ datasets.

Dataset	Mean		Median		Standard deviation	
	SDR [dB]	SIR [dB]	SDR [dB]	SIR [dB]	SDR [dB]	SIR [dB]
MUSDB	11.18	9.45	11.01	9.51	0.495	0.105
DSD100	10.71	9.43	10.66	9.51	0.309	0.190
MUSDB18-HQ	11.80	9.63	11.72	9.64	0.570	0.040

value (0.040). The combination of mean, median, and SD provides a detailed, multi-angle view of the model's ability to separate audio accurately across diverse datasets.

Utilizing SDR and SIR across four sources, Table 3 compares the ASNet's performance under various channel configurations (128, 256, and 512) on the MUSDB18-HQ dataset. At 11.80 dB average SDR, the 256-channel configuration produced the highest average SDR value, followed by the 512-channel configuration at the average SDR value of 6.86 dB and the 128-channel configuration at the average SDR value of 5.03. This represents an approximate 134.6 % improvement compared to the SDR using 128 channels, and approximately 71.9 % increase over the 512-channel setup. SIR values were comparable across all channels with slight variation, indicating no difference in interference suppression regardless of channel amount. These results demonstrate that a 256-channel configuration provides the most effective balance for source separation performance in ASNet.

Table 3. Implementation results of the ASNet architecture based on channel-wise configurations.

Channel count	SDR				SIR			
	B [dB]	D [dB]	V [dB]	O [dB]	B [dB]	D [dB]	V [dB]	O [dB]
128 channels	5.01	4.22	6.23	4.66	9.33	9.29	9.33	9.22
256 channels	12.01	12.63	11.42	11.14	9.57	9.61	9.66	9.67
512 channels	6.38	6.97	8.01	6.09	9.48	9.52	9.47	9.42

Table 4 compares the proposed ASNet model with several existing DL approaches for audio source separation on the MUSDB18-HQ dataset, based on SDR for four sources along with total network parameter count. Despite having a significantly smaller parameter count (0.31 M × 4), ASNet outperforms all other models in SDR across all separated source types, achieving 11.42 dB (V), 12.63 dB (D), 12.01 dB (B), and 11.14 dB (O). In contrast, larger models such as Res-U-Net (102 M × 4) and Mel-Roformer (84.2 M) deliver lower SDR, indicating less accurate source separation. Importantly, ASNet is an interesting case that is able to outperform heavily parameterized architectures, such as BSRNN, DTT, and Hybrid Demucs, when considering complexity and computational efficiency. This indicates that the specifications ASNet must give up for its lightweight design actually favor performance in source separation. All of this proves why ASNet is a better model, offering high performance and reduced computational complexity.

Table 4. Performance comparison of the proposed ASNet against existing DL approaches based on SDR and parameter count.

Model	Params	V	D	B	O
Res-U-Net (Kong *et al.*, 2021)	102.0 M ×4	8.98	6.62	6.04	5.29
Mel-Roformer (Wang *et al.*, 2023)	84.2 M	11.21	9.91	9.64	7.81
Hybrid Demucs (Defossez *et al.*, 2021)	83.6 M × 4	8.13	8.24	8.76	5.59
Band-split recurrent neural network (BSRNN)	37.6 M × 4	10.01	9.01	7.22	6.70
DTT (Chen *et al.*, 2024)	5.0 M × 4	10.12	7.74	7.45	6.92
ASNet (proposed)	0.31 M × 4	11.42	12.63	12.01	11.14

4.3. Discussion

To quantify the computational performance regarding execution speed, latency, throughput, and memory consumption of the audio source separation model, we performed tests using an NVIDIA RTX 3090 GPU and an Intel Xeon Silver 4210 CPU in conjunction with the Pytorch library. We also utilized Python libraries such as

Librosa and soundfile for the preprocessing and inference stages of the ASNet model. This approach allowed us to maintain consistency in input formatting and minimize I/O bottlenecks by providing high-quality audio data and an efficient method for loading audio files, resampling, and generating spectrograms. On average, ASNet processes a 5-second audio sample in 42.7 ms on the RTX 3090 and 315.4 ms on the Xeon 4210, yielding a throughput rate of 23.4 samples/second (audio signals) on the RTX 3090 vs. 3.1 samples/second on the Xeon 4210. Maximum memory usage for the audio source separation model was 2.6 GB on the RTX 3090 vs. 4.2 GB on the Xeon 4210, indicating that the increased speed, efficiency, and suitability for real-time music separation tasks provided by utilizing a GPU in conjunction with Librosa and soundfile for audio processing far outweigh the additional memory resources required by using a CPU.

The ASNet model is trained on three commonly used benchmark datasets: MUSDB18-HQ, MUSDB, and DSD100 with fixed training and validation splits. All models were validated within the same experimental configuration for fair comparison with the baselines. Source separation performance is measured using track-wise SDR and SIR, calculated using standard formulations obtained from the original and separated audio signals. Instead of using the museval framework, the evaluation process was managed through custom metric implementations to allow for complete control of the evaluation process, consistency across all datasets, and ease of integration into the proposed ASNet pipeline. The final performance scores are the average SDR and SIR for each track across the entire test dataset. To allow for comparison and replication of previous work, the baseline results are either reproduced or cited according to the same evaluation protocol established for the ASNet model evaluation compared with other models.

The proposed A2F framework outperforms baseline feature extraction methods by integrating both spectral (STFT, spectral centroid, spectral roll-off, MFCC) and temporal (ZCR, RMS) features. Unlike single-domain methods, A2F's hierarchical fusion captures fine-grained spectral-temporal cues and global acoustic patterns simultaneously. This multi-domain representation significantly enhances source separation accuracy compared to traditional feature pipelines. ASNet via a Bi-LSTM leverages bidirectional sequential modeling to capture temporal dependencies, improving continuity and accuracy in overlapping audio sources. Without the Bi-LSTM, the model relies solely on convolutional features, limiting its ability to learn long-range temporal relationships. The integration of the Bi-LSTM thus significantly enhances source separation performance compared to a purely CNN-based ASNet. Future work will focus on extending ASNet to real-time audio source separation by optimizing inference latency and reducing computational overhead. Model compression and lightweight deployment strategies will be explored to enable integration into streaming and live music applications.

The aim of performing a subjective evaluation, in addition to calculating objective SDR and SIR metrics, was to evaluate the quality of the separated audio recordings as judged by human listeners. In this evaluation, the proposed ASNet model was compared against two standard methods frequently used in music source separation: DTT and BSRNN. The results of the subjective evaluation indicate that the proposed ASNet model achieved substantially higher perceptual quality and naturalness than the two baseline methods on average. The majority of listeners reported fewer audible musical artifacts, better preservation of timbre, and improved temporal continuity in the output from ASNet when processing mixed audio samples consisting of many overlapping musical sources compared to the other two models. Therefore, the difference between the objective and subjective evaluations provides evidence of a strong correlation between objective improvements and perceived audio quality, indicating that the proposed method is well-suited for practical applications.

4.3.1. Limitations

The learning curves show that the validation loss fluctuated during the early and late stages of training, with the MUSDB and MUSDB18-HQ datasets showing the greatest fluctuation. It appears that these validation samples may have been somewhat sensitive or biased with respect to the training data. This could suggest that there was some degree of overfitting occurring despite applying regularization techniques such as dropout. Although the overall results still provide a good level of confidence in its accuracy, further advances can likely be made using more sophisticated methods, such as adaptive early stopping, curriculum learning, and dataset-aware augmentation.

4.3.2. Scope for future work

The goal of the strategies just mentioned is to improve the stability of convergence during model training and decrease the oscillation of the validation loss while maintaining the complexity of the model.

Another limitation of the current framework is that it relies only on using audio spectrogram representations. Although they work well, the use of audio-only representations may limit the ability of the model to separate sources that overlap heavily in very difficult audio environments. Future improvements may come from incorporating adaptive learning techniques and trying out different types of inputs, such as audio-visual fusion, to increase both accuracy and robustness. For instance, integrating information about the movements of instruments or the actions of performers could significantly help in separating overlapping audio sources.

Future research will extend beyond symbolic music representations to convert extracted source signals into MIDI format for the automated generation of improved musical scores. This expanded system will include pitch detection, onset-offset estimation, and note-segmentation models that work together with the separated source files to create a single pipeline for audio-to-score conversions. With this type of extension, there will be a seamless transition from audio source separation to automatic music transcription. Thus, the proposed framework could be utilized in various applications, such as providing educational opportunities for learning about music, digitally preserving music through archiving, analyzing music from an academic perspective, and providing assistance to musicians in compositional and performance endeavors.

In order to improve the fidelity of transcription and robustness, other avenues of exploration will involve incorporating transformer-based temporal modeling, perceptual loss functions, and the application of cross-domain approaches and learning techniques for joint audio and symbolic musical representation. Furthermore, one avenue of continued exploration is in the field of near real-time deployments and assessments of live and/or noisy musical/audio recordings.

5. Conclusion

The proposed ASNet, integrated with the A2F framework, demonstrates a significant advancement in audio source separation. By effectively combining spectral and temporal features, ASNet achieves superior separation quality. The architecture adopts a U-Net-inspired encoder-decoder structure to extract hierarchical feature representations and preserves spatial resolution through skip connections. This architecture allows the network to maintain crucial audio details while learning abstract features at multiple scales. Additionally, to enhance the modeling of long-range temporal dependencies, ASNet incorporates Bi-LSTM layers. These layers enable the network to capture both past and future contextual information. Notably, the 256-channel configuration of ASNet yields the highest SDR and SIR, despite having a considerably smaller parameter count. This effectiveness highlights ASNet's potential for implementation on devices with limited resources and applications that run in real time. Future improvements may also include incorporating adaptive learning strategies, as well as evaluating different modalities of input (e.g., adding audio-visual cues), to improve separation accuracy and robustness. Also, applying quantization-aware training could allow ASNet to be deployed on edge devices without a loss in performance.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' contributions

Shunmugalingam Parvathi Sakthidevi – data analysis, writing, review, and editing. Chandrasekar Divya – supervision.

Acknowledgments

This work was supported by the Manonmaniam Sundaranar University, Centre for Information Technology and Engineering, Tirunelveli, Tamil Nadu, India.

References

1. Chen J., Vekkot S., Shukla P. (2024), Music source separation based on a lightweight deep learning framework (DTTNET: DUAL-PATH TFC-TDF UNET), [in:] *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 656–660, https://doi.org/10.1109/ICASSP48485.2024.10448020.
2. Chuang S.Y., Wang H.M., Tsao Y. (2022), Improved lite audio-visual speech enhancement, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **30**: 1345–1359, https://doi.org/10.1109/TASLP.2022.3153265.
3. Défossez A. (2021), Hybrid spectrogram and waveform source separation, *arXiv*, https://doi.org/10.48550/arXiv.2111.03600.
4. Deng B., Yang H., Kim N.Y. (2024), A denoising autoencoder based on U-Net and bidirectional long short-term memory for multi-level random telegraph signal analysis, *Engineering Applications of Artificial Intelligence*, **135**: 108685, https://doi.org/10.1016/j.engappai.2024.108685.
5. Hamza A. *et al.* (2022), Deepfake audio detection via MFCC features using machine learning, *IEEE Access*, **10**: 134018–134028, https://doi.org/10.1109/ACCESS.2022.3231480.
6. Huber D.M., Caballero E., Runstein R. (2023), *Modern Recording Techniques: A Practical Guide to Modern Music Production*, Focal Press.
7. Issa R.J., Al-Irhaym Y.F. (2021), Audio source separation using supervised deep neural network, *Journal of Physics: Conference Series*, **1879**(2): 022077, https://doi.org/10.1088/1742-6596/1879/2/022077.
8. Kong Q., Cao Y., Liu H., Choi K., Wang Y. (2021), Decoupling magnitude and phase estimation with deep ResUNet for music source separation, *arXiv*, https://doi.org/10.48550/arXiv.2109.05418.
9. Luo Y., Yu J. (2023), Music source separation with band-split RNN, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **31**: 1893–1901, https://doi.org/10.1109/TASLP.2023.3271145.
10. O'Connell D.C., Kowal S. (2009), Transcription systems for spoken discourse, [in:] *The Pragmatics of Interaction*, pp. 240–254, John Benjamins Publishing Company.
11. Rafii Z., Liutkus A., Stöter F.R., Mimilakis S.I., Bittner R. (2019), MUSDB18-HQ-an uncompressed version of MUSDB18, *Zenodo*, https://doi.org/10.5281/zenodo.3338372.
12. Rajaby E., Sayedi S.M. (2022), A structured review of sparse fast Fourier transform algorithms, *Digital Signal Processing*, **123**: 103403, https://doi.org/10.1016/j.dsp.2022.103403.
13. Renukadevi P., John S., Shivani N. (2024), Forensic science: AI-powered image and audio analysis, [in:] *2024 5th International Conference on Smart Electronics and Communication (ICOSEC)*, pp. 1519–1525, https://doi.org/10.1109/ICOSEC61587.2024.10722068.
14. Rezaul K.M. *et al.* (2024), Enhancing audio classification through MFCC feature extraction and data augmentation with CNN and RNN models, *International Journal of Advanced Computer Science and Applications*, **15**(7): 37–53, https://doi.org/10.14569/IJACSA.2024.0150704.
15. Sanders M.E., Kant E., Smit A.L., Stegeman I. (2021), The effect of hearing aids on cognitive function: a systematic review, *PLoS One*, **16**(12): e0261207, https://doi.org/10.1371/journal.pone.0261207.
16. Sun C. *et al.* (2021), A convolutional recurrent neural network with attention framework for speech separation in monaural recordings, *Scientific Reports*, **11**(1): 1434, https://doi.org/10.1038/s41598-020-80713-3.

17. Takahashi N., Mitsufuji Y. (2020), D3Net: densely connected multidilated DenseNet for music source separation, *arXiv*, https://doi.org/10.48550/arXiv.2010.01733.

18. Tong W. *et al.* (2024), SCNet: sparse compression network for music source separation, [in:] *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1276–1280, https://doi.org/10.1109/ICASSP48485.2024.10446651.

19. Wang J.-C., Lu W.-T., Won M. (2023), Mel-band roformer for music source separation, *arXiv*, https://doi.org/10.48550/arXiv.2310.01809.

20. Wu X., Dang B., Zhang T., Wu X., Yang Y. (2024), Spatiotemporal audio feature extraction with dynamic memristor-based time-surface neurons, *Science Advances*, **10**(14): eadl2767, https://doi.org/10.1126/sciadv.adl2767.

21. Zhang Y. *et al.* (2025), Comparison of performance for cochlear-implant listeners using audio processing strategies based on short-time fast Fourier transform or spectral feature extraction, *Ear and Hearing*, **46**(1): 163–183, https://doi.org/10.1097/AUD.0000000000001565.

22. Zölzer U. (2022), *Digital Audio Signal Processing*, John Wiley & Sons.

Archives of Acoustics Vol. 51, No. 3, pp. 549–560 (2026), https://doi.org/10.24423/archacoust.2026.4416

Research Paper

Musical Key Detection Using a Triple Composite Signature of Fifths

Tomasz Łukaszewicz*, Dariusz Kania

Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology
Gliwice, Poland

*Corresponding Author: tlukaszewicz@polsl.pl

Received January 20, 2026; revised April 24, 2026; accepted April 22, 2026;
available online April 28, 2026; version of record September 14, 2026; published issue September 23, 2026.

This article introduces a novel algorithm for key detection in musical compositions represented in symbolic form, such as MIDI. The method is based on the analysis of a triple composite signature of fifths (TCSF), which is formed by combining the signatures of fifths obtained individually for the beginning, the end, and the entirety of the analyzed piece. The key is determined by the sign of the angle between the characteristic vector of the TCSF and the major/minor mode axis derived from that signature. To evaluate the algorithm's effectiveness, experiments were conducted using Chopin's Preludes, Op. 28, pieces from the Saarland Music Data: MIDI-Audio Piano Music dataset, as well as pieces from the Schubert Winterreise Dataset. As a reference method, a correlation-based algorithm implementing various tonal profiles was used. The proposed algorithm achieved the highest accuracy for Chopin's Preludes and the Schubert Winterreise Dataset (91.67 % in both cases), while for the Saarland Music Data, its accuracy was lower than that of the correlation-based approaches (72.34 %). The main advantages of the presented algorithm are low computational complexity, stable decision-making when extending the analytical window, as well as incorporation of expert analysis aspects, particularly by focusing on the beginning and end of the composition.

Keywords: key detection, tonality, music information retrieval, music classification.



Copyright © 2026 The Author(s).
This work is licensed under the Creative Commons Attribution 4.0 International CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The dynamic development of machine learning and artificial intelligence techniques in recent years has contributed to significant advances in computational music analysis, encompassing an increasingly broad range of applications – from acoustic signal processing to advanced structural and semantic analysis of musical pieces (Weiß *et al.*, 2019). Of particular importance in this context is the field of music information retrieval (MIR) and automatic genre classification, which form the foundation of many contemporary applications, such as intelligent recommendation systems, music search engines, and tools supporting artistic creation (Kostrzewa *et al.*, 2022; Rosner, Kostek, 2018; Schedl *et al.*, 2014; Sturm, 2013).

In music genre classification methods, sets of acoustic and musical features are commonly employed – such as spectral centroids, Mel-frequency cepstral coefficients (MFCCs), rhythm, and tempo – which enable effective differentiation of individual compositions and musical styles. These features are utilized not only for genre identification, but also for affective analyses, namely, in determining emotions expressed or evoked by music (Chapin *et al.*, 2020; Grekow, 2018; Juslin, Sloboda, 2010; Roig *et al.*, 2014; Yang *et al.*, 2023).

A very important and widely studied issue in MIR is the harmonic analysis and evaluation of musical compositions, encompassing automatic chord recognition, local key identification, and the identification of harmonic structures. Research in this area is of key importance for both music genre classification and emotional analysis,

as well as the formal segmentation of compositions (Cho, 2013; Harte *et al.*, 2006; Jiang *et al.*, 2024; Mauch, 2010; McFee *et al.*, 2018; Pauwels *et al.*, 2019). Precise determination of harmonic sequences allows for the extraction of characteristic tonal patterns and chord progressions, which form the foundation of many applications such as automatic accompaniment generators, recommendation systems, and tools supporting composition and style analysis (Pauwels *et al.*, 2019). Thanks to the development of machine learning methods, particularly sequential models and deep neural networks, it has become possible to precisely model increasingly complex harmonic dependencies in musical compositions (Bittner *et al.*, 2017; Briot *et al.*, 2020).

Research in the field of affective computing emphasizes the significance of tonality and harmonic structure as key elements in the emotional perception of a musical work. Therefore, in all the aforementioned applications, the precise algorithmic determination of the musical key is of particular importance, and above all, its tonal mode (major/minor), which constitutes one of the most significant indicators affecting the classification of music (Duan *et al.*, 2008). The integration of tonal analysis with modern deep learning methods represents one of the most intensively developing research directions in the field of MIR.

The literature presents many ways of modeling tonality. The most commonly encountered are 2D and 3D models (Chew, 2000; 2008; Longuet-Higgins, 1962a), although more complex, multidimensional models of tonality can also be found (Harte *et al.*, 2006; Bernardes *et al.*, 2016). The origins of 2D tonality representations date back to the work of Leonhard Euler, who proposed a harmonic network called the Tonnetz, depicting the relationships between the pitches of various chords (Euler, 1739). Similar intervallic dependencies were subsequently used in Longuet-Higgins (1962a; 1962b) maps and in various types of spiral spatial models (Chew, 2000; 2008; Shepard, 1982). Recently, approaches to tonal analysis have increasingly relied on artificial intelligence and machine learning algorithms (Dawson, 2018; Deng, Kwok, 2017; Korzeniowski, Widmer, 2017; Zhou, Lerch, 2015). Artificial intelligence methods, in combination with various tonality models, are used, for instance, in automated music composition systems (Huang *et al.*, 2016; Ji, Yang, 2024; Sabathé *et al.*, 2017; Wu *et al.*, 2023).

A very important element used in tonal analysis is tonal profiles, as they allow for the identification of the key of musical pieces (Albrecht, Shanahan, 2013; Herremans, Chew, 2019; Kania, Kania, 2019; Krumhansl, 1990; Temperley, Marvin, 2008). Although many different tonal profiles have been proposed in (Dawson, 2018; Albrecht, Shanahan, 2013; Krumhansl, 1990; Temperley, Marvin, 2008; Aarden, 2003; Bellmann, 2006), it remains unclear whether the problem of determining the key in musical pieces has been solved. Errors in key recognition often result in identifying the relative key or parallel key instead of the ground-truth one, leading to the misclassification of the tonal mode of the analyzed piece (Duan *et al.*, 2008; Łukaszewicz, Kania, 2025a). It is worth noting that alternative methods have also been developed that utilize the tonal relationships expressed in the structure of the circle of fifths (Kania, Kania, 2019; Kania *et al.*, 2024). These models became the starting point for research on the signature of fifths (Łukaszewicz, Kania, 2022; 2025a; 2025b; Kania *et al.*, 2021b).

Work on applying the signature of fifths to detect key signatures (Kania *et al.*, 2021b) and keys (Kania, Kania, 2019; Kania *et al.*, 2024), along with research on the temporal variability of the signature tracking via the trajectory of fifths (Łukaszewicz, Kania, 2025c; Kania *et al.*, 2021a; Kania, Kania, 2022) motivated the development of more refined methods. Kania *et al.* (2024) proposed a simple method of determining the key based on an analysis of the signature of fifths, which measures the deviation of the main directed axis of the signature relative to the major/minor mode axis. The obtained results became the basis for developing a very effective algorithm to classify the tonal mode, presented in a subsequent work (Łukaszewicz, Kania, 2025a). The effectiveness of this approach stems from a multi-criteria assessment of the tonal mode, consisting not only of analysis of the entire piece, but also its beginning and end. This inspired us to seek methods for improving the previously developed key determination method (Kania *et al.*, 2024) by analyzing the directions of characteristic vectors of the signatures of fifths, computed for different segments of an analyzed piece.

The purpose of this article is to present an algorithm for determining the key of a musical composition using a triple composite signature of fifths (TCSF), which is formed by combining the signatures of fifths obtained individually from the beginning, the end, and the entirety of the analyzed piece. The final classification decision

is determined by the sign of the angle between the characteristic vector of the TCSF and the major/minor mode axis derived from that signature.

2. Theoretical background

A musical composition constitutes an organized sequence of pitches in which it is typically possible to distinguish a melodic line and an accompaniment layer. The analysis of successive temporal intervals of a composition allows for the determination of changing chord configurations. Frequently, these chords contain notes separated by an octave or its multiples. The choice of the time interval for analysis can be linked to the metric foundation of the examined composition, which also influences the process of constructing the signatures of fifths (Kania, Kania, 2019; Kania *et al.*, 2024).

Let us consider the first quarter-note window of Prelude No. 2 in C minor (BWV 847) by J.S. Bach from the collection Das Wohltemperierte Klavier, Book I, which is shown in Fig. 1. The window contains eight notes: c0, f0, 2 x g0, d1, 2 x e♭1, and c2. Disregarding octave distinctions, the multiplicities of individual pitch classes, after normalization with respect to the maximum multiplicity (which equals 2), are, respectively: $\left|\overrightarrow{C}\right| = \left|\overrightarrow{E\flat}\right| = \left|\overrightarrow{G}\right| = 1$, and $\left|\overrightarrow{D}\right| = \left|\overrightarrow{F}\right| = 0.5$. Assuming a unit radius for the circle of fifths, the normalized multiplicities correspond to the lengths of the vectors forming the signature of fifths presented in Fig. 2.
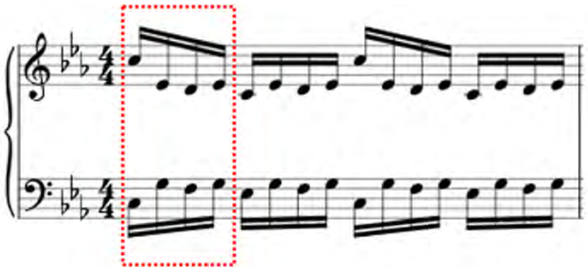


Fig. 1. Opening fragment of Prelude No. 2 in C minor by J.S. Bach from the collection Das Wohltemperierte Klavier, Book I (BWV 847).

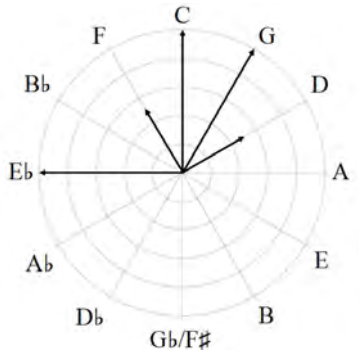


Fig. 2. Signature of fifths corresponding to the first quarter-note window of Prelude No. 2 in C minor by J.S. Bach from the collection Das Wohltemperierte Klavier, Book I (BWV 847), as shown in Fig. 1.

For the signature of fifths, it is possible to determine several characteristic elements that allow for the recognition of the key of the composition (Kania *et al.*, 2024). They include: the main directed axis of the signature of fifths (MDASF) (highlighted in red) – which is the axis for which the difference between the sum of the magnitudes of vectors on the right side and the sum of the magnitudes of vectors on the left side reaches its maximum value, perpendicular to it, the major/minor mode axis (shown in green), and the characteristic vector of the signature of fifths (shown in blue), which is the sum of the vectors forming the signature of fifths (Kania *et al.*, 2024). These elements, corresponding to the signature presented in Fig. 2, are illustrated in Fig. 3. Since only the direction of the characteristic vector of the signature of fifths is important in the key determination algorithm, its

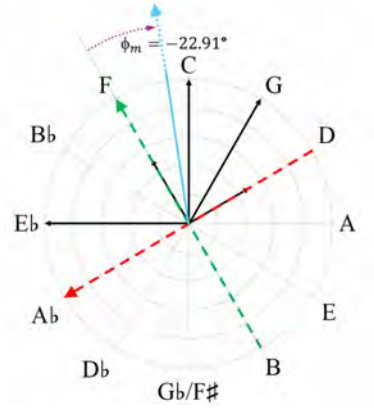


FIG. 3. Signature of fifths and its associated elements: the MDASF (red color), the major/minor mode axis (green color), and the characteristic vector of the signature of fifths (blue color).

length has been shortened, which is symbolically marked with a dashed line. From the perspective of determining the tonal mode, the angle between the characteristic vector of the signature of fifths and the major/minor mode axis is important, i.e., ϕ_m. Detailed information about the method of creating the aforementioned axes and the characteristic vector of the signature of fifths can be found in (KANIA *et al.*, 2024).

Each position on the circle of fifths can be associated not only with major keys but also with their corresponding relative minor keys. The direction of the MDASF enables the identification of a pair of relative keys (KANIA, KANIA, 2019). They are positioned 30° clockwise relative to the keys indicated by the MDASF (KANIA, KANIA, 2019; KANIA *et al.*, 2024). Determining the tonal mode boils down to the selection of one of those keys. For example, for the signature presented in Fig. 3, the pair of relative keys is E♭ major and C minor. The selection of one of the two relative keys can be reduced to determining the value of the angle ϕ_m shown in Fig. 4 (the angle between the characteristic vector of the signature of fifths and the major/minor mode axis) (KANIA *et al.*, 2024). In the case considered here, its value is −22.91°, which due to the negative sign indicates a minor key, namely C minor (KANIA *et al.*, 2024). It can be seen, therefore, that by analyzing the first quarter-note window of the piece, it was possible to correctly determine the key of the composition.

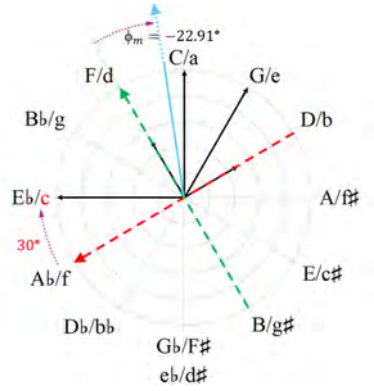


FIG. 4. Illustration of the elements used for key determination based on the signature of fifths (description in the text).

It should be noted that both the coefficients enabling the selection of the direction of the main directed axis and the value of the angle ϕ_m indicate the level of association of the analyzed fragment of the composition with a particular key. Knowing the orientation of the MDASF makes it possible to determine the pair of relative keys. The value of the angle ϕ_m, on the other hand, reflects the 'strength' of the association between the analyzed sample and a particular tonal mode.

Figure 5 illustrates several signatures of fifths corresponding to simple chords. In the first case (Fig. 5a), the signature corresponding to the C major chord is shown. The distribution of vectors in the signature clearly

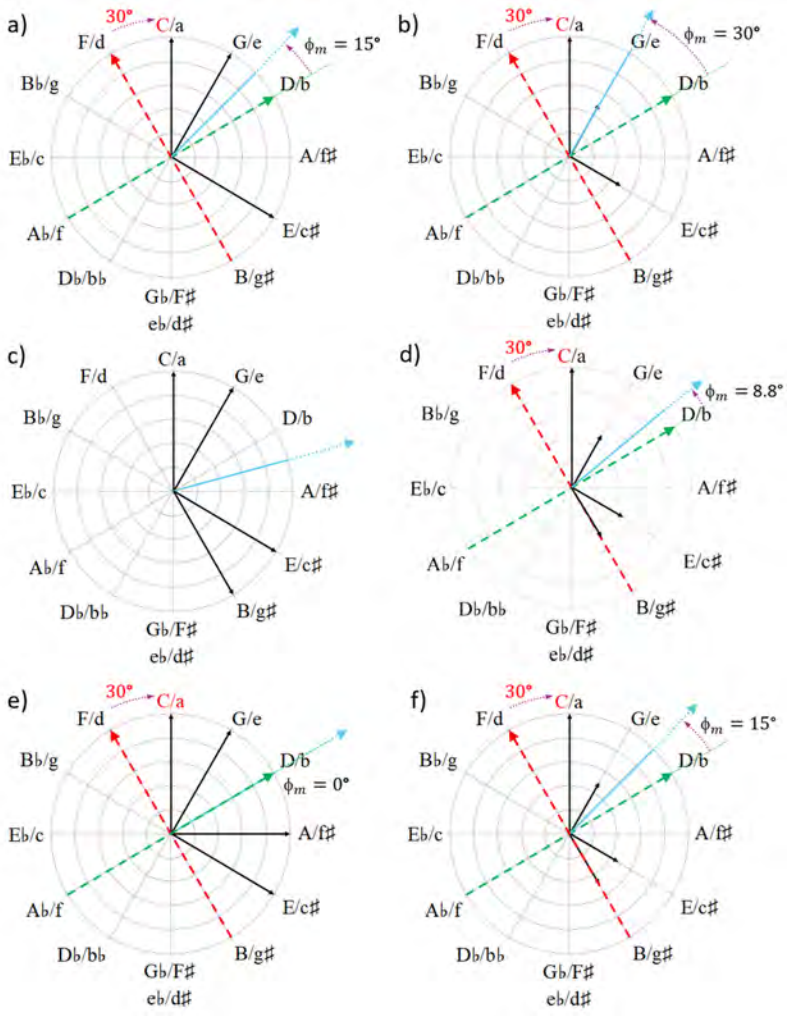


Fig. 5. Example signatures of fifths: a) C major chord, b) C major chord with doubled C note, c) C major seventh chord, d) C major seventh chord with doubled C note, e) C major sixth chord, f) C major sixth chord with doubled C note (description in the text).

indicates the MDASF as B→F. Algebraically, the MDASF is the one for which the difference between the sum of vector lengths on the right side of the axis and the sum of vector lengths on its left side is the greatest (Kania, Kania, 2019). In the considered case, its value, represented as the coefficient [B→F], equals 3. The next largest value (i.e., 2) was obtained for the axes: E♭ →B and G♭/F♯ →C. The direction of the MDAFS thus indicates the relative keys C major and A minor. The positive value of the angle ϕ_m, i.e., 15°, indicates the major key, namely C major. In Fig. 5b, the signature corresponding to the C major chord with a doubled tonic is shown. In this case, there is also no doubt regarding the determination of the main directed axis, which is the axis B→F. The larger value of angle ϕ_m, i.e., 30°, more clearly emphasizes the major character of the chord. The situation is different in the case of the C major chord with an added major seventh (C major seventh chord), whose signature is illustrated in Fig. 5c. In this case, there are two directed axes with a maximum value of the coefficient: [B→F] = [G♭/F♯ →C] = 3; hence, for such a sample, the tonality cannot be determined. This is not surprising, since the C major seventh chord contains two triads: C–E–G and E–G–B, of which the first has a clearly major character (C major) and the second a minor character (E minor). Additionally, they belong to two neighboring pairs of relative keys (C major / A minor; G major / E minor). Doubling the note C (Fig. 5d) shifts the balance toward the main axis B→F, indicating the relative keys C major and A minor, and in light of the positive value of the angle ϕ_m = 8.8°, the major key is chosen, i.e., C major. The situation is different in the case of the C major chord with added sixth (C6), whose signature of fifths is presented in Fig. 5e. As before, it contains two triads: C–E–G

and A–C–E, of which the first is major and the second is minor. In this case, the main directed axis is B→F, since these chords are associated with the same pair of relative keys (C major /A minor). The value of the angle $\phi_m = 0°$ does not, however, allow for determining the tonal mode, and therefore the key cannot be unambiguously determined based on the given sample. Doubling the tonic note (for example, C), shown in Fig. 5f, again allows for determining the key as C major.

3. Musical key detection algorithm based on a TCSF

The examples presented in Fig. 6 illustrate various scenarios that can be encountered when attempting to determine the key of a composition based on its selected fragment. Very often, similar ambiguities are observed when analyzing entire compositions. There are compositions that begin in one key and end in another, or while being clearly anchored in a minor key, end with a major chord, and vice versa. It is therefore clear that there is a need to draw conclusions based on several fragments of the analyzed piece. It is worth emphasizing that the method of determining the key using the signature of fifths is quite cautious in making a decision. This distinguishes it from methods utilizing tonal profiles, which tend to frequently change the indicated tonality when the analytical window expands (Kania *et al.*, 2022; 2024). Despite this caution, it would be worthwhile, while maintaining the simplicity of the previously developed algorithms (Kania, Kania, 2019; Kania *et al.*, 2024), to enrich it with information that can further improve the effectiveness of determining the key of a composition.

The results of numerous experiments conducted, including those presented in a previous work (Łukaszewicz, Kania, 2025a), clearly indicate that the best results are obtained when determining the key based on an entire composition. Sometimes the decision made for the entire composition is controversial; hence, it is worthwhile to additionally take into account the results of key detection for the beginning and end of the analyzed piece.

It is reasonable to propose a methodology in which the signature of fifths corresponding to a given piece of music would emphasize the influence of the beginning and ending fragments of the composition, since beginnings and endings of compositions are typically clearly anchored in their tonality.

Through extensive multifaceted work, a form of the signature of fifths was developed that contains the most important information from the perspective of correctly determining the key while maintaining the simplicity of the previously developed algorithms (Kania, Kania, 2019; Kania *et al.*, 2024).

The proposed form of the signature is the TCSF. It consists of three signatures of fifths, which correspond to the entire composition, its beginning, and its end. It is created by summing corresponding components of vectors comprising individual signatures of fifths, followed by normalization relative to the length of the longest vector.

The proposed algorithm for key determination based on the TCSF can be divided into the following general steps:

1. Determination of the signatures of fifths for the beginning and end of the composition (we search for the smallest number of initial/final notes for which it is possible to determine the direction of the MDASF).
2. Determination of the signature of fifths for the entire composition.
3. Determination of the TCSF by summing the corresponding component vectors of individual signatures obtained in steps 1 and 2.
4. Determination of the main directed axis of the TCSF, and identification of the pair of relative keys, as in (Kania, Kania, 2019; Kania *et al.*, 2021b; 2022; 2024).
5. Finding the major/minor mode axis of the analyzed composition, as in (Kania, Kania, 2019; Kania *et al.*, 2021b; 2022; 2024).
6. Determination of the value of the angle ϕ_m between the characteristic vector of the TCSF and the major/minor mode axis.
7. Selection of the tonal mode and one of the relative keys according to the rule:
   - major key when $\phi_m > 0°$,
   - minor key when $\phi_m < 0°$,
   - no decision when $\phi_m = 0°$.

A detailed algorithm is presented in Fig. 6.

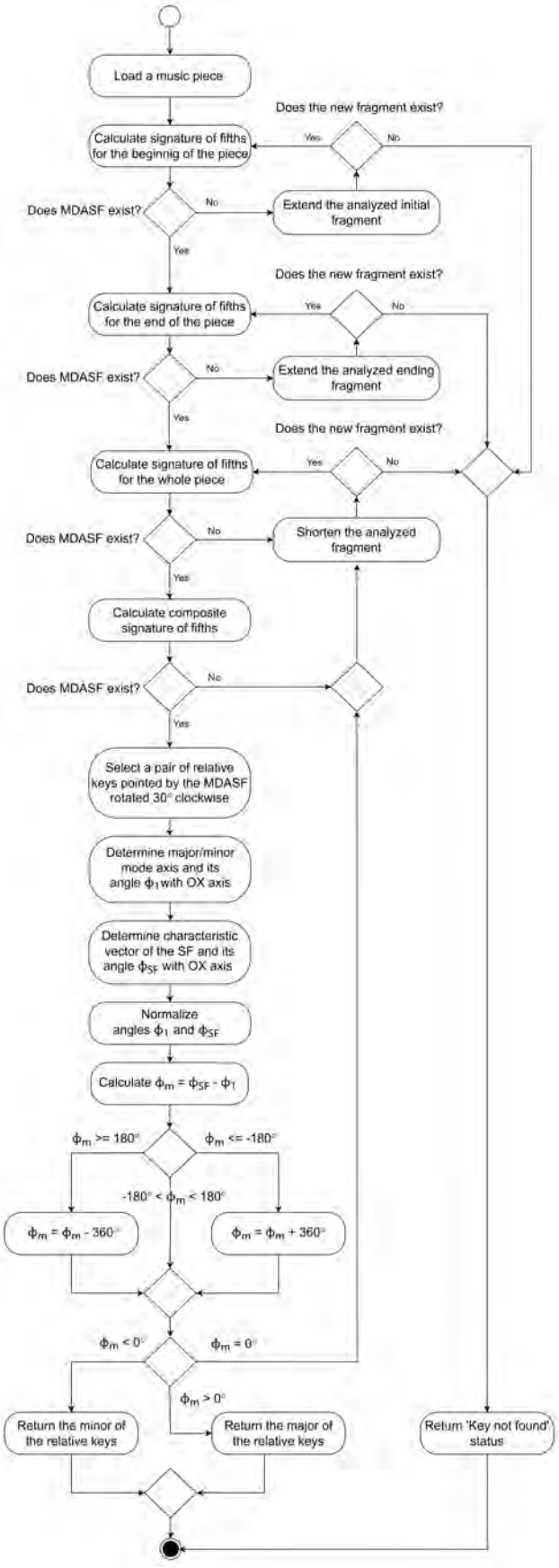


FIG. 6. Diagram presenting the algorithm for determining the key using the TCSF approach.

4. Experiments and discussion

To confirm the effectiveness of recognizing the key using the proposed algorithm, experiments were conducted on classical music compositions from the cycle of 24 Preludes op. 28 by F. Chopin, a collection of piano pieces contained in the Saarland Music Data: MIDI-Audio Piano Music dataset (Müller *et al.*, 2024), and the Winterreise collection, consisting of 24 songs by F. Schubert (Weiß *et al.*, 2020). All of the considered compositions were represented in symbolic form (MIDI files).

The set of Chopin's Preludes, Op. 28 contains compositions in all 24 possible keys, starting from those without key signatures (C major, A minor), then with one sharp (G major, E minor), with two sharps (D major, B minor), and so on, following the arrangement of keys on the circle of fifths. The Saarland Music Data collection contains 50 MIDI files of piano pieces by Bach, Beethoven, Haydn, Mozart, Chopin, Liszt, Bartók, Rachmaninoff, Scriabin, and Ravel. The analysis excluded Béla Bartók's piano sonata Sz. 80 (Müller *et al.*, 2024), as this composition uses the piano in a percussive manner, making it unsuitable for the comparison of key determination algorithms. In the collection of Schubert's Winterreise songs, one can find compositions written in diverse keys, featuring atypical modulations and numerous changes of local key, which make determining the global key of individual songs a non-trivial task. The results of the experiments conducted for the proposed algorithm (TCSF) and correlational approaches utilizing various tonal profiles (KK, T, AE, BB, AS) are presented in Table 1. The values in this table represent the accuracy of key recognition, expressed as the percentage of correctly detected keys in a given set of pieces.

Table 1. Comparison of key detection accuracies across different algorithms.

Algorithm	Chopin's Preludes, Op. 28 [%]	Saarland Music Data: MIDI-Audio Piano Music dataset [%]	Schubert Winterreise dataset [%]
TCSF	91.67	72.34	91.67
KK	79.17	74.47	75
T	83.33	91.49	79.17
AE	87.5	82.98	79.17
BB	87.5	89.36	83.33
AS	75	82.98	75

TCSF – the proposed algorithm using the triple composite signature of fifths approach; KK, T, AE, BB, AS – correlational algorithm implementing Krumhansl-Kessler (Krumhansl, 1990), Temperley (Temperley, 2008), Aarden-Essen (Aarden, 2003), Bellman-Budge (Bellmann, 2006), and Albrecht–Shanahan (Albrecht, Shanahan, 2013) key profiles, respectively.

The results of the conducted experiments showed that the proposed method (TCSF) achieved the highest key recognition accuracy, exceeding 90 %, for the set of Chopin's preludes and Schubert's songs. However, for the compositions contained in the Saarland dataset, significantly lower effectiveness was observed, only 72 %. In this case, the best results (91.49 %) were obtained using correlational method implementing Temperley profiles.

For the set of Chopin's Preludes, Op. 28, the proposed method incorrectly identified the key in two cases: Prelude No. 2 and Prelude No. 18. In the case of Prelude No. 2, the incorrect key identification arises from the specific structure of the composition, where the tonic chord appears only at the very end. Incorrect key determination for this prelude was also observed with the correlational method for all of tonal profiles considered.

An in-depth analysis of the partial results, with particular attention to cases of incorrect key determination, revealed many interesting nuances. In most cases, the classification result was unambiguous, although there were situations in which the detected key was not uniform across the beginning, end, and entirety of a composition. An example is Prelude No. 15, commonly known as the Raindrop Prelude, for which the correct key, i.e., D♭ major, was detected at the beginning and end, but the analysis of the entire composition indicated an incorrect key, i.e., G♯ minor. Analysis of the composition's structure shows that this occurs because the middle section of the piece is composed in C♯ minor. The detected key of G♯ minor likely results from the continuous presence of the note A♭/G♯, associated with the raindrop motif, which is written as A♭ in the first and final parts and as G♯ in the middle section.

An interesting case is Prelude No. 18, which is composed in F minor. However, the piece begins with the note A, which lies outside the F minor scale. This atypical opening causes key determination algorithms to misclassify the key, especially when relying on the initial segment of the composition. The structure of the signature of fifths at the beginning significantly influences the direction of the main axis in the composite signature, resulting in an incorrect outcome for the TCSF method. An in-depth analysis of partial results revealed that increasing the resolution of the analysis to eighth-note windows reduces the impact of the misleading signature of fifths at the beginning, yielding a correct final result. This demonstrates that proper adjustment of the analysis resolution can have a beneficial effect on the effectiveness of key recognition.

When comparing the accuracies of the considered methods for Chopin's Preludes, it is evident that the TCSF algorithm achieves significantly better results than approaches based on tonal profiles. For Prelude No. 2, the correlational method failed to identify the correct key across all profiles used. In the case of Prelude No. 21, only the proposed method correctly determined the key. This outcome is primarily due to the unique approach of the TCSF algorithm, which relies on the signature of fifths and employs 'cautious inference,' without a tendency to change earlier decisions when the analytical window is expanded. This behavior is discussed in detail in (Kania *et al.*, 2024).

In the case of compositions belonging to the Saarland dataset, the effectiveness of key recognition using the proposed method was significantly lower compared to the effectiveness observed for Chopin's preludes. In most cases of incorrect decisions, a key relative to the correct key was indicated (38.46 %). In 23.08 % of cases of incorrect detection, the parallel key was indicated, and in 30.77 %, a key a fifth apart from the correct key. Thus, the method of determining the tonal mode had the greatest impact on the reduction in detection accuracy. This observation is consistent with earlier experiments discussed in (Łukaszewicz, Kania, 2025b). Although it is possible to achieve better results in determining the tonal mode, using the method presented in a related study (Łukaszewicz, Kania, 2025a), it would require the creation and analysis of both the signature and trajectory of fifths (Kania *et al.*, 2021a), increasing the complexity of the algorithm.

In the case of Schubert's compositions, the proposed algorithm consistently outperformed approaches implementing tonal profiles. Only two pieces yielded incorrect results, specifically pieces No. 8 and No. 22. For these compositions, neither the proposed method nor the correlational approaches identified the correct key. These two pieces were composed in G minor, with significant portions in G major, resulting in a distinctly blurred key histogram, further intensified by frequent use of F♯, the third of the major dominant. Additionally, piece No. 8 ends with a G major chord, which unambiguously leads nearly all detection approaches considered (except the one using KK profiles) to identify G major as the key. A detailed analysis of partial results for signatures of fifths generated using the proposed algorithm indicates that, for many pieces, signatures of fifths indicate different keys for the beginning, end, or entirety of a piece. This is not surprising, given the frequent harmonic modulations and changes in local tonalities in Schubert's songs. In 11 pieces, non-uniform directions of the main directed axes in partial signatures of fifths were observed. However, this did not lead to an incorrect determination of the main directed axis of the TCSF, ensuring the correct key determination. For the correlational methods based on different tonal profiles, this variability of local keys, which occurs frequently in the analyzed Schubert's pieces, led to incorrect determination of the global key.

5. Conclusions

The article presented an original method for determining the key of a musical composition using the TCSF. The novelty of the proposed method lies in the way the signature is constructed, which is deeply rooted in the theoretical aspects of music. The concept of creating the TCSF was inspired by the results of experimental research presented in (Kania, Kania, 2019; 2022; Kania *et al.*, 2024; Łukaszewicz, Kania, 2025b).

The main goal of this article was to propose an effective method for determining the key of musical pieces, characterized by the simplicity of algorithms based on the signature of fifths and their inherent stability in the decision-making process as the analytical window expands, and incorporating expert analysis aspects – specifically, paying special attention to the beginning and end of the analyzed composition. The proposed algorithm using the

TCSF effectively fulfilled this goal. The experiments demonstrated the competitiveness of the proposed approach compared to key recognition based on tonal profiles.

An in-depth analysis of the obtained results indicated that the weakest point of the proposed method is the determination of the major/minor tonal mode. This is not surprising, as a similar conclusion emerged from the comparison of key identification algorithms presented in (Łukaszewicz, Kania, 2025c). It may be presumed that better results could be achieved by supplementing the key detection algorithm based on the TCSF with elements of statistical analysis, involving correlation of the signature of fifths obtained for the analyzed sample with tonal profiles, in order to select one of the relative keys indicated by the main directed axis of the signature. The disadvantage of such an approach, however, would be increased computational complexity. One could also use an even more effective method of determining the tonal mode, presented in related research (Łukaszewicz, Kania, 2025a), based on the analysis of the points forming the trajectory of fifths. It may be worthwhile to go even further and, disregarding the TCSF approach, attempt to determine the key of a musical composition based on the distribution of points forming the trajectory of fifths. This approach could be relatively simple and potentially more effective than existing methods.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' contributions

Both authors conceptualized the study, wrote the original draft, performed the analysis and contributed to data interpretation, reviewed and approved the final manuscript.

References

1. Aarden B.J. (2003), *Dynamic melodic expectancy*, Ph.D. Thesis, The Ohio State University.
2. Albrecht J., Shanahan D. (2013), The use of large corpora to train a new type of key-finding algorithm: an improved treatment of the minor mode, *Music Perception: An Interdisciplinary Journal*, **31**(1): 59–67, https://doi.org/10.1525/mp.2013.31.1.59.
3. Bernardes G., Cocharro D., Caetano M., Guedes C., Davies M.E.P. (2016), A multi-level tonal interval space for modelling pitch relatedness and musical dissonance, *Journal of New Music Research*, **45**(4): 281–294, https://doi.org/10.1080/09298215.2016.1182192.
4. Bellmann H. (2006), About the determination of key of a musical excerpt, [in:] *Computer Music Modeling and Retrieval (CMMR 2005)*, Kronland-Martinet R., Voinier T., Ystad S. [Eds], **3902**: 76–91, Springer, https://doi.org/10.1007/11751069_7.
5. Bittner R.M., McFee B., Salamon J., Li P., Bello J.P. (2017), Deep salience representations for F0 estimation in polyphonic music, [in:] *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 63–70.
6. Briot J.-P., Hadjeres G., Pachet F.-D. (2020), *Deep Learning Techniques for Music Generation*, Springer Cham, https://doi.org/10.1007/978-3-319-70163-9.
7. Chapin H., Jantzen K., Kelso J.A.S., Steinberg F., Large E. (2020), Dynamic emotional and neural responses to music depend on performance expression and listener experience, *PloS ONE*, **5**(12): e13812, https://doi.org/10.1371/journal.pone.0013812.
8. Chew E. (2000), *Towards a mathematical model of tonality*, Ph.D. Thesis, Massachusetts Institute of Technology.

9. Chew E. (2008), Out of the grid and into the spiral: geometric interpretations of and comparisons with the spiral-array model, *Computing in Musicology*, **15**: 51–72.
10. Cho T. (2013), *Improved techniques for automatic chord recognition from music audio signals*, Ph.D. Thesis, New York University, New York, USA.
11. Dawson M.R.W. (2018), *Connectionist Representations of Tonal Music. Discovering Musical Patterns by Interpreting Artificial Neural Networks*, AU Press, Athabasca University.
12. Deng J., Kwok Y.-K. (2017), Large vocabulary automatic chord estimation using deep neural nets: design framework, system variations and limitations, *arXiv*, https://doi.org/10.48550/arXiv.1709.07153.
13. Duan Z., Lu L., Zhang C. (2008), Audio tonality mode classification without tonic annotations, [in:] *IEEE International Conference on Multimedia and Expo*, pp. 1361–1364.
14. Euler L. (1739), *An Attempt at a New Theory of Music, Exposed in All Clearness, According to the Most Well-Founded Principles of Harmony* [in Latin: *Tentamen Novae Theoriae Musicae Ex Certissismis Harmoniae Principiis Dilucide Expositae*], Saint Petersburg Academy.
15. Grekow J. (2018), Musical performance analysis in terms of emotions it evokes, *Journal of Intelligent Information Systems*, **51**(2): 415–437, https://doi.org/10.1007/s10844-018-0510-y.
16. Harte C., Sandler M., Gasser M. (2006), Detecting harmonic change in musical audio, [in:] *Proceedings of Special 1st ACM Workshop on Audio and Music Computing Multimedia*, pp. 21–26, https://doi.org/10.1145/1178723.1178727.
17. Herremans D., Chew E. (2019), MorpheuS: generating structured music with constrained patterns and tension, *IEEE Transactions on Affective Computing*, **10**(4): 510–523, https://doi.org/10.1109/TAFFC.2017.2737984.
18. Huang C.-Z.A., Duvenaud D., Gajos K.Z. (2016), ChordRipple: recommending chords to help novice composers go beyond the ordinary, [in:] *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pp. 241–250, https://doi.org/10.1145/2856767.2856792.
19. Jiang X., Zhang Y., Lin G., Yu L. (2024), Music emotion recognition based on deep learning: a review, *IEEE Access*, **12**: 157716–157745, https://doi.org/10.1109/ACCESS.2024.3484470.
20. Ji S., Yang X. (2024), EmoMusicTv: emotion-conditioned symbolic music generation with hierarchical transformer VAE, *IEEE Transactions on Multimedia*, **26**: 1076–1088, https://doi.org/10.1109/TMM.2023.3276177.
21. Ji S., Yang X., Luo J. (2023), A survey on deep learning for symbolic music generation: representations, algorithms, evaluations, and challenges, *ACM Computing Surveys*, **56**(1): 1–39, https://doi.org/10.1145/3597493.
22. Juslin P.N., Sloboda J.A. [Eds] (2010), *Handbook of Music and Emotion: Theory, Research, Applications*, Oxford University Press, https://doi.org/10.1093/acprof:oso/9780199230143.001.0001.
23. Kania D., Kania P. (2019), A key-finding algorithm based on music signature, *Archives of Acoustics*, **44**(3): 447–457, https://doi.org/10.24425/aoa.2019.129260.
24. Kania D., Kania P., Łukaszewicz T. (2021a), Trajectory of fifths in music data mining, *IEEE Access*, **9**: 8751–8761, https://doi.org/10.1109/ACCESS.2021.3049266.
25. Kania M., Kania D. (2022), Trajectory of fifths – a two-dimensional representation of music [in Polish: Trajektoria kwintowa – dwuwymiarowa reprezentacja muzyki], *Przegląd Elektrotechniczny*, **98**(6): 70–73, http://doi.org/10.15199/48.2022.06.13.
26. Kania M., Łukaszewicz T., Kania D., Mościńska K., Kulisz J. (2022), A comparison of the music key detection approaches utilizing key-profiles with a new method based on the signature of fifths, *Applied Sciences*, **12**(21): 11261, https://doi.org/10.3390/app122111261
27. Kania P., Kania D., Łukaszewicz T. (2024), A low complexity key-finding algorithm based on the signature of fifths, *Archives of Acoustics*, **49**(4): 491–505, https://doi.org/10.24425/aoa.2024.148817.
28. Kania P., Kania D., Łukaszewicz T. (2021b), A hardware-oriented algorithm for real-time music key signature recognition, *Applied Sciences, Computing and Artificial Intelligence*, **11**(18): 8753, https://doi.org/10.3390/app11188753.
29. Korzeniowski F., Widmer G. (2017), End-to-end musical key estimation using a convolutional neural network, [in:] *Proceedings of the 25th European Signal Processing Conference (EUSIPCO)*, pp. 966–970.
30. Kostrzewa D., Mazur W., Brzeski R. (2022), Wide ensembles of neural networks in music genre classification, [in:] *Computational Science – ICCS 2022. ICCS 2022. Lecture Notes in Computer Science*, **13351**: 64–71, https://doi.org/10.1007/978-3-031-08754-7_9.

31. Krumhansl C.L. (1990), *Cognitive Foundations of Musical Pitch*, New York: Oxford University Press, pp. 77–110, https://doi.org/10.1093/acprof:oso/9780195148367.001.0001
32. Longuet-Higgins H.C. (1962a), Letter to a musical friend, *The Music Review*, **23**: 244–238.
33. Longuet-Higgins H.C. (1962b), Second letter to a musical friend, *The Music Review*, **23**: 271–280.
34. Łukaszewicz T., Kania D. (2022), A music classification approach based on the trajectory of fifths, *IEEE Access*, **10**: 73494–73502, https://doi.org/10.1109/ACCESS.2022.3190016
35. Łukaszewicz T., Kania D. (2025a), Trajectory of fifths in tonal mode detection, *IEEE Access*, **13**: 113763–113772, https://doi.org/10.1109/ACCESS.2025.3584400.
36. Łukaszewicz T., Kania D. (2025b), Automatic key detection in audio recordings using the signature of fifths, *IEEE Access*, **13**: 160149–160157, https://doi.org/10.1109/ACCESS.2025.3609536.
37. Łukaszewicz T., Kania D. (2025c), Trajectory of fifths based on chroma subbands extraction – a new approach to music representation, analysis, and classification, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **47**(3): 2157–2169, https://doi.org/10.1109/TPAMI.2024.3519420.
38. Mauch M. (2010), *Automatic chord transcription from audio using computational models of musical context*, Ph.D. Thesis, Queen Mary University of London.
39. McFee B., Salamon J., Bello J.P. (2018), Adaptive pooling operators for weakly labeled sound event detection, *IEEE Transactions on Audio, Speech, and Language Processing*, **26**(11): 2180–2193, https://doi.org/10.1109/TASLP.2018.2858559.
40. Müller M., Konz V., Arifi-Müller V., Bogler W., Zeitleri J. (2024), Saarland music data: MIDI-audio piano music, *Zenodo*, https://doi.org/10.5281/zenodo.13753319
41. Pauwels J., O'Hanlon K., Gómez E., Sandler M.B. (2019), 20 years of automatic chord recognition from audio, [in:] *20th International Society for Music Information Retrieval Conference*, pp. 54–63.
42. Roig C., Tardón L.J., Barbancho I., Barbancho A.M. (2014), Automatic melody composition based on a probabilistic model of music style and harmonic rules, *Knowledge-Based Systems*, **71**: 419–434, https://doi.org/10.1016/j.knosys.2014.08.018.
43. Rosner A., Kostek B. (2018), Automatic music genre classification based on musical instrument track separation, *Journal of Intelligent Information Systems*, **50**(2): 363–384, https://doi.org/10.1007/s10844-017-0464-5.
44. Sabathé R., Coutinho E., Schuller B. (2017), Deep recurrent music writer: memory-enhanced variational autoencoder-based musical score composition and an objective measure, [in:] *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 3467–3474, https://doi.org/10.1109/IJCNN.2017.7966292.
45. Schedl M., Gómez E., Urbano J. (2014), Music information retrieval: recent developments and applications, *Foundations and Trends in Information Retrieval*, **8**(2–3): 127–261, https://doi.org/10.1561/1500000042.
46. Shepard R.N. (1982), Geometrical approximations to the structure of musical pitch, *Psychological Review*, **89**(4): 305–333, https://doi.org/10.1037/0033-295X.89.4.305.
47. Sturm B.L. (2013), Classification accuracy is not enough: on the evaluation of music genre recognition systems, *Journal of Intelligent Information Systems*, **41**(3): 371–406, https://doi.org/10.1007/s10844-013-0250-y.
48. Temperley D., Marvin E.W. (2008), Pitch-class distribution and the identification of key, *Music Perception*, **25**(3): 193–212, https://doi.org/10.1525/mp.2008.25.3.193.
49. Weiß C., Brand F., Müller M. (2019), Mid-level chord transition features for musical style analysis, [in:] *ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 341–345, https://doi.org/10.1109/ICASSP.2019.8682293.
50. Weiß C. *et al.* (2020), Schubert Winterreise dataset, *Zenodo*, https://doi.org/10.5281/zenodo.4122060.
51. Wu X. *et al.* (2023), MelodyGLM: multi-task pre-training for symbolic melody generation, *arXiv*, https://doi.org/10.48550/arXiv.2309.10738.
52. Yang S., Reed C.N., Chew E., Barthet M. (2023), Examining emotion perception agreement in live music performance, *IEEE Transactions on Affective Computing*, **14**(2): 1442–1460, https://doi.org/10.1109/TAFFC.2021.3093787.
53. Zhou X., Lerch A. (2015), Chord detection using deep learning, [in:] *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015*, pp. 52–58.

Archives of Acoustics Vol. 51, No. 3, pp. 561–579 (2026), https://doi.org/10.24423/archacoust.2026.4395

Research Paper

Dispersion Characteristics of Shear Horizontal-Waves in a Rotating Micro-Structured Plate with Restrained Boundaries

Mandeep Kaur[(1)]*, Satish Kumar[(1)], Vikas Sharma[(2)]

[(1)] *Department of Mathematics, Thapar Institute of Engineering and Technology*
Patiala, India

[(2)] *Department of Mathematics, Lovely Professional University*
Phagwara, India

*Corresponding Author: mkaur2_phd23@thapar.edu

Received December 12, 2025; revised April 3, 2026; accepted April 7, 2026; available online April 16, 2026; version of record July 20, 2026; published issue September 23, 2026.

Shear horizontal (SH)-waves in plates are elastic stress disturbances that propagate along bounded surfaces, with in-plane shear motion perpendicular to the propagation direction, influenced by physical, geometric, and boundary conditions. Due to their sensitivity to surface and subsurface defects, SH-waves are widely employed in non-destructive testing (NDT), structural health monitoring (SHM), and material characterization. This study presents a theoretical investigation of SH wave propagation in a rotating, microstructural elastic plate within the framework of consistent couple stress theory (CCST), incorporating elastically restrained boundary conditions (ERBC). CCST introduces a characteristic length parameter to account for microstructural and size-dependent effects beyond classical elasticity. Restrained boundaries provide an intermediate model between idealized traction-free and rigid conditions, capturing the realistic boundary behavior encountered in practical applications. An analytical dispersion relation is derived to examine the combined effects of boundary restraints, microstructural parameters, and rotational speed on SH wave dispersion, with several special cases presented to validate the general formulation. Graphical results illustrate the influence of restraint stiffness, characteristic length, plate thickness, and rotation on SH wave dispersion characteristics. The findings may offer valuable insights for the design of advanced sensing devices, the development of non-destructive evaluation techniques, and the analysis of seismic wave propagation in rotating geophysical layers.

Keywords: shear hortizontal (SH)-waves, couple stress theory, stiffness coefficients, restrained boundary conditions, dispersion curves, rotating elastic plate.



Copyright © 2026 The Author(s).
This work is licensed under the Creative Commons Attribution 4.0 International CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Waves propagating in rotating elastic plates have garnered considerable attention due to their relevance in both engineering and geophysical domains. Rotation introduces Coriolis and centrifugal effects, which significantly alter the characteristics of wave motion – the Coriolis force modifies wave trajectories and dispersion, while the centrifugal force shapes the equilibrium state. Understanding these influences may aid the design, analysis, and health monitoring of rotating systems in various engineering applications, where stress waves and vibrations can affect performance and reliability. In broader contexts, wave propagation in rotating layers may offer insights into seismic wave behavior, earthquake mechanics, and planetary dynamics and may also support the development of advanced sensing and non-destructive evaluation techniques. Bhuta and Jones (1963) established the foundational theoretical framework for analyzing the influence of uniform rotation on the propagation characteristics of elastic waves. Their study was the first to reveal the occurrence of elliptical polarization in wave

motion, arising from the Coriolis-induced coupling between displacement components. Further, Schoenberg and Censor explored plane wave solutions in rotating elastic solids in two (Censor, Schoenberg, 1973) and three-dimensions (Schoenberg, Censor, 1973). They have found that these waves do not exhibit purely shear or purely compressional characteristics and showed that the wave speeds are governed by the ratio of the rotational frequency of the medium to the angular frequency of the waves. Extensive studies have been conducted and documented in (Advani, 1967; Advani, Bulkeley, 1969; Bulkeley, 1973; Hashemi, Richard, 2001; Huang, Wang, 2001; Lamb, Southwell, 1921; Luo, Mote Jr., 2000; Nowinski, 1981; Southwell, 1922; Auriault, 2004; Kumar, Chawla, 2011; Yu, 1996) on wave propagation in different rotating elastic media such as beams, disks, membranes, half-spaces and plates. Collectively, the findings demonstrate that rotation introduces strong coupling between different wave modes, profoundly influencing the stability, dispersion behavior, and dynamic response of the system. These insights significantly advance the understanding of wave–rotation interactions and the complex challenges associated with the analysis and design of rotating machinery.

Shear horizontal (SH)-waves represent a fundamental class of elastic wave motion with broad applicability across geophysics, material science, structural health monitoring (SHM), and ultrasonic diagnostics. SH-waves exhibit pure in-plane transverse particle motion perpendicular to the propagation direction, rendering them immune to mode conversion at free or bonded interfaces. This distinctive feature imparts high signal clarity, strong penetration through layered structures, and minimal dispersion, making SH-waves exceptionally effective for non-destructive testing (NDT), surface acoustic wave (SAW) sensors, and characterization of multilayered or anisotropic materials. When propagating in thin plates, SH-waves are confined and guided by the plate boundaries, resulting in well-defined dispersion characteristics and enhanced sensitivity to surface, subsurface, and interfacial discontinuities. Castaings and Hosten (2001) investigated the interaction between non-principal planes and the surrounding fluid in the propagation of quasi-shear horizontal (quasi-SH) modes, demonstrating that this coupling enables the efficient generation and detection of such modes using air-coupled transducers in solid plates. Subsequently, Lee *et al.* (2010) introduced a novel technique for damage detection in plates based on focused SH-waves, which significantly minimizes the number of transducers required while enhancing detection accuracy. Kuznetsov (2022) investigated the behavior of SH-waves in functionally graded plates, incorporating bifurcation effects to enhance the understanding of their dynamic response and potential engineering applications. SH-waves in thin plates find extensive applications across various branches of engineering, serving numerous practical purposes such as material characterization, design and optimization of multilayered composite structures, and damping analysis in automotive, aerospace, and research industries (Kuznetsov, 2006; Phan *et al.*, 2019; Djeran-Maigre, Kuznetsov, 2014; Simonetti, Cawley, 2004; Jiangong, 2011).

In practical engineering applications, plates are often influenced by interactions with their surrounding environments and boundary surfaces, which significantly affect their dynamic and mechanical responses. However, capturing these complex interactions with full physical accuracy is often difficult; hence, simplified boundary idealizations are introduced to make the analysis tractable. The most common boundary conditions include the stress-free boundary, representing surfaces devoid of external constraints and the fixed (clamped) boundary, corresponding to fully restrained conditions where both displacement and rotation are restricted. To address the practical scenarios that lie between the two extreme boundary conditions, Mindlin (1960) introduced the concept of elastically restrained boundary conditions (ERBC) in the normal direction, modeling the boundary as exhibiting spring-like behavior governed by Hooke's law. Subsequently, Graff (2012) elaborated on these conditions and referred to them as mixed boundary conditions, which provide a continuous transition between stress-free and rigidly fixed boundaries. Moukhomodiarov *et al.* (2010) further examined this transition in a linear isotropic elastic layer, establishing a framework for intermediate boundary modeling. Expanding upon this foundation, Moukhomodiarov and Rogerson (2012a; 2012b) employed various analytical approaches to investigate the influence of ERBC on long-wave motion, and later extended their study to analyze long-wave dispersion phenomena in pre-stressed, incompressible elastic layers under ERBC (Moukhomodiarov, Rogerson, 2013). More recently, Paimushin and Gazizullin (2018) have explored the effect of ERBC by introducing elastic spacers connecting the plate to a rigid frame, while Kuznetsov (2015) analyzed the propagation characteristics of Lamb waves in elastic layers constrained by clamped and partially clamped boundaries.

Wave propagation is governed not only by the geometry and boundary conditions of the structure but also by the intrinsic material properties. In essence, the propagation behavior – depending on the medium and the nature of the wave is determined by the complex interplay among material characteristics, structural geometry, and boundary constraints. Although the behavior of SH-waves has been extensively studied in the context of rotational effects, it has not yet been comprehensively investigated within the framework of consistent couple stress theory (CCST). Hadjesfandiari and Dargush (2011) developed CCST, following the work of Mindlin and Tiersten (1962) and Koiter (1969), generalising and extending the classical theory of elasticity. It addresses the limitations of earlier couple stress theories (Mindlin, Tiersten, 1962; Koiter, 1969; Toupin, 1962; Voigt, 1892; Cosserat, Cosserat, 1909; Yang *et al.*, 2002) by introducing the skew-symmetric couple stress tensor and refined couple stress coefficient ($\eta = \mu l^2$) determined by the bulk modulus (μ) and the characteristic length of the material (l). The skew-symmetric couple stress tensor captures internal rotations and moment stresses while the characteristic length scale parameter enables to study size-dependent effects and microstructural interactions. Development of size-dependent couple stress theory in Timoshenko beams results in an increased stiffness predictions compared to classical beam theories, highlighting the importance of size effects in structural analysis (Asghari *et al.*, 2011). Dargush *et al.* (2021) conducted 2D and 3D elastostatic analysis applying Ritz variational approach within the framework of size-dependent couple stress theory.

Several studies have also examined wave propagation in micro-structured plates using CCST models, focusing primarily on size-dependent effects and couple-stress elasticity. Ghodrati *et al.* (2018) applied CCST to study the Lamb waves propagation considering microstructure effects in very high frequencies. Huang *et al.* (2024) analysed Lamb waves in micro/nano-plates subjected to CCST. Their findings show that in anti-symmetric modes, the plate surface exhibits no in-plane displacement while at specific frequencies on the intersection points of symmetric and anti-symmetric dispersion curves out-of-plane displacement vanishes on the free surface. Wu and Hsu (2022) observed the mechanical behavior of functionally graded microplate accounting for size effects using the CCST. Sharma and Kumar (2015) not only studied the Lamb waves in a plate with the liquid loadings but also investigated the effects of rotation and thermal parameters on both plane and Rayleigh waves using CCST (Sharma, Kumar, 2023b). Wu and Hu (2021) proposed a unified size-dependent plate theory for static bending and free vibration analyses of micro- and nano-scale plates using the CCST.

The CCST has been extensively utilized in various structural analyses (Wang *et al.*, 2017; Kaur *et al.*, 2024; Deep, Sharma, 2023; Sharma, Kumar, 2014), its application to SH wave propagation in plates remains a relatively unexplored domain. The present study introduces a novel investigation into the propagation behavior of SH-waves in a rotating microstructural plate subjected to ERBC. The formulation employs CCST, which incorporates a characteristic length parameter linking the macroscopic response of the material to its intrinsic microstructure, thereby enhancing accuracy for micro- and meso-scale applications. Furthermore, the inclusion of additional terms in the equations of motion, corresponding to centripetal and Coriolis effects, enables a more realistic representation of rotational influences. ERBC, which interpolate between fully stress-free and fully clamped states, are critical for accurately modeling the behavior of plates in practical engineering systems. Comprehensive graphical analyses are presented to illustrate the effects of microstructural parameters, characteristic length, plate thickness, and boundary conditions on phase velocity under rotational motion. The outcomes are compared with non-rotating cases, providing useful insights into the influence of rotation on wave behavior. These findings may have potential relevance to a broad range of engineering applications, particularly in contexts where size-dependent effects and boundary constraints are significant.

2. Geometry and basic equations

Consider a thin micro-structured plate that is unbounded in the x- and z-directions and possesses a finite thickness of $2b$ along the y-axis ($-\infty < x < \infty, -b \leq y \leq b, -\infty < z < \infty$). The Cartesian coordinate system has an origin O located at the geometric center of the plate, equidistant from its upper and lower boundaries. The positive y-direction is taken vertically upward, through the thickness of the plate. The free surfaces of the plate are defined by the planes $y = b$ and $y = -b$. The medium (plate) is assumed to be rotating uniformly about an axis in the x-y plane with angular velocity $\mathbf{\Omega} = (\Omega_1, \Omega_2, \Omega_3)$ and the wave propagates along the x-direction.

For the 2D solution, all the physical quantities are supposed to be invariant along the z-direction, i.e., $\frac{\partial}{\partial z} \equiv 0$. The displacement vector $\mathbf{u}$ is expressed as

$$\mathbf{u} = (u(x,y,t), v(x,y,t), w(x,y,t)),$$

where u, v, and w denote the displacement components in the x-, y-, and z-directions, respectively, and t denotes time. According to the CCST, the equation of motion of a rotating body in the absence of body forces is given as (Schoenberg, Censor, 1973; Hadjesfandiari, Dargush, 2011):

$$(\lambda + \mu + \eta\nabla^2)\nabla(\nabla \cdot \mathbf{u}) + (\mu - \eta\nabla^2)\nabla^2\mathbf{u} = \rho\left[\frac{\partial^2 \mathbf{u}}{\partial t^2} + (\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{u})) + \left(2\boldsymbol{\Omega} \times \frac{\partial \mathbf{u}}{\partial t}\right)\right], \tag{1}$$

where $\mathbf{u} = (u, v, w)$ is the displacement vector, λ and μ are Lame's constants, $\eta = \mu l^2$ is the couple stress coefficient, l is the characteristic length, ρ is the density of material, $\boldsymbol{\Omega}$ is the rotation vector, and $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$. The term $\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{u})$ represents the centripetal acceleration resulting from time varying motion and the Coriolis acceleration is denoted by the term $2\boldsymbol{\Omega} \times \frac{\partial \mathbf{u}}{\partial t}$.

The constitutive relations for the CCST are given as

$$\sigma_{ji} = \lambda u_{k,k}\delta_{ij} + \mu(u_{i,j} + u_{j,i}) - \eta\nabla^2(u_{i,j} - u_{j,i}), \tag{2}$$

$$\mu_{ji} = 4\eta(\overline{\omega}_{i,j} - \overline{\omega}_{j,i}), \tag{3}$$

where $\overline{\omega}_i = \frac{1}{2}\epsilon_{ijk}u_{k,j}$, σ_{ji} is force stress tensor, μ_{ji} is couple stress tensor, δ_{ij} is Kronecker's delta, ϵ_{ijk} is the permutation tensor, $\overline{\omega}_i$ is the rotation vector, and $i, j, k = 1, 2, 3$.

3. Propagation of SH-waves under rotation

Consider SH-waves, propagating in a micro-structural plate under rotation. For the 2D plate in the x-y plane, the components of displacement and rotation are considered as $\mathbf{u} = (0, 0, w)$ and $\boldsymbol{\Omega} = (\Omega_1, \Omega_2, 0)$, respectively. Equation (1) is transformed into the single partial differential equation as follows:

$$C_2^2(1 - l^2\nabla^2)\nabla^2 w = \frac{\partial^2 w}{\partial t^2} - w(\Omega_1^2 + \Omega_2^2), \tag{4}$$

where $C_2 = \sqrt{\frac{\mu}{\rho}}$.

Let the solution to Eq. (4) is assumed as

$$w = h(y)e^{i\xi(x - Ct)}, \tag{5}$$

where $h(y)$ is the unknown function of y, ξ is the wave number, C is the phase speed, $i = \sqrt{-1}$. On substituting Eq. (5) into Eq. (4), Eq. (4) reduces to the following fourth-order differential equation:

$$\frac{d^4h(y)}{dy^4} - \left(\frac{1}{l^2} + 2\xi^2\right)\frac{d^2h(y)}{dy^2} + \left(\xi^4 + \frac{\xi^2}{l^2}\left(1 - \frac{C^2}{C_2^2}\right) - \left(\frac{\Omega_1^2 + \Omega_2^2}{C_2^2 l^2}\right)\right)h(y) = 0. \tag{6}$$

The solution to the differential Eq. (6) is obtained as

$$h(y) = D_1 \cosh(R_1 y) + D_2 \sinh(R_1 y) + D_3 \cos(R_2 y) + D_4 \sin(R_2 y), \tag{7}$$

where D_1, D_2, D_3, D_4 are arbitrary constants, and

$$R_1 = \sqrt{\xi^2 + \frac{1}{2l^2}\left(1 + \sqrt{1 + \frac{4\xi^2 C^2 l^2}{C_2^2} + \frac{4l^2}{C_2^2}(\Omega_1^2 + \Omega_2^2)}\right)}, \qquad R_2 = \sqrt{\frac{1}{2l^2}\left(\sqrt{1 + \frac{4\xi^2 C^2 l^2}{C_2^2} + \frac{4l^2}{C_2^2}(\Omega_1^2 + \Omega_2^2)} - 1\right) - \xi^2}.$$

Thus, the displacement component in Eq. (5) becomes:

$$w(y) = \left(D_1 \cosh\left(R_1 y\right) + D_2 \sinh\left(R_1 y\right) + D_3 \cos\left(R_2 y\right) + D_4 \sin\left(R_2 y\right)\right) e^{i\xi(x-Ct)}. \tag{8}$$

The non-zero component of stress and couple stress tensors are obtained as

$$\begin{aligned}\sigma_{yz} &= \mu\frac{\partial w}{\partial y} - \mu l^2\left(\frac{\partial^3 w}{\partial x^2 \partial y} + \frac{\partial^3 w}{\partial y^3}\right)\\ &= \left(D_1 L_1 \sinh\left(R_1 y\right) + D_2 L_1 \cosh\left(R_1 y\right) - D_3 L_2 \sin\left(R_2 y\right) + D_4 L_2 \cos\left(R_2 y\right)\right) e^{i\xi(x-Ct)},\end{aligned} \tag{9}$$

$$\begin{aligned}\mu_{yx} &= 2\mu l^2\left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2}\right)\\ &= \left(D_1 L_3 \cosh\left(R_1 y\right) + D_2 L_3 \sinh\left(R_1 y\right) - D_3 L_4 \cos\left(R_2 y\right) - D_4 L_4 \sin\left(R_2 y\right)\right) e^{i\xi(x-Ct)},\end{aligned} \tag{10}$$

where $L_1 = \mu R_1 + \mu l^2 \xi^2 R_1 - \mu l^2 R_1^3$, $L_2 = \mu R_2 + \mu l^2 \xi^2 R_2 + \mu l^2 R_2^3$, $L_3 = 2\mu l^2 (R_1^2 - \xi^2)$, $L_4 = 2\mu l^2 (R_2^2 + \xi^2)$.

4. Restrained boundary conditions

ERBC can be conceptualized as boundary surfaces embedded with distributed springs. These springs simulate elastic restraints by imposing resistance to displacements and rotations. To incorporate restrained boundary conditions within a couple stress elastic model, it is necessary to account not only for the classical stress–strain relations but also for additional stiffness coefficients and couple stress effects. These coefficients govern the relationship between force stresses and displacements, as well as between couple stresses and rotation gradients. Consequently, such conditions permit limited displacement and rotation at the boundaries. The shear spring constant (k_1) representing resistance to tangential (shear) deformation while (k_2) represents the rotational spring constant, indicating resistance to the rotational motion at the boundary. Here, 'rotation' refers to the angular motion of the plate itself, and its influence is incorporated into the dispersion relation of SH-waves by accounting for Coriolis and centrifugal forces within the framework of rotating couple stress theory. For SH wave propagation, the tangential and rotational stiffness coefficients, respectively, characterize the boundary's resistance to tangential displacement and rotational deformation. Boundary conditions at the two surfaces ($y = \pm b$) of the plate are as follows (Mindlin, 1960; Wang *et al.*, 2017):

1. $\sigma_{yz} = k_1 w$. This boundary condition leads to the following equation in four unknown arbitrary constants:

$$\begin{aligned}D_1 L_1 \sinh\left(R_1 y\right) + D_2 L_1 \cosh\left(R_1 y\right) - D_3 L_2 \sin\left(R_2 y\right) + D_4 L_2 \cos\left(R_2 y\right) - k_1 D_1 \cosh\left(R_1 y\right)\\ - k_1 D_2 \sinh\left(R_1 y\right) - k_1 D_3 \cos\left(R_2 y\right) - k_1 D_4 \sin\left(R_2 y\right) = 0.\end{aligned}$$

2. $\mu_{yx} = k_2 \overline{\omega}_1$, where $\overline{\omega}_1 = \frac{1}{2}\frac{\partial w}{\partial y}$. This boundary condition yields the following equation in four unknown arbitrary constants:

$$\begin{aligned}D_1 L_3 \cosh\left(R_1 y\right) + D_2 L_3 \sinh\left(R_1 y\right) - D_3 L_4 \cos\left(R_2 y\right) - D_4 L_4 \sin\left(R_2 y\right) - \frac{k_2}{2} D_1 R_1 \sinh\left(R_1 y\right)\\ - \frac{k_2}{2} D_2 R_1 \cosh\left(R_1 y\right) + \frac{k_2}{2} D_3 R_2 \sin\left(R_2 y\right) - \frac{k_2}{2} D_4 \cos\left(R_2 y\right) = 0,\end{aligned}$$

where σ_{yz} is the force stress, μ_{yx} is the couple stress, w is the displacement component, $\overline{\omega}_1$ is the rotation vector, k_1, k_2 are the tangential and rotational spring coefficients, respectively.

Enforcing the mentioned boundary conditions at the plate boundaries ($y = \pm b$) leads to the following set of four homogeneous equations:

$$D_1(L_1 S_{11} - k_1 C_{11}) + D_2(L_1 C_{11} - k_1 S_{11}) - D_3(L_2 S_{12} + k_1 C_{12}) + D_4(L_2 C_{12} - k_1 S_{12}) = 0, \tag{11}$$

$$D_1(-L_1S_{11} - k_1C_{11}) + D_2(L_1C_{11} + k_1S_{11}) - D_3(-L_2S_{12} + k_1C_{12}) + D_4(L_2C_{12} + k_1S_{12}) = 0, \quad (12)$$

$$D_1\left(L_3C_{11} - \frac{k_2}{2}R_1S_{11}\right) + D_2\left(L_3S_{11} - \frac{k_2}{2}R_1C_{11}\right) - D_3\left(L_4C_{12} - \frac{k_2}{2}R_2S_{12}\right) - D_4\left(L_4S_{12} + \frac{k_2}{2}R_2C_{12}\right) = 0, \quad (13)$$

$$D_1\left(L_3C_{11} + \frac{k_2}{2}R_1S_{11}\right) + D_2\left(-L_3S_{11} - \frac{k_2}{2}R_1C_{11}\right) - D_3\left(L_4C_{12} + \frac{k_2}{2}R_2S_{12}\right) - D_4\left(-L_4S_{12} + \frac{k_2}{2}R_2C_{12}\right) = 0, \quad (14)$$

here $S_{11} = \sinh(R_1b)$, $S_{12} = \sin(R_2b)$, $C_{11} = \cosh(R_1b)$, $C_{12} = \cos(R_2b)$.

For Eq. (11) to Eq. (14) to possess the non-trivial solution, the determinant of unknown coefficients must become zero:

$$|d_{ij}|_{4\times4} = 0, \quad (15)$$

where

$$d_{11} = -k_1C_{11}, \quad d_{12} = L_1C_{11}, \quad d_{13} = -k_1C_{12}, \quad d_{14} = L_2C_{12}, \quad d_{21} = L_1S_{11}, \quad d_{22} = -k_1S_{11},$$
$$d_{23} = -L_2S_{12}, \quad d_{24} = -k_1S_{12}, \quad d_{31} = 2L_3C_{11}, \quad d_{32} = -k_2R_1C_{11}, \quad d_{33} = -2L_4C_{12}, \quad d_{34} = -k_2R_2C_{12},$$
$$d_{41} = -k_2R_1S_{11}, \quad d_{42} = 2L_3S_{11}, \quad d_{43} = k_2R_2S_{12}, \quad d_{44} = -2L_4S_{12}.$$

The solution to Eq. (15) gives the dispersion relation written as

$$PX^2 + QX - P = 0, \quad (16)$$

where

$$X = \frac{\tanh(R_1b)}{\tan(R_2b)}, \qquad P = 2k_1k_2(L_2L_4R_1 - L_1L_3R_2) + k_1^2k_2^2R_1R_2 - 4L_1L_2L_3L_4,$$
$$Q = (k_1^2 - L_1^2)(k_2^2R_2^2 + 4L_4^2) + (k_1^2 + L_2^2)(-k_2^2R_1^2 + 4L_3^2) + 8k_1^2L_3L_4 - 4k_1k_2(L_2L_3R_2 + L_1L_4R_1).$$

The analysis of Eq. (16) leads to the derivation of two distinct dispersion relations associated with SH-waves:

$$\frac{\tanh(R_1b)}{\tan(R_2b)} = \frac{-Q - \sqrt{Q^2 + 4P^2}}{2P}, \quad (17)$$

$$\frac{\tanh(R_1b)}{\tan(R_2b)} = \frac{-Q + \sqrt{Q^2 + 4P^2}}{2P}. \quad (18)$$

5. Modes of SH-waves

The dispersion relation given in Eq. (16) comprises both symmetric and anti-symmetric components. Since, Eq. (17) and Eq. (18) are reduced forms of Eq. (16), it is pertinent to examine which of these equations corresponds to the symmetric or anti-symmetric modes.

Let the restraint parameters in the boundary conditions be relaxed, i.e., $k_1 = k_2 = 0$. Consequently, the boundary conditions discussed in Sec. 4 simplify to $\sigma_{yz} = 0$, $\mu_{yx} = 0$, corresponding to stress-free boundary conditions. Under these conditions, Eq. (15) reduces to the following two distinct dispersion relations for SH-waves, respectively:

$$\frac{\tanh(R_1b)}{\tan(R_2b)} = \frac{L_2L_3}{L_1L_4}, \quad (19)$$

$$\frac{\tanh(R_1b)}{\tan(R_2b)} = \frac{-L_1L_4}{L_2L_3}. \quad (20)$$

By imposing the symmetric conditions ($D_2 = D_4 = 0$) and anti-symmetric conditions ($D_1 = D_3 = 0$) as in (Yu, 1996) on Eq. (11) to Eq. (14), the derived Eq. (19) and Eq. (20) are directly obtained for the stress-free case. Accordingly, Eq. (17) and Eq. (18) can be identified as representing the symmetric and anti-symmetric modes of SH-waves.

6. Special cases

For stress-free boundary conditions, the surface tractions vanish, whereas rigid surface boundary conditions enforce zero displacements at the surface. Elastically restrained boundary conditions provide a transition between stress-free and rigid boundary conditions by introducing finite stiffness at the boundary. These three boundary conditions are discussed in this section.

Case 1. Stress-free boundary conditions

Stress free boundary conditions are achieved by specifying $k_1 = k_2 = 0$ in the boundary conditions discussed in Sec. 4. Upon using these conditions in Eq. (11) to Eq. (14), two dispersion relations (19) and (20) are attained in Sec. 5 which are the same as discussed by Sharma and Kumar (2023a).

Case 2. Mixed boundary conditions

In the boundary conditions mentioned in Sec. 4, if $k_1 \to \infty$ and $k_2 = 0$, then the conditions become $w = 0$, $\mu_{yx} = 0$. Thus, the solution to the determinant in Eq. (15) produces two dispersion relations for the SH-waves:

$$\cosh(R_1 b)\ \cos(R_2 b) = 0, \tag{21}$$

$$\sinh(R_1 b)\ \sin(R_2 b) = 0. \tag{22}$$

Using the mentioned mixed boundary conditions, Eq. (21) and Eq. (22) can also be derived from Eq. (17) and Eq. (18). Thus, Eq. (21) and Eq. (22) can be classified as the symmetric and anti-symmetric modes of vibrations of SH-waves, respectively.

Case 3. Rigid boundary conditions

In rigidly fixed boundary conditions, both displacement and rotation are totally restricted at the boundary. Thus, in this case $w = 0$, $\overline{\omega}_1 = 0$, obtained by putting $k_1 \to \infty$, $k_2 \to \infty$ in the boundary conditions stated in Sec. 4. The determinant in Eq. (15) reduces to the following two distinct dispersion relations:

$$\frac{\tanh(R_1 b)}{\tan(R_2 b)} = \frac{-R_2}{R_1}, \tag{23}$$

$$\frac{\tanh(R_1 b)}{\tan(R_2 b)} = \frac{R_1}{R_2}. \tag{24}$$

On the basis of the discussion in the case 1 and case 2 regarding obtaining dispersion relations directly by using these conditions in Eq. (17) and Eq. (18), we can group Eq. (23) and Eq. (24) to as symmetric and anti-symmetric modes of vibrations of SH-waves in a micro-structural plate subject to rigid boundary conditions.

7. Verification and comparative analysis of the model

7.1. Assessment of model consistency with classical theory

Within the framework of the CCST, the stress-strain behavior of a material is governed not only by the classical continuum parameters but also by an intrinsic material length scale parameter (l). This quantitatively characterizes size-dependent behavior resulting from the material's micro-structural effects.

In the vanishing length scale limit ($l \to 0$), higher order couple stress effects are suppressed. The differential Eq. (6) reduces to the following differential equation of the second-order:

$$\frac{d^2 h(y)}{dy^2} - \left(\xi^2\left(1 - \frac{C^2}{C_2^2}\right) - \left(\frac{\Omega_1^2 + \Omega_2^2}{C_2^2}\right)\right) h(y) = 0. \tag{25}$$

The solution to Eq. (25) is

$$h(y) = E_1 \cos(\beta y) + E_2 \sin(\beta y)$$

and thus

$$w(y) = (E_1 \cos(\beta y) + E_2 \sin(\beta y))\, e^{i\xi(x-Ct)}, \tag{26}$$

where E_1, E_2 are the arbitrary constants and $\beta^2 = \frac{\Omega_1^2+\Omega_2^2}{C_2^2} - \xi^2\left(1 - \frac{C^2}{C_2^2}\right)$.

The restrained boundary condition in Sec. 4 simplifies to $\sigma_{yz} = k_1 w$. Applying this condition at the two surfaces ($y = \pm b$) of the plate, following pair of homogeneous equations are obtained:

$$E_1(-k_1)\cos(\beta b) + E_2(\mu\beta)\cos(\beta b) = 0, \tag{27}$$

$$E_1(\mu\beta)\sin(\beta b) - E_2(k_1)\sin(\beta b) = 0. \tag{28}$$

Enforcing a non-trivial solution to Eq. (27) and Eq. (28) leads to the following equation:

$$\left(k_1^2 - (\mu\beta)^2\right)\sin(\beta b)\cos(\beta b) = 0. \tag{29}$$

Equation (29) is the dispersion relation for the propagation of SH-waves in a micro-structured rotating plate with restrained boundary condition.

On assuming the tangential spring coefficient (k_1) to be zero, the boundary condition becomes $\sigma_{yz} = 0$. Thus, Eq. (29) reduces to $\sin(\beta b)\cos(\beta b) = 0$, which can be further simplified to the resulting pair of dispersion relations:

$$\sin(\beta b) = 0, \quad \text{i.e.,} \quad \beta b = \frac{n\pi}{2}, \quad n = 0, 2, 4, \ldots, \tag{30}$$

$$\cos(\beta b) = 0, \quad \text{i.e.,} \quad \beta b = \frac{n\pi}{2}, \quad n = 1, 3, 5, \ldots, \tag{31}$$

where β is still having the rotational term in it. Thus, Eq. (30) and Eq. (31) correspond to the symmetric and anti-symmetric modes of SH-wave, respectively, in a rotating plate under classical theory of elasticity.

Furthermore, in the absence of rotation ($\Omega_1 = \Omega_2 = 0$), one obtains $\beta = \xi^2\left(\frac{C^2}{C_2^2} - 1\right)$. In this case, Eq. (30) and Eq. (31) represent the symmetric and anti-symmetric modes of SH-wave, respectively, in an elastic plate as derived in (Graff, 2012). This confirms that the proposed model is consistent with the classical elasticity theory and accurately reproduces the conventional behavior in the absence of micro-structural effects, rotational influences and restrained coefficients, thereby validating the formulation.

7.2. Comparison with the existing studies

A review of the literature indicates that numerous studies have examined the propagation of various elastic waves, investigating the influence of factors such as micro-structural parameter, rotation, stiffness, and boundary

Table 1. Comaprison of present model with existing studies.

Model	Rotation	Characteristic length	ERBC	Problem type	Relevant references
Classical elasticity	✗	✗	✗	SH-waves in plate	Graff (2012); Kuznetsov (2006)
	✓	✗	✗	Reflection and refraction of plane waves	Schoenberg, Censor (1973); Censor, Schoenberg (1973)
CCST	✗	✓	✗	Lamb waves in plate	Sharma, Kumar (2014)
	✗	✓	✗	SH-waves in coupled plates	Dua, Sharma (2024)
	✗	✗	✓	Waves in elastic media	Moukhomodiarov *et al.* (2012); Moukhomodiarov, Rogerson (2012a; 2012b)
	✗	✓	✗	SH-waves in plate	Sharma, Kumar (2023a)
	✓	✓	✓	Reflection and refraction of plane waves	Wang *et al.* (2017)
	✓	✓	✓	SH-waves in plate	Present study

conditions in different elastic media. However, when focus is restricted to plate structures, the available research remains comparatively limited. The present study investigates the combined effects of rotation, length scale parameter, and ERBC on SH-wave propagating in an elastic plate within the framework of CCST theory. This section presents a comparison of the proposed model with existing studies, as summarized in Table 1, along with relevant references.

8. Numerical results and discussion

SH-waves in a plate display multimodal behavior along with dispersion characteristics. One notable aspect of this study involves the combined effect of rotation and stiffness coefficients on the phase velocity dispersion. For graphical illustration of the results, the material parameters are taken (Wang *et al.*, 2017) as $\lambda = 2.2\times10^{10}\,\mathrm{N/m^2}$, $\mu = 1.1\times10^{10}\,\mathrm{N/m^2}$, $\rho = 2.6\times10^{3}\,\mathrm{kg/m^3}$. The standard values of stiffness constants are $k_1 = 10^{16}\,\mathrm{N/m^3}$ and $k_2 = 10^{4}\,\mathrm{N/m}$.

The characteristic length is chosen to be of the same order as the plate thickness $l \sim b$ to capture size-dependent effects. The stiffness parameters range from zero (stress-free) to infinity (perfectly constrained); however, finite values are adopted to represent realistic surface conditions in line with established literature. The rotational speed (10^8 rad/sec) though high for macroscopic systems, is appropriate for micro/nano-scale devices and enables capturing rotational effects on wave propagation. Moreover, to address this, dimensionless parameters have also been introduced. Since, similar behavior is observed in both symmetric and anti-symmetric modes, the graphs are presented only for the symmetric cases.

8.1. Symmetric and anti-symmetric modes of SH-waves

The dispersion relation expressed in Eq. (16), derived in Sec. 4 reduces to two distinct forms as presented in Eq. (17) and Eq. (18). These two relations correspond, respectively, to the symmetric and anti-symmetric cases of propagation, consistent with the classification outlined in Sec. 5. Figure 1 and Fig. 2 depict the variation of non-dimensional phase velocity with respect to the non-dimensional wavenumber for both cases, highlighting the influence of the rotational parameter. These figures plot the non-dimensional phase velocity against non-dimensional wavenumber for mode numbers, $n = 1, 2, 3$. These profiles clearly show two robust features. First, for both the non-rotating and rotating cases the phase-velocity curves for successive mode numbers are shifted upward (higher modes propagate with higher phase velocity for a given non-dimensional wave number), indicating the classical higher-mode stiffening. Second, the introduction of rotation systematically reduces the phase velocity of every mode. This reflects the additional effective inertia introduced by rotation (rotational inertia and Coriolis-type coupling) mitigates the stiffening effect associated with couple stress. A modal asym-

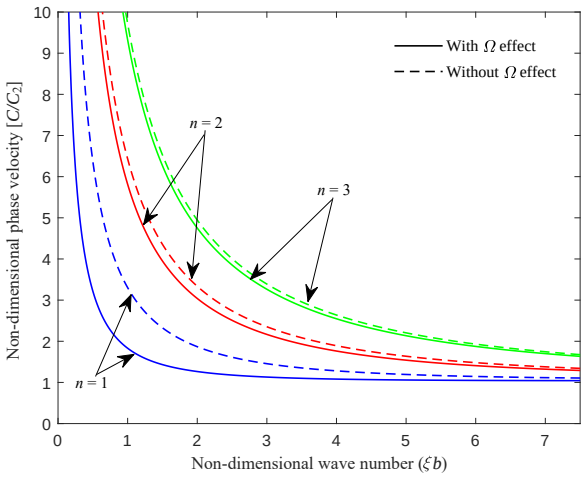


Fig. 1. Phase velocity profiles for illustrating modes of symmetric SH-waves.

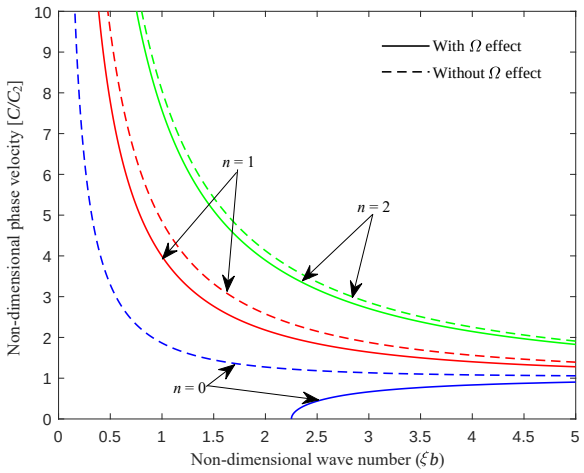


FIG. 2. Phase velocity profiles for illustrating modes of anti-symmetric SH-waves.

metry is also evident: the symmetric branch does not admit a fundamental (low-frequency) propagating mode under the present couple-stress boundary conditions, whereas the anti-symmetric branch supports a fundamental mode that exhibits a trend opposite to that of the higher anti-symmetric branches and is characterized by a cutoff frequency/wavenumber below which wave propagation does not occur. The cutoff and the reversed low-frequency trend are controlled by the non-dimensional characteristic length and the rotational parameter, demonstrating a sensitive competition between microstructural stiffening and rotational softening.

8.2. Effect of dimensionless rotation parameter (Ω_2^*)

In a rotating couple-stress elastic medium, the propagation of SH-waves is strongly influenced by the inertial effects associated with the rotational components, which directly modify the governing equations of motion and thereby alter the dispersion characteristics. The two components of the rotation vector, Ω_1 and Ω_2, appear in the governing relations through their sum of squares, $\Omega_1^2 + \Omega_2^2$, as evident from the values of R_1 and R_2 in Sec. 3.

Dimensionless parameters play a crucial role in characterizing the wave propagation behavior by encapsulating the combined effects of material properties and characteristic length scales. In particular, $\Omega_2^* = (\Omega_2 l)/C_2$ quantifies the relative influence of characteristic length, rotation and shear wave velocity. Accordingly, the dispersion behavior of symmetric and anti-symmetric SH-waves is examined under varying values of Ω_2^*. Figure 3 shows that for a fixed wave number (ξb), an increase in the value of Ω_2^* leads to the systematic reduction in phase velocity. This effect is more pronounced at smaller wave numbers. The significant variation in phase velocity at

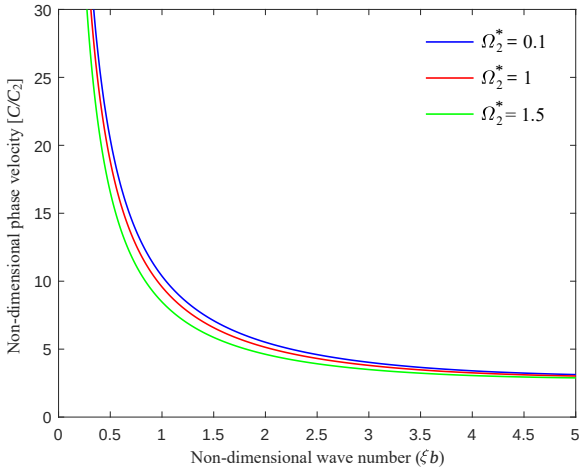


FIG. 3. Phase velocity profiles for illustrating impact of dimensionless rotation parameter Ω_2^* on symmetric SH-waves.

small wave numbers reflects strong dispersive effects that progressively diminish as (ξb) increases. At large wave numbers, the dispersion curves coalesce due to the diminishing influence of coupling and micro-structural effects.

Since, Ω_1 and Ω_2 enter the governing equations through the combined quadratic form $\Omega_1^2 + \Omega_2^2$; consequently, the dimensionless rotation parameter Ω_1^* contributes identically to the phase velocity of SH-waves.

8.3. Effects of plate thickness

Geometric parameters play a crucial role in determining the propagation characteristics of waves in solid media. In the present case, the finite thickness of the plate governs the boundary conditions and, consequently, the modal structure of the SH-waves. Figure 4 illustrates the variation of non-dimensional phase velocity with respect to non-dimensional wavenumber for different plate thicknesses.

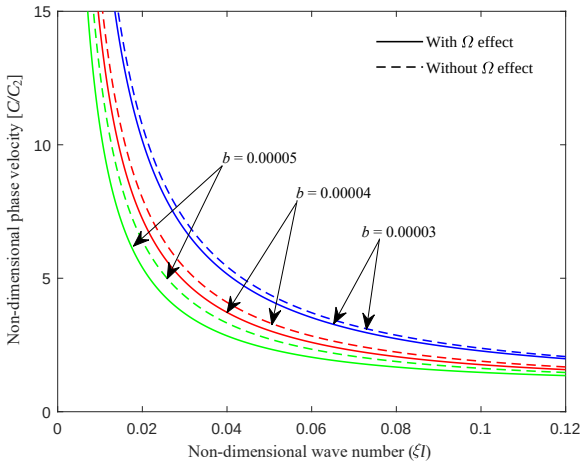


Fig. 4. Phase velocity profiles for illustrating impact of varying plate thickness on symmetric SH-waves.

It is observed that the phase velocity exhibits a negative correlation with the plate thickness, whereby as the thickness increases, the phase velocity decreases. Both the non-rotating and rotating cases display this decreasing trend. However, when rotation is incorporated, the phase velocity profiles are slightly suppressed. Overall, the results emphasize that geometric scaling, in conjunction with rotational dynamics and couple stress effects, plays a decisive role in defining the dispersive response of SH-waves in finite plates. Increasing plate thickness tends to reduce the phase velocity. Rotational effects further enhance this trend, leading to additional suppression of wave propagation.

8.4. Effects of characteristic length parameter (l)

The characteristic length parameter (l) is a fundamental micro-structural quantity in couple stress elasticity, representing the internal material length scale or the average size of the microstructural cells. Typically of the order of 10^{-6} m, this parameter governs the extent to which micro-rotations and couple stresses influence the deformation field. To examine its impact on SH wave propagation under constrained plate conditions, phase velocity profiles were computed for different values of l. As shown in Fig. 5, an increase in the characteristic length parameter leads to a significant enhancement in phase velocity. This behavior can be attributed to the strengthening effect introduced by the microstructural interactions, which effectively increase the material's resistance to shear deformation. As l increases, the contribution of couple stresses to the total stress field becomes more dominant, resulting in an apparent stiffening of the medium and hence higher phase velocities. The trend clearly indicates that the internal length parameter acts as a stiffening factor, counteracting the dispersive softening associated with geometric and rotational effects. Consistent with earlier observations, the inclusion of rotational parameters slightly reduces the overall magnitude of phase velocity for each value of l. This demonstrates a competing interplay between microstructural stiffening (through l) and rotational softening (through Ω_1, Ω_2). At higher values of l,

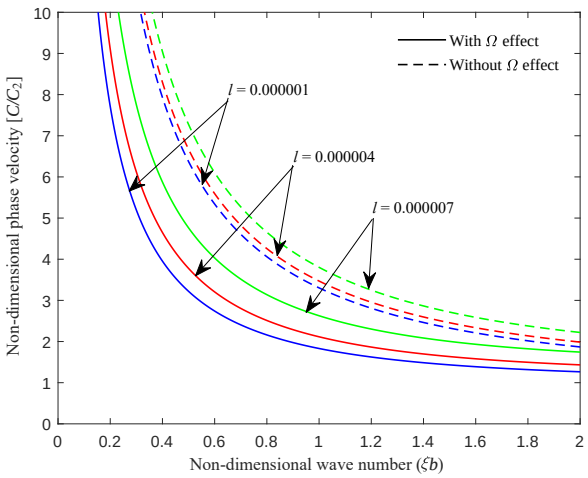


FIG. 5. Phase velocity profiles for illustrating impact of varying characteristic length (l) on symmetric SH-waves.

however, the microstructural effects tend to dominate, leading to comparatively higher phase velocity even in the presence of rotation.

8.5. Effect of dimensionless parameters (l/b)

By analyzing the pronounced effects of the characteristic length scale and plate thickness on the dispersion behavior, the phase velocity is examined under rotational effects as a function of the dimensionless ratio l/b in Fig. 6. It is observed that increasing this ratio leads to a notable amplification of phase velocity, reflecting the enhanced resistance of the material to shear deformation. This effect is due to the additional stiffness contributed by microstructural effects. Proper consideration of the length-to-thickness ratio is therefore essential for accurately predicting deformation, natural frequencies, and stress distribution in microstructures or thin-scale structures.

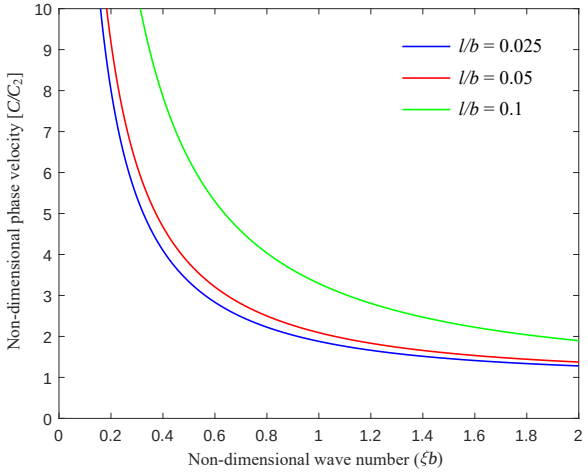


FIG. 6. Phase velocity profiles for illustrating impact of dimensionless parameter (l/b) on symmetric SH-waves.

8.6. Effect of dimensionless stiffness parameters (k_1^* and k_2^*)

Wave propagation in microstructured solids are governed by both, the geometrical configuration and intrinsic physical properties. The introduction of dimensionless parameters $k_1^* = (k_1 l)/\mu$ and $k_2^* = k_2/(l\mu)$ consolidate the effects of material constant (μ), stiffness coefficients (k_1 and k_2), and the length scale parameter (l).

Figure 7 indicates that increasing k_1^* enhances the contribution of the material's microstructure, thereby increasing the effective shear stiffness of the plate and, consequently, the phase velocity. At constant wave number,

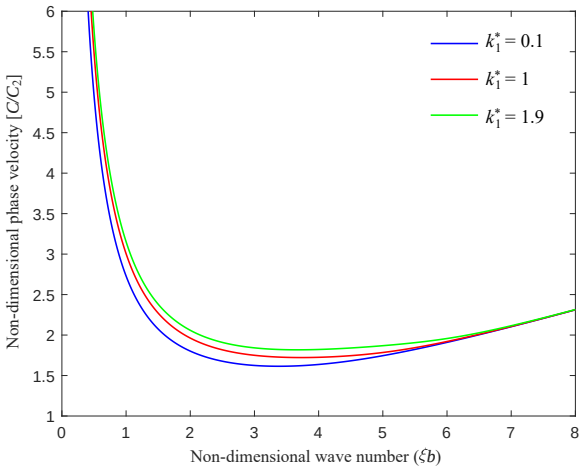


FIG. 7. Phase velocity profiles for illustrating impact of dimensionless shear stiffness parameter (k_1^*) on symmetric SH-waves.

profiles exhibit a direct relationship between the phase velocity and the dimensionless stiffness parameter, k_1^*. A pronounced dispersive effect is observed, particularly at short wavelengths, where the phase velocity exhibits a more rapid increase.

Contrary to the effect of k_1^*, the dimensionless quantity k_2^* in Fig. 8 is inversely related to the phase velocity. Higher values of k_2^* enhances the system's resistance to dynamic deformation and strongly constrain microrotations, consequently oppose wave propagation despite the apparent increase in boundary stiffness.

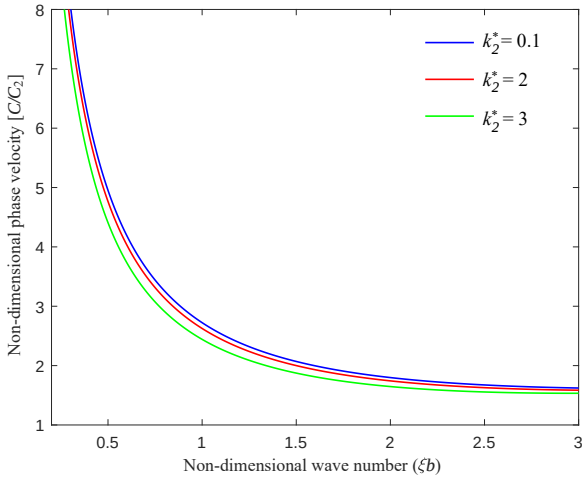


FIG. 8. Phase velocity profiles for illustrating impact of dimensionless rotational stiffness parameter (k_2^*) on symmetric SH-waves.

8.7. Effect of boundary conditions

The dispersion relations labeled as Eq. (19) and Eq. (20), Eq. (21) and Eq. (22), and Eq. (23) and Eq. (24) correspond to three distinct types of boundary conditions: stress-free, mixed, and fixed, respectively. These relations have been derived under both rotating and non-rotating conditions and are graphically represented in Fig. 9 and Fig. 10 for both symmetric and anti-symmetric SH wave modes, respectively.

The stress-free boundary condition implies that both the normal and shear stress components at the boundary are zero, representing a completely traction-free surface (ideal for free surfaces like air-solid interfaces). The fixed boundary condition assumes zero displacement at the surface, meaning the boundary is perfectly clamped or rigidly constrained, often used to model bonding with rigid substrates. The mixed boundary condition represents

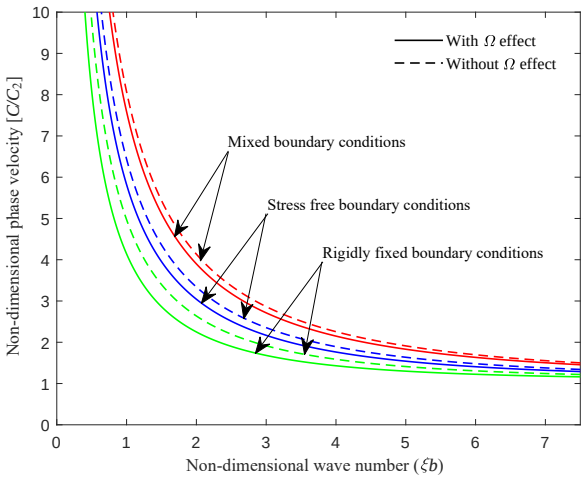


FIG. 9. Phase velocity profiles for illustrating impact of boundary conditions on symmetric SH-waves.

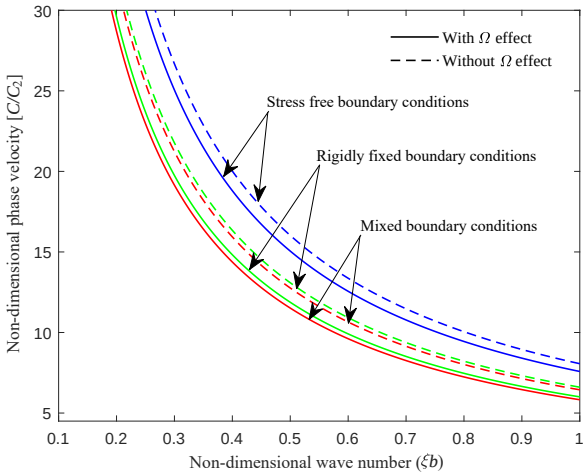


FIG. 10. Phase velocity profiles for illustrating impact of boundary conditions on anti-symmetric SH-waves.

an intermediate case where one stress component and one displacement component are set to zero. For example, tangential stress might be zero while normal displacement is constrained, modeling partial bonding or imperfect adhesion.

From the dispersion curves in Fig. 9, it is observed that for the symmetric mode of SH-waves, the phase velocity corresponding to the stress-free boundary condition lies between those obtained for the mixed and fixed boundary conditions. In contrast, for the anti-symmetric mode shown in Fig. 10, a different ordering of phase velocity profiles is evident. In this case, the phase velocity is highest for the stress-free boundary condition and lowest for the mixed condition. Overall, the results indicate that the influence of boundary conditions on phase velocity is mode-dependent, leading to distinct trends for symmetric and anti-symmetric wave propagation.

8.8. SH WAVE PROPAGATION UNDER DIFFERENT MODELING ASSUMPTIONS

SH wave propagation in a micro-structured plate is governed by size-dependent effects, rotational motion, and boundary constraints, and the present model captures their combined influence. In the absence of rotation and boundary constraints, the formulation reduces to the CCST model, while further neglect of microstructural effects recovers the classical continuum model. The results show that the incorporation of the characteristic length parameter within the classical framework leads to an increase in phase velocity, as illustrated by curves 1 (present model), 2 (CCST), and 3 (classical) in Fig. 11. However, the inclusion of rotational effects reduce the phase velocity, resulting in the present model exhibiting an intermediate response between the classical and CCST predictions.

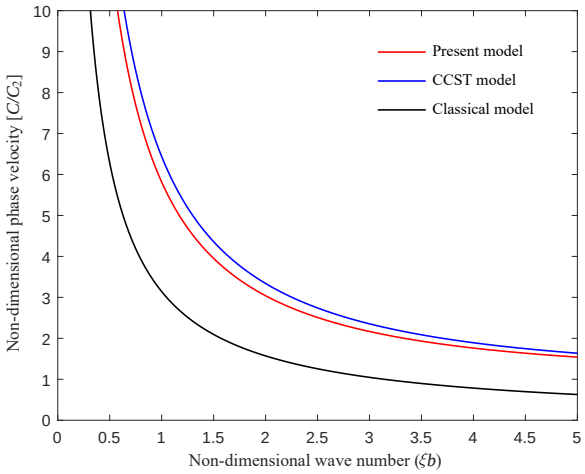


FIG. 11. Phase velocity profiles for illustrating symmetric SH-waves behavior under different models.

9. Conclusion

The present study has thoroughly examined the dispersion behavior of SH-waves in a micro-structured couple stress elastic plate subjected to rotational effects induced by Coriolis and centrifugal forces. The primary objective was to investigate the role of restrained boundary conditions, micro-structural, and geometric features such as characteristic length and plate thickness as well as the dimensionless parameters on the wave propagation characteristics.

9.1. Plate rotation effects on SH wave modes

A detailed comparison between cases with and without plate rotation reveals that rotation leads to a consistent reduction in phase velocity across both symmetric and skew-symmetric SH wave modes. While symmetric modes exhibit minimal sensitivity, anti-symmetric modes show a more pronounced effect. Interestingly, the fundamental mode in skew-symmetric profiles undergoes a polarity reversal and exhibits a cutoff frequency, a distinct behavior not seen in non-rotating plates.

9.2. Role of dimensionless parameters (Ω_1^* and Ω_2^*)

The analysis indicates that both the rotation parameters (Ω_1^*) and (Ω_2^*) influence wave dispersion symmetrically due to the appearance of Ω_1 and Ω_2 as additive and squared terms in the governing equations. An increase any of these dimensionless parameters causes a decrease in phase velocity. This reduction is consistently observed in both symmetric and anti-symmetric SH wave modes.

9.3. Effect of plate thickness

The study reveals a clear inverse relationship between the plate thickness and phase velocity, showing that an increase in thickness leads to a consistent reduction in phase velocity. This behavior highlights the sensitivity of wave propagation to geometric parameters, with plates of greater thickness exhibiting comparatively slower wave transmission.

9.4. Impact of characteristic length parameter (l)

As a core parameter of the couple stress theory, the characteristic length significantly influences wave behavior. Increasing l enhances phase velocity by amplifying micro-structural effects, thus enabling more efficient wave energy transmission. This effect highlights the importance of considering material length scales in micro-structured wave-guide designs.

9.5. Effect of dimensionless quantity (l/b)

The results reveals that the phase velocity pattern is significantly influenced by variations in the ratio l/b. Changes in this parameter lead to noticeable modifications in the propagation characteristics of SH-waves, highlighting the sensitivity of phase velocity to geometric scaling. An increase in l/b results in an amplification of the phase velocity and enhanced dispersive behavior. This suggests that the structural configuration plays a crucial role in governing wave dynamics and is essential for optimizing design parameters in applications involving guided wave propagation.

9.6. Impact of dimensionless parameters (k_1^* and k_2^*)

The shear stiffness parameter (k_1) governs resistance to tangential deformation, whereas the rotational stiffness (k_2) regulates angular restraint. The phase velocity exhibits contrasting dependence on the non-dimensional parameters (k_1^*) and (k_2^*). Specifically, an increase in (k_1^*) leads to an increase in phase velocity, attributable to the enhanced stiffness that facilitates faster wave propagation. In contrast, an increase in (k_2^*) is associated with a reduction in phase velocity. Thus, these parameters influence wave speed in distinct ways and may be tuned to achieve desired propagation characteristics, as is often explored in wave-control technologies where analogous stiffness parameters are adjusted to tailor wave speed, attenuation, or confinement.

9.7. Implications of boundary conditions

The study also provides a comparative analysis of three different boundary condition types – stress-free, fixed, and mixed. By examining all three cases, the study highlights how boundary constraints can modify phase velocity profiles and wave behavior, providing insights into the sensitivity of wave propagation to edge conditions.

9.8. Assessment of SH wave propagation across different models

SH wave propagation in a microstructured plate is influenced by size-dependent effects, the rotational motion, and boundary conditions, all of which are incorporated in the present model. The result shows that introducing a characteristic length scale increases phase velocity, while rotation decreases it, resulting in phase velocity curve of the present model to lie between the classical theory and CCST predictions.

This analysis highlights the nuanced interplay between microstructure, boundary conditions, rotation, and geometric parameters in influencing the behavior of SH-waves. The results contribute to a broader theoretical understanding of wave propagation in micro-structured rotating plates and may offer useful guidance for design considerations in certain related engineering applications.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' contributions

All authors have contributed equally in the preparation of different sections of the manuscript. All authors reviewed and approved the final manuscript.

References

1. Advani S.H. (1967), Stationary waves in a thin spinning disk, *International Journal of Mechanical Sciences*, **9**(5): 307–313, https://doi.org/10.1016/0020-7403(67)90023-9.
2. Advani S.H., Bulkeley P.Z. (1969), Nonlinear transverse vibrations and waves in spinning membrane discs, *International Journal of Non-Linear Mechanics*, **4**(2): 123–127, https://doi.org/10.1016/0020-7462(69)90021-3.
3. Asghari M., Kahrobaiyan M.H., Rahaeifard M., Ahmadian M.T. (2011), Investigation of the size effects in Timoshenko beams based on the couple stress theory, *Archive of Applied Mechanics*, **81**: 863–874, https://doi.org/10.1007/s00419-010-0452-5.
4. Auriault J.-L. (2004), Body wave propagation in rotating elastic media, *Mechanics Research Communications*, **31**(1): 21–27, https://doi.org/10.1016/j.mechrescom.2003.07.002.
5. Bhuta P.G., Jones J.P. (1963), Symmetric planar vibrations of a rotating disk, *Journal of the Acoustical Society of America*, **35**(7): 982–989, https://doi.org/10.1121/1.1918643.
6. Bulkeley P.Z. (1973), Stability of transverse waves in a spinning membrane disk, *Journal of Applied Mechanics*, **40**(1): 133–136, https://doi.org/10.1115/1.3422911.
7. Castaings M., Hosten B. (2001), Lamb and SH waves generated and detected by air-coupled ultrasonic transducers in composite material plates, *NDT & E International*, **34**(4): 249–258, https://doi.org/10.1016/S0963-8695(00)00065-7.
8. Censor D., Schoenberg M. (1973), Two-dimensional wave problems in rotating elastic media, *Applied Scientific Research*, **27**: 401–414, https://doi.org/10.1007/BF00382503.
9. Cosserat E., Cosserat F. (1909), *Theory of Deformable Bodies*, National Aeronautics and Space Administration, Washington, D.C.
10. Dargush G.F., Apostolakis G., Hadjesfandiari A.R. (2021), Two- and three-dimensional size-dependent couple stress response using a displacement-based variational method, *European Journal of Mechanics-A/Solids*, **88**: 104268, https://doi.org/10.1016/j.euromechsol.2021.104268.
11. Deep S., Sharma V. (2023), Effects of microstructures, heterogeneity, and imperfectness on propagation of SH-waves in a fiber-reinforced layer sandwiched between two microstructural half-spaces, *Iranian Journal of Science and Technology, Transactions of Mechanical Engineering*, **47**: 1161–1176, https://doi.org/10.1007/s40997-022-00570-5.
12. Djeran-Maigre I., Kuznetsov S.V. (2014), Velocities, dispersion, and energy of SH-waves in anisotropic laminated plates, *Acoustical Physics*, **60**: 200–207, https://doi.org/10.1134/S106377101402002X.
13. Dua N., Sharma V. (2024), Analysis of shear horizontal waves in heterogeneous/microstructural coupled plates: exploring the influence of interfacial bonding and the boundary conditions, *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, **46**(11): 643, https://doi.org/10.1007/s40430-024-05231-z.
14. Ghodrati B., Yaghootian A., Ghanbar Zadeh A., Mohammad-Sedighi H. (2018), Lamb wave extraction of dispersion curves in micro/nano-plates using couple stress theories, *Waves in Random and Complex Media*, **28**(1): 15–34, https://doi.org/10.1080/17455030.2017.1308582.
15. Graff K.F. (2012), *Waves Motion in Elastic Solids*, Courier Corporation.
16. Hadjesfandiari A.R., Dargush G.F. (2011), Couple stress theory for solids, *International Journal of Solids and Structures*, **48**(18): 2496–2510, https://doi.org/10.1016/j.ijsolstr.2011.05.002.
17. Hashemi S.M., Richard M.J. (2001), Natural frequencies of rotating uniform beams with Coriolis effects, *Journal of Vibration and Acoustics*, **123**(4): 444–455, https://doi.org/10.1115/1.1383969.
18. Huang H., Guan W., He X. (2024), Modal displacement analyses of Lamb waves in micro/nano-plates based on the consistent couple stress theory, *Ultrasonics*, **138**: 107272, https://doi.org/10.1016/j.ultras.2024.107272.
19. Huang Y.M., Wang C.-M. (2001), Combined methodology for analysis of rotary systems, *Journal of Vibration and Acoustics*, **123**(4): 428–434, https://doi.org/10.1115/1.1385204.
20. Jiangong Y. (2011), Viscoelastic shear horizontal wave in graded and layered plates, *International Journal of Solids and Structures*, **48**(16–17): 2361–2372, https://doi.org/10.1016/j.ijsolstr.2011.04.011.
21. Kaur M., Kumar S., Sharma V. (2024), Surface waves in a microstructural couple stress half space under the extended Mindlin's restrained boundary conditions, *Mechanics of Solids*, **59**(1): 483–495, https://doi.org/10.1134/S0025654423602720.

22. Koiter W.T. (1969), Couple stresses in the theory of elasticity, I and II, *Philosophical Transactions of the Royal Society of London B*, **67**: 17–44.
23. Kumar R., Chawla V. (2011), Surface wave propagation in a elastic layer lying over a generalized thermodiffusive elastic half-space with imperfect boundary, *Mechanics of Advanced Materials and Structures*, **18**(5): 352–363, https://doi.org/10.1080/15376494.2010.517617.
24. Kuznetsov S.V. (2006), SH-waves in laminated plates, *Quarterly of Applied Mathematics*, **64**(1): 153–165, https://doi.org/10.1090/S0033-569X-06-00992-1.
25. Kuznetsov S.V. (2015), Lamb waves in a clamped and a partially clamped elastic layer, *Mechanics of Solids*, **50**: 81–95, https://doi.org/10.3103/S0025654415010082.
26. Kuznetsov S.V. (2022), On bifurcation of guided wave in functionally graded plates, *The European Physical Journal Plus*, **137**(10): 1198, https://doi.org/10.1140/epjp/s13360-022-03435-7.
27. Lamb H., Southwell R.V. (1921), The vibrations of a spinning disk, *Proceedings of the Royal Society of London. Series A*, **99**(699): 272–280, https://doi.org/10.1098/rspa.1921.0041.
28. Lee J.S., Kim H.W., Jeon B.C., Cho S.H., Kim Y.Y. (2010), Damage detection in a plate using beam focused shear-horizontal wave magnetostrictive patch transducers, *AIAA Journal*, **48**(3): 654–663, https://doi.org/10.2514/1.44895.
29. Luo A.C.J., Mote Jr. C.D. (2000), Nonlinear vibration of rotating thin disks, *Journal of Vibration and Acoustics*, **122**(4): 376–383, https://doi.org/10.1115/1.1310363.
30. Mindlin R.D. (1960), Waves and vibrations in isotropic, elastic plates, [in:] *Structural Mechanics*, pp. 199–232, Pergamon Press, New York.
31. Mindlin R.D., Tiersten H.F. (1962), Effects of couple-stresses in linear elasticity, *Archive for Rational Mechanics and Analysis*, **11**: 415–448, https://doi.org/10.1007/BF00253946.
32. Moukhomodiarov R.R., Pichugin A.V., Rogerson G.A. (2010), The transition between Neumann and Dirichlet boundary conditions in isotropic elastic plates, *Mathematics and Mechanics of Solids*, **15**(4): 462–490, https://doi.org/10.1177/1081286509103781.
33. Moukhomodiarov R.R., Rogerson G.A. (2012a), Generalisation of elastic models for a layer with elastically restrained boundaries, *International Journal of Engineering Science*, **57**: 79–89, https://doi.org/10.1016/j.ijengsci.2012.04.004.
34. Moukhomodiarov R.R., Rogerson G.A. (2012b), Long-wave dispersion phenomena in a layer subject to elastically restrained boundary conditions, *Zeitschrift für angewandte Mathematik und Physik*, **63**: 171–188, https://doi.org/10.1007/s00033-011-0161-0.
35. Moukhomodiarov R.R., Rogerson G.A. (2013), Asymptotic long wave models for a pre-stressed elastic layer with elastically restrained boundaries, *International Journal of Solids and Structures*, **50**(11–12): 1944–1953, https://doi.org/10.1016/j.ijsolstr.2013.02.014.
36. Nowinski J.L. (1981), Stability of thermoelastic waves in membrane-like spinning disks, *Journal of Thermal Stresses*, **4**(1): 1–11, https://doi.org/10.1080/01495738108909948.
37. Paimushin V.N., Gazizullin R.K. (2018), Acoustic wave propagation through a plate fixed on a rigid frame via elastic spacers and located between two barriers, *Journal of Applied Mechanics and Technical Physics*, **59**: 733–746, https://doi.org/10.1134/S0021894418040211.
38. Phan H., Cho Y., Pham C.V., Nguyen H., Bui T.Q. (2019), A theoretical approach for guided waves in layered structures, [in:] *AIP Conference Proceedings*, **2102**(1): 050011, https://doi.org/10.1063/1.5099777.
39. Schoenberg M., Censor D. (1973), Elastic waves in rotating media, *Quarterly of Applied Mathematics*, **31**(1): 115–125, https://doi.org/10.1090/qam/99708.
40. Sharma V., Kumar S. (2014), Velocity dispersion in an elastic plate with microstructure: effects of characteristic length in a couple stress model, *Meccanica*, **49**: 1083–1090, https://doi.org/10.1007/s11012-013-9854-0.
41. Sharma V., Kumar S. (2015), Effects of liquid loadings on lamb waves in context of size dependent couple stress theory, *Journal of Theoretical and Applied Mechanics*, **53**(4): 925–934, http://dx.doi.org/10.15632/jtam-pl.53.4.925.
42. Sharma V., Kumar S. (2023a), A comprehensive analysis of horizontally polarized shear waves in a thin microstructural plate, *Structural Engineering and Mechanics*, **85**(4): 501–510, https://doi.org/10.12989/sem.2023.85.4.501.

43. Sharma V., Kumar S. (2023b), A study of plane and Rayleigh waves in a microstructural medium: the role of size dependency and thermal effects, *Mechanics of Solids*, **58**: 1335–1350, https://doi.org/10.3103/S0025654423600599.

44. Simonetti F., Cawley P. (2004), On the nature of shear horizontal wave propagation in elastic plates coated with viscoelastic materials, *Proceedings of the Royal Society of London. Series A*, **460**(2048): 2197–2221, https://doi.org/10.1098/rspa.2004.1284.

45. Southwell R.V. (1922), On the free transverse vibrations of a uniform circular disc clamped at its centre; and on the effects of rotation, *Proceedings of the Royal Society of London. Series A*, **101**(709): 133–153, https://doi.org/10.1098/rspa.1922.0032.

46. Toupin R.A. (1962), Elastic materials with couple-stresses, *Archive for Rational Mechanics and Analysis*, **11**(1): 385–414, https://doi.org/10.1007/BF00253945.

47. Voigt W. (1892), Theoretical studies of the elastic behaviour of crystals, *Presented at the session of the Royal Society of Science on 2 July 1887*.

48. Wang C., Chen X., Wei P., Li Y. (2017), Reflection of elastic waves at the elastically supported boundary of a couple stress elastic half-space, *Acta Mechanica Solida Sinica*, **30**(2): 154–164, https://doi.org/10.1016/j.camss.2017.03.004.

49. Wu C.-P., Hsu C.-H. (2022), A three-dimensional weak formulation for stress, deformation, and free vibration analyses of functionally graded microscale plates based on the consistent couple stress theory, *Composite Structures*, **296**: 115829, https://doi.org/10.1016/j.compstruct.2022.115829.

50. Wu C.-P., Hu H.-X. (2021), A unified size-dependent plate theory for static bending and free vibration analyses of micro- and nano-scale plates based on the consistent couple stress theory, *Mechanics of Materials*, **162**: 104085, https://doi.org/10.1016/j.mechmat.2021.104085.

51. Yang F., Chong A.C.M., Lam D.C.C., Tong P. (2002), Couple stress based strain gradient theory for elasticity, *International Journal of Solids and Structures*, **39**(10): 2731–2743, https://doi.org/10.1016/S0020-7683(02)00152-X.

52. Yu Y.Y. (1996), Linear vibrations of plates based on elasticity theory, [in:] *Vibrations of elastic plates*, pp. 31–55, Springer, New York, https://doi.org/10.1007/978-1-4612-2338-2_2.

Archives of Acoustics Vol. 51, No. 3, pp. 581–600 (2026), https://doi.org/10.24423/archacoust.2026.4371

Technical Note

Implementation of a COVAREP-Integrated Application for Voice Analysis Using LLMs

Hubert Daniszewski[(1)], Krzysztof Szklanny[(1)*], Piotr Wrzeciono[(2)]

[(1)] *Multimedia Department, Polish-Japanese Academy of Information Technology* Warsaw, Poland

[(2)] *Institute of Information Technology, Warsaw University of Life Sciences* Warsaw, Poland

*Corresponding Author: kszklanny@pjwstk.edu.pl

Received November 7, 2025; revised April 26, 2026; accepted May 24, 2026; version of record August 25, 2026; published issue September 23, 2026.

This paper presents VocalCheck, a mobile application designed to monitor voice conditions using recordings made with a smartphone. The initial phase involved a literature review and an evaluation of existing voice processing applications, which informed the project's design requirements. Based on this analysis, an application was developed to record and preliminarily analyze voice samples using parameters provided by the Collaborative Voice Analysis Repository (COVAREP) for the Speech Technologies Toolkit. Additionally, an experimental study was conducted to assess the effectiveness of large language models (LLMs), specifically Microsoft Copilot and ChatGPT, in generating MATLAB scripts for computing voice quality parameters. Although these tools proved helpful for general programming tasks, their performance in speech signal processing was unsatisfactory.

Keywords: voice quality, voice diagnostics, mobile app, speech processing, LLM.

Copyright © 2026 The Author(s).
This work is licensed under the Creative Commons Attribution 4.0 International CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Voice disorders affect individuals across various professions. Teachers, in particular, are highly susceptible to excessive strain on their vocal apparatus. Similar challenges can also affect choristers, singers, actors, public speakers, lecturers, and fitness instructors. Health factors, such as gastroesophageal reflux disease and rare genetic conditions like Pompe disease or Morquio syndrome type A (MPS IVA) (Hendriksz *et al.*, 2014), also affect the quality of the voice. Children may experience voice problems caused by excessive crying or screaming, which can lead to the development of vocal nodules. The project aimed to develop a tool to detect voice abnormalities, enabling the maintenance of vocal hygiene through preventive measures. As part of this goal, the mobile application VocalCheck was created, inspired by a prototype described in (Szklanny, 2019).

2. Literature review and existing solutions

The topic of using smartphones in voice therapy has garnered significant interest, which aided in the decision to choose the type of device for the VocalCheck application. The available hardware solutions included desktop computers, laptops, tablets, smartphones, and smartwatches. According to Google Scholar, approximately 15 800 scientific papers on this topic were published between 2015 and 2025 (Fig. 1).

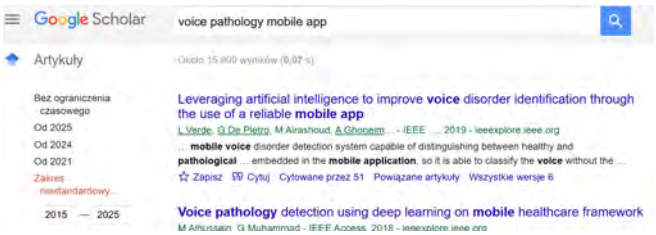


FIG. 1. Search results for 'voice pathology mobile app' on Google Scholar.

Descriptions of research on this subject are drawn from scientific journals such as the *Journal of Voice*, *European Archives of Oto-Rhino-Laryngology*, *IEEE Access*, and *Archives of Acoustics*. Many of the articles presented in these journals address issues related to devices and conditions for voice recording, as well as the support of machine learning tools. In the following, we describe some papers that utilize various devices in human voice diagnostics.

In (MANFREDI *et al.*, 2017), the usefulness of smartphone recordings was compared with results obtained under laboratory conditions. The experiment compared two phone models: HTC One and Wiko CINK SLIM2. Recordings from these devices were compared with recordings from the Sennheiser MD421U microphone, which, according to the mentioned article, is commonly used in acoustic laboratories. Measurements of jitter, shimmer (BOERSMA, 2001), and HNR (YUMOTO *et al.*, 1982) parameters for the phonation of the vowel /: a/ were strongly correlated with each other for all devices. The study demonstrated that even inexpensive smartphones can reliably record voices with pathological changes, comparable to those of professional laboratory microphones.

The authors of the article (ULOZA *et al.*, 2015) used natural phonation to compare values of fundamental frequency (f_0) (COOPER, SORENSEN, 1981), normalized noise energy (NNE) (KASUYA *et al.*, 1986), and signal-to-noise ratio (SNR) (KLINGHOLZ, 1987) obtained for recordings from a Samsung Galaxy Note 3 phone and an AKG Perception 220 microphone. The calculated parameters, except for f_0 and shimmer differed statistically between phone and microphone recordings. The study described in the mentioned article used a classifier based on a random forest model (HO, 1995), which distinguished between healthy and pathological voices with an accuracy of 73.7 % for phone recordings and 79.5 % for microphone recordings.

Similar binary classifiers were described in (VERDE *et al.*, 2019) (accuracy at 84.5 %) and (ASCI *et al.*, 2020). The authors of the latter achieved an 88.5 % accuracy in distinguishing the age of the speaker regardless of gender, using a set of 20 most significant acoustic parameters, including SNR, Mel-frequency cepstrum coefficients (MFCC) (ZHENG *et al.*, 2001), jitter, shimmer, and smoothed cepstral peak prominence (CPPs) (HILLENBRAND, HOUDE, 1996), selected from over 6000 parameters describing the voice. The mentioned article noted that age affects the human voice and that speech analysis is possible using recordings made in home conditions, without specialized equipment or acoustically adapted rooms.

In another study (ALHUSSEIN, MUHAMMAD, 2019), a voice pathology detection system based on parallel convolutional neural networks (CNNs) was presented, achieving 94.9 % accuracy when trained on the Saarbrücken Voice Database (SVD) and was tested in the Massachusetts Eye and Ear Infirmary (MEEI) database. When both the training and test sets were drawn from SVD, the accuracy was 95.5 %. The system was also able to classify three types of pathologies (unilateral paralysis, polyps, and cysts of the vocal folds) with accuracies of 95.1 %, 95.9 %, and 99.7 %, respectively.

In addition to diagnosing changes in the vocal folds, such as polyps and cysts, studies also focus on other voice disorders. SZKLANNY *et al.* (2016) described the process of monitoring the voice condition of individuals with late-onset Pompe disease. This disease, in its advanced form, often causes speech disorders by negatively affecting the work of nerves and muscles in the respiratory system. Electroglottographic (EGG) recordings were analyzed using the Collaborative Voice Analysis Repository (COVAREP) (DEGOTTEX *et al.*, 2014) repository and PRAAT (BOERSMA, 2001). These analyses provided detailed information on the functioning of the vocal tract, consistent with the phoniatric diagnosis. All nine participants with the juvenile form of the disease were diagnosed with voice abnormalities, and 7 out of 10 patients with the mature form of the disease were found to have dysphonia.

Dysphonia is characterized by hoarseness, abnormal timbre, loudness, and duration of the utterance (Kosztyła-Hojna *et al.*, 2014).

The continuation of the study on Pompe disease patients was described in (Szklanny, Tylki-Szymańska, 2018). Fifteen patients with the mature form of the disease underwent re-evaluation three years after the initial study. All patients participated in enzyme replacement therapy. Analysis of acoustic and EGG measurements revealed that the patient's voice condition had deteriorated. An increased degree of vocal fold insufficiency and increased vocal tension were also observed. Three years after the previous study, the average value of the closed quotient (CQ H) (Howard, 1995) parameter decreased from 0.345 to 0.307, which, with a p-value of 0.025, was considered statistically significant. Similarly, the average value of peak slope (PS) among patients decreased from −0.304 to −0.372 (Kane, Gobl, 2011). The article indicates that parameters such as CQ H and PS not only allow for the detection of voice disorders but also for tracking the development of diseases that negatively affect the voice.

The same signal analysis methods were applied in (Szklanny *et al.*, 2018). This rare genetic disease negatively affects the larynx and vocal tract. Due to the accumulation of glycosaminoglycans (GAG), MPS IV A can cause acute obstruction of the upper airways and deformities of the laryngeal vestibule, vocal folds, and posterior laryngeal mucosa. In individuals with MPS IV A, an increased value of the PS parameter was observed compared to the control group, indicating a breathy voice, a decreased value of the CPP parameter, characteristic of dysphonia, and a low value of HNR, which is common in a hoarse voice. Shimmer and jitter were elevated compared to the control group, indicating instability in vocal fold vibration.

Szklanny *et al.* (2019) described a similar process of analyzing EGG recordings in the diagnosis of vocal nodules in children. Vocal fold nodules are typically caused by voice overuse or vocal hyperfunction, and their symptoms include persistent hoarseness, a disturbance in vocal fold vibrations that results in turbulent airflow in the glottis, manifesting itself as a raspy, rough voice. The study involved 57 patients aged 5 to 14 years who had been diagnosed with vocal nodules. CQ H, PS, normalized amplitude quotient (NAQ), and CPP were used to determine voice quality. An elevated PS value of −0.19, compared with to –0.28 in the control group consisting of 37 healthy children aged 7.5 to 16 years, indicated a breathy voice. Tension in the voice was detected in 14 patients, signaled by an elevated NAQ value of 0.14. For the control group, NAQ was 0.12. CPP allowed for the detection of dysphonia in 30 subjects, with an average value of 11.64 among the sick and 11.66 in the control group. The average CQ H was 0.31 in sick children and 0.41 in the control group, indicating vocal fold insufficiency in patients with vocal nodules.

The study of vocal nodules in children was expanded with the RBH scale in (Szklanny, Wrzeciono, 2019b). It described the Voice Pathology Analysis Database (VPADB), which was created as part of this study. It contains recordings of voices with vocal nodules, data obtained using an electroglottograph, and perceptual assessments of roughness, breathiness, and hoarseness (RBH) scale (Nawka *et al.*, 1994). Data from healthy voices from the SVD database were used for the control group. The assessments were made based on the opinions of two independent experts in voice analysis. To determine the RBH scale values for a given voice sample, the parameters PS, NAQ, and CQ H were used. Each of the three RBH features was analyzed for three verdicts: 0 for healthy individuals (denoted as R0*, B0*, and H0*), 0 for individuals with vocal nodules (R0, B0, H0), and 1 for individuals with vocal nodules (R1, B1, H1). Using ANOVA, it was found that CQ H and NAQ allow distinguishing among R0*, R0, and R1, as well as B0*, B0, and B1. For the distinction between H0*, H0, and H1 values, the PS parameter was also needed, although it was considered significantly less effective. The results of the voice quality tests agreed with the assessments of both experts. A further supplement to the study of vocal nodules was described in (Szklanny, Wrzeciono, 2019a).

As part of this manuscript, a classifier based on a genetic algorithm (Holland, 1975) was prepared, which diagnosed the degree of vocal fold insufficiency based on voice samples and electroglottograph recordings. The final form of the classifier used the following parameters: CQ H, PS, NAQ, and RDS. The accuracy of the classifier was evaluated at 92 % for the training set and 94.73 % for the validation set. The high effectiveness of the obtained classifier, combined with the possibility of its representation using arithmetic functions understandable to humans, enabled its implementation in the LibreOffice package using Visual Basic for Applications (VBA), making it available, for example, in spreadsheets.

The literature also describes the impact of voice pitch on subglottal pressure values and the NAQ parameter. In (Björkner *et al.*, 2005), professional opera artists were asked to perform a diminuendo, gradually softening their voice by reducing the sound level while singing the syllable /pae:/. The recording was repeated three times, each performed at 25 % and 75 % of the vocal range of the subjects, expressed in the number of semitones. Airflow calculations were performed using the Rothenberg mask. The subglottal pressure was measured by connecting a plastic tube, held in the corner of the singers' mouths, to a subglottal pressure sensor. It was observed that while subglottal pressure varies among singers, it is strongly correlated with their fundamental frequency component, f_0. In the case of high tones, it was about twice as high as in recordings where the same singers used lower tones. After normalizing the subglottal pressure relative to the main component, it was observed that as the normalized measurement increased, the NAQ value decreased. This parameter decreases significantly in the case of low pressure. It was concluded that this illustrates the manner of phonation, which in the case of professional opera artists is independent of the sound level and pitch of the sung tones.

The aforementioned acoustic parameters, part of the COVAREP toolkit, are reliable and enable a detailed analysis of voice quality. The COVAREP toolkit is currently not being developed; its maintenance is important for conducting further experiments, which is part of our article.

Many experiments are conducted in acoustic cabins to minimize the influence of the acoustic background. The acceptable level of ambient noise to obtain reliable measurements was described in (Lebacq *et al.*, 2017). In this experiment, the maximum acceptable level of ambient noise that does not affect the values of jitter [%] and NHR was determined. Voice samples were generated by a synthesizer described in (Manfredi *et al.*, 2017). Artificial recordings differed in terms of phonation type, f_0 frequency, and the level of jitter and noise to simulate the speaker's voice condition. The ambient noise used in the experiment was described as 'shopping small ambient.' In the conclusions, the researchers stated that using smartphones to measure NHR and f_0 perturbations is possible, provided that the maximum level of ambient sounds does not exceed 50 dBA and that voice samples are limited to the phonation of the vowel /a/. In relation to the article, it should be noted that parameters such as PS, NAQ, and H1–H2, H1–A3 are weakly sensitive to low SNR levels. According to the mentioned publication, the required SNR level is 10 dB. This corresponds to a situation where the ambient noise level is 50 dBA, and the phonation level is 60 dBA.

In (Uloza *et al.*, 2023), the influence of ambient noise and the device used on the values read by PRAAT was presented. The level of noise added to the recordings averaged 29.61 dB, and its SNR, compared to voice recordings, was 38.11 dB. To compare the acoustic voice quality index (AVQI) calculated for recordings made with an AKG Perception 220 microphone and an iPhone 6s smartphone, a Bland–Altman plot was used. This is a scatter plot of two variables, where the horizontal axis represents the arithmetic mean of both measurements, and the vertical axis represents the difference between these measurements. It aims to illustrate the difference between the two measurements relative to the size of the measured values (Kaur, Stoltzfus, 2017). Despite a p-value of 0.07, the study's authors considered the differences in recordings from the microphone and the phone with added noise to be not statistically significant. Additionally, they emphasized that the obtained Bland–Altman plots for all three versions of the recordings did not demonstrate a systematic difference in the calculated AVQI. Therefore, according to the study's authors, smartphones can be an acceptable tool for recording voice for therapeutic purposes.

In the literature on voice diagnostics using smartphones, there are also studies examining the usefulness of such applications. Angadi *et al.* (2023) described the MyVocalHealth application, which was only available during the study described in that paper. The application aimed to motivate users to participate in voice therapy exercises on a regular basis. At the end of the study, data showed that application users performed about 50 % more than the control group, which did not use the application. In addition, the study's authors placed special emphasis on usability tests, i.e., tests in which users had to perform a given task using the application independently. Seventeen participants took part in them, of whom over 95 % performed them correctly. The most common mistake was skipping or failing to pause the instructional video attached to the exercise. Thirteen of the test participants stated that the application was easy to use, and seven noted that the instructional videos were helpful. The most significant reservations regarding usability were problems with the button leading to the next

task and the video. 12 respondents answered a survey enabling the calculation of the application's score on the system usability score (SUS) scale. The application scored 78.13 points, exceeding the average rating of 68 points for web systems considered usable. After conducting the survey, six testers participated in a group discussion, during which the following suggestions for improving the quality of the application were formulated:

1. The application should contain a clear indication showing that the exercise has been completed.
2. A modifiable system of reminders and snooze functions would help in regularly participating in exercises.
3. The best result in the exercises should be displayed at the top of the screen. The application only displayed a chart of chronologically performed exercises over the last 7 and 30 days.

The respondents stated that the application was most often used at home, in the car, in the parking lot, and in restaurants. Most respondents performed the exercises according to the recommendations, in the morning and evening, but some also performed them during their lunch break. The MyVocalHealth application did not calculate any voice-related parameters; it only measured the phonation duration. However, because for the study of the application's usability, it serves as a good reference point when designing applications that support voice diagnostics.

In summary, numerous studies have been conducted on the use of recordings in voice analysis. Many of them allow for a comparison of the effectiveness of using different types of recording devices, ranging from specialized ones such as electroglottographs to professional microphones and commonly available mobile phones. In analyzing this topic, a variety of measures are used to assess voice condition. Some of the parameters used, such as CQ H, require the use of an electroglottograph; however, there are also many others, including PS, NAQ, NHR, CPP, AVQI, jitter, and shimmer, which can be calculated based on audio recordings. They enable detection of vocal fold insufficiency and allow observations of changes in the voice caused by various factors, such as vocal strain or aging. Thanks to the PRAAT program and the COVAREP toolkit (Degottex *et al.*, 2014), there are already ready-made implementations of algorithms calculating the values of useful parameters. The conditions under which the recording is made are also important, as well as what is being recorded. Some studies utilize soundproof rooms or soundproof booths, while other recordings to be made in standard rooms or public settings, such as restaurants or hospitals. Attention should be paid to the distance between the device and the speaker, the angle of its inclination, and the level of ambient noise. It is important that the phone is far enough from the speaker's mouth and that the level of background noise does not exceed 50 dB. Study participants were often asked to phonate the vowel /a/ or read a sentence that allowed the analysis of intonation changes in the voice. Phonation, due to its simplicity, is understandable to many user groups and is effectively processed by existing algorithms analyzing audio signals. Many of the cited articles affirm the use of mobile phones for diagnostic purposes and present real applications supporting voice therapy. Both machine learning and traditional statistical methods achieve high accuracy in distinguishing between healthy and pathological voices. Artificial intelligence methods, such as machine learning, have already been used in voice diagnosis in studies by Verikas *et al.* (2006), Crowson *et al.* (2020), and Suvvari (2023). Furthermore, Compton *et al.* (2022) employed a deep learning approach, and Tirronen *et al.* (2023) utilized transfer learning.

The literature review in this area also provided important information on how to design such tools to be user-friendly and useful for the target audience. Attention was paid, among other things, to the clarity of instructions and the visualization of changes in voice condition assessment over time, which allows tracking the therapy progress.

In summary, numerous studies confirm the usefulness of audio recordings as a tool for analyzing and assessing voice conditions. The literature compares the effectiveness of various types of recording devices – from specialized devices such as electroglottographs to high-quality laboratory microphones and commonly available smartphones. The use of these devices enables the determination of a wide range of acoustic parameters, such as NAQ, PS, NHR, CPP, AVQI, which provide significant information on the quality of phonation and enable the detection of abnormalities such as vocal fold insufficiency. Some parameters, such as closed quotient, require an electroglottograph. Still, many of them can be calculated solely based on acoustic recordings using available tools such as the COVAREP (Degottex *et al.*, 2014) computational package or the PRAAT program.

However, the effectiveness of the analysis depends not only on the selection of parameters and algorithms but also on the conditions under which the recordings are made. Key factors include the microphone's distance from the speaker's mouth, its angle, and the level of background noise – it is recommended that background noise not exceed 50 dB. Depending on the research design, recordings are made in both controlled conditions (e.g., soundproof booths) and natural environments, such as living rooms, restaurants, or medical facilities. Such a variety of conditions enables the assessment of the resistance of analytical algorithms to interference, bringing them closer to clinical applications and daily therapeutic practice.

In studies, tasks that involve prolonged phonation of the vowel /:a/ or reading a short sentence are most often used. Prolonged phonation is a straightforward and intuitive method for obtaining voice data, that is understandable and feasible for a wide range of users, while simultaneously providing a signal of sufficient length and stability for a reliable acoustic analysis.

The results of numerous scientific papers suggest that both classical statistical methods and modern machine learning techniques achieve high accuracy in distinguishing healthy from pathological voices, based on recordings made using a mobile phone. Importantly, the literature emphasizes the need to design applications that support diagnosis and therapy, including a clear user interface, comprehensible instructions, and visualization of the patient's progress. These elements are crucial to the acceptability and usability of such tools in clinical practice and daily monitoring of voice conditions.

3. Existing applications

The process of preparing for the project involved testing selected applications thematically related to voice processing. The choice of applications was limited to those that are functional and have support. This section describes five applications, including the functionalities they provide, along with a list of identified advantages and disadvantages for each solution. The first of the tested applications is the Sound Meter (Decibel Meter)[1]. The central part of the application's screen displays the sound level in dB. Below it, depending on the settings, there is a detailed graph of the signal level readings over time or a list of sound level thresholds with examples of corresponding sources. At the bottom of the screen, control buttons are available to perform various functions, including calibration, resetting and pausing the measurement, and changing the color scheme. The main advantage of this tool is its simple and intuitive interface; however, it does not provide functions sufficient for assessing the condition of the voice.

The Vocal Pitch Monitor[2], offers a more advanced method of measurement. This program is designed for individuals involved in music, so some of its features may be incomprehensible to the average user. The main panel of the tool displays the received pitch and the voice pitch in real time. The application enables recording, playback, and sharing of recordings in the WAV format. However, this functionality does not always work correctly, as the application cuts off fragments of recordings and does not allow for their free scrolling. The application offers a wide range of options, allowing users to adjust, among other things, the graph zoom, select the format of note names (letters and solmization), display frequencies, and change the colors of interface elements on the RGB scale.

The Voice Tools application[3], is designed for transgender individuals, and its main purpose is to help them practice speech to affirm their gender identity. However, the exercises and tools available in it can allow for a deeper analysis of audio signals than the previous solutions. It consists of five modules: pitch, which allows to see a graph of the frequency reading of a sample sentence; tone, containing a list of buttons playing synthesized sounds in different tones; a panel similar to pitch, volume, with a graph of the sound level; analyse, which shows a longer fragment of text to be read aloud, and then displays statistical data regarding frequency, signal level, and ambient noise; and the 'more' button, which provides access to statistics from the entire time of using the tool, application settings, and a spectrogram generated in real time.

[1]Available on Google Play at: https://play.google.com/store/apps/details?id=com.gamebasic.decibel.

[2]Available on Google Play at: https://play.google.com/store/apps/details?id=com.tadaoyamaoka.vocalpitchmonitor&hl=en_IN&gl=US.

[3]Available on Google Play at: https://play.google.com/store/apps/details?id=com.DevExtras.VoiceTools&hl=en_US.

The aforementioned options – in addition to visual customization – include settings that help maintain vocal health, such as a reminder to drink water before vocal exertion and support for pronunciation practice. The user can choose the difficulty level of the sentences that will be suggested by the application and add his or her own sentences. An advantage of the program is also the presence of a tutorial in the form of dialog boxes with clear instructions. The biggest limitation of Voice Tools is the lack of support for analyzing pre-recorded audio, it supports only live recording, with a maximum duration of 120 s.

Loud and Clear for Parkinson's[4] is a solution designed for practicing diction and is specifically aimed at individuals with Parkinson's disease. Its creator is a speech therapist and graduate of the University of Texas at Dallas. He also obtained a certificate from the American Speech-Language-Hearing Association (ASHA). The therapy involves doing warm-up exercises, daily exercises, and homework assignments. The tasks in the application are varied, ranging from simple phonations of individual sounds to reading simple sentences, and include language puzzles and answering questions that simulate conversation (e.g., about a favorite ice cream flavor). Before the exercises, a short video is displayed showing how to perform them. The exercise itself is accompanied by a meter that indicates whether the user should speak louder, softer, or maintain the current sound level. The application also contains educational resources on Parkinson's disease, suggested therapists, information on the length of the daily exercise series with achievements, and sequences of commands with a voice recording function, which allows the user to monitor progress in diction training. The tool uses gamification elements, including a series of daily exercises and achievements, to encourage regular therapy. The gamification process involves introducing elements that make a given experience similar to a game. Such an approach to application design is explored, among others, in education. Among the negative aspects of the experience with this program, one can mention the lengthy module load times compared to other solutions and the random requests to access sensitive data that the application does not require for proper operation, such as multimedia files stored on the device or location data. It also does not provide any precise data, for example, regarding the voice level in dB.

The last application tested as part of preparing to create our tool is AI SoundLab[5]. It requires registration, during which the user can provide several medical data, including information regarding voice and respiratory system diseases. AI SoundLab consists of submodules called 'studies,' access to which requires entering codes provided by authorized entities. During the tests, the publicly available COVID-19 study, as well as the Health AI Demo and Respiratory AI Demo studies, were reviewed. The first of these is a survey during which, in addition to answering questions about health status, the user performs several voice tasks, including recording their cough, laughter, phonations of sounds, and reading text. Health AI Demo involves making a 30 s to 60 s voice recording of any content, after which the application evaluates the detected sentiment in it. The result is presented as a point on a graph, where the horizontal axis is labeled 'positivity' and the vertical axis is labeled 'activation.' The Respiratory AI Demo is a study that evaluates voice health. After recording coughs, reading a text fragment, and a five-second phonation of the vowel /i/, the user receives an assessment based on four metrics:

- respiratory health: detection of respiratory system disease symptoms, such as wet and dry cough and sore throat, calculated by a deep neural network,
- vocal strength: voice strength measured by the energy of high frequencies of the signal,
- speech rate: articulation speed, measured by the average distance between speech spectra adjacent in time,
- voicing control: the speaker's control over the vocal cords, measured by the average length of speech fragments.

In summary, AI SoundLab is the most complex of all the tools tested as part of this project, allowing for the acquisition of valuable information regarding the design of applications focused on voice analysis. The application highlights the importance of clear instructions for performing tests and serves as a reference point for the user interface's graphic design. The quality of solutions in AI SoundLab largely depends on the person designing the study. Still, the application also contains noteworthy universal elements, such as the ability to record and analyze audio samples.

[4]Available on Google Play at: https://play.google.com/store/apps/details?id=com.ateamo.loud_and_clear&hl=en_US.
[5]The application is available online at: https://aisoundlab.audeering.com/welcome.

All the tested applications contributed to determining the key functions of the target project and to preparing the user interface based on the expectations of the target group.

4. VocalCheck application

After testing applications on similar topics and establishing the project's main assumptions, a mobile application called VocalCheck was created. The application's backend was written in Java with the Spring Boot framework, designed for building such applications. The frontend of the project was implemented using HTML pages, Cascading Style Sheets (CSS), and JavaScript. When attempting to access the application without logging in, it displays a panel with a button that allows login using a Google account (Fig. 2). The primary advantage of using the Google platform is that the application administrator does not need to provide a registration panel or manage user data independently, which may be considered sensitive under certain countries' legal regulations. For example, Article 4(1) of the General Data Protection Regulation (GDPR), adopted on April 27, 2016, defines personal data as information that allows direct or indirect identification of a natural person. Data related to voice may be classified as a special category of personal data under Article 9 of the regulation, as it may reveal the ethnic origin of a person or the diseases they are struggling with. Thanks to the Google Cloud API, all data regarding the voice test history and personal user data, such as gender, is stored directly in the application's files on the logged-in user's Google Drive. This avoids complications related to storing sensitive information on one's servers.



Fig. 2. VocalCheck application login panel.

The screen for recording a voice sample (Fig. 3), which will be evaluated, allows for two ways of providing the recording:

– recording a five-second phonation of the vowel /a/ in real time,
– providing a wave file with a voice recording from the device.

Rather than regular speech, sustained vowel phonation has been chosen since it is the traditional form of voice testing (Askenfelt, Hammarberg, 1986). The drawback of this approach is that it does not represent the natural form of speech in a standard spontaneous setting (Parsa, Jamieson, 2001).

Pressing the microphone icon starts the recording. Before the signal is saved, the user must grant the web browser access to the microphone; otherwise, an error message will be displayed. During recording, the sound wave is shown in the center of the screen as a waveform. Below the graph, buttons are available to cancel the recording (red cross) or end it before the allotted time expires (green checkmark) (Fig. 4). The application checks whether the recording is clipped and whether it has the appropriate signal strength. Clipping occurs when the input signal voltage exceeds the range of values that a given device can process. This results in the excess signal value being 'cut off,' introducing additional harmonic components that are audible as a characteristic distortion in the sound. Using the Web Audio API in JavaScript, the application detects the time-domain moment at which

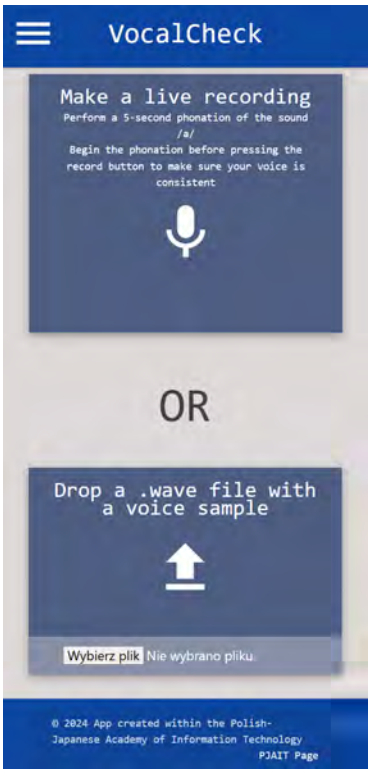


FIG. 3. Voice sample collection panel in the VocalCheck application.



FIG. 4. Recording user voice in the VocalCheck application.

the signal amplitude exceeds its maximum value and displays an error message prompting the user to check their microphone settings.

To additionally detect whether the signal strength corresponds to the sufficient strength needed for calculations, the power of each sampled signal is also counted. This was done using Eq. (1). The total signal level is compared with the −20 dB threshold. This is the minimum signal level required to obtain correct calculation results. This level was determined empirically during application tests:

$$P = \frac{1}{N}\sum_{n=0}^{N-1}|f(n)|^2, \tag{1}$$

$$L = 10\log_{10} P. \tag{2}$$

After submitting the voice sample in the form of a live recording or a wave file, functions from the COVAREP (Degottex *et al.*, 2014) repository are launched, calculating the following parameters:

- PS: calculated by a function from the file *glottalsource/peakslope.m*,
- CPP: calculated by a function from the file *glottalsource/cpp.m*,
- NAQ: calculated by a function from the file *glottalsource/get_vq_params.m*.

The PS is defined as the average decrease in amplitude (in dB) between successive spectral peaks within a specified frequency range of the magnitude spectrum of a speech frame. It reflects the spectral sharpness and correlates with voice quality features such as breathiness and tenseness. It allows for the detection of breathy and strained voice. The formula for calculating PS was taken from (Kane, Gobl, 2017). The PS is calculated by observing the wavelet decomposition, given the following formula for the mother wavelet:

$$g(t) = -\cos(2\pi f_n t) \cdot \exp\left(\frac{t^2}{2\tau^2}\right), \tag{3}$$

where f_n is $\frac{f_s}{2}$ for sampling frequency $f_s = 16\,\text{kHz}$, τ is $\frac{1}{2f_n}$.

The CPP is a measure calculated based on the cepstrum, which is the spectrum obtained by calculating the inverse Fourier transform for the logarithm of the spectrum of the audio signal. The cepstrum in the real number domain was described as follows:

$$C_r(q) = \mathcal{F}\left\{\log|S(f)|^2\right\}, \tag{4}$$

where $\mathcal{F}$ is the Fourier transform, $S^2(f)$ is the power spectrum of the signal $s(t) : S^2(f) = \mathcal{F}\{E[s(t) \cdot s(t-\tau)]\}$.

CPP is the difference between the maximum value on the cepstrum graph and the corresponding point of linear regression for the entire cepstrum. It is defined as the spectrum calculated from the spectrum, whose low value is usually correlated with a greater intensity of dysphonia (Patel *et al.*, 2018).

The NAQ is defined as the ratio between the maximum amplitude of the differentiated glottal flow and the product of the maximum negative amplitude of the glottal flow derivative and the fundamental period. It reflects how abrupt or smooth the glottal closing phase is during vocal fold vibration. Its ability to express the phonation type has been confirmed in the study by Airas and Alku (2007). NAQ was described by the following formula:

$$\text{NAQ} = \frac{\text{AQ}}{T}, \tag{5}$$

where AQ is the amplitude quotient, T is the fundamental period of the flow.

The transmitted audio signal is temporarily saved on the server's disk. Then, using the open-source environment Octave, which allows running scripts in the MATLAB language, the code is executed to compute the values of selected voice parameters. The PS and NAQ parameters allow for the classification of the processed voice into one of three categories:

1. Modal voice – neutral, healthy voice.
2. Strained voice (tense voice or pressed voice) – a voice condition resulting from increased tension in the vocal tract. It is characterized by complete and sharp closure of the vocal folds, accompanied by more intense articulation.
3. Breathy voice – the opposite of a strained voice. It results from low tension in all elements of the vocal tract and is characterized by incomplete closure of the vocal folds, accompanied by an audible inhalation sound.

Other parameters that could be used to distinguish between those categories are maxima dispersion quotient (MDQ) (Kane, Gobl, 2013), parabolic spectral parameter (PSP) (Alku *et al.*, 1997), harmonic richness factor (HRF) (Childers, Lee, 1991), H1H2 (Hanson, 1997), quasi-open quotient (QOQ) (Hacki, 1989).

In (Kane, Gobl, 2011) box plots were presented regarding the statistical values of PS and NAQ parameters for each of the three voice types. Based on the mentioned article, it was assumed that for modal voice, PS values should be in the range PS $\in [-0.19; -0.5]$. PS below -0.35 indicates a strained voice, while values above -0.19 indicate a breathy voice. The same article also contains a range of NAQ values. However, in (Alku *et al.*, 2002),

healthy ranges of values for prolonged phonation of the vowel /a/ are additionally distinguished depending on the gender of the speaker. Based on data on the average NAQ value for each voice along with their standard deviations, the range of modal voice for women was set to NAQ $\in [0.14; 0.20]$, and for men to NAQ $\in [0.12; 0.25]$. Exceeding the upper limit of the range indicates a breathy voice, and exceeding the lower limit indicates a strained voice. The CPP parameter, as described in (Szklanny *et al.*, 2019), enables the detection of dysphonia at an early stage of development. In patients without dysphonia, the CPP value was recorded as 11.64 dB, and for patients exhibiting dysphonia, as 11.66 dB. Based on this, a binary assessment was adopted, according to which dysphonia can occur for CPP < 11.6 dB, and otherwise, it does not happen for this study group. To simplify the interpretation of results and make them understandable to a user unfamiliar with voice analysis, six different possible test results were adopted:

– modal voice (marked in green in app):

$$\text{PS} \in [-0.19; -0.35] \wedge \text{NAQ} \in [\text{NAQ\_MIN}; \text{NAQ\_MAX}],$$

– moderate breathy voice (marked in light blue):

$$(\text{PS} > -0.19 \wedge \text{NAQ} \in [\text{NAQ\_MIN}; \text{NAQ\_MAX}]) \vee (\text{PS} \in [-0.19; -0.35] \wedge \text{NAQ} > \text{NAQ\_MAX}),$$

– severe breathy voice (marked in dark blue):

$$\text{PS} > -0.19 \wedge \text{NAQ} > \text{NAQ\_MAX},$$

– moderate strained voice (marked in light red):

$$(\text{PS} < -0.35 \wedge \text{NAQ} \in [\text{NAQ\_MIN}; \text{NAQ\_MAX}]) \vee (\text{PS} \in [-0.19; -0.35] \wedge \text{NAQ} < \text{NAQ\_MIN}),$$

– severe strained voice (marked in dark red):

$$\text{PS} < -0.35 \wedge \text{NAQ} < \text{NAQ\_MIN},$$

– uncertain result (marked in gray):

$$(\text{PS} > -0.19 \wedge \text{NAQ} < \text{NAQ\_MIN}) \vee (\text{PS} < -0.35 \wedge \text{NAQ} > \text{NAQ\_MAX}),$$

where NAQ_MIN is 0.14 for women and 0.12 for men, and NAQ_MAX is 0.20 for women and 0.25 for men. In cases where only one parameter indicates a given voice type, the voice type is classified as a moderate variant. When PS and NAQ indicate it simultaneously, it is treated as a severe form. The result is uncertain if one parameter indicates a strained voice and the other a breathy voice.

All six possible classifications were presented on a 2D board in the voice test result panel in the application (Fig. 5). The shadow cast by the ball representing the phonation indicates its type. The verdict regarding dysphonia depends on CPP and is visualized through the color of the ball and its height relative to the board. The lower the ball is, the higher the CPP value. When the ball is positioned below the board and is red, it indicates



Fig. 5. Voice classification in the VocalCheck application.

that the application has detected the risk of dysphonia in the user's voice. The exact values of all parameters can be read by hovering the cursor over the ball. The backgrounds of the data blocks correspond to the voice conditions indicated by the displayed value.

In addition to the login panel, voice sample panel, and the results panel, VocalCheck includes a panel for viewing the history of voice tests. The data is stored in the form of a .csv file in the application's configuration files on the user's Google Drive and is updated each time the user conducts a test. After entering the Results History tab, the user is redirected to a page displaying a table showing PS, CPP, and NAQ values and the date of the tests (Fig. 6). The colors of the individual cells in the table correspond to the verdicts regarding the voice condition described in Sec. 3. Pressing a row in the table will open a visualization of the result.



FIG. 6. VocalCheck voice sample test history.

The last tab, gender (Fig. 7), allows setting data regarding the user's gender. This is required because different NAQ ranges for the modal voice vary depending on gender. The gender panel is always open when the application cannot read this data from the configuration files on the user's Google Drive.



FIG. 7. VocalCheck gender selection panel.

Graphical elements, such as buttons, tables, and cards containing individual elements for providing voice samples and displaying results, were generated in the MaterializeCSS framework[6]. Its basis is the language and

[6]Data from the repository: https://github.com/dogfalo/materialize.

style of Material Design, an application design language created by Google. The goal of Material Design is to introduce a set of guidelines for designing mobile and web applications, including adding the depth to objects and using animations to create a responsive user interface. The use of MaterializeCSS enabled the preparation of a user interface that adapts to screen dimensions, allowing the application to be displayed on both desktop and mobile screens. The colors in VocalCheck were also selected in accordance with Material Design guidelines[7]. The final color palette was chosen based on usability tests, selecting color with the hexadecimal code #0d47a1, provided by MaterializeCSS under the name 'blue darken-4,' as the primary color.

The wavesurfer.js library, along with the wavesurfer.js/dist/plugins/record.esm.js plugin, was used to display the sound wave visualization, and to draw it during recording. This plugin requires manual modification to avoid automatic lowering of the volume. Prolonged phonation was, by default, detected by the getUserMedia() function from the MediaDevices web API class in JavaScript as noise. An attempt was made to use generative models, such as Microsoft Copilot and ChatGPT in version gpt-4o, to verify whether they can generate code that allows for the calculation of PS, CPP, and NAQ parameters.

VocalCheck has been tested by two experts in human-computer interaction, who also specialize in voice analysis. The test has been conducted by students of PJAIT and members of the author's personal network. In total, 15 individuals took part in usability tests, during which they were tasked to perform three actions:

1. Record a live voice sample and interpret the results.
2. Send a prerecorded sample and interpret the results.
3. Check the evaluation history to find the results from a specific date.

Test participants had to complete the tasks with minimal guidance, and each step they performed was recorded, timed, and further analyzed. After performing the tests, the participants were interviewed and asked for feedback. The tests lead to the following conclusion:

1. Recording a live voice sample took, on average, slightly longer than sending a prerecorded sample: 27.68 s versus 20.71 s:
   - during the interviews, 7 participants preferred recording their voices live, 3 preferred sending a prerecorded sample, and 5 had no strong opinion one way or the other,
   - only 2 participants used the drag-and-drop method to upload files, while the others used the system browser,
   - thirteen participants (those who were not familiar with the voice quality parameters) did not understand what each of the three parameters represents, and 7 asked about the meaning of terms such as dysphonia and modal voice.
2. Checking the evaluation history took an average of 16.26 s:
   - one participant suggested adding the ability to filter the result list by voice classification.

ChatGPT is a large language model developed by OpenAI, trained on a vast corpus of texts to provide accurate information on a wide range of topics. The GPT-4o version, where 'o' stands for 'omni,' is the latest version of this model available at the time of writing. It was released in May 2024 and introduces the ability to transmit combinations of text, audio, images, and recordings as input data to a single model. The Microsoft Copilot tool is powered by three AI components, including an older version of the ChatGPT model – GPT-4. In addition, Copilot utilizes the Prometheus module, a proprietary Microsoft set of techniques for working with machine learning models from OpenAI, designed to provide accurate, up-to-date, and targeted results while enhancing security, as well as the DALL-E 3 model for generating images from text. The ChatGPT-4o, ChatGPT-4, and DALL-E 3 models are based on the transformer architecture described in the paper 'Attention is all you need.'

As part of the experiment, both models were asked to generate scripts in Octave that would allow for the calculation of PS, CPP, and NAQ parameters using queries in the following format: How to calculate X when analyzing a voice recording in Octave? Where X is the full name of one of the three parameters, each query was performed as the first in a separate English chat. In case of errors during script execution, subsequent messages describing the problems were sent to the model until a functional code was generated. Additionally, after obtaining scripts returning values for the given parameters, information was provided that the result was different from

[7]The tool used for this purpose available at: https://material-foundation.github.io/material-theme-builder/.

the results provided by COVAREP, with a request to analyze the scripts from this repository, explain the cause of the differences in results, and modify the generated code so that the parameters were calculated according to the actual implementation. These queries were written according to the following format:

```
The output of the script is different from the COVAREP's implementation of the X calculation
algorithm. Please take a look at the script at: <LINK TO THE SCRIPT CALCULATING THE PARAMETER
IN COVAREP> and analyze what may be the cause of the output differences. If possible, adjust
your code so that it matches COVAREP's calculations.
```

All errors encountered during script execution, including those related to COVAREP, were also addressed in subsequent queries to the model. Changes were made to the generated code regarding the names of input files, functions that display results in the console, etc., to enhance the readability of the results. The values calculated by the scripts before and after considering the COVAREP repository were collected, compared with those calculated by the COVAREP scripts, and compiled together in Table 1. There are situations where the Microsoft tool can generate working code that gives the desired results. This occurs when the machine learning model has encountered the given problem multiple times during the training data processing.

Table 1. Comparison of voice parameter values depending on the source of implementation of counting functions.

	COVAREP	Microsoft Copilot	Microsoft Copilot after taking into account COVAREP	ChatGPT	ChatGPT after taking into account COVAREP
PS	–0.395743	0.0903	0.0062304	0.0002711	0
CPP	7.11601	0.78335	6.3649	0.045087	0.8329
NAQ	0.0946473	–0.077603	–0.30935	0.066899	0.27388

The operation of Microsoft Copilot was tested to generate a Java Swing fragment of code. The obtained code runs a graphical application with a text field and the text 'Hello World!'. As part of the query, specific guidelines were provided regarding the displayed window, including requirements to use a JTextArea component, place the text in an H1 header, bold it, use a red font color, and wrap it. The functional code was obtained after two queries were made. Attempting to place HTML tags in a component of the JTextArea class caused Microsoft Copilot to use the non-existent setEditorKit() method in this class. This method does exist, however, in the JEditorPane class, which the Microsoft system uses after receiving the compilation error. Running the final version of the code displays the window configured as expected (Fig. 8).

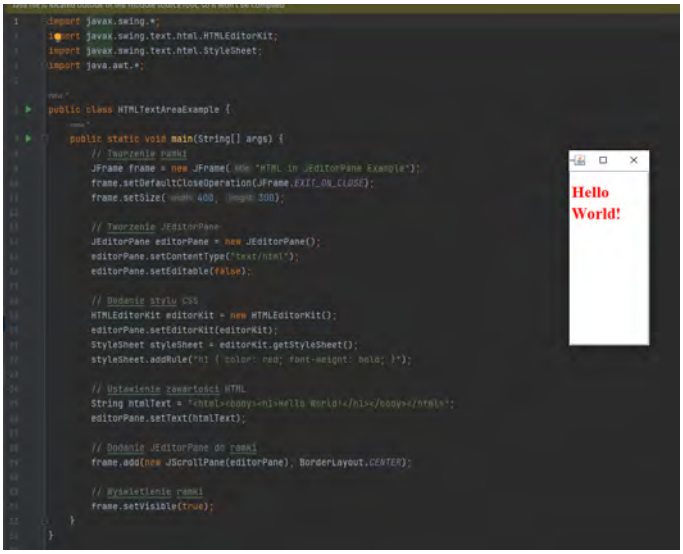


Fig. 8. Output of Java Swing code generated in Microsoft Copilot.

Repeating the experiment with the ChatGPT-4o model gave a different result. In response to the first query, the model correctly identified the inability to render HTML data and printed only the appropriate tags, using the text-wrapping function (Fig. 9). At the same time, the response concluded by explaining why using JTextArea in this case is impossible and suggested replacing it with JEditorPane or JTextPane. The second query, requesting the creation of a version of the program that renders the HTML code, resulted in the functional code using JEditorPane (Fig. 10). Potentially, the greater insightfulness of ChatGPT-4o may stem from it being a newer, more developed model than the ChatGPT-4 used in Microsoft Copilot.

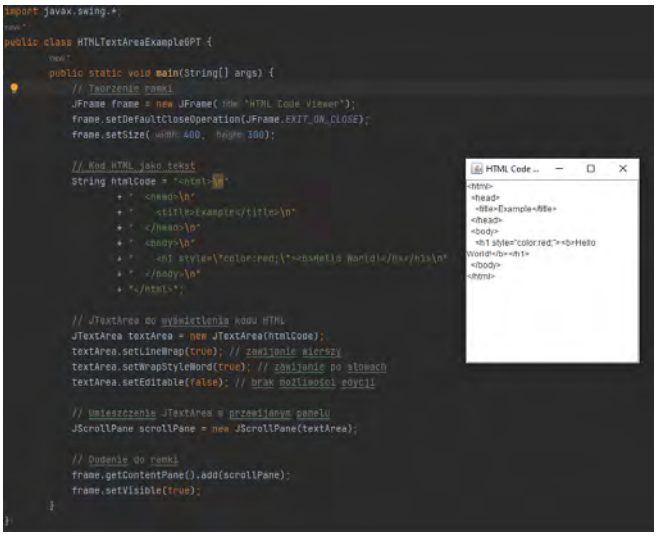


Fig. 9. Code generated by ChatGPT after the first query.



Fig. 10. Code generated by ChatGPT after requesting the HTML rendering version.

5. Discussion

5.1. Using the mobile diagnostics application

The application presented in this study, VocalCheck, was developed as a tool for the preliminary assessment of vocal health. It is intended to support early-stage voice diagnostics by providing accessible, smartphone-based analysis of vocal parameters. Contemporary approaches in computational voice assessment shape its design and aim to serve as a practical aid for singers, vocal coaches, and researchers in the field of voice science.

The underlying approach to automatic diagnostics in VocalCheck draws on several scientific studies that utilize quantitative voice features derived using the COVAREP library. These works provide the empirical basis for identifying indicators of vocal strain or dysfunction and have been instrumental in shaping the diagnostic logic embedded in the application.

A key component of the research involved validating the results produced by VocalCheck through comparative analysis with previously published experimental studies, notably that of (Szklanny, 2019). The outcomes generated by VocalCheck closely mirrored those reported in that study, suggesting that the application's analytical mechanisms provide a comparable level of diagnostic resolution. This fact reinforces VocalCheck's credibility as a practical alternative or supplement to more established laboratory-grade voice assessment tools.

The application underwent empirical testing among members of the PJAIT choir. This group was chosen because of their regular vocal activity, which makes them particularly sensitive subjects for voice monitoring. Notably, the application identified signs of vocal strain among male singers – an observation that led to improved vocal hygiene practices within the ensemble. In contrast, no significant vocal anomalies were detected in the female participants, indicating a generally healthy vocal condition within this subgroup. The tests were conducted using the choir members' personal smartphones, encompassing a variety of device models with varying microphone specifications and audio processing capabilities. Despite these hardware variations, VocalCheck demonstrated a high degree of usability. Its consistent performance across different devices suggests that the application is not only technically sound but also adaptable to real-world conditions, making it a promising tool for non-invasive widely accessible voice screening in diverse user populations.

5.2. Software development using large language models

As part of the project, a fully functional application for assessing voice condition from recordings was created. The latest articles suggest that, under appropriate conditions, mobile phones can be used effectively to obtain recordings from which sufficiently precise calculations of various voice parameters are possible. The application's interface facilitates the interpretation of parameter values, and the recording of test history allows tracking changes in the user's voice over time. VocalCheck takes into account widely used standards for graphic interface design, which positively influences users' feelings when using the application compared to competing solutions.

Ultimately, the scripts generated by the language models do not correctly calculate the PS, CPP, and NAQ parameters. Considering COVAREP in the queries did not have a significant impact on the quality of the results. The reasons for this may be as follows: a high level of complexity in the algorithms calculating the parameters – GPT models are trained comprehensively. Even with the use of fine-tuning methods, they are still susceptible to 'hallucinations,' i.e., situations where the system fabricates information, thereby risking misleading the user. Insufficient query quality – despite using good practices, such as writing specific commands, expanding abbreviations, and guiding the model through a 'chain of thought,' we are uncertain that the generated responses will be accurate. Errors may result, among other things, from the model's inability to verify the transmitted information, overfitting, or bias in the training data.

The second example of using large language model (LLM) in code production shows the potential and development of generative models in this area. Reproducing methods from the COVAREP repository may have been too niche a task for these systems; however, solving more popular problems, such as creating simple graphical interfaces, is susceptible to automation. To achieve greater precision in voice condition evaluation, VocalCheck can be expanded with additional parameters, such as those available in plugins for the PRAAT tool. Another direction for the development of research is the use of LLM for processing functions in the MATLAB language. Generative machine learning models demonstrate promising results in performing simple programming tasks. In addition to ChatGPT and Microsoft Copilot, a potential candidate for further testing is the DeepSeek model from China. Version R1 of this model is publicly available on the GitHub server. Other competitive LLM models include, for example, the Grok AI model, integrated with the X platform (formerly Twitter), and the Claude AI model, created by former OpenAI employees. With more precise construction of queries, there is a chance of successfully generating functions that calculate voice parameters in MATLAB, as is the case with COVAREP.

An attempt can then be made to rewrite them in other programming languages, such as Python or Java. The last COVAREP release dates back to 2016, and since then, this project has not been updated. The high complexity of the functions implemented there makes it difficult to support this tool, and machine learning techniques offer a chance to reproduce them in a more accessible programming environment.

The naming convention for scripts generated by the LLMs is as follows: name of the LLM and the name of the calculated parameter, separated by an underscore. Versions of the scripts generated after the prompt that points to the COVAREP implementations are marked by the text covarep_mentioned at the end of the filename.

The tests conducted using artificial intelligence tools demonstrate that software development can be effectively supported by such technologies, highlighting their potential to automate and accelerate the implementation of computational solutions, particularly for general programming tasks. However, for highly specialized systems with a single established implementation, such as COVAREP, the results remain suboptimal, as AI-generated code often falls short of the precision and performance required for such domain-specific frameworks. It can nevertheless be assumed that, when sufficiently detailed prompts are provided, the generated code could function correctly, as evidenced by operational examples presented in this study. In the case of extensive and intricate projects like COVAREP, the challenge of regenerating computational code that reproduces identical results constitutes a distinct research problem within computer science, encompassing issues such as reproducibility, algorithmic fidelity, and the validation of AI-assisted implementations.

6. Conclusion

The application VocalCheck, presented in this study, has demonstrated its utility as a tool for preliminary voice diagnostics. Designed with accessibility and practical deployment in mind, it offers an effective means of assessing vocal health using widely available mobile devices.

Usability tests conducted as part of the study confirmed the application's ease of use, as evidenced by positive UX (user experience) evaluations. These findings indicate that VocalCheck can be easily used by users without specialized training, making it a practical option for both individual and group voice monitoring contexts.

In addition, exploratory experiments were conducted to evaluate the potential of LLMs for generating diagnostic software code. These tests provided valuable insights into how contemporary AI tools can be used to support the development of applications aimed at human voice assessment.

Unlike many existing mobile applications for vocal analysis, VocalCheck incorporates state-of-the-art methodologies from modern voice diagnostics. Its analytical framework is aligned with current scientific standards, ensuring precision and reliability that distinguish it from simpler or more narrowly focused tools.

Data availability statement

The following supporting information can be downloaded at: https://github.com/s19110/VocalCheck. The materials contain code generated by chatGPT and MS Copilot.

The link to the VocalCheck application: https://vocalcheck.pjwstk.edu.pl.

The application supporting the findings of this study is available from the corresponding author upon reasonable request.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that there are no known competing financial interests or personal relationships that could have influenced the work described in this paper.

Authors' contributions

Hubert Daniszewski – software development, validation, formal analysis, investigation, resources, writing – original draft preparation, visualization. Krzysztof Szklanny – conceptualization, methodology, resources, writing – review and editing, supervision. Piotr Wrzeciono – conceptualization, methodology, resources, writing – review and editing, supervision. All authors reviewed and approved the final manuscript.

References

1. Alhussein M., Muhammad G. (2019), Automatic voice pathology monitoring using parallel deep models for smart healthcare, *IEEE Access*, **7**: 46474–46479, https://doi.org/10.1109/ACCESS.2019.2905597.
2. Alku P., Strik H., Vilkman E. (1997), Parabolic spectral parameter – a new method for quantification of the glottal flow, *Speech Communication*, **22**(1): 67–79, https://doi.org/10.1016/s0167-6393(97)00020-4.
3. Alku P., Bäckström T., Vilkman E. (2002), Normalized amplitude quotient for parametrization of the glottal flow, *The Journal of the Acoustical Society of America*, **112**(2): 701–710, https://doi.org/10.1121/1.1490365.
4. Airas M., Alku P. (2007), Comparison of multiple voice source parameters in different phonation types, [in:] *Proceedings Interspeech 2007*, https://doi.org/10.21437/interspeech.2007-28.
5. Angadi V., Chih M.-Y., Stemple J. (2023), Developing and testing a smartphone application to enhance adherence to voice therapy: a pilot study, *International Journal of Environmental Research and Public Health*, **20**(3): 2436, https://doi.org/10.3390/ijerph20032436.
6. Askenfelt A.G., Hammarberg B. (1986), Speech waveform perturbation analysis: a perceptual-acoustical comparison of seven measures, *Journal of Speech, Language, and Hearing Research*, **29**(1): 50–64, https://doi.org/10.1044/jshr.2901.50.
7. Asci F. *et al.* (2020), Machine-learning analysis of voice samples recorded through smartphones: the combined effect of ageing and gender, *Sensors*, **20**(18): 5022, https://doi.org/10.3390/s20185022.
8. Björkner E., Sundberg J., Alku P. (2005), Subglottal pressure and NAQ variation in voice production of classically trained baritone singers, [in:] *Proceedings Interspeech 2005*, pp. 1057–1060, https://doi.org/10.21437/interspeech.2005-424.
9. Boersma P. (2001), Praat, a system for doing phonetics by computer, *Glot International*, **5**(9/10): 341–345.
10. Childers D.G., Lee C.K. (1991), Vocal quality factors: analysis, synthesis, and perception, *The Journal of the Acoustical Society of America*, **90**(5): 2394–2410, https://doi.org/10.1121/1.402044.
11. Compton E.C. *et al.* (2022), Developing an Artificial Intelligence tool to predict vocal cord pathology in primary care settings, *The Laryngoscope*, **133**(8): 1952–1960, https://doi.org/10.1002/lary.30432.
12. Cooper W.E., Sorensen J.M. (1981), *Fundamental Frequency in Sentence Production*, Springer Science & Business Media.
13. Crowson M.G. *et al.* (2020), A contemporary review of machine learning in otolaryngology–head and neck surgery, *The Laryngoscope*, **130**(1): 45–51, https://doi.org/10.1002/lary.27850.
14. Degottex G., Kane J., Drugman T., Raitio T., Scherer S. (2014), COVAREP – a collaborative voice analysis repository for speech technologies, [in:] *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 960–964, https://doi.org/10.1109/icassp.2014.6853739.
15. Hacki T. (1989), Classification of glottal dysfunctions on the basis of electroglottography [in German: Klassifizierung von glottiscysfunktionen mit hilfe der elektroglottographie], *Folia Phoniatrica*, **41**(1): 43–48, https://doi.org/10.1159/000265931.
16. Hanson H.M. (1997), Glottal characteristics of female speakers: acoustic correlates, *The Journal of the Acoustical Society of America*, **101**(1): 466–481, https://doi.org/10.1121/1.417991.
17. Hendriksz C.J. *et al.* (2014), Burden of disease in patients with Morquio A syndrome: results from an international patient-reported outcomes survey, *Orphanet Journal of Rare Diseases*, **9**(1): 32, https://doi.org/10.1186/1750-1172-9-32.

18. Hillenbrand J., Houde R.A. (1996), Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech, *Journal of Speech, Language, and Hearing Research*, **39**(2): 311–321, https://doi.org/10.1044/jshr.3902.311.
19. Ho T.K. (1995), Random decision forests, [in:] *Proceedings of 3rd International Conference on Document Analysis and Recognition*, **1**: 278–282, https://doi.org/10.1109/ICDAR.1995.598994.
20. Holland J.H. (1975), *Adaptation in Natural and Artificial Systems. An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, The MIT Press.
21. Howard D.M. (1995), Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers, *Journal of Voice*, **9**(2): 163–172, https://doi.org/10.1016/s0892-1997(05)80250-4.
22. Kane J., Gobl C. (2011), Identifying regions of nonmodal phonation using features of the wavelet transform, [in:] *Proceedings Interspeech 2011*, pp. 177–180, https://doi.org/10.21437/interspeech.2011-76.
23. Kane J., Gobl C. (2013), Wavelet maxima dispersion for breathy to tense voice discrimination, [in:] *IEEE Transactions on Audio, Speech, and Language Processing*, **21**(6): 1170–1179, https://doi.org/10.1109/tasl.2013.2245653.
24. Kasuya H., Shigeki O., Mashima K., Ebihara S. (1986), Normalized noise energy as an acoustic measure to evaluate pathologic voice, *The Journal of the Acoustical Society of America*, **80**(5): 1329–1334, https://doi.org/10.1121/1.394384.
25. Kaur P., Stoltzfus, J.C. (2017), Bland–Altman plot: a brief overview, *International Journal of Academic Medicine*, **3**(1): 110–111, https://doi.org/10.4103/IJAM.IJAM_54_17.
26. Klingholz R. (1987), The measurement of the signal-to-noise ratio (SNR) in continuous speech, *Speech Communication*, **6**(1): 15–26, https://doi.org/10.1016/0167-6393(87)90066-5.
27. Kosztyła-Hojna B., Moskal D., Kuryliszyn-Moskal A., Rutkowski R. (2014), Visual assessment of voice disorders in patients with occupational dysphonia, *Annals of Agricultural and Environmental Medicine*, **21**(4): 898–902, https://doi.org/10.5604/12321966.1129955.
28. Lebacq J., Schoentgen J., Cantarella G., Bruss F.T., Manfredi C., DeJonckere P. (2017), Maximal ambient noise levels and type of voice material required for valid use of smartphones in clinical voice research, *Journal of Voice*, **31**(5): 550–556, https://doi.org/10.1016/j.jvoice.2017.02.017.
29. Manfredi C. *et al.* (2017), Smartphones offer new opportunities in clinical voice research, *Journal of Voice*, **31**(1): 111, https://doi.org/10.1016/j.jvoice.2015.12.020.
30. Nawka T., Anders L., Wendler J. (1994), The auditory assessment of hoarse voices according to the RBH system [in German], *Sprache Stimme, Gehör*, **18**: 130–133.
31. Parsa V., Jamieson D.G. (2001), Acoustic discrimination of pathological voice: sustained vowels versus continuous speech, *Journal of Speech, Language, and Hearing Research*, **44**(2): 327–339, https://doi.org/10.1044/1092-4388(2001/027).
32. Patel R.R. *et al.* (2018), Recommended protocols for instrumental assessment of voice: American Speech Language-Hearing Association expert panel to develop a protocol for instrumental assessment of vocal function, *American Journal of Speech-Language Pathology*, **27**(3): 887–905, https://doi.org/10.1044/2018_ajslp-17-0009.
33. Suvvari T.K. (2023), The role of Artificial Intelligence in diagnosis and management of laryngeal disorders, *Ear, Nose & Throat Journal*, **104**(11), https://doi.org/10.1177/01455613231175053.
34. Szklanny K. (2019), Acoustic parameters in the evaluation of voice quality of choral singers. Prototype of mobile application for voice quality evaluation, *Archives of Acoustics*, **44**(3): 439–446, https://doi.org/10.24425/aoa.2019.129257.
35. Szklanny K., Gubrynowicz R., Iwanicka-Pronicka K., Tylki-Szymańska A. (2016), Analysis of voice quality in patients with late-onset Pompe disease, *Orphanet Journal of Rare Diseases*, **11**(1): 99, https://doi.org/10.1186/s13023-016-0480-5.
36. Szklanny K., Gubrynowicz R., Ratyńska J., Chojnacka-Wądołowska D. (2019), Electroglottographic and acoustic analysis of voice in children with vocal nodules, *International Journal of Pediatric Otorhinolaryngology*, **122**: 82–88, https://doi.org/10.1016/j.ijporl.2019.03.030.
37. Szklanny K., Gubrynowicz R., Tylki-Szymańska A. (2018), Voice alterations in patients with Morquio A syndrome, *Journal of Applied Genetics*, **59**(1): 73–80, https://doi.org/10.1007/s13353-017-0421-6.

38. Szklanny K., Tylki-Szymańska A. (2018), Follow-up analysis of voice quality in patients with late-onset Pompe disease, *Orphanet Journal of Rare Diseases*, **13**(1): 189, https://doi.org/10.1186/s13023-018-0932-1.

39. Szklanny K., Wrzeciono P. (2019a), Relation of RBH auditory-perceptual scale to acoustic and electroglottographic voice analysis in children with vocal nodules, *IEEE Access*, **7**: 41647–41658, https://doi.org/10.1109/ACCESS.2019.2907397.

40. Szklanny K., Wrzeciono P. (2019b), The application of a genetic algorithm in the noninvasive assessment of vocal nodules in children, *IEEE Access*, **7**: 44966–44976, https://doi.org/10.1109/ACCESS.2019.2908313.

41. Tirronen S., Javanmardi F., Kodali M., Reddy Kadiri S., Alku P. (2023), Utilizing Wav2Vec in database-independent voice disorder detection, [in:] *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, https://doi.org/10.1109/ICASSP49357.2023.10094798.

42. Uloza V. *et al.* (2015), Exploring the feasibility of smart phone microphone for measurement of acoustic voice parameters and voice pathology screening, *European Archives of Oto-rhino-laryngology*, **272**(11): 3391–3399, https://doi.org/10.1007/s00405-015-3708-4.

43. Uloza V. *et al.* (2023), Accuracy of acoustic voice quality index captured with a smartphone – measurements with added ambient noise, *Journal of Voice*, **37**(3): 465, https://doi.org/10.1016/j.jvoice.2021.01.025.

44. Verde L., De Petro G., Alrashoud M., Ghoneim A., Al-Mutib K.N., Sannino G. (2019), Leveraging Artificial Intelligence to improve voice disorder identification through the use of a reliable mobile app, *IEEE Access*, **7**: 124048–124054, https://doi.org/10.1109/ACCESS.2019.2938265.

45. Verikas A., Gelzinis A., Bacauskiene M., Uloza V. (2006), Towards a computer-aided diagnosis system for vocal cord diseases, *Artificial Intelligence in Medicine*, **36**(1): 71–84, https://doi.org/10.1016/j.artmed.2004.11.001.

46. Yumoto E., Gould W.J., Baer T. (1982), Harmonics-to-noise ratio as an index of the degree of hoarseness, *The Journal of the Acoustical Society of America*, **71**(6): 1544, https://doi.org/10.1121/1.387808.

47. Zheng F., Zhang G., Song Z. (2001), Comparison of different implementations of MFCC, *Journal of Computer science and Technology*, **16**(6): 582–589, https://doi.org/10.1007/BF02943243.

Archives of Acoustics Vol. 51, No. 3, pp. 601–616 (2026), https://doi.org/10.24423/archacoust.2026.4434

Technical Note

From Speech to Underwater Acoustics: A Transfer Learning Framework for Real-Time Passive Diver Detection Using Keyword Spotting Models

Osama Deeb[(1)], Saier Mahmoud[(2)*], Louay Saleh[(2)], Assef Jafar[(3)], Oumayma Al Dakkak[(1)], Ibrahim Chouaib[(2)]

[(1)] *Department of Telecommunications*
[(2)] *Department of Electronic and Mechanical Systems*
[(3)] *Department of Informatics*

Higher Institute for Applied Sciences and Technology
Damascus, Syria

*Corresponding Author: saier.mahmoud@gmail.com

Received February 9, 2026; revised April 9, 2026; accepted April 21, 2026; available online May 11, 2026; version of record July 22, 2026; published issue September 23, 2026.

Passive acoustic detection of divers faces challenges such as low signal-to-noise ratios (SNRs), data scarcity, and latency in conventional methods. This paper proposes keyword spotting for diver detection (KWS-DD) – a transfer learning framework that repurposes speech-oriented KWS models for data-efficient diver detection. Diver inhalation signatures are treated as acoustic 'keywords,' enabling adaptation of the transformer-based HuBERT architecture (pre-trained on speech) to identify quasi-periodic respiratory events in underwater audio. The core innovation of this work lies in adapting the state-of-the-art speech model HuBERT for accurate diver detection via non-speech inhalation acoustics. This approach eliminates the need for accumulation during the respiratory cycle, enabling real-time detection using a minimal amount of domain-specific data (120 inhalation samples). The solution, deployed in diverse marine conditions, achieved 94.4 % accuracy and 94.6 % $F1$-score for inhalation sounds. This represents a more than 50 % range extension over conventional methods, which proved unreliable at distances greater than 10 meters in low-SNR environments. The framework reduces false alarms caused by boat noise and generalizes to external datasets, validating cross-domain transferability. This work bridges AI-based speech processing and passive sonar signal processing, offering a resource-efficient solution for real-time underwater surveillance.

Keywords: passive diver detection, keyword spotting, transfer learning, real-time detection, passive sonar, hydrophone.



Copyright © 2026 The Author(s).
This work is licensed under the Creative Commons Attribution 4.0 International CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The acoustic signal generated by a scuba diver breathing through a regulator serves as the basis for passive sonar detection. This signal exhibits a broadband frequency spectrum (Johansson *et al.*, 2010), ranging from several hundreds of hertz to 75 kHz (Tu *et al.*, 2020), and characterized by quasi-periodicity (Gorovoy *et al.*, 2015). The repetition interval varies with an activity level, ranging from approximately 7.09 s at rest to 2.44 s during heavy flapping (Donskoy *et al.*, 2008), directly reflecting respiratory rates from 0.14 Hz to 0.41 Hz. These rates and the associated emitted power vary based on the diver's experience, exertion level, and equipment (Donskoy *et al.*, 2008). The respiratory cycle acoustically bifurcates: inhalation manifests predominantly above 2 kHz, whereas bubble-generating exhalation manifests predominantly below 2 kHz (Sun *et al.*, 2022). Both phases are useful for detection (Sun *et al.*, 2022). However, inhalation signals are often considered preferable due to their pulse characteristics (Tu *et al.*, 2020).

Detectability depends on both the signal power within relevant frequency bands and the repetition rate (Stolkin *et al.*, 2006). Research focus diverges between the high-frequency band (>2 kHz) (Tu *et al.*, 2020; Jin, Xu, 2020) and the low-frequency band (<2 kHz) (Mahmoud *et al.*, 2025; Radford *et al.*, 2005; Hari *et al.*, 2015), with notable challenges in detecting a closed-circuit scuba (rebreather) which emits significantly lower acoustic energy than an open-circuit scuba (Gorovoy *et al.*, 2015; Radford *et al.*, 2005).

Two primary methodologies are employed for detecting diver-generated acoustic signals: traditional signal processing techniques and artificial intelligence AI approaches. The first method follows a sequential processing chain. Initial band-pass filtering enhances the signal-to-noise ratio (SNR) by focusing on frequency bands relevant to the diver's spectral signature, which varies with the diving apparatus (Chung *et al.*, 2007; Hari *et al.*, 2015). This results in the use of different bands, such as 25 kHz to 75 kHz (Hari *et al.*, 2015), 200 Hz to 500 Hz (Gorovoy *et al.*, 2014), or 13 kHz to 18 kHz (Tu *et al.*, 2020). Subsequently, periodicity detection leverages energy estimation (Gorovoy *et al.*, 2014), envelope detection (Tu *et al.*, 2020), or matched filtering (Gorovoy *et al.*, 2014; Chen *et al.*, 2006). Finally, the fast Fourier transform (FFT) analysis quantifies power around the breathing rate and a threshold-based trigger is launched when results exceed a predetermined level (Stolkin *et al.*, 2006; Lennartsson *et al.*, 2009). These approaches have been augmented through adaptive noise subtraction for range extension (20 m to 40 m) (Tu *et al.*, 2020), background noise whitening to suppress transient interference (Johansson *et al.*, 2010), and complementary cross-correlation techniques applied to dual-hydrophone signals for enhanced detection ranges (Chung *et al.*, 2007; Korenbaum *et al.*, 2020).

Traditional signal processing techniques have limited capabilities when it comes to detecting divers and, more broadly, detecting underwater objects, particularly in detection accuracy and latency. These limitations have motivated the adoption of artificial intelligence approaches (the second method), with machine learning (ML) and deep learning emerging as promising solutions (Feng *et al.*, 2024). Like most AI-based solutions, the ML pipeline for this application typically comprises four critical stages as shown in Fig. 1:

1. Data acquisition, where generative adversarial networks (GANs) are used for both synthetic data generation and augmentation (Fuchs *et al.*, 2019; Jin *et al.*, 2020; Karjalainen *et al.*, 2019; Phung *et al.*, 2019; Zhao *et al.*, 2023).
2. Preprocessing, where autoencoder architectures have been successfully applied to critical preprocessing tasks including noise reduction, resampling, and signal enhancement (Dong *et al.*, 2022; Huang *et al.*, 2020; Vincent *et al.*, 2010; Wang *et al.*, 2014).
3. Feature extraction, where learned features derived through autoencoders and self-supervised learning techniques demonstrate the ability to automatically extract high-level representations by exploiting inherent data structures (Chen, Shang, 2019; Wang *et al.*, 2022).
4. Classification which has evolved through successive generations of ML approaches including but not limited to support vector machines (SVMs) (Zhang *et al.*, 2003), K-nearest neighbors (KNN) (Alvaro *et al.*, 2021), recurrent neural networks (RNNs), convolutional neural networks (CNNs), attention-based mechanisms and transformers (Hewamalage *et al.*, 2021; Wang *et al.*, 2021).

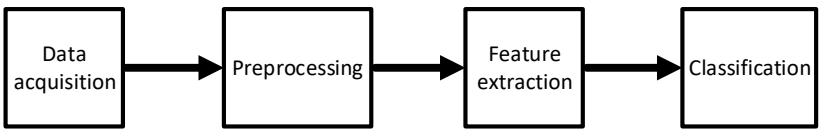


Fig. 1. General ML-based pipeline for underwater acoustic detection task.

Among all of these types, transformers, with their multi-head self-attention (MHSA) mechanisms, have shown particular promise despite computational challenges, prompting the development of specialized training protocols incorporating transfer learning and advanced data augmentation techniques to mitigate these constraints (Gong *et al.*, 2022; Li *et al.*, 2022).

Diver detection constitutes a specialized application of underwater acoustic signal processing, distinguished by the unique temporal and spectral characteristics of respiratory signatures. Of particular diagnostic value are

inhalation patterns, which exhibit distinctive pulsed waveforms (Tu *et al.*, 2020). SVM classifier, employing energy characteristics across eight distinct non-overlapping sub-bands (49 kHz to 51 kHz) as discriminative features, demonstrated superior capability in processing irregular respiratory patterns that challenge conventional matched filter approaches (Zhao *et al.*, 2016). However, it should be noted that these promising results were exclusively validated in controlled pool environments, potentially limiting generalizability to real-world settings. Subsequent research demonstrates the superiority of frequency-domain multi-sub-band energy (FMSE) features over both Mel frequency cepstral coefficient filtering (MFCC) and gamma tone frequency cepstral coefficient filtering (GFCC) within the 2 kHz to 8 kHz band when processed through SVM classifiers (Sun *et al.*, 2024), with validation extended to diverse environments including pools, lakes, and shallow sea. While these approaches operate on 1D acoustic data, CNNs that process 2D spectrogram, demonstrate robust detection capabilities in marine environments, achieving 93 % correct classification rates at ranges up to 12 m (Cole, 2019).

The aforementioned methods (both traditional and AI-based) rely heavily on respiratory periodicity – a feature consistently leveraged in existing literature. This dependence introduces inherent latency of up to 10 s due to the requisite observation window for capturing multiple respiratory cycles (Jin, Xu, 2020) which is considered a long time for some applications especially those related to protecting strategic maritime facilities (military and civilian) such as aircraft carriers and oil platforms against terrorist attacks. While approaches using energy exclusively within dominant diver frequencies eliminate this delay, they exhibit proportionally heightened vulnerability to noise interference, compromising detection reliability (Stolkin *et al.*, 2006). Jin *et al.* (2020) addressed this limitation through a deep learning framework that uses time-frequency features and treats the diver's signal as a keyword; however, their framework relies on a synthesized and augmented dataset (Jin, Xu, 2020).

Recent breakthroughs in speech processing using deep learning present significant opportunities for cross-domain applications. Keyword spotting (KWS) systems exemplify this potential, achieving >99 % accuracy in detecting target keywords within continuous speech. This success suggests promising applicability to diver detection by treating characteristic acoustic signatures (e.g., inhalation/exhalation patterns) as analogous to spoken keywords. Extending the KWS architecture of (Deeb *et al.*, 2025), which leverages a pretrained hidden unit bidirectional encoder representations from the transformer (HuBERT) model to achieve 99.7 % classification accuracy with only 15 samples per keyword, this work adapts the framework for diver detection via respiratory acoustic pattern recognition. Although HuBERT was originally trained on speech data, its transformer-based architecture, employing multi-head self-attention, exhibits robust representation learning capabilities. Through fine-tuning, the model can discern fundamental sound units within divers' non-speech vocal signals and align them with latent representations acquired during pretraining. This adaptation facilitates effective detection and discrimination of acoustic patterns in diver respiratory signals. This extension delivers three principal advantages:

1. Enhanced detection accuracy through physiologically distinctive acoustic signatures.
2. Real-time processing capability eliminating detection latency.
3. Data-efficient through employing transfer learning from the speech domain that makes it suitable for resource-constrained scenarios.

Collectively, this approach directly addresses the critical challenge of limited training data availability inherent in diver detection applications.

Beyond this introductory section, the paper is divided into eight parts. Section 2 details the characteristics of acoustic signals produced by divers. Section 3 describes the experimental setup, including location, conditions, and instrumentation used for data acquisition. Section 4 reviews the traditional signal processing algorithm employed for diver detection using acoustic signatures. Section 5 introduces the novel transfer learning algorithm designed for keyword spotting, and its adaptation for detecting divers based on inhalation signatures. Section 6 presents the experimental tests performed, providing a comparative analysis of the traditional method versus the proposed AI-based approach. Section 7 analyzes the obtained experimental results and examines the shortcomings of the proposed algorithm. Section 8 concludes the paper, summarizing the system's principal advantages and outlining potential avenues for future enhancement.

2. Acoustic signals emitted by diver

Diver breathing generates two distinct acoustic signatures: an inhalation signal characterized by energy predominantly above 2 kHz, and an exhalation signal concentrated below 2 kHz. Each respiratory cycle exhibits a duration of approximately 3 s, as illustrated in Fig. 2. This figure presents the time-domain waveform and the corresponding spectrogram of acoustic data recorded from a diver swimming within a ~3-meter radius of the hydrophone used in this research. The spectrogram was computed using a 1024-sample Hamming window, 50 % overlap, and 1024-point FFT.

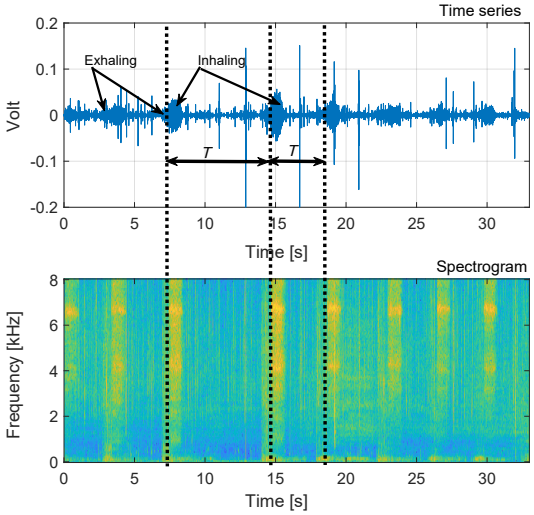


Fig. 2. Time-domain waveform and spectrogram of acoustic emissions from a diver breathing while swimming within a 3-meter radius of the hydrophone. The spectrogram was computed using a 1024-sample Hamming window, 50 % overlap, and 1024-point FFT.

An analysis of Fig. 2 reveals that inhalation manifests as a series of quasi-periodic transient events, visible as vertical striations in the spectrogram, though the periodicity is not strictly constant. The exhalation signal is less prominent in the time-domain waveform but appears in the spectrogram as low-frequency energy band preceding each inhalation event.

As the diver's distance from the hydrophone increases, the signal attenuation increases significantly, particularly for the higher-frequency inhalation components due to greater absorption losses in the water column (Tu *et al.*, 2020). Consequently, robust detection of the diver presence using conventional signal processing techniques presents significant challenges. Figure 3 illustrates this difficulty, showing the time-domain waveform and

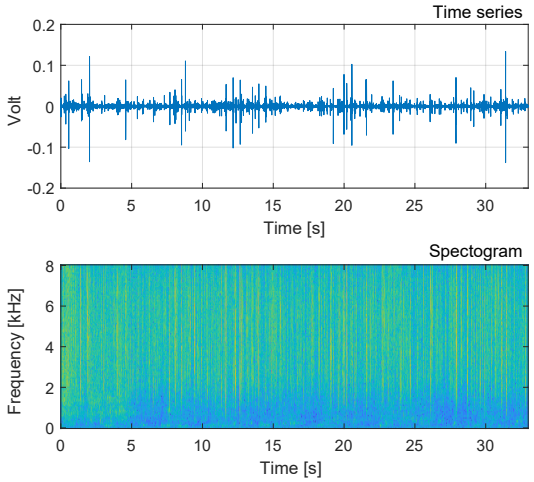


Fig. 3. Time-domain waveform and spectrogram of diver acoustic emissions at extended range (10 m). Spectrogram computed using a 96 kHz, 1024-sample Hamming window, 50 % overlap, and 1024-point FFT.

the corresponding spectrogram of acoustic data recorded from a diver swimming within a ~10-meter radius of the hydrophone. The spectrogram was computed using a 1024-sample Hamming window, 50 % overlap, and 1024-point FFT. Critically, the effect of inhalation is not clearly discernible either in the time-domain signal or in the spectrogram representation.

3. Data acquisition

Acoustic signatures were acquired during three collection rounds (August 2–3 and October 2, 2024) using two omnidirectional hydrophones mounted 70 cm apart on a rigid stationary platform positioned 60 cm above the seabed. An acoustic vector sensor (AVS) was co-located at the platform center but remained unused in this study, as shown in Fig. 4. Critically, each hydrophone channel was processed in strict isolation with zero cross-sensor techniques applied – including beamforming, time-difference-of-arrival estimation, coherence analysis, or other array-based methods. This dual-hydrophone configuration provided exclusively time-synchronized redundant measurements.



Fig. 4. Experimental setup showing diver, sensor platform (hydrophones + AVS), and support boat.

Experiments were conducted in a large seawater basin (350 m × 250 m × 5 m depth), where two divers, equipped with an open-circuit SCUBA, sequentially executed controlled maneuvers: circular traversals at 1 m to 30 m radial distances, radial approaches/moves away, and swimming-to-stationary state transitions. Activities spanned azimuthal bearings from 0° to 360° relative to the sensor platform to capture representative single-channel signatures across diverse source–receiver geometries and motion dynamics. Environmental conditions were deliberately varied: days 1–2 (August) featured calm, sunny conditions, while the third day (of October) was marked by sustained winds that generated surface wave disturbance. This deliberate diversity in diver activities, motion profiles, and environmental conditions yields a rich source for building a diverse dataset that closely mirrors real-world operational variability, making it particularly valuable for training and validating robust AI-based detection models. All data was recorded at 44.1 kHz with 24-bit resolution per channel.

To rigorously evaluate both traditional diver detection methods and the proposed transfer learning-based model, a diverse set of acoustic signals was curated. These signals encompass varying environmental conditions, SNRs, motion patterns, and interference sources. The test suite includes both locally recorded data and externally sourced signals to assess robustness and generalization capability. Table 1 describes these signals.

These signals were reserved for testing purposes and never used for model training.

Table 1. Evaluation signal specifications.

ID	Description
Diver_1	Diver moves within 3 m (high SNR (~10 dB))
Diver_2	Stationary diver at 10 m (low SNR <−8 dB)
Diver_3	Moving diver: starts at 25 m, approaches up to 15 m at $t = 22$ s, stays in place for 25 s, moves away at $t = 47$ s
Noise	Ambient noise (no divers/boats)
Boat	Moving boat (0 m to 150 m) approaching and departing
Diver_4	Externally sourced diver signal (Davi-sh, 2022)
Sonar	Externally sourced sonar signal (VoiceBosch, 2024)

4. Conventional diver detection algorithm

The conventional passive diver detection algorithm follows the processing chain illustrated in Fig. 5, consisting of these sequential stages (Mahmoud *et al.*, 2025):

1. Acoustic signal acquisition: raw hydrophone signals are captured for processing.
2. Bandpass filtering: signals are filtered within the dominant breathing frequency band to enhance the SNR by attenuating out-of-band 2 kHz to 7 kHz.
3. Periodicity enhancement: an envelope detector highlights breathing periodicity by a computing signal power within 100-ms sliding windows with 50 % overlap.
4. Spectral analysis: the FFT is applied to the envelope signal.
5. Diver index calculation: the detection energy is quantified by integrating the FFT power within the 0.14 Hz to 0.42 Hz band, corresponding to expected breathing rates.
6. Detection decision: the computed diver index is compared against a predefined threshold to determine the diver presence.

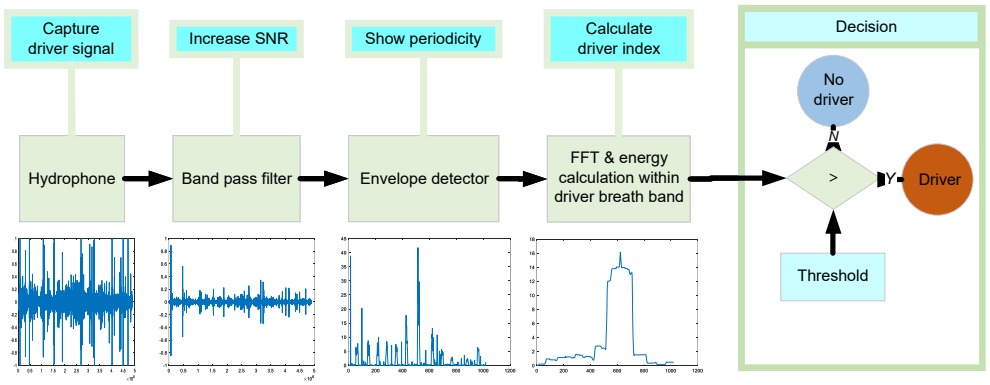


Fig. 5. Diver detection algorithm.

A critical implementation challenge involves the non-stationary nature of ambient noise, which necessitates adaptive threshold determination. This may be achieved through either:

- a reference hydrophone at a suitable distance from the primary sensor, or
- real-time noise power estimation algorithms applied to the acquired signal.

Figure 6 demonstrates the performance of the algorithm when applied to the signals described in Sec. 3 (Table 1). The analysis uses a 10-second window, with power estimation windows (100-ms duration, advanced in 50-ms increments). Key observations include:

- Diver_1: the diver index significantly exceeds noise levels more than 100 times, enabling reliable detection.
- Diver_2: the reduced diver index amplitude degrades detection reliability.
- Boat: the diver index for boat noise reaches observed levels for distant divers, highlighting susceptibility to false alarms from high-energy sources.

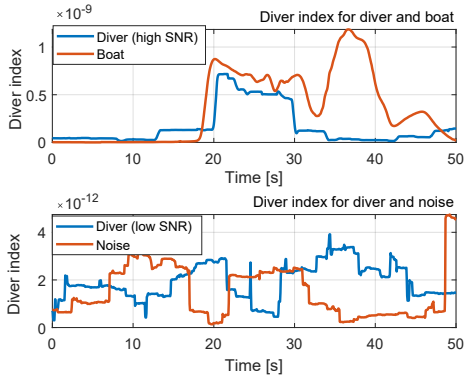


Fig. 6. Diver index estimation for a close-proximity diver (<3 m), distant diver (>10 m), boat noise, and ambient noise.

5. AI-based diver detection

This section introduces the proposed method that leverages transfer learning to adopt an AI-based KWS model to solve the diver detection task.

5.1. Adapting keyword spotting for diver detection

The primary challenge in applying KWS systems to diver detection lies in the substantial training data requirements. Conventional KWS frameworks typically demand approximately 4000 positive and negative samples per keyword to achieve sufficient accuracy (Lin *et al.*, 2020). This requirement proves particularly problematic for diver detection given the inherent difficulties in acquiring sufficient sonar acoustic data, which has limited the adoption of advanced ML architectures such as transformers. While data augmentation and synthesis techniques offer partial solutions, transfer learning presents a more fundamental approach by enabling knowledge transfer between domains. This paradigm has demonstrated particular effectiveness in KWS applications, where models pre-trained on English speech could be adapted to other languages with minimal target-language samples through fine-tuning (Seo *et al.*, 2021). Building on this foundation, a transfer learning framework is proposed, where a general speech representation model – using transformers in its architecture – is fine-tuned to detect diver respiratory sounds (particularly inhalation patterns), significantly reducing the required training data while maintaining the performance advantages of transformer architectures, as demonstrated in underwater object detection tasks (Feng *et al.*, 2024; Domingos *et al.*, 2022).

5.2. Model architecture

The proposed framework adapts and modifies a previously established KWS architecture (Deeb *et al.*, 2025) for the diver detection downstream task through fine-tuning on underwater acoustic recordings of diver sounds. As illustrated in Fig. 7, the architecture comprises two fundamental components: (1) a HuBERT encoder block (Hsu *et al.*, 2021) and (2) a classifier block (this architecture is denoted as KWS-DD). The HuBERT block employs a multi-head self-attention mechanism and has demonstrated state-of-the-art performance in speech processing since its introduction in 2021. Pretrained using masked prediction self-supervised learning, this component generates discriminative contextual representations of input signals that effectively separate distinct acoustic patterns in the embedding space.

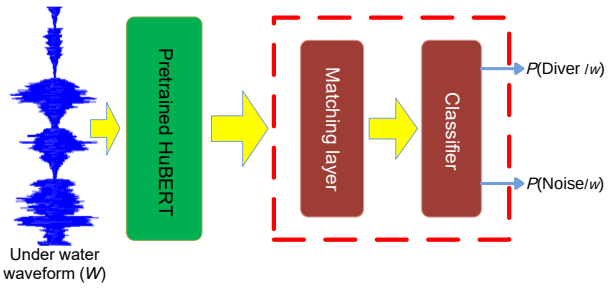


Fig. 7. KWS-based diver detection model.

The classifier block subsequently learns to map these representations to predefined target classes during fine-tuning. Notably, the HuBERT architecture processes raw waveform inputs without requiring manual feature engineering, as its integrated feature extractor – comprising 7-layer convolutional network – automatically extracts optimal acoustic features. This end-to-end design preserves signal integrity while allowing feature extraction parameters to be jointly optimized during training, yielding superior representational capacity compared to conventional preprocessing pipelines.

Upon receiving a raw underwater waveform lasting approximately one second, the pretrained HuBERT model segments it into 20 ms frames, each frame is represented as a (1×1024)-dimensional tensor at the output of HuBERT. These tensors are fed to the matching layer which computes the average of the tensor across the time,

resulting in a condensed (1×1024)-tensor for the whole waveform, which is subsequently fed into the classifier. The classifier is designed to categorize the output tensors into noise or diver classes. A softmax layer at the final stage of the classifier generates the posterior probabilities $P(c_i/w) : i \in [\text{Diver, Noise}]$, which quantify the likelihood of observing the class c_i at the output of the model when a tensor representing the input word w is applied to its input. The final classification decision is made based on these posteriors following the formula provided in Eq. (1):

$$\text{ChosenClass} = \underset{c_i}{\text{Argmax}} \left(P(c_i/w) \right). \tag{1}$$

The framework employs HuBERT-Large, a pretrained version of HuBERT with 317 million parameters, to extract contextual embeddings from raw acoustic waveforms, followed by the KWS head. The final diver detection model contains (317.02 million parameter), it was trained by fine-tuning the whole model's parameters to adapt the system for diver detection. The fine-tuning process was performed using a specialized dataset containing: (1) positive samples of authentic diver-generated acoustic signatures (inhaling signals), and (2) negative samples encompassing various underwater acoustic interference sources. This approach enables effective transfer learning while maintaining the model's ability to discriminate the subtle acoustic patterns characteristic of a diver presence.

5.3. Upgrading model for continuous waveform

In operational scenarios, the exact timing of diver inhalation events is inherently unpredictable, necessitating continuous analysis of the hydrophone's time-domain acoustic signal. To reconcile this streaming data requirement with the model's fixed-length input specification (1.0 ±0.2 s segments), a sliding window algorithm with 100 ms temporal increments was implemented (this temporal resolution was carefully selected to optimize the trade-off between computational efficiency and signal capture reliability, ensuring complete coverage of critical acoustic features while enabling a near-real-time operation). Each 1-second window undergoes probabilistic classification, generating posterior estimates $P(c_i/w) : i \in [\text{Diver, Noise}]$ through the trained model. As the window advances across the continuous waveform, this process produces discrete-time probability functions (Eq. (2)) that quantitatively characterize the evolving likelihood of diver presence versus ambient noise (Fig. 8), enabling real-time detection without compromising accuracy;

$$f_{c_i}(k) = P(c_i/w_k) \quad @k^{\text{th}}\text{window}. \tag{2}$$

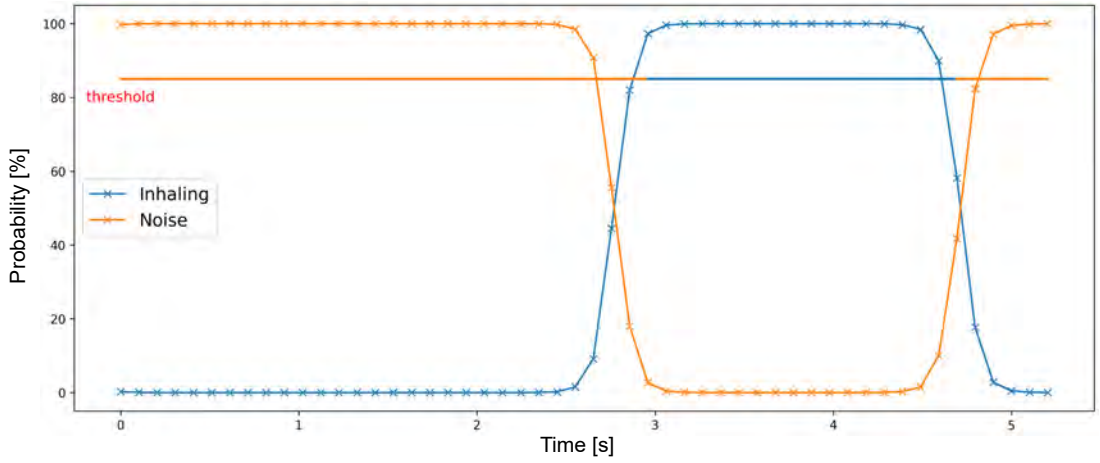


Fig. 8. Applying the sliding window approach on a waveform containing a single inhaling signal, the blue curve indicates the discrete probability function of diver's inhaling signal, the orange curve represents the discrete probability function of the noise.

The discrete probability functions enable robust diver detection through temporal analysis of the posterior probability evolution. A detection event is triggered when $P(\text{Diver}|w)$ exceeds a predefined threshold θ for a min-

imum duration Δt, ensuring protection against transient false positives. While not strictly required, observed periodicity in detection events (typically 0.14 Hz to 0.42 Hz for human respiration) provides secondary validation of a diver presence. In the following, the continuous-processing variant of KWS-DD is denoted as CKWS-DD.

6. Experiments and results

6.1. Data preparation

The most distinctive acoustic signatures associated with divers and their equipment are generated by respiratory activity, particularly inhalation and exhalation processes. An air bubble movement during exhalation produces unique signatures that aid detection. However, since similar acoustic patterns may be produced by marine creatures or equipment, the inhalation signal was selected as the primary discriminative feature due to its high specificity to human divers.

For model fine-tuning, representative acoustic samples were required for each target class (inhalation signatures and ambient noise). Sample extraction involved segmentation of hydrophone recordings with different SNRs into 1-second intervals, corresponding to the mean inhalation duration (1.0 ±0.2 s). Two qualified experts independently performed the segmentation through aural and visual inspection of spectrographic representations. Inter-rater reliability was enforced, with only concordantly verified samples retained for the final dataset. The curated collection ultimately comprised four distinct marine acoustic categories relevant for coastal surveillance:

1. Inhalation signatures: isolated inhalation events.
2. Exhalation signatures: isolated exhalation events.
3. Vessel signatures: propeller cavitation and engine harmonics.
4. Ambient noise: site-specific hydrodynamic and biological noise.

While the present study employs binary classification, this categorical taxonomy enables future transition to multiclass detection frameworks. For the current objective of diver detection, samples were aggregated into two classes: the diver class encompasses verified inhalation events and the noise class encompassing composite non-target acoustics (exhalation, vessels, ambient noise).

The acquired acoustic samples underwent standard preprocessing, including down sampling to 16 kHz to ensure compatibility with the HuBERT architecture's input specifications. The curated dataset consisted of 120 validated inhalation exemplars (the diver class) and 344 interference samples (the noise class). The dataset was partitioned into three subsets while preserving class distributions: a training set (70 %), validation set (15 %), and test set (15 %). The dataset's limited scale, particularly the test portion containing fewer than 70 samples, restricts its ability to provide a comprehensive evaluation of model performance. While this constrained benchmark may not fully represent the model's generalization capabilities, it serves as an initial performance indicator. Instead, the definitive evaluation utilizes extended, continuous audio samples incorporating diverse acoustic conditions, as presented in the experiments section.

6.2. Implementation and training details

The proposed architecture was implemented in Python using PyTorch, incorporating the pre-trained Hubert-large model (facebook/hubert-large-ls960-ft) as described in Sec. 5. During fine-tuning, the 7-layer convolutional feature extractor parameters remained frozen to preserve learned acoustic representations. The model optimization was performed on a workstation equipped with the Intel Core i7-11370H processor and NVIDIA GTX 1650 GPU (16GB RAM) using the training partition of the dataset. Training ran for 30 epochs, converging in ~110 min. Figure 9 illustrates the training dynamics through the evolution of both training and validation loss curves throughout the optimization process.

The trained model underwent comprehensive performance assessment using the reserved test dataset subset. Quantitative evaluation employed four standard classification metrics: accuracy, precision, recall, and $F1$-score, providing complementary perspectives on detection capability. These metrics collectively characterize: (1) overall prediction correctness (accuracy), (2) positive prediction reliability (precision), (3) target detection sensitivity

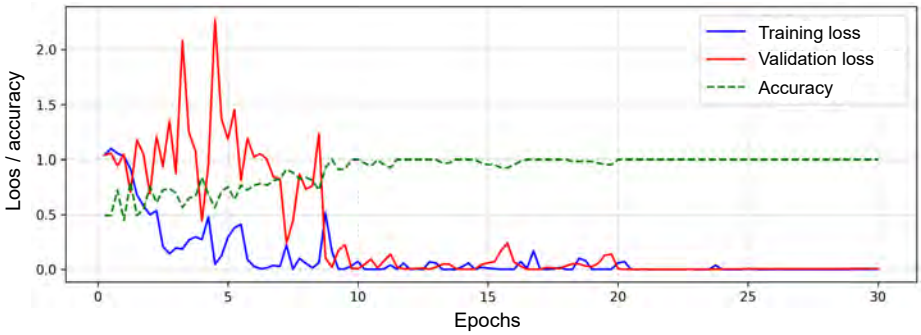


FIG. 9. Training metrics of the KWS-DD model through the fine-tuning phase.

(recall), and (4) their harmonic balance ($F1$-score), enabling robust assessment of the system's operational viability. Results are shown in Table 2.

TABLE 2. Testing result of the KWS-DD model using the curated dataset.

Precision	Recall	$F1$-score	Accuracy
0.955	0.944	0.946	0.944

6.3. Results of continuous signals

Following the evaluation of the model on the curated dataset containing isolated samples of various underwater sounds, a comprehensive evaluation is conducted on evaluation signals presented in Table 1. These results are compared with the traditional detection methods presented in Sec. 4.

6.3.1. Diver with high SNR 'Diver_1'

Figure 10 shows the detection results of the CKWS-DD model when applying 'Diver_1' to its input. The model detected nearly all inhalations with high probabilistic value, and could successfully distinguish it from the ambient noise. The periodicity is also clear, and it enhances the detection decision. Both the CKWS-DD model and the traditional method successfully detect the diver, demonstrating robust performance under high-SNR conditions.

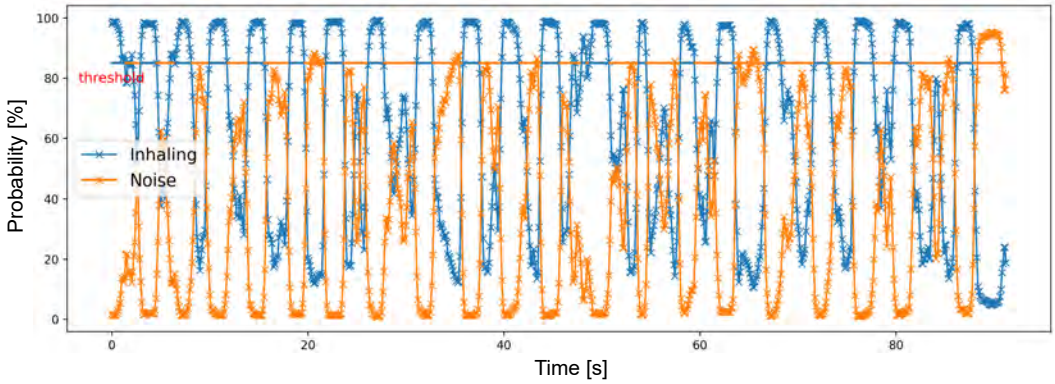


FIG. 10. Model prediction for diver signal (high SNR).

6.3.2. Diver with low SNR 'Diver_2', 'Diver_3'

In this case, 'Diver_2' was applied to the CKWS-DD model, and results are shown in Fig. 11, the traditional method failed to distinguish the diver's signals from the ambient noise in this signal, as depicted in Sec. 4 (Fig. 6). In contrast, the CKWS-DD model exhibited reliable detection. To demonstrate the model's performance

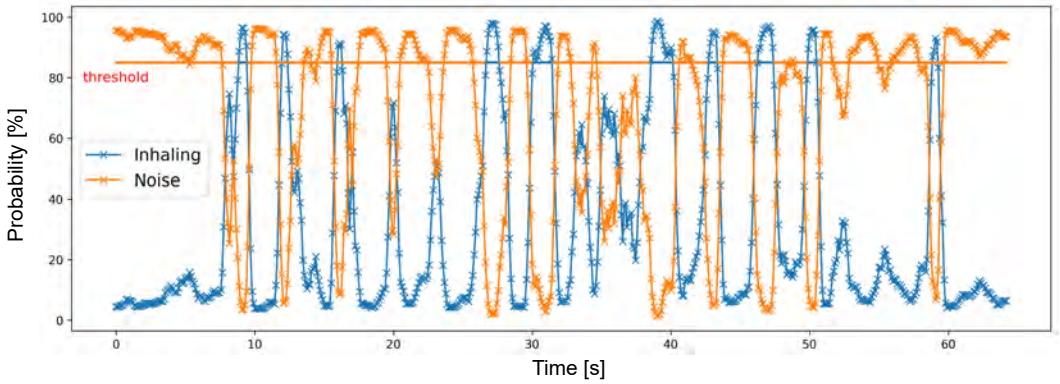


FIG. 11. Model prediction for diver signal 'Diver_2.'

in a moving target, 'Diver_3' was applied to the system input. Figure 12 shows the result. It can be seen that the detection becomes unstable when the distance between the diver and the sensor exceeds 15 m, regardless of whether the diver is approaching or moving away from the sensor.

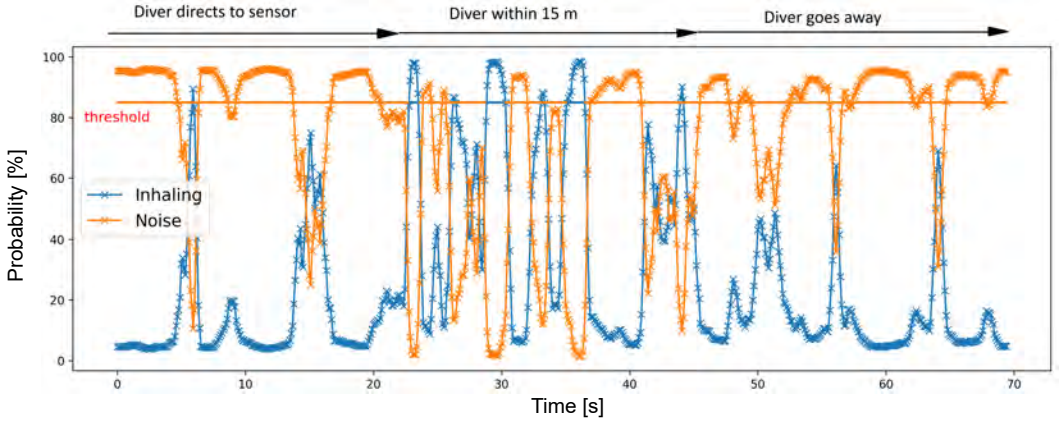


FIG. 12. Model prediction for diver signal 'Diver_3.'

6.3.3. AMBIENT NOISE DETECTION 'NOISE'

The model accurately identifies the ambient noise with a minimal number of false alarms, confirming strong noise-rejection capabilities, as shown in Fig. 13, obtained after applying 'Noise' to the model.

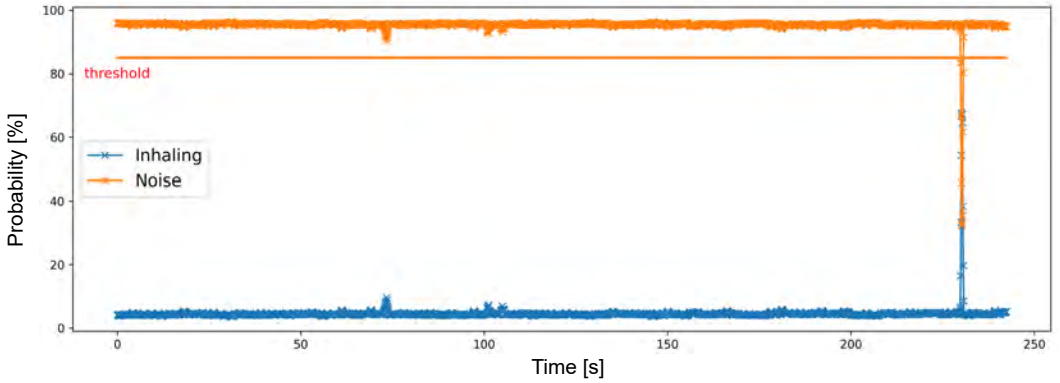


FIG. 13. Model prediction for ambient noise 'Noise.'

6.3.4. BOAT SIGNAL DETECTION 'BOAT'

Boat signals are classified as noise but exhibit sporadic false alarms (Fig. 14, 'Boat'). This stems from non-stationary motor noise and dynamic electrical status variations.

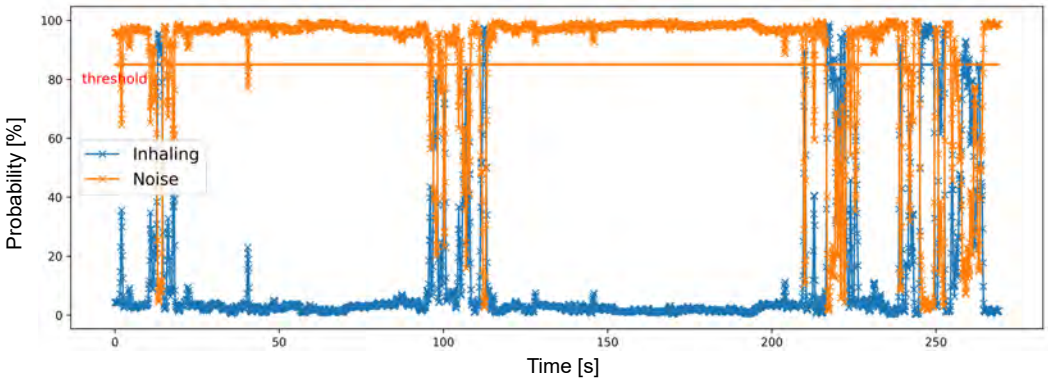


FIG. 14. Model prediction for boat noise 'Boat.'

6.4. CROSS-PLATFORM GENERALIZABILITY ASSESSMENT

To evaluate model robustness across different acquisition systems, we conducted validation testing using externally-collected recordings comprising: diver respiratory and active sonar signal captured with alternative hydrophone. This independent verification protocol assesses the system's ability to maintain detection performance when deployed with varying sensor configurations and environmental conditions. The model generalizes effectively to real-world data:

– diver detection: robust inhaling signal identification when evaluating (Fig. 15, 'Diver_4').
– sonar rejection: sonar signals are correctly classified as noise when evaluating (Fig. 16, 'Sonar').

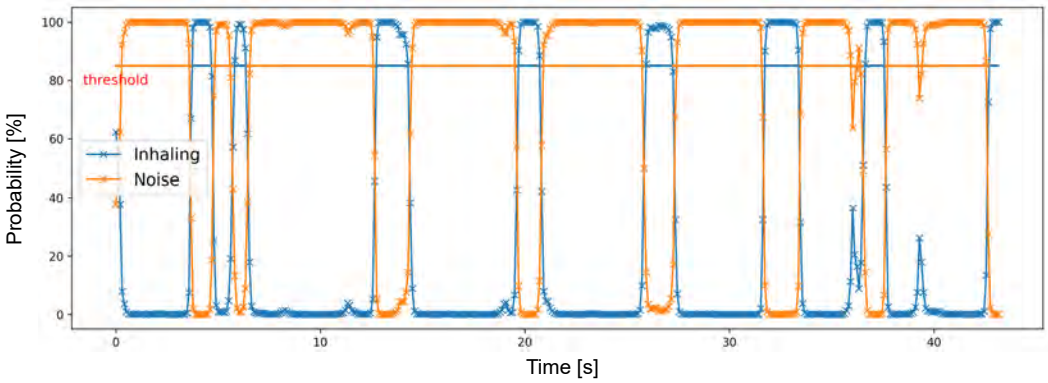


FIG. 15. External diver signal 'Diver_4.'

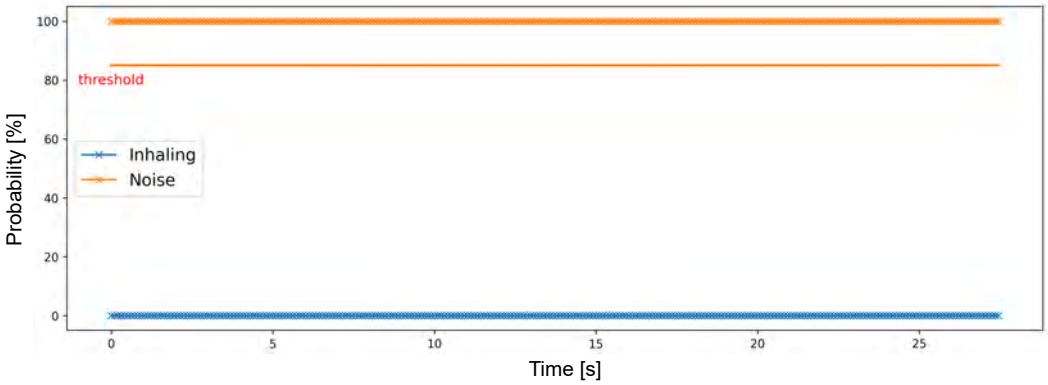


FIG. 16. Sonar signal misclassified as noise 'Sonar.'

As illustrated, the model discriminated the sonar signal from the diver's inhalation signal even though the sonar data were acquired using different equipment and under different environmental conditions – highlighting the trained model's high generalization ability.

7. Discussion

The results demonstrate that the CKWS-DD model represents a significant advancement in diver signal detection by addressing critical limitations of traditional methods. Most notably, the system extends a reliable diver detection range up to 15 m, outperforming conventional approaches that typically fail beyond 10 m in low-SNR conditions. This enhanced capability originates from the model's noise-resilient architecture, which effectively extracts faint inhaling signatures obscured by ambient noise.

The proposed system achieves an inference time of less than 0.45 s on a standard laptop computer (see Subsec. 6.2). Although the model requires a 1-second audio input window for reliable detection, the total decision latency from the start of a recording segment is approximately 1.45 s. This is substantially shorter than traditional periodicity-based methods, which often need more than 10 s to analyze multiple breathing cycles. By making frame-wise predictions on 1-second windows without waiting for a periodic structure to emerge, our model enables near-instantaneous detection.

Furthermore, transfer learning facilitates effective model training, even with limited data while maintains high detection accuracy. Validation using external real-world data, including publicly available diver and interference recordings, confirms robust generalization to unknown environments, underscoring practical deploy ability for underwater surveillance systems.

Despite these advances, two key limitations require consideration. First, false alarms during boat detection occur due to non-stationary motor noise and dynamic electrical variations related to the motor type and the operation mode, suggesting a need for additional dataset samples covering diverse boat noise profiles. Second, detection instability beyond 15 m appears linked to hydrophone sensitivity limitations.

The model depicts a clear performance dichotomy: while demonstrating performance comparable to that of traditional methods in high-SNR scenarios, it provides distinct advantages in noisy environments where conventional approaches prove ineffective.

While the current model demonstrates robust performance in distinguishing diver breathing sounds from various non-biological transient noises – including boat engine sounds and sonar pulses – we acknowledge that further enhancements are possible. In future work, we plan to enrich the training and testing dataset by incorporating a wider variety of marine animal sounds (e.g., snapping shrimp, fish vocalizations, and marine mammal calls) as well as additional artificial noise sources such as different types of boat engines, ship propellers, and other mechanical underwater equipment. We believe that exposing the model to a more diverse and challenging acoustic environment will further improve its generalization capability and reduce the risk of false positives in real-world deployment scenarios. This dataset expansion, along with corresponding model retraining and evaluation, constitutes a key direction for our ongoing research.

Another thing that should be mentioned is that water temperature and salinity gradients create sound speed profiles that can refract acoustic waves, thereby affecting the SNR of a diver's regulator sound over distance. For example, negative thermoclines, may bend sound downward and reduce the amount of surface-detected energy. Our experiments were conducted in a shallow, well-mixed environment where such gradients were minimal. For deeper or stratified waters, these effects could impact the detection range and should be considered in the system deployment.

8. Conclusion

This study establishes the viability of repurposing speech-focused KWS models for passive diver detection. Distinctive inhalation signatures are treated as acoustic 'keywords' in CKWS-DD, a transfer learning framework leveraging the HuBERT transformer architecture pre-trained on speech data. When fine-tuned with only 120 curated inhalation samples and 344 noise samples, the model achieves 94.4 % accuracy and 94.6 % $F1$-score, demonstrating significant data efficiency compared to conventional deep learning approaches (Jin, Xu, 2020). Key advantages include:

1. Real-time capability: detection latency reduces to <1 s through analysis of 1 s audio segments, eliminating the 10 s observation windows required by periodicity-based methods.

2. Noise resilience: reliable detection is maintained at SNRs as low as −8 dB (at ranges ≤15 m), outperforming traditional methods that fail beyond 10 m.
3. Interference rejection: false alarms from boat noise decrease compared to spectral energy thresholding approaches.

Current limitations included performance degradation beyond 15 m and occasional false positives from non-stationary motor noise. Cross-validation using externally sourced data confirms strong generalizability. Future research directions include dataset expansion with diverse interference signals, incorporation of marine animal sounds, and investigation of multi-diver scenarios. The CKWS-DD framework provides a scalable solution for resource-constrained underwater monitoring systems by enabling accurate, low-latency diver identification with minimal training data requirements.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that there are no known competing financial interests or personal relationships that could have influenced the work described in this paper.

Authors' contributions

Osama Deeb and Saier Mahmoud: writing – original draft, writing – review and editing, visualization, validation, resources, software, methodology, investigation, data curation, conceptualization. Louay Saleh and Assef Jafar: validation, methodology, supervision, project administration, conceptualization. Oumayma Al-Dakkak and Ibrahim Chouaib: validation, methodology, supervision, conceptualization. All authors reviewed and approved the final manuscript.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Alvaro A., Schwock F., Ragland J., Abadi S. (2021), Ship detection from passive underwater acoustic recordings using machine learning, *The Journal of the Acoustical Society of America*, **150**(4_Supplement): A124, https://doi.org/10.1121/10.0007848.
2. Chen X., Wang R., Tureli U. (2006), Passive acoustic detection of divers under strong interference, [in:] *OCEANS 2006*, https://doi.org/10.1109/OCEANS.2006.306869.
3. Chen Y., Shang J. (2019), Underwater target recognition method based on convolution autoencoder, [in:] *2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*, https://doi.org/10.1109/ICSIDP47821.2019.9173362.
4. Chung K.W., Li H., Sutin A. (2007), A frequency-domain multi-band matched-filter approach to passive diver detection, [in:] *2007 Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers*, https://doi.org/10.1109/ACSSC.2007.4487426.
5. Cole A.M. (2019), *Automated open circuit scuba diver detection with low cost passive sonar and machine learning*, MA Thesis, Massachusetts Institute of Technology, https://hdl.handle.net/1721.1/122269.
6. Davi-sh (2022), Underwater scuba diver, Pixabay, https://pixabay.com/sound-effects/underwater-scuba-diver-28467/ (access: 21.06.2025).

7. Deeb O., Jafar A., Al Dakkak O. (2025), A deep learning framework for Arabic continuous speech keyword spotting in low-resource settings using isolated-word keyword spotting and posterior probability functions, *Advances in Artificial Intelligence and Machine Learning*, **5**(3): 4074–4093, https://doi.org/10.54364/AAIML.2025.53229.
8. Domingos L.C.F., Santos P.E., Skelton P.S.M., Brinkworth R.S.A., Sammut K. (2022), A survey of underwater acoustic data classification methods using deep learning for shoreline surveillance, *Sensors*, **22**(6): 2181, https://doi.org/10.3390/s22062181.
9. Dong Y., Shen X., Wang H. (2022), Bidirectional denoising autoencoders-based robust representation learning for underwater acoustic target signal denoising, *IEEE Transactions on Instrumentation and Measurement*, **71**: 1–8, https://doi.org/10.1109/TIM.2022.3210979.
10. Donskoy D.M., Sedunov N.A., Sedunov A.N., Tsionskiy M.A. (2008), Variability of SCUBA diver's acoustic emission, [in:] *Proceedings of SPIE – The International Society for Optical Engineering*, **6945**, https://doi.org/10.1117/12.783500.
11. Feng S., Ma S., Zhu X., Yan M. (2024), Artificial intelligence-based underwater acoustic target recognition: a survey, *Remote Sensing*, **16**(17): 3333, https://doi.org/10.3390/rs16173333.
12. Fuchs L.R., Larsson C., Gällström A. (2019), Deep learning based technique for enhanced sonar imaging, [in:] *Underwater Acoustic Conference and Exhibition Series*.
13. Gong Y., Lai C.-I., Chung Y.-A., Glass J. (2022), SSAST: self-supervised audio spectrogram transformer, [in:] *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**(10): 10699–10709, https://doi.org/10.1609/AAAI.V36I10.21315.
14. Gorovoy S. *et al.* (2014), A possibility to use respiratory noises for diver detection and monitoring physiologic status, *The Journal of the Acoustical Society of America*, **135** (4_Supplement): 2303, https://doi.org/10.1121/1.4877584.
15. Gorovoy S. *et al.* (2015), Detecting respiratory noises of diver equipped with rebreather in water, [in:] *Proceedings of Meetings on Acoustics*, **24**(1): 070020, https://doi.org/10.1121/2.0000171.
16. Hari V.N., Chitre M., Too Y.M., Pallayil V. (2015), Robust passive diver detection in shallow ocean, [in:] *OCEANS 2015 – Genova*, https://doi.org/10.1109/OCEANS-Genova.2015.7271656.
17. Hewamalage H., Bergmeir C., Bandara K. (2021), Recurrent neural networks for time series forecasting: current status and future directions, *International Journal of Forecasting*, **37**(1): 388–427, https://doi.org/10.1016/J.IJFORECAST.2020.06.008.
18. Hsu W.-N., Bolte B., Tsai Y.-H.H., Lakhotia K., Salakhutdinov R., Mohamed A. (2021), HuBERT: self-supervised speech representation learning by masked prediction of hidden units, *IEEE/ACM Transactions on Audio Speech and Language Processing*, **29**: 3451–3460, https://doi.org/10.1109/TASLP.2021.3122291.
19. Huang F., Zhang J., Zhou C., Wang Y., Huang J., Zhu L. (2020), A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction, *Landslides*, **17**(1): 217–229, https://doi.org/10.1007/s10346-019-01274-9.
20. Jin B., Xu G. (2020), A passive detection method of divers based on deep learning, [in:] *2020 IEEE 3rd International Conference on Electronics Technology (ICET)*, https://doi.org/10.1109/ICET49382.2020.9119556.
21. Jin G., Liu F., Wu H., Song Q. (2020), Deep learning-based framework for expansion, recognition and classification of underwater acoustic signal, *Journal of Experimental and Theoretical Artificial Intelligence*, **32**(2): 205–218, https://doi.org/10.1080/0952813X.2019.1647560.
22. Johansson A.T., Lennartsson R.K., Nolander E., Petrović S. (2010), Improved passive acoustic detection of divers in harbor environments using pre-whitening, [in:] *OCEANS 2010 MTS/IEEE SEATTLE*, https://doi.org/10.1109/OCEANS.2010.5664549.
23. Karjalainen A.I., Mitchell R., Vazquez J. (2019), Training and validation of automatic target recognition systems using generative adversarial networks, [in:] *2019 Sensor Signal Processing for Defence Conference (SSPD 2019)*, https://doi.org/10.1109/SSPD.2019.8751666.
24. Korenbaum V., Kostiv A., Gorovoy S., Dorozhko V., Shiryaev A. (2020), Underwater noises of open-circuit scuba diver, *Archives of Acoustics*, **45**(2): 349–357, https://doi.org/10.24425/aoa.2020.133155.
25. Lennartsson R.K., Dalberg E., Persson L., Petrović S. (2009), Passive acoustic detection and classification of divers in Harbor environments, [in:] *OCEANS 2009*, https://doi.org/10.23919/OCEANS.2009.5422407.

26. Li P., Wu J., Wang Y., Lan Q., Xiao W. (2022), STM: spectrogram transformer model for underwater acoustic target recognition, *Journal of Marine Science and Engineering*, **10**(10): 1428, https://doi.org/10.3390/JMSE10101428.
27. Lin J., Kilgour K., Roblek D., Sharifi M. (2020), Training keyword spotters with limited and synthesized speech data, [in:] *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings*, https://doi.org/10.1109/ICASSP40776.2020.9053193.
28. Mahmoud S., Saleh L., Chouaib I. (2025), Experimental results of diver detection in Harbor environments using single acoustic vector sensor, *Archives of Acoustics*, **50**(2): 173–185, https://doi.org/10.24425/aoa.2025.153663.
29. Phung S.L. *et al.* (2019), Mine-like object sensing in sonar imagery with a compact deep learning architecture for scarce data, [in:] *2019 Digital Image Computing: Techniques and Applications (DICTA 2019)*, https://doi.org/10.1109/DICTA47822.2019.8945982.
30. Radford C.A., Jeffs A.G., Tindle C.T., Cole R.G., Montgomery J.C. (2005), Bubbled waters: the noise generated by underwater breathing apparatus, *Marine and Freshwater Behaviour and Physiology*, **38**(4): 259–267, https://doi.org/10.1080/10236240500333908.
31. Seo D., Oh H.-S., Jung Y. (2021), Wav2KWS: transfer learning from speech representations for keyword spotting, *IEEE Access*, **9**: 80682–80691, https://doi.org/10.1109/ACCESS.2021.3078715.
32. Stolkin R. *et al.* (2006), Feature based passive acoustic detection of underwater threats, [in:] *Proceedings SPIE 6204, Photonics for Port and Harbor Security II*, https://doi.org/10.1117/12.663651.
33. Sun Y. *et al.* (2022), Multi-resonance flextensional hydrophone for open-circuit scuba diver detection, *AIP Advances*, **12**(11): 115310, https://doi.org/10.1063/5.0101999.
34. Sun Y. *et al.* (2024), Feature extraction methods for underwater acoustic target recognition of divers, *Sensors*, **24**(13): 4412, https://doi.org/10.3390/s24134412.
35. Sutin A., Salloum H., DeLorme M., Sedunov N., Sedunov A., Tsionskiy M. (2013), Stevens passive acoustic system for surface and underwater threat detection, [in:] *2013 IEEE International Conference on Technologies for Homeland Security (HST)*, https://doi.org/10.1109/THS.2013.6698999.
36. Tu Q., Yuan F., Yang W., Cheng E. (2020), An approach for diver passive detection based on the established model of breathing sound emission, *Journal of Marine Science and Engineering*, **8**(1): 44, https://doi.org/10.3390/JMSE8010044.
37. Vincent P., Larochelle H., Lajoie I., Bengio Y., Manzagol P.-A. (2010), Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion, *Journal of Machine Learning Research*, **11**: 3371–3408.
38. VoiceBosch (2024), Sonar, Pixabay, https://pixabay.com/sound-effects/sonar-183436/ (access: 21.05.2025).
39. Wang W., Huang Y., Wang Y., Wang L. (2014), Generalized autoencoder: a neural network framework for dimensionality reduction, [in:] *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, https://doi.org/10.1109/CVPRW.2014.79.
40. Wang X., Meng J., Liu Y., Zhan G., Tian Z. (2022), Self-supervised acoustic representation learning via acoustic-embedding memory unit modified space autoencoder for underwater target recognition, *The Journal of the Acoustical Society of America*, **152**(5): 2905–2915, https://doi.org/10.1121/10.0015138.
41. Wang Y. *et al.* (2021), Underwater communication signal recognition using sequence convolutional network, *IEEE Access*, **9**: 46886–46899, https://doi.org/10.1109/ACCESS.2021.3067070.
42. Zhang X., Lu Z., Kang C. (2003), Underwater acoustic targets classification using support vector machine, [in:] *Proceedings of 2003 International Conference on Neural Networks and Signal Processing*, https://doi.org/10.1109/ICNNSP.2003.1280753.
43. Zhao J. *et al.* (2023), Underwater target perception algorithm based on pressure sequence generative adversarial network, *Ocean Engineering*, **286**(Part 1): 115547, https://doi.org/10.1016/J.OCEANENG.2023.115547.
44. Zhao W. *et al.* (2016), Passive acoustic detection of diver based on SVM, [in:] *2016 IEEE International Conference on Mechatronics and Automation*, https://doi.org/10.1109/ICMA.2016.7558635.