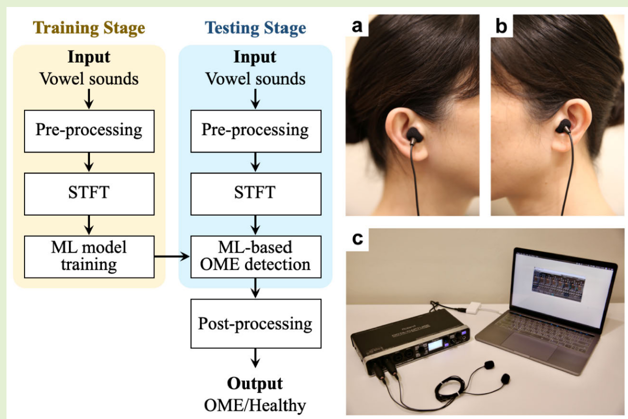
Detection of Otitis Media With Effusion Using In-Ear Microphones and Machine Learning

Kuan-Chung Ting, Syu-Siang Wang, You-Jin Li, Chii-Yuan Huang, Tzong-Yang Tu, Chun-Che Shih, Kai-Chun Liu, *Member, IEEE*, and Yu Tsao, *Senior Member, IEEE*

***Abstract*—The diagnostic accuracy (ACC) of otitis media with effusion (OME) depends on a clinician's experience and evaluation tools. Various assessment technologies have been applied to support clinical diagnosis, such as digital otoscopy and tympanometry. However, several challenges and issues limit the capabilities and usability of these assessment technologies, including high costs and needing to rely on specialists' interpretations. In this work, we designed and validated OME detection using a machine learning (ML) model and in-ear microphones. Two off-the-shelf microphones were placed in the bilateral ear canals to record the voice when participants pronounced five 3-s sustained vowel sounds. Various signal processing and ML techniques were applied to the recordings, and the magnitude spectrograms of the vowel sound recording from in-ear microphones can distinguish ears with OME from healthy ears according to the differences in high-frequency response. Our results using in-ear microphones and ML algorithms had an ACC of 80.65% in detecting OME, similar to that of typical OME detection approaches. This work demonstrates the potential to provide healthcare practitioners with a simple, safe, and more reliable expert-level diagnostic tool.**



***Index Terms*—In-ear microphones, machine learning (ML), otitis media with effusion (OME).**

Manuscript received 28 April 2023; accepted 22 July 2023. Date of publication 6 October 2023; date of current version 14 November 2023. This work was supported in part by the Ministry of Science and Technology of Taiwan under Grant MOST 110-2222-E-075-001-MY2, in part by the Taipei Veterans General Hospital under Grant V112B-011, and in part by the Academia Sinica under Grant AS-GC-111-M01. The associate editor coordinating the review of this article and approving it for publication was Dr. Ravibabu Mulaveesala. *(Corresponding authors: Chun-Che Shih; Kai-Chun Liu.)*

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Taipei Veterans General Hospital Review Boards under IRB-TPEVGH No. 2021-02-011BC.

Kuan-Chung Ting is with the Department of Otolaryngology-Head and Neck Surgery, Taipei Veterans General Hospital, Taipei 112, Taiwan, also with the School of Medicine, National Yang Ming Chiao Tung University, Taipei 30010, Taiwan, and also with the Institute of Clinical Medicine, National Yang Ming Chiao Tung University, Taipei 30010, Taiwan.

Syu-Siang Wang, You-Jin Li, Kai-Chun Liu, and Yu Tsao are with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 30010, Taiwan (e-mail: t22302856@gmail.com).

Chii-Yuan Huang and Tzong-Yang Tu are with the Department of Otolaryngology-Head and Neck Surgery, Taipei Veterans General Hospital, Taipei 112, Taiwan, and also with the School of Medicine, National Yang Ming Chiao Tung University, Taipei 30010, Taiwan.

Chun-Che Shih is with the Institute of Clinical Medicine, National Yang Ming Chiao Tung University, Taipei 30010, Taiwan, also with the Division of Cardiovascular Surgery, Taipei Municipal Wan-Fang Hospital, Taipei 116, Taiwan, and also with the Taipei Heart Institute, Taipei Medical University, Taipei 110, Taiwan (e-mail: profccshih@gmail.com).

Digital Object Identifier 10.1109/JSEN.2023.3321093

I. Introduction

OTITIS media (OM) is a common disease and infection of the middle ear in children [1]. A previous report showed that the annual medical cost of OM treatment was approximately $4 billion in the United States [2]. OM with effusion (OME) is a specific type of OM characterized by middle ear effusion but lacks the signs of acute infection [3]. Persistent OME for three months or more, also known as chronic OME [4], may lead to hearing loss, sleep disruption, and balance issues [3], [5]. Additionally, children may have decreased learning efficiency, signs of inattention, and delayed speech and developmental skills [4], [6], [7]. Chronic OME without spontaneous resolution may also cause eardrum structural damage [4]. A typical examination tool is pneumatic otoscopy, which is used to visually assess the tympanic membrane and middle ear. However, a clinician using otoscopy has an accuracy (ACC) of 50%–70% for diagnosing OM [3], [8], [9], [10]. In particular, OME is often underdiagnosed in clinical environments [3].

Various assessment technologies have been applied to support clinical diagnosis, including image-based and acoustic-based approaches. Telescopy with video cameras, an image-based approach, is the most common assistive method in clinical environments [11], [12], [13], [14]. It can enhance the diagnostic ACC [15], but it still requires professional interpretation [16], and the high-cost video system is not affordable for all primary clinics. Tympanometry, the acoustic-based approach, is a dynamic documentation of middle ear impedance, using increased and decreased air pressure in the external auditory canal and then measuring the associated reflected response recorded by a microphone [17]. However, an airtight seal is required between the probe and the ear canal. This may cause manual errors or bias and decrease usability in clinical settings. Moreover, tympanometry cannot be used alone as a diagnostic tool. For example, the result will differ if cerumen or a perforation is present. In summary, several challenges and issues limit the capabilities and usability of these assessment technologies. First, OM assessment tools rely on specialists' interpretation. The diagnostic ACC of otolaryngologists, pediatricians, and general practitioners is different and depends on their experience and skills. The second challenge is that the high-cost equipment decreases the penetration rate and availability in primary clinics: cost can be the main barrier for primary clinics to use tympanometry in diagnosis [18]. For example, only 7%–33% of caregivers have pneumatic otoscopy [3]. Therefore, it is essential to develop low-cost, easy-to-use, and objective approaches that help primary clinics detect and identify OM.

In recent years, machine learning (ML) techniques have been widely applied to support clinical assessment [19], [20] and objectively classify the status of the middle ear. Clinically, ML can enhance current predictions and be a novel tool for helping clinicians assess diseases [21], [22], [23].

The primary aim of this study is to propose an OME detection system using in-ear microphones and ML techniques to assist clinicians in diagnoses. The system records the voices of the user using in-ear microphones, extracting acoustic features with short-time Fourier transform (STFT) and estimating the health status of the middle ear with ML classifiers and weighted-threshold postprocessing. The developed low-cost, easy-to-use, and objective approaches have the potential to help clinicians detect and identify OME in home-based environments.

In summary, the main contributions of this work are as follows.

1) The proposed system using in-ear microphones directly records the voices of the user transmitted through the middle ear, in contrast to typical acoustic-based assessment tools, which require additional speakers to generate a specific tone for the recording and analysis.

2) Off-the-shelf microphones are used for voice recording instead of customized designs, which decreases the cost of the equipment and lowers the barrier to accessing OME screening tools for caregivers and physicians.

3) This work explores the effectiveness of various ML classification models in automatic OME detection, including support vector machine (SVM), Gaussian Naive Bayes (GNB), adaptive boosting (AdaBoost), random forest (RF), and convolutional neural networks (CNNs).

The rest of this article is organized as follows: the related works are presented in Section II. In Section III, we introduce the experimental setups and the processes of the OME detection systems. Section IV presents the experimental results using the proposed approach. The performance analysis, limitations, and futures are discussed in Section V. Section VI concludes this article.

II. Related Works

A. OM Detection Based on Images

Several assessments based on image detection have been proposed for OM. Optical coherence tomography, a cross-sectional imaging technique, combined with ML methods may provide a promising platform with more than 90% ACC [24]. Viscaino et al. [25] successfully applied ML approaches for external and middle ear assessment with computer vision (digital otoscopy). Their experimental results show that diagnosing external and middle ear conditions using an SVM can achieve 93.9% ACC. Crowson et al. [26] reported that they employed ML to assist with the diagnosis of OM using tympanic images obtained from children. The proposed model achieved an ACC of 83.8%. Kashani et al. [27] integrated shortwave infrared into an otoscope along with an ML-based model. The RF method yielded the highest ACC of 90.3%. In general, the ACC of image-based detection with ML has satisfactory results. The ACC of ML approaches for diagnosing middle ear disorders using tympanic membrane images has ranged from 76.00% to 98.26% [28]. However, these image-based approaches require high-cost equipment.

B. OM Detection Based on Acoustics

Compared to image-based detection of OME, research on acoustics-based OME detection is less common. Chan et al. [29] developed an acoustic-based OME classification system using smartphones and logistic regression classifiers. A funnel was designed and placed between the smartphone's speaker/microphone and the participant's ear canal, and the raw acoustic waveform was obtained from the ear canal after chirps were played into an ear, with or without middle ear fluid. The system used to assess eardrum mobility can achieve 85% sensitivity (SEN) and 82% specificity (SPE). Binol et al. [30] combined otoscopy imaging and tympanogram to detect OME automatically. The proposed decision fusion method using the RF classifier has an ACC of 84.9%. Grais et al. [31] employed ML techniques to examine the differences between normal ears and those with OME using wideband tympanometry analysis. They used an RF model to create class activation maps for interpreting diagnostic decisions. These works demonstrated the superiority of ML models in supporting OME assessment in medical practice.

III. Materials and Methods

The framework of the proposed OME detection systems is shown in Fig. 1. Initially, several signal preprocessing and feature extraction approaches were applied to the recorded vowel sounds, including Z score normalization and STFT. The OME detection models were trained and developed with various ML techniques and training vowel sounds. Then, in the testing stage, the testing data were input into the trained ML

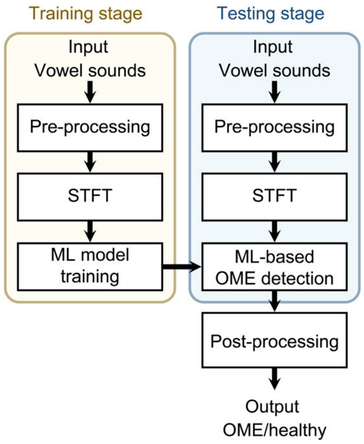


Fig. 1. Framework of the proposed OME detection systems.

TABLE I
PATIENT DEMOGRAPHICS

Variable	
Age, mean (range)	59.65 (26-79) years old
Gender (F:M)	11:20
OME	
Right ear	15
Left ear	10
Bilateral ears	6
Pure tone audiogram (ears with OME)	
Air-bone gap, mean (SD)	25.81 (13.26) dB HL
Tympanogram (ears with OME)	
Type B	27 (72.97%)
Type C	10 (27.03%)

models, and preliminary results were obtained. Finally, the weighted threshold function of postprocessing was employed to make the final OME decision.

A. Participants and Experimental Protocol

This cross-sectional study was completed in a tertiary and academic medical center, and data were collected between November 2020 and August 2021. Adults above 20 years of age who were diagnosed with unilateral or bilateral OME at outpatient departments were enrolled. All patients underwent pure tone audiometry, tympanometry, video telescopy, and nasopharyngoscopy. Patients with a history of head and neck cancer or middle ear surgery were excluded. The study was conducted under a protocol approved by the Taipei Veterans General Hospital Review Boards (IRB-TPEVGH No. 2021-02-011BC). All participants provided written informed consent.

Thirty-one adults diagnosed with OME [20 men and 11 women with a mean age of 60 years (range 26–79 years)] participated in this study. Twenty-five patients had unilateral OME (15 right ear and ten left ear), and six had bilateral OME. In summary, there were 25 healthy ears and 37 ears with OME. Of the 37 ears with OME, 27 had type B tympanograms and ten had type C tympanograms, with a mean air-bone gap of 25.81-dB hearing level (HL) (SD 13.26 dB) in pure tone audiograms (see Appendix). The patient demographics are summarized in Table I.

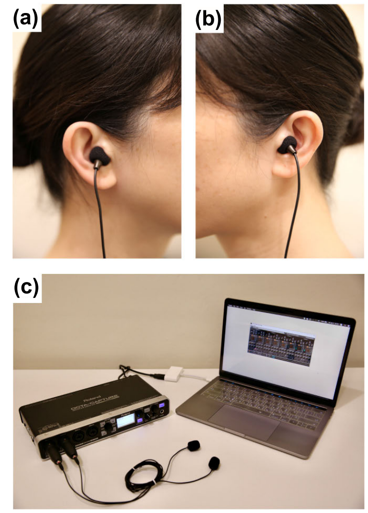


Fig. 2. Placements of the microphones in (a) right ear canal and (b) left ear canal. (c) Hardware setup for voice recording.

B. Hardware and Data Recording

Each participant was invited to pronounce five 3-s sustained vowel sounds, namely, /e:/, /i:/, /aj/, /o:/, and /a:/. Two off-the-shelf microphones (Sony Electronic Inc., Tokyo, Japan) sampled at 44.1 kHz were placed at the ear canal entrance to record the voice in the ears. Roland OCTA-CAPTURE (Roland Corporation Shizuoka, Japan) was applied to synchronize the two microphone signals and transmit the data to a laptop via a USB connection. The participants sat in a quiet office room with an average noise level of 30 dBA and were asked to wear earmuffs (3M[1] PELTOR[1] X3A) to reduce the environmental noise from the recordings and ensure the acoustic signal quality. An illustration of the hardware and microphone placements is shown in Fig. 2.

C. Data Preprocessing and Acoustic Feature Extraction

A series of data processing steps were employed to obtain satisfactory acoustic features for OME detection. In the data preprocessing, the raw signals were resampled first at 16 kHz (the frequency of a human speaking voice is typically up to 4 kHz). Then, the researchers labeled the starting and ending points for each vowel sound recording. Finally, because of the unstable voice quality, we removed the first and last 0.70 s from the recorded waveform.

This study applied STFT to extract acoustic features (Fig. 3). The STFT has been widely used in many speech processing applications [32]. For each utterance in the U recording set, the framing technique was first applied to split the input waveform into a series of temporal frames, where the frame size and the hop length for the framing process were 64 and 32 ms, respectively.

In addition, the Hamming window function was used to smooth the frame boundary. Then, a Fourier transform was applied to provide magnitude spectra for each temporal frame signal. Next, we collected and stacked all the magnitude spectra to generate the acoustic features. Finally, we normalized

[1]Trademarked.

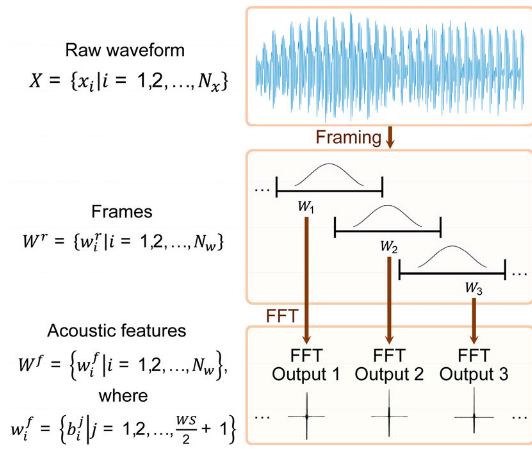


Fig. 3. Illustration of STFT to the raw waveform.

this spectrogram to zero mean and identity variance to enhance the energy in the high-frequency range.

D. ML-Based OME Detection

ML classification models were applied to acoustic feature vectors to detect OME. ML-based OME detection involves two stages: training and testing stages. In the training stage, the training feature set ϕ^{train} was composed of N_{train} pairs of a feature frame vector F_{train} and the associated OEM label c_{train}, which was a binary value. Then, a classification model $\mathcal{H}$ was leveraged to predict the OEM label $\hat{c}_{\text{train}}$ with respect to the input F_{train}, i.e., $\hat{c}_{\text{train}} = \mathcal{H}\{F_{\text{train}}\}$. The error between the predicted $\hat{c}_{\text{train}}$ and c_{train} was then minimized to optimize $\mathcal{H}$. In the testing stage, a testing acoustic feature was placed at the input side of $\mathcal{H}$, and the classification output was obtained.

In this study, five popular ML classification models were applied to acoustic feature vectors to detect OME: SVM, GNB, AdaBoost, RF, and CNN classification models. These models have been widely used in acoustic-based medical applications, achieving reliable performance [33], [34]. A brief introduction and the important parameters of the applied models are as follows.

SVM: SVM is a general classification method in many healthcare and medical diagnosis applications [33], [35]. The effectiveness and usability of SVM have been validated in previous studies [36]. The main objective of SVM is to observe a hyperplane in the feature space to make decisions. In the training stage, the hyperplane is optimized with the maximum margin between two classes. The trained SVM classifies the testing data according to the decision hyperplane. This work applied a linear kernel function and Bayesian optimization to train the SVM model.

GNB: GNB is a typical identification approach based on the Bayes theorem [37], which assumes that the input features are independent. In the training stage, GNB calculates the mean and the standard deviation of the training data and applies them to the conditional likelihood function. GNB observes the input feature with the largest posterior probability and outputs the class label.

AdaBoost: AdaBoost [38] aims to build an ensemble of weak classifiers and apply them to final decision making. During the training, each weak classifier literately fits the training data in a specific feature domain to minimize weighted errors. Then, more weights are updated for the classifier that has a lower false prediction rate. The final classification model is a linear combination of the trained weak classifiers. In this study, we investigated the number of weak classifiers in a range of 80–120 with a step of 10. The best detection performance with 100 weak classifiers is shown in the final results.

RF: RF [39] is an ensemble learning approach for classification. This approach trains k decision trees with k different training subsets, where the training subset data are randomly sampled from the training set by the bootstrap method. The RF classification model is then constructed based on these trained k decision trees. In the testing stage, the testing data are input into each decision tree, and the final classification output is generated by the majority voting of k decision trees. In this work, a range of k from 30 to 60 with a step of 10 was tested, and the best performance using $k = 50$ is presented in the results.

CNN: CNN is a common type of neural network that has been applied in different applications, including image recognition [40] and audio processing [41]. The CNN-based classification model mainly contains 1-D convolution, batch normalization, dropout, rectified linear unit (ReLU), max pooling, full connection, and softmax layers. In the beginning, we apply two CNN blocks to process the inputs, where a CNN block is composed of a 1-D convolution layer with a kernel number of 32, kernel size of 30, stride size of 1, batch normalization, ReLU, max pooling with a size of 2, and dropout ($p = 0.2$). After the process of the CNN blocks, we employ the flattening process to the feature maps and connect to one full connection with 32 filters and softmax layers for the detection. The softmax layer calculates the possibility of the target classes. Finally, the classification model outputs the class with the highest probability.

The Adam optimizer [42] is utilized during the training process. The loss function is cross-entropy. The learning weight and weight decay are 0.0001 and 0.000001, respectively. A total of 40 epochs are applied for model training, and the batch size is 32. The CNN model is implemented on Python 3.9, PyTorch 1.9.1, and CUDA 1.1.1.

E. Postprocessing

Weighted threshold postprocessing was applied to the outputs of the ML classifiers for the final OME decision. An efficient postprocessing mechanism can enhance performance in the field of computer-aided diagnostic systems, such as for the detection of Parkinson's disease [43] and lung function diagnosis [44]. The weighted threshold function Γ was achieved via the following steps: 1) the number of positive labels in an utterance was represented as the score S_{c_p} and 2) the threshold β, which varied from 5 to 40, was compared with S_{c_p} to assess the positive OME hypothesis of $S_{c_p} \geq \beta$.

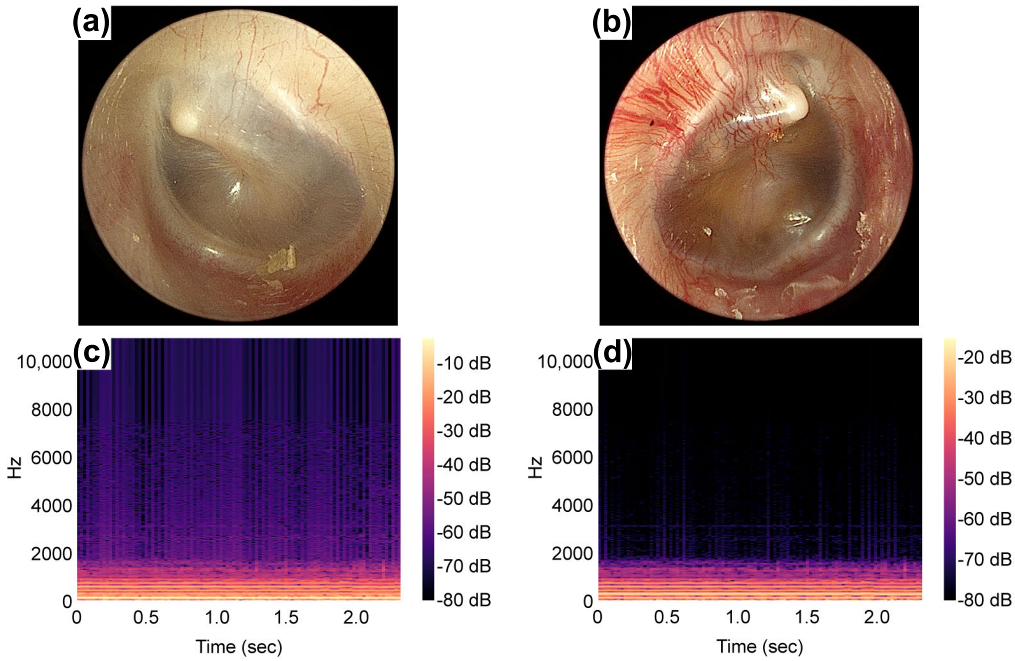


Fig. 4. Example of otoscopy and the magnitude spectrograms of vowel sound recording. A participant with right side OME pronounced a 3-s sustained vowel sound /a:/. (a) Left healthy ear without middle ear effusion. The tympanic membrane is translucent with a light reflex on otoscopic examination. (b) Right ear with OME. The noninfected fluid accumulates in the middle ear space. The tympanic membrane is retracted with an amber appearance. (c) Magnitude spectrogram of the left healthy ear. High-frequency responses to voice are retained from 2000 to 10 000 Hz. (d) Magnitude spectrogram of the right ear with OME. High-frequency features of sounds are eliminated over 2000 Hz.

F. Performance Evaluation

Several evaluation metrics were used to evaluate the performance of the proposed OME detection systems, such as ACC, SEN, precision (PRE), and SPE. The definitions of these metrics are as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{1}$$

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{2}$$

$$\mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{3}$$

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{4}$$

where TP and TN represent that the number of OME and the healthy ears are correctly identified, while FP and FN determine the misidentification number of OME and healthy ears.

This study applied the leave-one-subject-out (LOSO) cross-validation approach to validate the effectiveness of the proposed OME detection approach. LOSO divided all of the data into κ subsets based on κ subjects involved in the experiments. Then, one subset and the remaining $\kappa - 1$ subsets were used to test and train the OME detection model. This validation approach was repeated κ times until all subsets were tested. Finally, the testing results were averaged over all folds/subjects.

The proposed OME detection approach was implemented in the Python environment. The data preprocessing [e.g., fast Fourier transform (FFT) and feature extraction] and ML models were realized with librosa 0.8.0 music and audio analysis package [45] and scikit-learn 0.24.1 ML toolkit [46].

IV. Experimental Results

The magnitude spectrograms of the vowel sound recording from a unilateral OME patient are shown in Fig. 4. The results showed that the frequency response >2 kHz was lost in OME patients, while it was maintained in healthy ears. Thus, the voices recorded from the healthy ear retained most high-frequency responses to voice, as shown in Fig. 4(a) and (c).

The high-frequency features of sounds were eliminated in the ear with OME, as shown in Fig. 4(b) and (d), because effusion in the middle ear space had a considerable impact on voice transmission, which led to the loss of consonant components. Such distinguished differences in high-frequency regions enabled the proposed system to distinguish ears with OME from healthy ears.

The ACC, SEN, PRE, and SPE of the proposed OME detection approach using ML models and vowel recordings with different β values are presented in Fig. 5. In general, most ML models, including SVM, AdaBoost, RF, and CNN, using β from 10 to 28 achieved the best ACC, and their SEN increased when β was greater than 5. In contrast, their PRE and SPE decreased notably with $\beta \geq 5$, particularly for SVM and AdaBoost. For the GNB model, the ACC and SEN improved when β was larger than 20.

Table II presents the best ACC with the corresponding β values of the ML models using all vowel sounds and five single vowel sounds (/e:/, /i:/, /aj/, /o:/, and /a:/). The proposed OME detection approach using the CNN model with

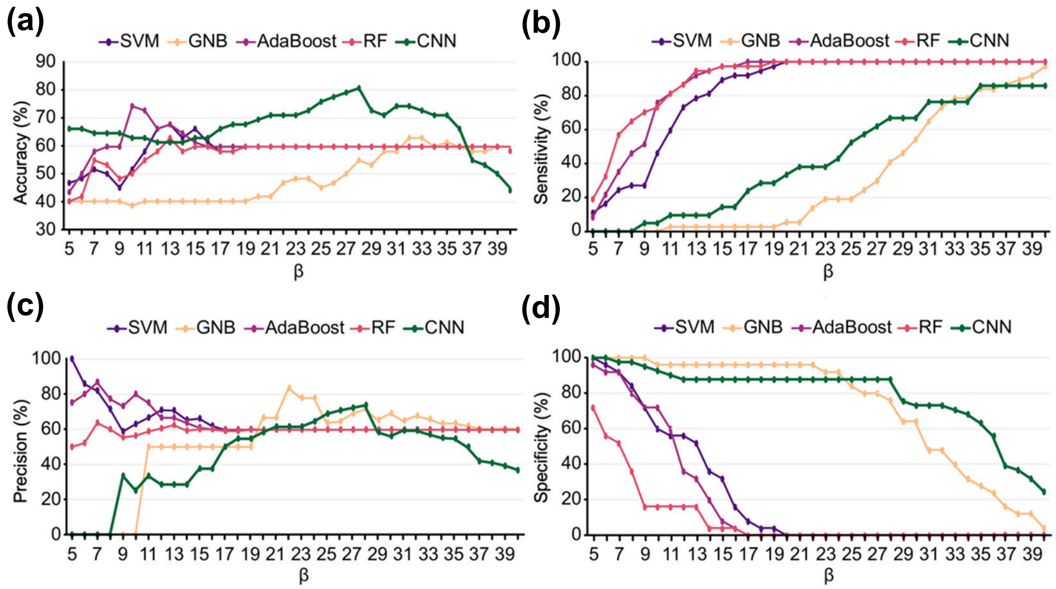


Fig. 5. Performance analysis of ML classification versus different threshold weights (β) in (a) ACC, (b) SEN, (c) PRE, and (d) SPE.

$\beta = 28$ achieved the best ACC of 80.65%, while those achieved using typical ML models were lower than 75.00%. Similarly, the CNN-based OME detection system achieved the best SPE of 87.8%. The RF model with $\beta = 13$ achieved the best SEN of 94.59% but the worst SPE of 16.32%. Additionally, the detection system achieved the best ACC of 77.42% with the single vowels /**i**:/. For the OME detection systems using single vowels, the CNN model had the best performance. Overall, most ML models using single vowel sounds had high SEN and moderate PRE, while their SPE was relatively low.

V. Discussion

The accurate diagnosis of OM is important; however, the diagnostic ACC depends on the experience of the clinician and evaluation tools. Otolaryngologists had a better correct diagnostic rate than pediatricians and general practitioners [16]. The correct diagnosis rates of video-presented OME, acute OM, and retracted tympanic membrane among otolaryngologists, pediatricians, and general practitioners in the US were 74%, 51%, and 46%, respectively [16]. Pichichero and Poole [10] evaluated 188 otolaryngologists and 514 pediatricians for their diagnostic ACC of OM, and the correct diagnosis rates by otolaryngologists and pediatricians were 73% and 50%, respectively. Jones and Kaleida [8] reported that pneumatic otoscopy ACC for OME was 76% and that static otoscopy ACC for OME was 61% in 34 pediatric residents, four pediatricians, and two pediatric otolaryngologists. Both pneumatic otoscopy and video-telescopy require professional interpretation. Our results using in-ear microphones and ML algorithms had an ACC of 80.65% in detecting OME, similar to that of typical OME otoscopy detection approaches. Such detection ACC shows the potential of the proposed approach in clinical practice.

Tympanometry is a rapid and simple examination to assess the middle ear condition. According to the American Academy of Otolaryngology-Head and Neck Surgery clinical practice guidelines, tympanometry should be performed when the diagnosis of OME is uncertain after performing pneumatic otoscopy [3]. However, the equipment for tympanometry is relatively expensive and requires a referral to a clinician. Our system records the voices of the user using commercial in-ear microphones, extracts acoustic features with STFT, and estimates the health status of the middle ear with ML classifiers and rule-based postprocessing. This low-cost and easy-to-use OME detection system is usable and beneficial for nonmedical caregivers and primary clinicians.

To support the clinical assessment of OME, this study proposed an OME detection system using in-ear microphones and ML models. Our analysis showed that the proposed OME detection system using an ensemble learning approach could achieve better detection performance than that using a single strong classification model. For example, the system using the CNN model and all vowel sounds achieved the best ACC of 80.65%. Similar trends showed that the CNN model was suitable for the system using single vowel records. Several previous studies have shown the superiority of ensemble learning methods for classification problems [47], [48]. The effectiveness of ensemble models has been validated in different applications [49], [50]. Additionally, we demonstrated that the weighted threshold function can improve the detection performance since the detection system is sensitive to OME diseases and avoids misclassifying healthy ears as OME. Suitable weight tuning can enhance the correct rejection ability of the system.

To the best of our knowledge, this is the first work to propose an OME detection system using in-ear microphones and ML techniques to support clinical diagnosis and assessment. Previous works have demonstrated the feasibility of ML models for acoustic-based OME detection [17], [51], [52], [53], [54]. However, they had to use both microphones and speakers for disease detection. In contrast to previous works,

TABLE II
BEST DETECTION PERFORMANCE OF DIFFERENT ML MODELS

Detection Technique	Performance Measure (%)				
	ACC	SEN	PRE	SPE	β[a]
ML models using all vowel sounds					
SVM	67.74	78.38	70.73	52.73	13
GNB	62.90	72.97	67.50	48.08	32
AdaBoost	74.19	75.68	80.00	72.67	10
RF	62.90	94.59	62.50	16.32	13
CNN	80.65	66.67	73.68	87.80	28
ML models using a single vowel sound of /e:/					
SVM	59.68	100.00	59.68	0.00	16
GNB	62.90	59.46	73.33	68.24	12
AdaBoost	61.29	75.68	65.12	40.56	13
RF	64.52	97.30	63.16	16.32	11
CNN	75.81	57.14	63.16	82.93	23
ML models using a single vowel sound of /i:/					
SVM	61.29	78.38	64.44	36.98	9
GNB	59.68	100.00	59.68	0.00	12
AdaBoost	61.29	97.30	61.02	8.06	8
RF	62.90	100.00	61.67	8.06	6
CNN	77.42	90.48	39.58	29.27	17
ML models using a single vowel sound of /aj/					
SVM	59.68	100.00	59.68	0.00	17
GNB	59.68	91.89	60.71	12.12	15
AdaBoost	61.29	97.30	61.02	8.06	13
RF	64.52	100.00	62.71	12.73	11
CNN	75.81	71.43	46.88	58.54	16
ML models using a single vowel sound of /o:/					
SVM	59.68	100.00	59.68	0.00	16
GNB	59.68	91.89	60.71	12.73	14
AdaBoost	61.29	94.59	61.40	12.12	11
RF	62.90	100.00	61.67	8.06	10
CNN	74.19	57.14	50	70.73	20
ML models using a single vowel sound of /a:/					
SVM	59.68	100.00	59.68	0.00	15
GNB	59.68	100.00	59.68	0.00	17
AdaBoost	59.68	100.00	59.68	0.00	14
RF	62.90	97.30	62.07	12.12	10
CNN	75.81	80.95	60.71	73.17	35

[a]: the determined weighted threshold

the proposed approach required microphones only for voice recording and analysis. Furthermore, similar to previous work [29], this study preliminarily validated the feasibility of off-the-shelf microphones in assessing the health status of the middle ear. These advantages and experimental results show that the proposed detection system has the potential to provide a low-cost, easy-to-use, and user-friendly screening tool for clinical evaluation outside of the examination room.

Currently, the proposed OME detection system is a proof-of-concept prototype and still has several limitations. First, individual differences in pronunciation, sex, and age caused high variance and diversity in the voice recording. For example, the fundamental frequency was different for females and males. This challenge limits the OME detection performance. Second, previous works have demonstrated that different design and environmental factors affect system reliability, including background noise and wearing position [55]. Furthermore, it is necessary to validate the intra- and interrater reliability for OME detection. Finally, the proposed system did not classify the types of middle ear fluid, such as purulent, serous, and mucoid. These detailed disease types would support clinical professionals for more robust assessment and evaluation.

In future work, we will aim to recruit more patients from different age groups to validate the feasibility of the proposed detection systems. More types of OMs (e.g., acute otitis media and chronic otitis media) and their subtypes will be involved in developing a robust disease detection system. Various acoustic feature extraction techniques (e.g., wavelet transform [56] and empirical mode decomposition [57]) and advanced deep learning techniques (e.g., transformers [58] and self-supervised learning [59]) will be applied to improve the detection ACC and address the technical issues of individual differences. Furthermore, we plan to explore the impacts of extrinsic factors on the system performance, including wearing positions, background noise, and microphone type.

Telehealth increased during the coronavirus disease 2019 (COVID-19) pandemic, with many clinicians using video or telephone visits to provide medical consultation and to reduce the risk of exposure to the virus [60], [61], [62]. Our ultimate goal is to develop a smartphone-based OME detection system using derived algorithms and models. The use of smartphone-based telemedicine can provide healthcare practitioners with a simple, safe, and reliable tool to detect middle ear disease.

VI. CONCLUSION

To develop OME diagnostic tools for caregivers and primary clinicians in home/clinical environments, this study proposes a low-cost, easy-to-use, and reliable OME detection system using off-the-shelf microphones and ML models. The experimental results show that the proposed OME detection system can achieve 80.65% ACC, 66.67% SEN, 73.68% PRE, and 87.8% SPE. We demonstrate that using in-ear microphones is sufficient to support OME detection, and it can help to develop telemedicine in pandemic and postpandemic healthcare delivery.

APPENDIX

1) Tympanometry with a 226-Hz stimulus is a standard audiological procedure to evaluate middle ear function [63]. In the absence of any objects in the ear canal and with an intact eardrum, there are three main types of tympanograms: types A–C [64].
   a) A normal, or type A, tympanogram typically displays a peak in compliance that falls within the range of 0–100 dPa in the ear canal. Both the location and height of this peak must fall within the normal range on the pressure dimension for the recording to be considered normal.
   b) Type B tympanogram exhibits a flat pattern with little or no apparent change in compliance as pressure is applied to the ear canal, as opposed to a distinct peak in compliance. This pattern is frequently indicative of fluid within the middle ear space, known as OM.
   c) Type C tympanogram exhibits a compliance peak similar to that of type A recordings, except that

the peak falls within the negative pressure region beyond approximately 100 dPa. This pattern is typically observed in patients with eustachian tube dysfunction and insufficient ventilation of the middle ear space. It often precedes the acquisition of a type B tympanogram during the development of OM.

2) The mean of pure tone frequencies refers to the average of hearing threshold levels at a specific set of frequencies, typically including 500, 1000, 2000, and 4000 Hz.

References

[1] G. A. Zielhuis, G. H. Rach, A. Bosch, and P. Broek, "The prevalence of otitis media with effusion: A critical review of the literature," *Clin. Otolaryngol.*, vol. 15, no. 3, pp. 283–288, Jun. 1990.

[2] P. Shekelle et al., "Diagnosis, natural history, and late effects of otitis media with effusion," *Evidence Rep./Technol. Assessment*, vol. 10, no. 55, pp. 1–5, Jun. 2002.

[3] R. M. Rosenfeld et al., "Clinical practice guideline: Otitis media with effusion (update)," *Otolaryngol.-Head Neck Surg.*, vol. 154, no. 1, pp. 1–41, 2016.

[4] R. M. Rosenfeld and D. Kay, "Natural history of untreated otitis media," *Laryngoscope*, vol. 113, no. 10, pp. 1645–1657, Oct. 2003.

[5] V. A. Holm and L. H. Kunze, "Effect of chronic otitis media on language and speech development," *Pediatrics*, vol. 43, no. 5, pp. 833–839, May 1969.

[6] S. D. Emmett, J. Kokesh, and D. Kaylie, "Chronic ear disease," *Med. Clinics North Amer.*, vol. 102, no. 6, pp. 1063–1079, Nov. 2018.

[7] R. M. Rosenfeld et al., "Clinical practice guideline: Tympanostomy tubes in children," *Otolaryngol. Head Neck Surg.*, vol. 149, no. 1, pp. 1–35, Jul. 2013.

[8] W. S. Jones and P. H. Kaleida, "How helpful is pneumatic otoscopy in improving diagnostic accuracy?" *Pediatrics*, vol. 112, no. 3, pp. 510–513, Sep. 2003.

[9] W. J. Steinbach et al., "Pediatric residents' clinical diagnostic accuracy of otitis media," *Pediatrics*, vol. 109, no. 6, pp. 993–998, Jun. 2002.

[10] M. E. Pichichero and M. D. Poole, "Assessing diagnostic accuracy and tympanocentesis skills in the management of otitis media," *Arch. Pediatrics Adolescent Med.*, vol. 155, no. 10, pp. 1137–1142, 2001.

[11] P. H. Karma, M. A. Penttilä, M. M. Sipilä, and M. J. Kataja, "Otoscopic diagnosis of middle ear effusion in acute and non-acute otitis media. I. The value of different otoscopic findings," *Int. J. Pediatric Otorhinolaryngol.*, vol. 17, no. 1, pp. 37–49, Feb. 1989.

[12] C. Gonzalez and C. Bluestone, "Visualization of a retraction pocket/cholesteatoma: Indications for use of the middle ear telescope in children," *Laryngoscope*, vol. 96, pp. 109–110, Apr. 2009.

[13] S. Preyer, "Endoscopic ear surgery—A complement to microscopic ear surgery," *HNO*, vol. 65, no. S1, pp. 29–34, Jan. 2017.

[14] K.-C. Ting and T.-Y. Tu, "The application of a 70° endoscope in performing transcanal middle ear surgery," *J. Chin. Med. Assoc.*, vol. 84, no. 3, pp. 309–313, 2021.

[15] A. S. Shiao and Y. C. Guo, "A comparison assessment of videotelescopy for diagnosis of pediatric otitis media with effusion," *Int. J. Pediatric Otorhinolaryngol.*, vol. 69, no. 11, pp. 1497–1502, Nov. 2005.

[16] M. E. Pichichero and M. D. Poole, "Comparison of performance by otolaryngologists, pediatricians, and general practioners on an otoendoscopic diagnostic video examination," *Int. J. Pediatric Otorhinolaryngol.*, vol. 69, no. 3, pp. 361–366, Mar. 2005.

[17] P. Marchisio et al., "Acute otitis media: From diagnosis to prevention. Summary of the Italian guideline," *Int. J. Pediatric Otorhinolaryngol.*, vol. 74, no. 11, pp. 1209–1216, Nov. 2010.

[18] P. Abbott, S. Rosenkranz, W. Hu, H. Gunasekera, and J. Reath, "The effect and acceptability of tympanometry and pneumatic otoscopy in general practitioner diagnosis and management of childhood ear disease," *BMC Family Pract.*, vol. 15, no. 1, p. 181, Dec. 2014.

[19] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Inf.*, vol. 2, Jan. 2006, Art. no. 117693510600200030.

[20] J. H. Thrall et al., "Artificial intelligence and machine learning in radiology: Opportunities, challenges, pitfalls, and criteria for success," *J. Amer. College Radiol.*, vol. 15, no. 3, pp. 504–508, Mar. 2018.

[21] S. Gupta et al., "Machine-learning prediction of cancer survival: A retrospective study using electronic administrative records and a cancer registry," *BMJ Open*, vol. 4, no. 3, Mar. 2014, Art. no. e004007.

[22] Z. Obermeyer and E. J. Emanuel, "Predicting the future—Big data, machine learning, and clinical medicine," *New England J. Med.*, vol. 375, no. 13, pp. 1216–1219, Sep. 2016.

[23] M. Shew, J. New, and A. M. Bur, "Machine learning to predict delays in adjuvant radiation following surgery for head and neck cancer," *Otolaryngol.-Head Neck Surg.*, vol. 160, no. 6, pp. 1058–1064, Jun. 2019.

[24] G. L. Monroy et al., "Automated classification platform for the identification of otitis media using optical coherence tomography," *NPJ Digit. Med.*, vol. 2, no. 1, p. 22, Mar. 2019.

[25] M. Viscaino, J. C. Maass, P. H. Delano, M. Torrente, C. Stott, and F. A. Cheein, "Computer-aided diagnosis of external and middle ear conditions: A machine learning approach," *PLoS ONE*, vol. 15, no. 3, Mar. 2020, Art. no. e0229226.

[26] M. G. Crowson et al., "Machine learning for accurate intraoperative pediatric middle ear effusion diagnosis," *Pediatrics*, vol. 147, no. 4, pp. 1–15, Apr. 2021.

[27] R. G. Kashani et al., "Shortwave infrared otoscopy for diagnosis of middle ear effusions: A machine-learning-based approach," *Sci. Rep.*, vol. 11, no. 1, p. 12509, Jun. 2021.

[28] Z. Cao et al., "Machine learning in diagnosing middle ear disorders using tympanic membrane images: A meta-analysis," *Laryngoscope*, vol. 133, no. 4, pp. 732–741, Apr. 2023.

[29] J. Chan, S. Raju, R. Nandakumar, R. Bly, and S. Gollakota, "Detecting middle ear fluid using smartphones," *Sci. Transl. Med.*, vol. 11, no. 492, May 2019, Art. no. eaav1102.

[30] H. Binol et al., "Decision fusion on image analysis and tympanometry to detect eardrum abnormalities," *Proc. SPIE*, vol. 11314, pp. 375–382, Mar. 2020.

[31] E. M. Grais et al., "Analysing wideband absorbance immittance in normal and ears with otitis media with effusion using machine learning," *Sci. Rep.*, vol. 11, no. 1, p. 10643, May 2021.

[32] Y. Tsao and Y.-H. Lai, "Generalized maximum a posteriori spectral amplitude estimation for speech enhancement," *Speech Commun.*, vol. 76, pp. 112–126, Feb. 2016.

[33] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 370–379, Feb. 2011.

[34] G. Costantini et al., "Deep learning and machine learning-based voice analysis for the detection of COVID-19: A proposal and comparison of architectures," *Knowl.-Based Syst.*, vol. 253, Oct. 2022, Art. no. 109539.

[35] K.-C. Liu, K.-H. Hung, C.-Y. Hsieh, H.-Y. Huang, C.-T. Chan, and Y. Tsao, "Deep-learning-based signal enhancement of low-resolution accelerometer for fall detection systems," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 3, pp. 1270–1281, Sep. 2022.

[36] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020.

[37] M. Dahmani and M. Guerti, "Vocal folds pathologies classification using Naïve Bayes networks," in *Proc. 6th Int. Conf. Syst. Control (ICSC)*, 2017, pp. 426–432.

[38] Y. Cao, Q.-G. Miao, J.-C. Liu, and L. Gao, "Advance and prospects of AdaBoost algorithm," *Acta Automatica Sinica*, vol. 39, no. 6, pp. 745–758, Mar. 2014.

[39] S. Raval, R. Balar, and V. Patel, "A comparative study of early detection of Parkinson's disease using machine learning techniques," in *Proc. 4th Int. Conf. Trends Electron. Informat.*, 2020, pp. 509–516.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[41] J.-K. Wang et al., "Automatic recognition of murmurs of ventricular septal defect using convolutional recurrent neural networks with temporal attentive pooling," *Sci. Rep.*, vol. 10, no. 1, p. 21797, Dec. 2020.

[42] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[43] T. Tuncer, S. Dogan, and U. R. Acharya, "Automated detection of Parkinson's disease using minimum average maximum tree and singular value decomposition method with vowels," *Biocybernetics Biomed. Eng.*, vol. 40, no. 1, pp. 211–220, Jan. 2020.

[44] W. Xu et al., "mCOPD: Mobile phone based lung function diagnosis and exercise system for COPD," in *Proc. 6th Int. Conf. Pervasive Technol. Rel. Assistive Environ.*, 2013, pp. 1–8.

[45] B. McFee et al., "librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, 2015, pp. 18–25.

[46] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 10, pp. 2825–2830, Jul. 2017.

[47] G. Brown, "Ensemble learning," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA, USA: Springer, 2010, pp. 312–320.

[48] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken, NJ, USA: Wiley, 2014.

[49] R. Das and A. Sengur, "Evaluation of ensemble methods for diagnosing of valvular heart disease," *Expert Syst. Appl.*, vol. 37, no. 7, pp. 5110–5115, Jul. 2010.

[50] L. I. Kuncheva, J. J. Rodriguez, C. O. Plumpton, D. E. J. Linden, and S. J. Johnston, "Random subspace ensembles for fMRI classification," *IEEE Trans. Med. Imag.*, vol. 29, no. 2, pp. 531–542, Feb. 2010.

[51] E. Chiappini, B. Bortone, M. Doria, M. Landi, G. Di Mauro, and P. Marchisio, "What not to do in acute otitis media: The top five recommendations proposed by the Italian society of preventive and social pediatrics," *Expert Rev. Anti-Infective Therapy*, vol. 15, no. 10, pp. 897–902, Oct. 2017.

[52] N. Erkkola-Anttinen, M. K. Laine, P. A. Tähtinen, and A. Ruohola, "Can changes in parentally measured acoustic reflectometry levels predict the middle ear status?" *Int. J. Pediatric Otorhinolaryngol.*, vol. 95, pp. 72–74, Apr. 2017.

[53] H. Y. Lee, M. S. Choi, N. S. Park, and C. S. Cho, "Applicability of spectral gradient acoustic reflectometry (EarCheck): Screening patients who need surgical treatment for chronic otitis media with effusion," *Int. J. Pediatric Otorhinolaryngol.*, vol. 78, no. 1, pp. 107–111, Jan. 2014.

[54] J. Abbasi, "Smartphone-based detection of middle ear fluid," *JAMA*, vol. 322, no. 2, p. 107, 2019.

[55] Y. Gao, W. Wang, V. V. Phoha, W. Sun, and Z. Jin, "EarEcho: Using ear canal echo for wearable authentication," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–24, Sep. 2019.

[56] S.-S. Wang, P. Lin, Y. Tsao, J.-W. Hung, and B. Su, "Suppression by selecting wavelets for feature compression in distributed speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 3, pp. 564–579, Mar. 2018.

[57] Y. D. Lin, Y. K. Tan, and B. Tian, "A novel approach for decomposition of biomedical signals in different applications based on data-adaptive Gaussian average filtering," *Biomed Signal Process Control*, vol. 71, Jan. 2022, Art. no. 103104.

[58] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[59] R. E. Zezario, T. Hussain, X. Lu, H.-M. Wang, and Y. Tsao, "Self-supervised denoising autoencoder with linear regression decoder for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6669–6673.

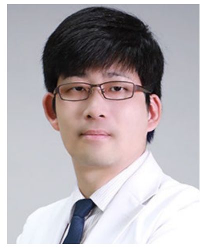
[60] Z. S. Predmore, E. Roth, J. Breslau, S. H. Fischer, and L. Uscher-Pines, "Assessment of patient preferences for telehealth in post–COVID-19 pandemic health care," *JAMA Netw. Open*, vol. 4, no. 12, Dec. 2021, Art. no. e2136405.

[61] L. M. Koonin et al., "Trends in the use of telehealth during the emergence of the COVID-19 pandemic—United States, January–March 2020," *Morbidity Mortality Weekly Rep.*, vol. 69, no. 43, pp. 1595–1599, Oct. 2020.

[62] L. Hsueh et al., "Disparities in use of video telemedicine among patients with limited English proficiency during the COVID-19 pandemic," *JAMA Netw. Open*, vol. 4, no. 11, Nov. 2021, Art. no. e2133129.

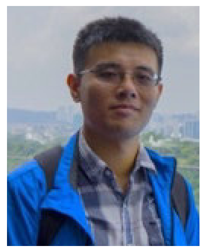
[63] Y. W. Liu et al., "Wideband absorbance tympanometry using pressure sweeps: System development and results on adults with normal hearing," *J. Acoust. Soc. Amer.*, vol. 124, no. 6, pp. 3708–3719, Dec. 2008.

[64] E. Onusko, "Tympanometry," *Amer. Family Physician*, vol. 70, no. 9, pp. 1713–1720, Nov. 2004.



Kuan-Chung Ting received the M.D. degree from National Defense Medical Center, Taipei, Taiwan, in 2011.

He currently serves as an Otologist with the Taipei Veterans General Hospital, Taipei, specializing in diagnosing and treating various ear-related disorders. His clinical expertise encompasses a wide range of ear conditions and is committed to provide high-quality patient care. His research endeavors focus on the intersection of otology and artificial intelligence, where he explores innovative applications of AI in the field of ear sciences. By leveraging advanced AI techniques, he aims to enhance the accuracy of diagnostics, optimize treatment strategies, and contribute to the overall advancement of ear healthcare.



Syu-Siang Wang received the bachelor's degree in electrical engineering from National Changhua Normal University, Changhua, Taiwan, in 2008, the master's degree in electrical engineering from National Jinan University, Nantou, Taiwan, in 2010, and the Ph.D. degree from the Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan, in 2018.

During the Ph.D. studies, he served as a Research Assistant with the Information Technology Innovation Research Center, Academia Sinica, Taipei. From 2012 to 2018, focusing on robust speech feature extraction and enhancement. After completing the Ph.D., he conducted Postdoctoral Research with the NTU Joint Research Center for AI Technology and All Vista Healthcare, Singapore, where he worked on hearing aid signal processing technology. Currently, he holds the position of Assistant Professor with the Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan. His research interests include encompass speech signal processing and biological signal processing.

You-Jin Li received the B.S. degree from the Department of Electronic Engineering, National Ilan University, Ilan, Taiwan, in 2014, and the M.S. degree from the Department of Electrical Engineering With Communications, National Ilan University, in 2016. He is currently pursuing the Ph.D. degree with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan.

His research interests include signal processing, speech enhancement, beamforming, deep learning, and multichannel compression.

Chii-Yuan Huang received the M.D. degree from Taipei Medical University, Taipei, Taiwan, in 1991, and the Ph.D. degree in biomedical engineering from National Cheng Kung University, Tainan, Taiwan, in 2006.

Then, he fully trained as an Otolaryngologist Surgeon with Taipei Veteran General Hospital and continuing his clinical research interests. He was an Assistant Professor with the Medical School of National Yang-Ming University for many years. His research interests include developing of digitalized-interactive evaluation system platform for Mandarin speech and background noise.

Dr. Huang is the Chairperson of Taiwan Tinnitus Association.

Tzong-Yang Tu received the M.D. degree in medicine from Taipei Medical University, Taipei, Taiwan, in 1987, and the Ph.D. degree from the Institute of Clinical Medicine, Yang Ming University, Taipei, in 1998.

He is a Senior Surgeon and the former Chief in Otology with Taipei Veterans General Hospital, Taipei, and National Yang-Ming University, Taipei. His laboratory research interests include the ion transport of the stria vascularis in the cochlea and the processes of osteochondrogenesis and ossification in the cartilage obliterated in the mastoid. He has achieved several important results in these fields. His significant clinical experiences include minimally invasive cholesteatoma surgery and microscopic-endoscopic dual-system middle ear surgery.

Chun-Che Shih received the M.D. degree from China Medical University, Taichung, Taiwan, in 1987, and the Ph.D. degree from the Institute of Clinical Medicine, National Yang Ming Jiao Tong University, Taipei, Taiwan, in 2000.

From 1989 to 2019, he worked as a Cardiovascular Surgeon and served as the Chief of the Cardiovascular Division, from 2012 to 2018. He also held positions at the Institute of Clinical Medicine as an Assistant Professor, from 2001 to 2004 and an Associate Professor from 2004 to 2007, and was promoted as a Professor, 2007. In 2018, he transitioned to Taipei Medical School as the Director of the Institute of Clinical Medicine and Medical Deputy Superintendent. Since then, he has been a Professorship at the School of Medicine at Taipei Medical University, Taipei. His research interests include in nanoparticle and regenerative medicine in the context of cardiovascular and aortic diseases.

Kai-Chun Liu (Member, IEEE) received the M.S. and Ph.D. degrees in biomedical engineering from National Yang-Ming University, Taipei, Taiwan, in 2015 and 2019, respectively.

He is currently a Postdoctoral Scholar with the Research Center for Information Technology Innovation, Academia Sinica, Taipei. His research interests include pervasive healthcare, wearable computing, machine learning, and biosignal processing.

Yu Tsao (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1999 and 2001, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2008.

From 2009 to 2011, he was a Researcher with the National Institute of Information and Communications Technology, Tokyo, Japan, where he engaged in research and product development in automatic speech recognition for multilingual speech-to-speech translation. He is currently a Research Fellow (Professor) and the Deputy Director with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. He is also a Jointly Appointed Professor with the Department of Electrical Engineering, Chung Yuan Christian University, Taoyuan, Taiwan. His research interests include assistive oral communication technologies, audio coding, and biosignal processing.

Dr. Tsao was a recipient of the Academia Sinica Career Development Award in 2017, National Innovation Awards from 2018 to 2021, Future Tech Breakthrough Award 2019, Outstanding Elite Award, Chung Hwa Rotary Educational Foundation from 2019 to 2020, NSTC Future Tech Award 2022, and Young Author, Best Paper Award. He is the corresponding author of a paper that received the 2021 IEEE Signal Processing Society (SPS). He is currently an Associate Editor for the IEEE/ACM Transactions on Audio, Speech, and Language Processing and IEEE Signal Processing Letters.