IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 May 2023, accepted 19 May 2023, date of publication 24 May 2023, date of current version 6 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3279502

RESEARCH ARTICLE

Person Identification Using Bronchial Breath Sounds Recorded by Mobile Devices

VAN-THUAN TRAN[1], YIH-LON LIN[2], AND WEI-HO TSAI[1], (Member, IEEE)
[1]Department of Electronic Engineering, National Taipei University of Technology, Taipei City 10608, Taiwan
[2]Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology, Douliu City 64002, Taiwan

Corresponding author: Wei-Ho Tsai (whtsai@ntut.edu.tw)

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 111-2221-E-027-136.

ABSTRACT This study examines the use of breath sounds intrusively recorded by mobile devices for person identification (PID), which is referred to as mobile-sensed BreathPID. A custom dataset of breath sounds from 21 volunteers is prepared for investigation and analysis. To overcome the problem of spare training data, we incorporate various audio data augmentation (DA) methods with the self-supervised learning (SSL) approach to train the BreathPID's learning models. SSL-based models for BreathPID are developed in two phases: firstly, solving proposed pretext task(s) without identity information to effectively learn core characteristics of data; then further finetuning the models on the labeled data for the downstream BreathPID task. Several types of pretext or auxiliary tasks are investigated. First, when considering each DA technique, the pretext task is defined as the detection of augmentation levels, for instance, the levels of noise added to original data samples. When utilizing multiple DA techniques, the identification of DA types is defined as the pretext task. In addition, various issues in developing robust BreathPID systems are taken into consideration, including network design, changes in input length, and the ability of noise resistance. From the experimental results, we find that SSL-based BreathPID with the combined use of four DA techniques (i.e., noise addition, speed changing, time shifting, and spectrogram masking) achieves promising results which are higher than those of SSL-based models using single DA technique and those of typical supervised models. Also, the proposed system shows good resistance to noise effects and changes in the input size. Mobile-sensed BreathPID achieves the equivalent or superior results compared to stethoscope-sensed BreathPID (where breath sounds are sensed using specialty stethoscopes). The proposed approach can be applied to the authentication function or health monitoring applications on mobile devices.

INDEX TERMS Audio-based identification, breath sounds, biometrics, data augmentation, mobile devices, neural networks, person identification, representation learning, self-supervised learning.

I. INTRODUCTION

Person identification (PID), which refers to the process of recognizing and verifying the identity of an individual, plays a critical role in the digital world. In recent years, there has been a significant increase in the demand for PID systems due to the growing concerns over security threats and the need for access control in various fields such as banking, healthcare, and law enforcement. To date, various methods for PID have been developed, ranging from traditional knowledge-based methods (e.g., password and personal identification number) and token-based methods (e.g., ID card, driving license, and member card) to more advanced biometric methods like voice and iris recognition. These approaches are widely used to identify individuals with an acceptable degree of accuracy and speed, making them an essential tool in maintaining security and protecting sensitive information. However, existing methods also face several challenges such as security threats, privacy concerns, and robustness. Also, each method may not be flexibly applied for all applications. Thus, there is still room for the improvement of existing systems or the development of brand-new

The associate editor coordinating the review of this manuscript and approving it for publication was Xuebo Zhang.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

approaches. This work examines biometric-based PID using breath sounds intrusively recorded by mobile devices, rather than using widely-used biometrics like voice, face, signature, or fingerprint.

Breath sounds are the noises produced by the movement of air as it flows in and out of the lungs during breathing. Depending on the measurement positions, breath sounds may have particular characteristics and can be categorized into vesicular breath sounds (VBS) and bronchial breath sounds (BBS). VBS is soft and low-pitched and can be heard over the chest wall, while BBS is heard when air moves through the larger airways of the lungs, such as the bronchi, and the sound is louder and higher-pitched [1]. BBS can be heard over the upper part of the chest, near the trachea. Conventionally, breath sounds are sensed with the use of specialty stethoscopes and can provide important information about the functioning of the respiratory system, which is especially useful for medical examination and treatment activities. Recently, stethoscope-sensed bronchial breath sounds have been found as a new biometric trait for PID due to its unique and stable characteristics that can be used to identify individuals [2], [3]. The stethoscope-based BreathPID is particularly suitable to use by specialists or doctors in medical applications. In some situations, using stethoscopes to capture breath sounds could be inconvenient, thus limiting the deployment of the stethoscope-sensed BreathPID to practical use. This work aims to investigate the use of bronchial breath sounds sensed by mobile devices for PID, from which the applications of BreathPID can be extended significantly.

Mobile-sensed BreathPID is a convenient, cost-effective, and secure method. Mobile devices such as smartphones are ubiquitous, and most people carry them at all times. Using mobile devices to record breath sounds for identification purposes can be very convenient for the individual being identified, as it eliminates the need to carry any additional devices or equipment like stethoscopes and signal processing units. Generally, BreathPID offers several advantages compared to other biometric-based methods. Since breath sounds are the most ubiquitous BreathPID can be used for everyone. The invasive recording of breath sounds can significantly reduce the effects of external factors like ambient noise, so the signal could be more consistent over time. BreathPID is also secure because breath sounds are almost inaudible to non-intrusive devices and difficult to replicate. By contrast, knowledge-based and token-based PID systems are less secure as information like passwords and things like smart cards can be stolen or lost. Meanwhile, other biometric-based systems require certain conditions to be applicable and to well operate. For example, fingerprint recognition could be unstable for subjects with unclear fingerprints or even disabilities, voice recognition could be severely affected by loud noise, and low image quality can degrade the accuracy of facial recognition.

In summary, this works makes the following major contributions:

- We examine BreathPID with bronchial breath sounds recorded by mobile devices rather than specialty devices like stethoscopes, improving the applicability of BreathPID and opening up new avenues for further research in the field of biometric-based PID to develop secure and privacy-preserving PID systems.
- Due to the lack of published datasets, custom datasets are collected for experiments and analysis. It is widely acknowledged that collecting a large amount of labeled data for training performant deep networks is a challenging task. Thus, we propose to apply self-supervised learning (SSL) approaches with audio data augmentation (DA) to resolve the problem of spare training data. Pretext tasks, including the identification of DA methods and identification of augmentation levels, are designed for self-supervised representation learning, in which DA techniques can be employed independently or in combination. Then, the resulting models are finetuned on labeled breath sounds to solve the downstream PID task.
- Different experimental aspects (e.g., accuracy and robustness of the proposed PID approach, and effectiveness of data augmentation techniques) are taken into consideration. The experiment results show that the proposed SSL-based mobile-sensed BreathPID outperforms the standard supervised frameworks, in which the SSL-based system with a smaller size yields better accuracies and robustness compared to the performances of baseline models. Also, mobile-sensed BreathPID achieves equivalent or superior results compared to stethoscope-sensed BreathPID.
- The results of this work can bring about potential applications, such as in biometric authentication and healthcare service. By analyzing the unique patterns in a person's breath sounds, mobile devices could be used to identify and authenticate individuals for various purposes like accessing secure applications or making financial transactions. In healthcare, mobile-sensed BreathPID could be used to monitor patients with respiratory diseases, track their progress, and adjust their treatment plans accordingly.

The remainder of this paper is organized as follows. Section II provides an overview of related works. Section III analyzes the methods we use to develop the mobile-sensed BreathPID system. Then, we present the experimental results in Section IV and provide a conclusion in Section V.

II. RELATED WORKS

Biometric-based PID [4] refers to the identification of an individual based on the unique biometric data captured from that person. There are two major types of biometrics, including physiological and behavioral biometrics. The former refers to the analysis of a person's physical characteristics such as fingerprints, iris or retina scans, face, and DNA. The latter, on the other hand, includes the analysis of patterns in human

activities like signature, voice, gait, keystroke, mouse activity, and lip motion. Depending on the types of biometric data used for analysis, biometric-based PID systems can be also categorized into vision-based biometrics (i.e., using fingerprint [5], face [6], [7], or iris [8] images), audio-based biometrics (e.g., using voice recordings [9]), audio-visual biometrics (e.g., using face images and voice recording in combination [10], [11]), bioelectrical biometrics (e.g., ECG [12], and EEG [13]), and others. BreathPID belongs to the category of audio-based biometric systems as we aim to employ breath sounds recorded by the mobile device as the input data for analysis.

To date, most of the works in audio-based PID focus on the use of verbal voices, including speech (e.g., in [14], [15], [16]) and singing voices (e.g., in [17], [18], [19]) to analyze and establish identity. A few works have investigated PID based on non-verbal voices, which are the sounds produced by speakers using their vocal organs that do not have linguistic content. For example, Bachorowski et al. [20] showed that individual identity is conveyed in laugh acoustics as they employed laughter to classify speakers using an automatic approach and achieved above-chance level accuracy. Engelberg et al. [21] found that participants were able to discriminate between speakers from scream stimuli. It was found that some other sources of audio, such as heart sounds and breath sounds, can be also useful for identification. Reference [22] confirmed the biometric properties of heart sound signals, which can thus be included among the physiological signs used by an automatic identification system.

Recently, some studies have shown sounds of human breath as a potential biometric [2], [3], [23], [24], [25], [26]. Reference [23] extracted breath sounds during inhalations from speech corpus for breath-based PID experiments with various learning models, in which conventional machine learning ones achieved accuracies of less than 75% while CNN-RNN yielded a higher result of 91.3%. Reference [24] proposed a sophisticated scheme comprises of breath demarcation, feature extraction, and feature matching, which achieved better performance compared to [23]. However, systems in [23] and [24] required the step of breath demarcation from original speech recordings, resulting in extra complexity. Also, those works conducted experiments on clean datasets without the report of robustness evaluation. Chauhan et al. [25], [26] proposed to sense the sounds near an individual's nose (i.e., measure the sounds unintrusively 1-2cm under the nose) by mobile phones for identification using Gaussian mixture models (GMM) [25] and recurrent neural networks (RNN) [26]. The stability of systems in [25] and [26] is compromised, as their performance experienced a significant degradation when the measurement distance exceeded 2cm and when functioning under noisy conditions. In [2] and [3], respectively, we proposed the use of intrusive stethoscope-sensed bronchial breath sounds (BBS) [2] as well as the combined use of BBS and speech [3] for PID. This work investigates BreathPID based on BBS sensed intrusively by mobile devices rather than specialty stethoscopes, thus extending the applicability of breath-based PID. BreathPID based on mobile-sensed BBS is more flexible than the stethoscope-sensed approach [2], [3], while intrusive measurement can help improve the security, mitigate the influence of ambient noise, and avoid the problem of performance degradation due to changes in the distance of unintrusive measurement [25], [26].

In prior works, various automatic methods have been developed for audio-based PID, including similarity-based schemes [24], traditional machine learning methods (e.g., GMM [25], [27], [28] and support vector machines (SVM) [3], [23], [29]), and deep learning methods (e.g., convolutional neural networks (CNN) [30], recurrent neural networks (RNN) [26], [31], and CNN-RNN [23]). These methods utilized handcrafted features like Mel-frequency cepstral coefficients (MFCCs) and spectrograms to represent audio inputs and further processed them to extract discriminative information for identification tasks. Compared to traditional machine learning-based systems, deep learning-based PIDs offer several advantages in terms of accuracy, robustness, scalability, adaptability, and efficiency. This is because deep networks can learn complex patterns and relationships in data that traditional machine-learning methods may struggle to detect. Additionally, deep networks can learn to identify the most important features from the data and eliminate irrelevant information. However, developing efficient deep-leaning-based systems typically requires a large amount of training data, which may not always be feasible due to the cost and time involved in data collection, particularly for emerging research topics.

To address the issue of limited training data, various techniques have been developed, including data augmentation, transfer learning, regularization, generative models, ensemble learning, self-supervised learning, and contrastive learning. Self-supervised learning (SSL) is a technique in which a model learns from the data without explicit supervision. In SSL, the model learns to predict a specific aspect of the data, such as the missing word within a sentence [32], the next sentences based on the current one [32], or the rotation angle of an image [33], without being given explicit labels for these tasks. SSL is particularly useful when labeled data is scarce or expensive to obtain. By leveraging the inherent structure and patterns in the data, SSL can enable models to learn useful representations that can be applied to downstream tasks such as classification or regression. SSL has achieved significant success in natural language processing [32], vision domain [34], [35], and audio domain [19], [36] as well. In this stage of investigation on mobile-sensed bronchial breath sounds for PID, the goal is to collect a custom dataset of moderate size for experiments and analysis. Thus, in addition to typical supervised training of deep networks (i.e., which is similar to the prior works), we apply the combined use of various techniques consisting of data augmentation,

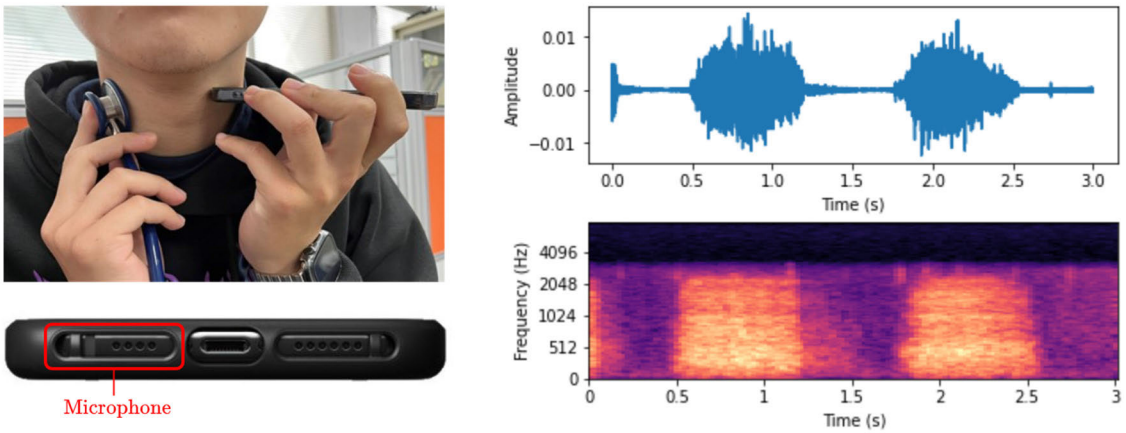

FIGURE 1. (left) Acquisition of breath sounds using mobile device and stethoscope; (right) waveform and spectrogram of a breath sound sample recorded by mobile phone.

regularization, and self-supervised learning to enhance identification accuracy.

III. METHODOLOGY

A. DATA COLLECTION

Recall that bronchial breath sounds (BBS), which are loud and high-pitched, can be heard along the large airway. Thus, to collect experimental data, we choose to sense BBS from the front side of the subject's neck which belongs to the recommended BBS measurement areas [1], [37] and brings about convenience for the collection process. In contrast to sensing vesicular breath sounds (VBS) over the chest wall, sensing data nearby the neck can be done without any obstacles, such as clothing. The selected measurement position is also well-suitable for practical use.

Figure 1 (left) illustrates the collection of BBSs. It is worth noting that the microphones of smartphones are commonly hiding behind a collection of small holes on the bottom frame. Thus, we can place the bottom side of phones on a subject's neck to capture BBSs easily. We also collect parallel data using stethoscopes, in which the chest-piece of the stethoscope will be used to sense breath sounds from the other side of the subject's neck, and signals are recorded by a microphone plugged into one of the stethoscope's earpieces. The detailed setup for data collection using stethoscopes can be found in [2], [3]. The mobile-sensed and stethoscope-sensed recordings are collected simultaneously, resulting in the mobile-sensed BBS dataset (M-BBS-DS) and stethoscope-sensed dataset (S-BBS-DS), respectively. The S-BBS-DS dataset is utilized for comparative analysis regarding the performances of mobile-sensed BreathPID and stethoscope-sensed BreathPID. The right-hand side of Figure 1 shows the waveform and spectrogram of a mobile-sensed BBS recording which is 3 seconds in length.

We invited 21 volunteers, including 11 female and 10 male subjects in different age groups, to participate in the collection of our experimental data. It should be noted that all participants were in their typical state of health during the time of data collection. Each volunteer provided 40 BBS recordings which could be recorded in dis-continuous days and after different physical activities such as walking, running, and stair climbing. Each recording lasted for around 15 seconds. The detail of experimental data is provided in Table 1. For each subject, 30 and 10 out of 40 recordings were utilized for the training and testing phases, respectively. We further split original long recordings into smaller non-overlapping samples of 3 seconds which is equivalent to the average length of a breathing cycle and is a more suitable length for practical applications. As a result, the entire M-BBS-DS dataset contains 4,801 samples consisting of 3,686 and 1,115 samples for development and evaluation, respectively. Using a similar data preparation procedure, training and testing subsets of the S-BBS-DS dataset has 3,537 and 1,034 samples, respectively.

TABLE 1. Detail of experimental datasets.

Subset	#Recordings	Lengths	Sampling rates	Number of 3s samples	
				M-BBS-DS	S-BBS-DS
Train	30 × 21 = 630	around 15s	16-48kHz	3,686	3,537
Test	10 × 21 = 210	around 15s	16-48kHz	1,115	1,034
All	40 × 21 = 840	around 15s	16-48kHz	4,801	4,571

B. BUILDING BREATHPID SYSTEM USING THE SELF-SUPERVISED LEARNING APPROACH

1) SELF-SUPERVISED LEARNING (SSL)

Typical deep network training requires large-scale labeled data to learn general features and achieve better performance and generality. With more and more sophisticated architectures and huge datasets, recent networks keep outperforming state-of-the-art counterparts for almost all domains. However, it is an undeniable fact that collecting and annotating a huge amount of data is expensive and sometimes

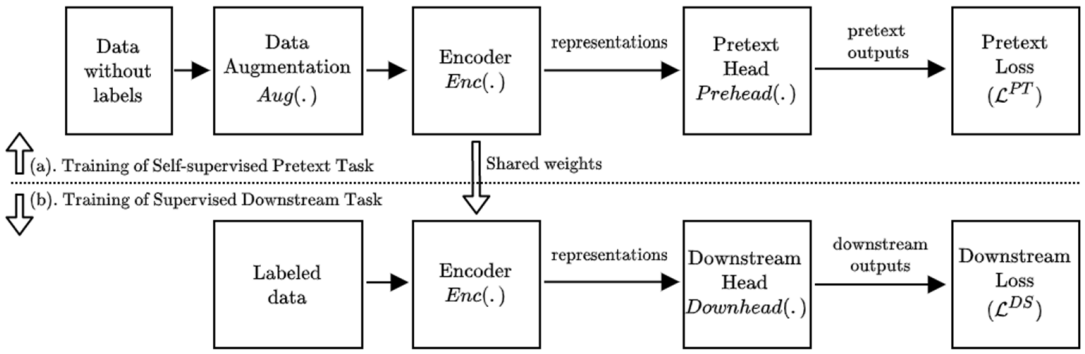

FIGURE 2. The general concept of SSL.

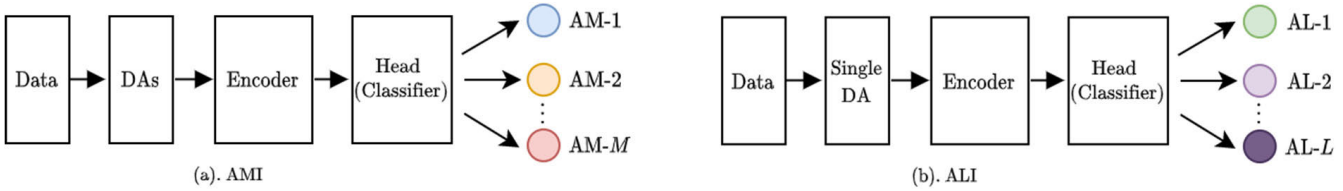

FIGURE 3. Two SSL pretext tasks: (a) augmentation method identification (AMI); (b) augmentation level identification (ALI). DA, AM, and AL stand for data augmentation, augmentation method, and augmentation level, respectively. *M* denotes the total number of DA methods, and *L* indicates the number of DA levels.

infeasible. As a subset of unsupervised learning methods, self-supervised learning (SSL) was proposed to learn prototypical representation from unlabeled data, thus eliminating the time-consuming process of data collection and annotation. In SSL, the common solution for feature learning is to propose pretext or auxiliary tasks for networks to solve, and features are learned while networks are trained with corresponding objective functions of pretext tasks.

Figure 2 illustrates the general concept of SSL, in which the development process includes 2 training phases. For the self-supervised training phase (Figure 2. a), we define a pretext task for a deep network to solve, and the pseudo labels for this task are generated based on some attributes of data or using some hand-designed methods (e.g., transformations) on unlabeled data. It is worth noting that the generation of pseudo-labels does not require any human annotation. An objective function $\mathcal{L}^{PT}$ for the pretext task is also defined, and the network is trained to learn this objective function. Assuming that we utilize data augmentation $Aug\,(\cdot)$ to transform an original sample into an augmented version corresponding to a pseudo label. Given a batch of N data instances $B \equiv \{X_i\}_{i=1}^{N}$, the pretext task provides a set of N pseudo labels $O = \{P_i\}_{i=1}^{N}$, in which P_i is the pseudo label associated with the augmented signal $X_i^{aug} = Aug(X_i)$ of the i-th data instance X_i in B. The deep network for the pretext task is comprised of a feature extraction structure called encoder $Enc\,(\cdot)$ followed by a head namely $Prehead\,(\cdot)$. Let θ^{PT} be parameters of the pretext task network, we train the network to minimize the loss defined by (1). The pre-trained encoder received from the self-supervised training step is transferred to the training of supervised downstream task (Figure 2. b), in which a separate loss function $\mathcal{L}^{DS}$ for the downstream task is utilized.

$$\begin{aligned}\mathcal{L}^{PT}(B) &= \min_{\theta^{PT}} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}^{PT}\left(F\left(X_i^{aug}, \theta^{PT}\right), P_i\right) \\ &= \min_{\theta^{PT}} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}^{PT}\left(Prehead\left(Enc\left(Aug\left(X_i\right)\right)\right), P_i\right)\end{aligned} \tag{1}$$

The network for downstream task also consists of an encoder, which is identical to the encoder $Enc\,(\cdot)$ in the first training step, and a head. Since the main task of this work is BreathPID, the downstream head $Downhead\,(\cdot)$ is a classifier with several fully-connected (FC) layers added to the end, among which the last FC layer has C nodes corresponding to the number of classes or subjects. The downstream head/classifier is trained on the outputs of the frozen pre-trained encoder $Enc\,(\cdot)$ using the cross-entropy loss $\mathcal{L}^{CE}$ (i.e., $\mathcal{L}^{DS} \equiv \mathcal{L}^{CE}$). The loss for each input sample X_j is calculated using the probability distribution vector $\hat{\mathbf{y}}_j = Downhead(Enc(X_j))$ generated by the softmax activation on the outputs of the last FC layer and the corresponding true probability distribution vector $\mathbf{y}_j$. Let $\mathbf{y}_j$ be the one-hot vector from the true label, the cross-entropy loss for a training sample X_j can be expressed by equation (2). Here, $\hat{y}_{j,c}$, the c^{th} element of $\hat{\mathbf{y}}_j$, indicates the probability that the training sample belongs to the c^{th} class. The c^{th} element $y_{j,c}$ of the one-hot vector $\mathbf{y}_j$ is 0 or 1, indicating whether the c^{th} class is the correct label or not. C denotes the number of classes or

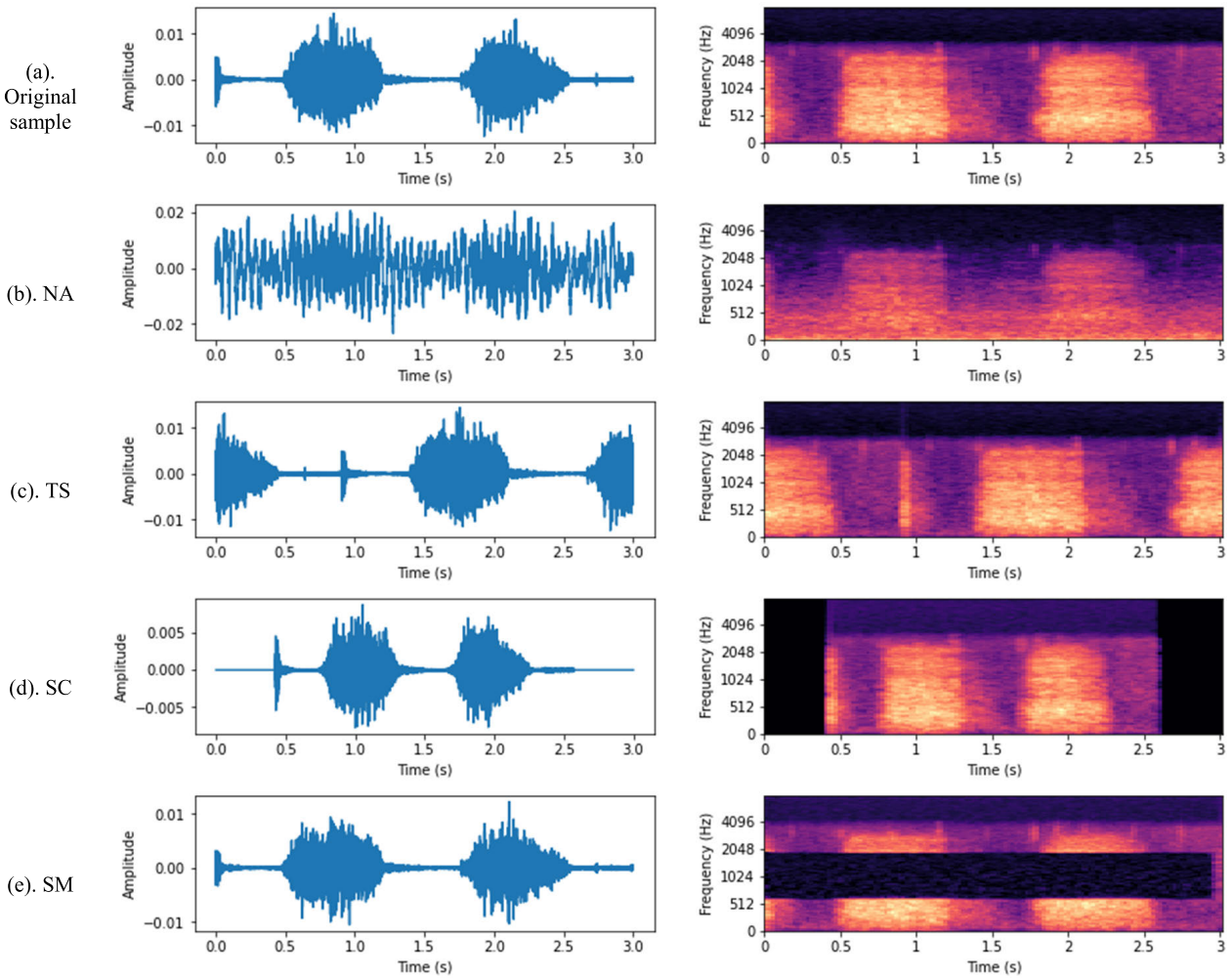

FIGURE 4. Waveforms and spectrograms of a breath sound file and augmented signals.

subjects in the dataset.

$$\mathcal{L}^{CE} = -\sum_{c=1}^{C} y_{j,c} \log \hat{y}_{j,c} \tag{2}$$

2) PRETEXT TASKS FOR SSL-BASED BREATHPID

Inspiring by the pretext task proposed in [38], which is to identify the correct orientation of images (i.e., training a network to recognize the 2D rotation applied to the image that the network gets as input), we apply a similar idea to the audio domain, in which four data augmentation (DA) methods consisting of noise addition (NA), time shifting (TS), speed changing (SC), and spectrogram masking (SM) are utilized. Specifically, we investigate two pretext tasks for SSL-based BreathPID, including augmentation method identification (AMI) and augmentation level identification (ALI). Here, the augmentation level is referred to as the parameter used to perform a certain augmentation. For example, the signal-to-noise (SNR) ratio we utilize to mix a data instance with a noise recording in noise-addition augmentation, or ''faster'' and ''slower'' speeds in speed-changing augmentation.

The AMI pretext task (Figure 3. a) is useful to evaluate the efficiency of using various DA methods in combination. During self-supervised training based on the AMI task, pseudo labels (i.e., NA, TS, SC, or SM) are automatically generated for input samples using four aforementioned DA methods, and the network is trained to maximize the rate of correct identifications. The second type (Figure 3. b) of pretext task (i.e., ALI) is applied separately to each DA method. Thus, we can evaluate the effectiveness of every DA method in SSL-based BreathPID, which could provide us with good suggestions regarding the selection of the DA method combination. It is worth noting that both AMI and ALI are considered as classification tasks, so the pretext heads are classifiers and we employ cross-entropy as the loss function for the training of those classifiers. The benefits of defining pretext tasks based on data augmentation can be listed as follows. It is useful when the amount of training data is small or not large enough because applying DAs results

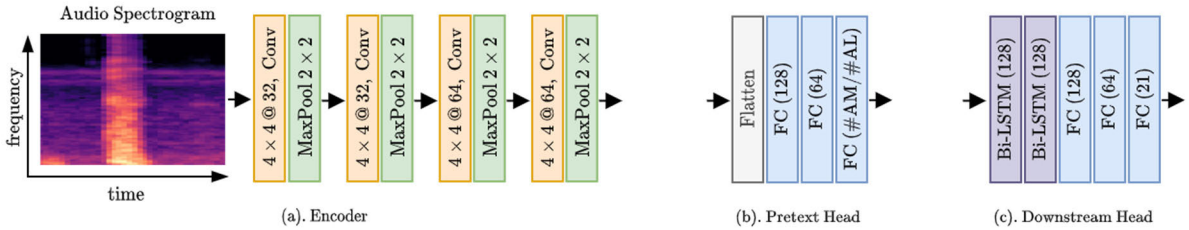

FIGURE 5. Structures of the encoder (a), pretext head (b), and downstream head (c).

in more data for training, which helps to avoid overfitting. By solving pretext tasks, the network learns to extract useful features of the sound and identify possible transformations applied to the signal, so the network can detect high-level semantics of the data. Also, augmentation provides a more diverse set of training data, so the trained network can become invariant to certain changes made to the input signal.

3) DATA AUGMENTATION

Figure 4 shows the waveforms and spectrograms of an original audio file and its augmented versions using noise addition (NA), time shifting (TS), speed changing (SC), and spectrogram masking (SM) augmentation techniques which are described as follows:

Noise Addition (NA): an original sample is mixed with a random noise recording according to a signal-to-noise (SNR) ratio, producing an augmented signal. The noise sources can come from soundscapes that are closely related to BreathPID applications, such as at offices, supermarkets, and on the street. The pseudo labels for NA augmentation can be denoted by NA-dB where dB represents the SNR ratio. For example, NA-m5 and NA-p5 are the pseudo labels for augmented signals of -5dB and $+5$dB, respectively.

Time Shifting (TS): we shift the values of a signal either right or left by a random number (i.e., K) of data points, ranging from 30% to 70% of the signal size (i.e., S). Two types of shifting are examined, including cycling TS and zero-padding TS. If a signal is shifted to the right by K data points, it is first separated into two parts of $(S-K)$ and K data points, respectively. In cycling TS, the second part of K data points is placed in front of the first part to form the augmented sample. On the other hand, in the zero-padding TS, we pad K zeros to the left of the first part to achieve an augmented signal of size S. The procedure for left shifting is reversed. The set of pseudo labels for TS augmentation includes TS-left and TS-right.

Speed Changing (SC): In this method, the speed of the audio sample is changed according to a random rate. As a result, the augmented sound is faster or slower than the original one. It is worth noting that the length of the augmented signal could be different from that of the original one, so we perform zero padding or cropping on the augmented file to obtain the same size as the original signal, which facilitates the phase of model training. The set of pseudo labels for SC augmentation includes SC-fast and SC-slow.

Spectrogram Masking (SM): The audio spectrogram is randomly masked along the time and/or frequency dimensions. This method provides three types of deformations consisting of time masking, frequency masking, and time-frequency masking. Thus, the set of pseudo labels for SM augmentation includes SM-time, SC-freq, and SM-timefreq.

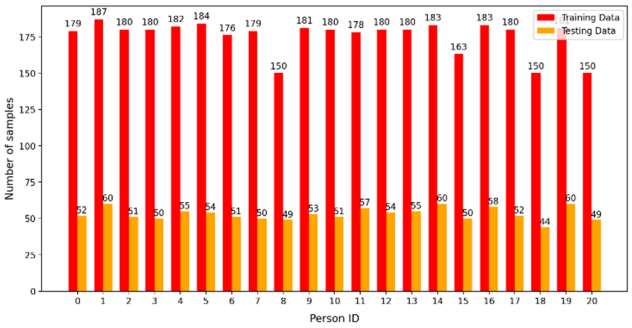

FIGURE 6. Distribution of training and testing data in the M-BBS-DS dataset.

IV. EXPERIMENTS AND RESULTS

A. EXPERIMENT SETUP

The experimental dataset was collected by different mobile devices, and the sampling rate of an audio recording can be different from that of one another, so we resample original recordings to 16 kHz and mono channel. To prepare 2D representations for training 2D convolutional neural networks (2DCNNs), we convert the audio signal into Mel-spectrogram, the time-frequency representation, using 128 filterbanks, window size of 25ms (i.e., 400 data points at a sampling rate of 16 kHz), and hop length of half window size or 200 data points. Thus, for an audio sample of 3 seconds, we receive the spectrogram of shape 128×241. Here, $241 = \lceil (3 \times 16,000)/200 \rceil$, where $\lceil \cdot \rceil$ is the ceiling function, is the number of 50% overlapping frames per 3-second recording.

We develop and assess the effectiveness of the CNN-based encoder, pretext head, and downstream head, which are illustrated in Figure 5. The encoder comprises four convolutional blocks, each of which has a 2D-Conv layer with a kernel size of 4×4 and a max-pooling layer with a size of 2×2. The four 2D-Conv layers have 32, 32, 64, and 64 filters, respectively.

TABLE 2. Results of SSL pretext tasks with 2DCNN model. The "Org" pseudo label indicates the original sample (i.e., w/o augmentation).

Model	Pretext task	Augmentation	Pseudo labels	Accuracy (%)	F1-score	Precision	Recall
2DCNN	Augmentation level identification (ALI)	Cycling time shifting (TS)	TS-left, TS-right, Org	53.99	48.61	63.66	39.59
2DCNN	ALI	Zero-padding TS	TS-left, TS-right, Org	89.75	89.64	91.32	88.06
2DCNN	ALI	Speed change (SC)	SC-fast, SC-slow, Org	91.06	90.32	92.78	88.07
2DCNN	ALI	Spectrogram masking (SM)	SM-time, SM-freq, SM-timefreq, Org	95.58	95.27	96.60	94.02
2DCNN	ALI	Noise addition (NA)	NA-m10, NA-m5, NA-0, NA-p5, NA-p10, Org	91.79	91.48	92.47	90.55
2DCNN	Augmentation method identification (AMI)	Zero-padding TS, SC, SM, NA	TS, SC, SM, NA, Org	90.42	90.59	92.27	89.03

TABLE 3. Results of the mobile-sensed BreathPID downstream task with CRNN model.

Model	#Parameters (in M)	SSL	Pretext task	Augmentation for pretext task	Accuracy (%)	F1-score	Precision	Recall
CRNN	1.21	without (w/o)	No	No	92.55	92.86	93.64	92.12
VGG-16	18.41	w/o	No	No	94.52	94.53	94.62	94.44
AlextNet	48.36	w/o	No	No	89.59	65.95	99.84	49.79
CRNN	1.21	with (w/)	ALI	TS	95.96	96.13	96.66	95.62
CRNN	1.21	w/	ALI	SC	96.41	96.42	96.42	96.42
CRNN	1.21	w/	ALI	SM	97.84	97.85	97.85	97.85
CRNN	1.21	w/	ALI	NA	97.48	97.54	97.58	97.50
CRNN	1.21	w/	AMI	TS, SC, SM, NA	98.38	98.45	98.54	98.37

The pretext head or classifier (i.e., for AMI or ALI tasks) is a multi-layer perceptron (MLP) with three fully-connected layers (FCs), where the number of nodes in the final FC is set according to the number of augmentation methods (i.e., #AM) in AMI task or the number of augmentation levels (i.e., #AL) in ALI tasks. The downstream classifier contains two bi-directional long-short-term memory (Bi-LSTM) layers with 128 cells followed by an MLP of three FC layers, in which the last FC layer has 21 nodes which is the number of person IDs in the training dataset. The network formed by the encoder and pretext head is called 2DCNN, while the network composed of the encoder and downstream head is referred to as the CRNN model. In other words, 2DCNN and CRNN are networks for pretext and downstream tasks, respectively. For comparative analysis, we also consider several baseline models including VGG [39] and AlexNet [40].

The amount of data samples for each data class (i.e., person ID) is almost comparable in both training and testing sets, except for person ID 18 (Figure 6). In other words, the experimental dataset is relatively balanced. Thus, accuracy is employed as the main metric for performance evaluation. The computation of identification accuracy is expressed by equation (3), where *TP*, *TN*, *FP*, and *FN* denote true positive, true negative, false positive, and false negative, respectively. Results on other metrics including F1 score, precision, and recall are also provided.

$$Acc(\text{in}\%) = \frac{\#Correctly\ Identified\ Samples}{\#Testing\ Samples} \times 100\% = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (3)$$

The TensorFlow framework is utilized to execute the proposed architectures and reproduce baseline models. All models are trained on an NVIDIA RTX-2080 GPU employing the Adam optimizer [41], with a starting learning rate of 0.0001, and batch size of 128. We additionally utilize batch normalization [42] for all layers to speed up the training process, and dropout regularization [43] is applied to alleviate overfitting.

B. RESULTS OF SSL-BASED BREATHPID

Table 2 provides the results of the augmentation level identification (ALI) pretext task across four types of data augmentation as well as performance on the augmentation method identification (AMI) task. All results are reported using the mobile-sensed BBS dataset (M-BBS-DS) where 80% and 20% of the original training set were utilized for training and validating networks, respectively. For the ALI tasks, except for time shifting with the cycling approach, the pretext tasks with the remaining augmentation methods obtained promising results, in which 2DCNN yielded accuracies ranging from 89.75% (zero-padding TS) to 95.58% (SM). The ALI task on cycling TS obtained an accuracy of 53.99% which is only slightly above the chance-level accuracy and 35.58% smaller than the result of the ALI task on zero-padding TS. This poor performance indicates that it is challenging to recognize the difference between the original data sample and augmented signals received by left-cycling TS and right-cycling TS. On the other hand, zero-padding TS results in a clearer indication of the shifting directions. Among examined pretext tasks and augmentation methods, the ALI task on spectrogram masking achieved the highest accuracy of 95.58%, which is

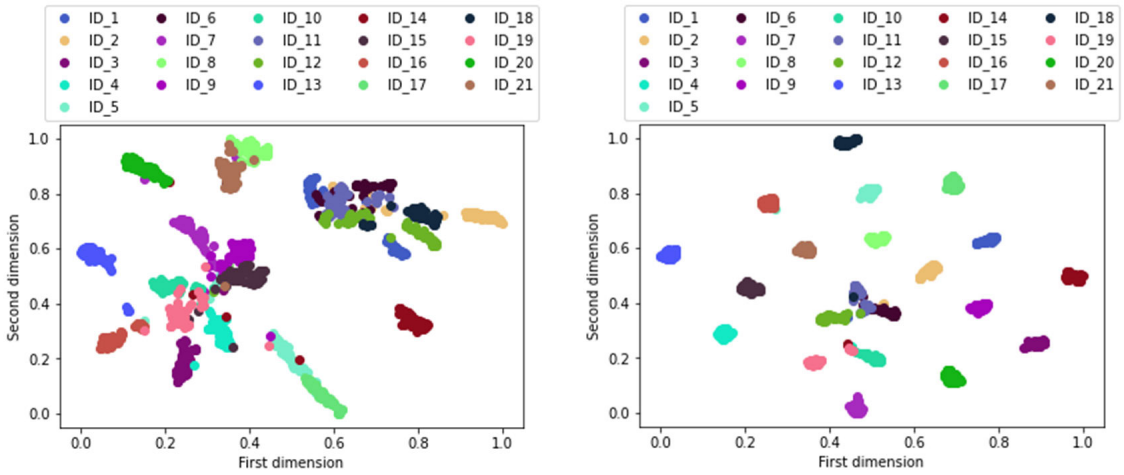

FIGURE 7. t-SNE plots of features extracted by the downstream model trained without SSL (left) and by the downstream model trained with SSL (right) of AMI pretext task.

TABLE 4. Robustness evaluation of mobile-sensed BreathPID.

Models	SSL	Pretext task	Augmentation for pretext task	Accuracy (%) on each SNR							
				-15dB	-10dB	-5dB	0dB	+5dB	+10dB	+15dB	Original data
CRNN	w/o	No	No	55.15	73.63	82.60	86.72	88.34	90.49	91.03	92.55
CRNN	w/	AMI	SM, SC, NA, TS	75.34	87.89	93.27	95.70	97.30	97.39	97.76	98.38
CRNN	w/	ALI	SM	70.04	84.12	91.47	93.99	95.96	96.86	97.13	97.84
CRNN	w/	ALI	SC	61.61	76.50	87.17	91.12	93.81	94.43	94.88	96.41
CRNN	w/	ALI	TS	65.30	77.13	86.81	91.56	95.33	95.87	96.50	97.75
CRNN	w/	ALI	NA	66.91	81.25	89.32	92.64	94.97	95.78	96.68	97.48

higher than the results of the other tasks by at least 3.79%. For the AMI pretext task, we only consider augmentation methods on which the ALI pretext task produced promising performances. Thus, the AMI task was evaluated with the combined use of zero-padding TS, SC, SM, and NA, resulting in a classification accuracy of 90.42%. This result shows that the 2DCNN model can learn useful features to distinguish signals which were transformed with different augmentation methods.

Table 3 provides the results of the mobile-sensed BreathPID downstream task based on various pretext tasks. We conducted experiments on the mobile-sensed BBS dataset (M-BBS-DS) and reported results on the testing set of this dataset. In this table and for the rest of this article, Zero-padding TS is referred to as TS for short. Recall that the CRNN model for the downstream task and the 2DCNN model for pretext tasks share the same architecture of the encoder (i.e., four 2D-Conv blocks). Firstly, we examined the performance of CRNN without using self-supervised learning (SSL), which means the entire model was trained from scratch without a frozen pre-trained encoder. In this case, CRNN yielded an accuracy of 92.55%. It is worth mentioning that in the training of CRNN (w/o SSL), all augmentation methods, TS, SC, SM, and NA were randomly utilized to increase the number of training data. Next, we evaluated the SSL-based BreathPID downstream task across four ALI and one AMI pretext tasks, in which the classifier of the CRNN model was trained on the output of the frozen pre-trained encoder received from the corresponding pretext task. CRNN produced accuracies of 95.96%, 96.41%, 97.84%, and 97.48% based on four ALI tasks, respectively, showing accuracy improvements by 3.41% to 5.29% compared to the result of the typical supervised training. CRNN based on the AMI task received the highest accuracy of 98.38%. The results of this experiment show the effectiveness of the SSL training approach based on proposed pretext tasks for the performance of BreathPID using mobile-sensed bronchial breath sounds. The features learned by dealing with pretext tasks are useful for identification purposes. Among the five pretext tasks, AMI based on the combination of TS, SC, SM, and NA augmentation methods is the most effective one. Figure 7 provides t-distributed stochastic neighbor embedding (t-SNE) plots of features extracted by CRNN trained with and without self-supervised learning. It is shown that SSL-based CRNN can build a clear boundary to classify subjects.

In addition to the proposed CRNN architecture, we also performed experiments with two baseline models (i.e., VGG [39] and AlexNet [40]) which were trained without the step of self-supervised representation learning. As shown in Table 3, VGG and AlexNet yielded accuracies of 94.52% and 89.59%, respectively. It is worth noting that VGG and AlexNet have more than 18 and 48 million parameters, respectively, which are much larger than the figure for CRNN

TABLE 5. Results of mobile-sensed BreathPID with different input sizes.

Model	SSL	Pretext task	Augmentation for pretext task	Input size	Accuracy (%) on each SNR							
					-15dB	-10dB	-5dB	0dB	+5dB	+10dB	+15dB	Original data
CRNN	w/o	No	No	3s	55.15	73.63	82.60	86.72	88.34	90.49	91.03	92.55
				2s	44.84	64.93	77.30	81.97	84.48	86.90	87.80	90.49
				1s	39.91	55.69	67.53	76.32	79.01	82.51	84.12	87.71
CRNN	w/	AMI	SM, SC, NA, TS	3s	75.34	87.89	93.27	95.70	97.30	97.39	97.76	98.38
				2s	62.78	79.55	91.03	94.34	95.78	95.87	96.05	96.95
				1s	57.39	73.00	84.93	89.59	91.83	92.73	93.54	94.43

TABLE 6. Results of stethoscope-sensed BreathPID.

Model	SSL	Pretext task	Augmentation for pretext task	Input size	Accuracy (%)	
					Stethoscope-sensed data (S-BBS-DS)	Mobile-sensed data (M-BBS-DS)
CRNN	w/o	No	No	3s	88.97	92.55
				2s	84.81	90.49
				1s	82.01	87.71
CRNN	w	AMI	SM, SC, NA, TS	3s	97.19	98.38
				2s	94.97	96.95
				1s	90.03	94.43

(1.21 million parameters). Although VGG produced higher accuracy compared to that of CRNN, VGG is less efficient in terms of network size. Equally important, by applying SSL in advance CRNN outperformed VGG with significant differences in accuracies, by 1.44% to 3.86%. In contrast to VGG, AlexNet had an even larger capacity but underperformed the CRNN. It appears that AlexNet is an oversized network for BreathPID based on the current amount of experimental data. These comparative results further confirm the effectiveness of the proposed approaches for mobile-based BreathPID. When compared to the outcomes of related studies [23], [24], [25], [26], the results demonstrate that the SSL-based PID using mobile-sensed bronchial breath sounds achieved identification accuracies that were either on par or superior.

C. ROBUSTNESS EVALUATION

To evaluate the robustness of the examined models, we analyze their performances on testing sets of different signal-to-noise (SNR) ratios consisting of +15dB, +10dB, +5dB, 0dB, −5dB, −10dB, and −15dB. The creation of noisy testing sets was conducted by artificially adding noises to the original testing data according to the aforementioned SNRs, in which indoor and outdoor background sounds collected in offices, supermarkets, and on the street were utilized as the noise sources (i.e., noise database). For example, to create the -5dB testing set, every testing sample in the M-BBS-DS dataset was mixed with a noise recording randomly selected from the noise database at −5dB.

Table 4 shows the performances across different noise levels of the CRNN model without and with self-supervised representation learning, trained on different pretext tasks. The results suggest that using SSL with proposed pretext tasks can significantly improve the model's resistance to noise, especially at low SNRs. By contrast, the model trained without SSL is highly sensitive to noise as its performance degrades considerably across the increase of noise levels. The best performance is achieved when training the CRNN model with SSL and AMI pretext task using SM, SC, NA, and TS augmentations in combination. From moderate noise condition of +5dB, the accuracy of CRNN (w/o SSL) starts reducing significantly to less than 90%, while the figures for SSL-based models remain above 90% at the SNR of −5dB. For more challenging conditions of −10dB and −15dB, the accuracy gaps between models become larger. At the S NR of −10dB, the accuracy of the AMI-based model is 87.89% while that of the model trained without SSL is 14.26% smaller and those of ALI-based models are smaller by 3.77% to 11.39%. Similarly, at the noisiest condition of −15dB, the accuracy difference between the AMI-based model and the model trained without SSL increases to almost 20%. Among downstream models based on ALI pretext tasks, the model finetuned with the pre-trained encoder of SM-based ALI task yields the best performance which is worse than that of the CRNN (AMI) model but much better than those of models based on the remaining ALI tasks. Thus, spectrogram masking (SM) is among the most useful augmentation method for representation learning in mobile-sensed BreathPID.

D. PERFORMANCES ON DIFFERENT INPUT SIZES

Recall that we choose the input size of 3s for BreathPID because this size is close to the length of a normal breathing cycle which includes an inhalation period (i.e., 1s to 1.5s) and an exhalation period (i.e., 1.5s to 2s). In this experiment, we analyze the effects of the reduction in input sizes on the performance of the proposed SSL-based BreathPID, to evaluate the feasibility of achieving acceptable identification accuracies with shorter inputs, which may bring about better user experience in practical applications.

TABLE 7. Robustness evaluation of stethoscope-sensed BreathPID.

Models	SSL	Pretext task	Augmentation for pretext task	Data	Accuracy (%) on each SNR							
					-15dB	-10dB	-5dB	0dB	+5dB	+10dB	+15dB	Original data
CRNN	w/o	No	No	Stethoscope -sensed	34.23	52.41	68.37	76.88	82.78	86.46	87.42	88.97
				Mobile-sensed	55.15	73.63	82.60	86.72	88.34	90.49	91.03	92.55
CRNN	w/	AMI	SM, SC, NA, TS	Stethoscope -sensed	44.48	61.89	78.23	88.39	91.78	94.39	95.65	97.19
				Mobile-sensed	75.34	87.89	93.27	95.70	97.30	97.39	97.76	98.38

Table 5 summarizes the performances of CRNN trained with two experimental settings across three cases of input lengths (i.e., 3s, 2s, and 1s) and various levels of noise. It is shown that the proposed CRNN model trained on the frozen pre-trained encoder of the AMI pretext task provides well-acceptable accuracies when the input size is decreased. Compared to the result of 3s input, the decreases in accuracy are 1.43% and 3.95% for clean data of 2s and 1s lengths, respectively. Among the three input sizes, results on 1s data are the lowest and experience more considerable decreases across noisier conditions, especially at 0dB, −5dB, −10dB, and −15dB. When the noises have the same power as the breath sounds (i.e., SNR or 0dB), accuracies obtained on 3s and 2s data remain above 94%, while the figure for 1s data is smaller than 90%. The difference in accuracies becomes greater at higher levels of noise. Although at SNRs of above −5dB results on 2s data and 3s data are almost comparable, accuracies for the former are much smaller than those for the latter at −10dB and −15dB. The results of this experiment show that 3s and 2s are more favorable input lengths for mobile-sensed BreathPID. In contrast to the proposed SSL-based method, CRNN trained without SSL is more sensitive to reduction of input size, especially at high levels of noise. Starting from a moderate noise condition, at SNR of +5dB, accuracies of CRNN (w/o SSL) across all cases of input lengths are lower than 90%, as shown in Table 5.

E. COMPARING RESULTS ON MOBILE-SENSED DATA AND STETHOSCOPE-SENSED DATA

This experiment is to compare the results of BreathPID using bronchial breath sounds sensed by mobile devices and specialty stethoscopes. Recall that by simultaneous collection we have prepared two datasets, namely M-BBS-DS and S-BBS-DS, which are collected using mobile devices and stethoscopes, respectively. For the comparative purpose, the general experimental setting for stethoscope-sensed data (S-BBS-DS) is the same as that for mobile-sensed data (M-BBS-DS). However, we only conducted experiments with two model settings, including SSL-based CRNN with AMI pretext task and CRNN without SSL. Table 6 compares the results based on two data collection approaches across different input sizes, while Table 7 provides a comparison regarding the robustness.

We can see from the statistics in Table 6 that for the input size of 3 seconds, the SSL-based CRNN yields almost comparable results on both data collection approaches, with 97.19% accuracy on stethoscope-sensed data, which is only 1.19% lower than the result on mobile-sensed data. However, larger accuracy differences are observed for shorter inputs (i.e., 2s and 1s) in which results on stethoscope-sensed data are lower. For example, with 1s input, SSL-based CRNN produces an accuracy of 90.03% on stethoscope-sensed data, which is 4.40% smaller than the result on mobile-sensed data. For models trained without SSL, results for stethoscope-sensed data across all cases of input lengths are much lower than those of mobile-sensed data. Similar to experiments on the M-BBS-DS dataset, for the S-BBS-DS dataset, the SSL training approach also yields much better performance compared to the traditional supervised training approach. As shown in Table 7, across all noise levels applied to stethoscope-sensed data, similar trends in superior performances of SSL-based CRNN are observed. Both training methods achieve superior results on mobile-sensed data over stethoscope-sensed data. By comparing the results based on two data collection approaches, we can observe the potential of mobile-sensed BreathPID. It yields comparable or superior accuracies while also providing flexibility for data collection, thereby extending the applicability of BreathPID.

V. CONCLUSION

This work examines mobile-sensed BreathPID which is a person identification system based on bronchial breath sounds intrusively sensed by mobile devices. Based on the experiment results of this work, it can be concluded that mobile-sensed BreathPID can be effectively developed using self-supervised learning (SSL) approaches and audio data augmentation (DA) techniques. Different pretext tasks for self-supervised representation learning were investigated with the use of four DA methods, including noise addition (NA), time shifting (TS), speech changing (SC), and spectrogram masking (SM). The proposed SSL-based systems achieved promising results, surpassing those of typical supervised models and showing good resistance to noise effects, in which the proposed CRNN model trained on frozen pre-trained encoder of AMI (i.e., augmentation method identification) pretext task yielded the best accuracy

of 98.38% on the clean testing data. In addition, we found that the input sizes of 3 seconds or 2 seconds are recommended for BreathPID, while shorter input lengths (e.g., 1 second) caused significant degradation in identification performance. Furthermore, the mobile-sensed BreathPID system achieved results equivalent or superior to those of stethoscope-sensed BreathPID, showing that capturing bronchial breath sounds using mobile devices rather than using specialty stethoscopes not only still guarantee well-acceptable identification accuracy but also provides flexibility for the development of practical mobile-based PID applications.

This work has demonstrated promising results on mobile-sensed BreathPID which can be potentially applied for authentication or health monitoring applications on mobile devices. However, it is still at the early stage of investigation and several future works can be listed as follows. First, we have conducted experiments on a moderate custom dataset of 21 subjects, so one of the directions for future works could be to expand the dataset and test the proposed system with a larger and more diverse group of participants to evaluate its real-world effectiveness. Second, the optimization of network architectures could be further considered, and the deployment of the BreathPID system on mobile devices could be conducted, from which we can evaluate the efficiency of the proposed methods more comprehensively. Third, since the main objectives of this study were to show the feasibility of mobile-sensed BreathPID and the efficiency of the SSL approach, we only proposed simple SSL-based methods with small-size models. In future works, more sophisticated SSL approaches and advanced deep learning techniques could be considered to explore further improvement in identification accuracy. Lastly, it is crucial to explore the potential of integrating BreathPID with other audio-based PID methods, such as speech-based PID and singing-based PID, to enhance the overall security and flexibility of identification systems.

REFERENCES

[1] M. Sarkar, I. Madabhavi, N. Niranjan, and M. Dogra, "Auscultation of the respiratory system," *Ann. Thorac. Med.*, vol. 10, no. 3, p. 168, Jul. 2015, doi: 10.4103/1817-1737.160831.

[2] V.-T. Tran, Y.-C. Lin, and W.-H. Tsai, "On the use of bronchial breath sounds for person identification," *J. Inf. Sci. Eng.*, vol. 37, no. 1, pp. 219–241, 2021.

[3] V. Tran and W. Tsai, "Stethoscope-sensed speech and breath-sounds for person identification with sparse training data," *IEEE Sensors J.*, vol. 20, no. 2, pp. 848–859, Jan. 2020.

[4] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 4–20, Jan. 2004, doi: 10.1109/TCSVT.2003.818349.

[5] A. K. Jain, L. Hong, S. Pankanti, and R. Bolle, "An identity-authentication system using fingerprints," *Proc. IEEE*, vol. 85, no. 9, pp. 1365–1388, Sep. 1997, doi: 10.1109/5.628674.

[6] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey," *Pattern Recognit.*, vol. 25, no. 1, pp. 65–77, Jan. 1992, doi: 10.1016/0031-3203(92)90007-6.

[7] R. Brunelli and T. Poggio, "Face recognition: Features versus templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 10, pp. 1042–1052, Oct. 1993, doi: 10.1109/34.254061.

[8] R. P. Wildes, "Iris recognition: An emerging biometric technology," *Proc. IEEE*, vol. 85, no. 9, pp. 1348–1363, Sep. 1997, doi: 10.1109/5.628669.

[9] G. R. Doddington, "Speaker recognition—Identifying people by their voices," *Proc. IEEE*, vol. 73, no. 11, pp. 1651–1664, Nov. 1985, doi: 10.1109/PROC.1985.13345.

[10] R. Brunelli and D. Falavigna, "Person identification using multiple cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 10, pp. 955–966, Oct. 1995, doi: 10.1109/34.464560.

[11] P. S. Aleksic and A. K. Katsaggelos, "Audio-visual biometrics," *Proc. IEEE*, vol. 94, no. 11, pp. 2025–2044, Nov. 2006, doi: 10.1109/JPROC.2006.886017.

[12] L. Biel, O. Pettersson, L. Philipson, and P. Wide, "ECG analysis: A new approach in human identification," *IEEE Trans. Instrum. Meas.*, vol. 50, no. 3, pp. 808–812, Jun. 2001, doi: 10.1109/19.930458.

[13] S. Marcel and J. D. R. Millan, "Person authentication using brainwaves (EEG) and maximum A posteriori model adaptation," *IEEE Comput. Soc.*, vol. 29, no. 4, pp. 743–748, Apr. 2007, doi: 10.1109/TPAMI.2007.1012.

[14] Z. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 2, pp. 457–468, Feb. 2019, doi: 10.1109/TASLP.2018.2881912.

[15] A. Chowdhury and A. Ross, "Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1616–1629, 2020.

[16] H. Taherian, Z. Wang, J. Chang, and D. Wang, "Robust speaker recognition based on single-channel and multi-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1293–1302, 2020, doi: 10.1109/TASLP.2020.2986896.

[17] Y. Hu and G. Liu, "Separation of singing voice using nonnegative matrix partial co-factorization for singer identification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 4, pp. 643–653, Apr. 2015, doi: 10.1109/TASLP.2015.2396681.

[18] Z. Shen, B. Yong, G. Zhang, R. Zhou, and Q. Zhou, "A deep learning method for Chinese singer identification," *Tsinghua Sci. Technol.*, vol. 24, no. 4, pp. 371–378, Aug. 2019, doi: 10.26599/TST.2018.9010121.

[19] H. Yakura, K. Watanabe, and M. Goto, "Self-supervised contrastive learning for singing voices," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 1614–1623, 2022, doi: 10.1109/TASLP.2022.3169627.

[20] J. W. M. Engelberg, J. W. Schwartz, and H. Gouzoules, "Do human screams permit individual recognition?" *PeerJ*, vol. 7, p. e7087, Jun. 2019, doi: 10.7717/peerj.7087.

[21] J.-A. Bachorowski, M. J. Smoski, and M. J. Owren, "The acoustic features of human laughter," *J. Acoust. Soc. Amer.*, vol. 110, no. 3, pp. 1581–1597, Sep. 2001, doi: 10.1121/1.1391244.

[22] F. Beritelli and S. Serrano, "Biometric identification based on frequency analysis of cardiac sounds," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 3, pp. 596–604, Sep. 2007, doi: 10.1109/TIFS.2007.902922.

[23] W. Zhao, Y. Gao, and R. Singh, "Speaker identification from the sound of the human breath," 2017, *arXiv:1712.00171*.

[24] L. Lu, L. Liu, M. J. Hussain, and Y. Liu, "I sense you by breath: Speaker recognition via breath biometrics," *IEEE Trans. Dependable Secure Comput.*, vol. 17, no. 2, pp. 306–319, Mar. 2020, doi: 10.1109/TDSC.2017.2767587.

[25] J. Chauhan, Y. Hu, S. Seneviratne, A. Misra, A. Seneviratne, and Y. Lee, "BreathPrint: Breathing acoustics-based user authentication," in *Proc. 15th Annu. Int. Conf. Mobile Syst., Appl., Services*, Jun. 2017, pp. 278–291, doi: 10.1145/3081333.3081355.

[26] J. Chauhan, S. Seneviratne, Y. Hu, A. Misra, A. Seneviratne, and Y. Lee, "Breathing-based authentication on resource-constrained IoT devices using recurrent neural networks," *Computer*, vol. 51, no. 5, pp. 60–67, May 2018, doi: 10.1109/MC.2018.2381119.

[27] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1085–1095, May 2012, doi: 10.1109/TASL.2011.2172422.

[28] N. Iliev, A. Gianelli, and A. R. Trivedi, "Low power speaker identification by integrated clustering and Gaussian mixture model scoring," *IEEE Embedded Syst. Lett.*, vol. 12, no. 1, pp. 9–12, Mar. 2020, doi: 10.1109/LES.2019.2915953.

[29] N. A. Al Hindawi, I. Shahin, and A. B. Nassif, "Speaker identification for disguised voices based on modified SVM classifier," in *Proc. 18th Int. Multi-Conf. Syst., Signals Devices (SSD)*, Mar. 2021, pp. 687–691, doi: 10.1109/SSD52085.2021.9429403.

[30] A. Jati and P. Georgiou, "Neural predictive coding using convolutional neural networks toward unsupervised learning of speaker characteristics," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 10, pp. 1577–1589, Oct. 2019, doi: 10.1109/TASLP.2019.2921890.
[31] M. B. Andra and T. Usagawa, "Improved transcription and speaker identification system for concurrent speech in Bahasa Indonesia using recurrent neural network," *IEEE Access*, vol. 9, pp. 70758–70774, 2021, doi: 10.1109/ACCESS.2021.3077441.
[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, Jun. 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
[33] S. Gidaris, P. Singh, and N. Komodakis. (2018). *Unsupervised Representation Learning by Predicting Image Rotations*. [Online]. Available: https://openreview.net/forum?id=S1v4N2l0-
[34] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021, doi: 10.1109/TPAMI.2020.2992393.
[35] X. Liu, J. V. D. Weijer, and A. D. Bagdanov, "Exploiting unlabeled data in CNNs by self-supervised learning to rank," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1862–1878, Aug. 2019, doi: 10.1109/TPAMI.2019.2899857.
[36] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3451–3460, 2021, doi: 10.1109/TASLP.2021.3122291.
[37] P. Corbishley and E. Rodriguez-Villegas, "Breathing detection: Towards a miniaturized, wearable, battery-operated monitoring system," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 1, pp. 196–204, Jan. 2008, doi: 10.1109/TBME.2007.910679.
[38] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–16.
[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
[41] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
[42] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[43] N. S. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2019.

VAN-THUAN TRAN received the M.S. degree in electrical engineering and computer science and the Ph.D. degree in electronic engineering from the National Taipei University of Technology (NTUT), Taipei, Taiwan, in 2018 and 2022, respectively. From 2015 to 2016, he was an Engineer with the JGCS Consortium, NSRP Project, Vietnam. He is currently a Postdoctoral Researcher with the Department of Electronic Engineering, NTUT. His research interests include multimedia signal processing and artificial intelligence.

YIH-LON LIN received the B.S. degree in electrical engineering from the Department of Electronic Engineering, I-Shou University, Kaohsiung, Taiwan, in 1997, and the M.S. and Ph.D. degrees in electrical engineering from the Department of Electrical Engineering, National Sun Yat-sen University, Kaohsiung, in 1999 and 2006, respectively. He is currently an Associate Professor with the Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology, Yunlin, Taiwan. His research interests include medical image processing, natural language processing, computer vision, and deep learning.

WEI-HO TSAI (Member, IEEE) received the Ph.D. degree in communication engineering from National Chiao-Tung University, Hsinchu, Taiwan, in 2001. From 2001 to 2003, he was with Philips Research East Asia, Taipei, Taiwan, where he worked on speech processing problems in embedded systems. From 2003 to 2005, he was a Postdoctoral Fellow with the Institute of Information Science, Academia Sinica, Taipei. He is currently a Professor with the Department of Electronic Engineering, National Taipei University of Technology, Taiwan. His research interests include spoken language processing and music information retrieval.