Toward Real-World Voice Disorder Classification

Heng-Cheng Kuo, Yu-Peng Hsieh, Huan-Hsin Tseng, Chi-Te Wang, Shih-Hau Fang, *Senior Member, IEEE*, and Yu Tsao, *Senior Member, IEEE*

***Abstract*—*Objective:* Voice disorders significantly compromise individuals' ability to speak in their daily lives. Without early diagnosis and treatment, these disorders may deteriorate drastically. Thus, automatic classification systems at home are desirable for people who are inaccessible to clinical disease assessments. However, the performance of such systems may be weakened due to the constrained resources and domain mismatch between the clinical data and noisy real-world data. *Methods:* This study develops a compact and domain-robust voice disorder classification system to identify the utterances of health, neoplasm, and benign structural diseases. Our proposed system utilizes a feature extractor model composed of factorized convolutional neural networks and subsequently deploys domain adversarial training to reconcile the domain mismatch by extracting domain-invariant features. *Results:* The results show that the unweighted average recall in the noisy real-world domain improved by 13% and remained at 80% in the clinic domain with only slight degradation. The domain mismatch was effectively eliminated. Moreover, the proposed system reduced the usage of both memory and computation by over 73.9%. *Conclusion:* By deploying factorized convolutional neural networks and domain adversarial training, domain-invariant features can be derived for voice disorder classification with limited resources. The promising results confirm that the proposed system can significantly reduce resource consumption and improve classification accuracy by considering the domain mismatch. *Significance:* To the best of our knowledge, this is the first study that jointly considers real-world model compression and noise-robustness issues in voice disorder classification. The proposed system is intended for application to embedded systems with limited resources.**

***Index Terms*—Voice disorder classification, model compression, domain adaptation, real-world application.**

Manuscript received 2 May 2022; revised 3 October 2022, 30 November 2022, and 21 March 2023; accepted 13 April 2023. Date of publication 26 April 2023; date of current version 28 September 2023. This work was supported by Far Eastern Memorial Hospital under Grant FEMH2021-C-029. *(Corresponding author: Yu Tsao.)*

Heng-Cheng Kuo, Yu-Peng Hsieh, and Huan-Hsin Tseng are with the Research Center for Information Technology Innovation, Academia Sinica, Taiwan.

Chi-Te Wang is with the Department of Electrical Engineering, Yuan Ze University, Taiwan and also with the Department of Otolaryngology Head and Neck Surgery, Far Eastern Memorial Hospital, Taiwan.

Shih-Hau Fang is with the Department of Electrical Engineering and AI Research Center, Yuan Ze University, Taiwan.

Yu Tsao is with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 11529, Taiwan and also with the Department of Electrical Engineering, Chung Yuan Christian University, Taoyuan 32023, Taiwan (e-mail: yu.tsao@citi.sinica.edu.tw).

Digital Object Identifier 10.1109/TBME.2023.3270532

I. Introduction

EARLY epidemiological studies reported varying estimates of the prevalence of voice disorders, ranging from 3.85% to 15% [1], [2]. A later report estimated the prevalence among the US to be approximately 3% to 9% [3]. More recently, a regional telephone survey of 1326 random subjects revealed a current voice disorders prevalence of 6.6% and a lifetime prevalence of 29.9% in adults aged less or equal to 65 years [4]. Another study based on primary care physicians also demonstrated similar results on the lifetime prevalence (4.3% to 29.1%) and current prevalence (7.5%) of voice disorders [5]. Two large-scale claims data-based epidemiological studies revealed that the prevalence rate of voice disorders ranges from 0.26% to 0.98% [6], [7]. All studies have indicated that the overall prevalence of voice disorders is quite alarming. For such diseases, accurate diagnosis requires experienced specialists and expensive equipment. Patients without health insurance or other medical resources would face a few months of waiting times for specialist appointments. Appropriate and instant disease assessment may be inaccessible to people in need. Therefore, this study proposes a non-invasive self-screening classification system that allows individuals to diagnose pathological voices (health, neoplasm, and benign structural diseases) at home to help schedule the priority of medical resource allocation. For example, if a patient is diagnosed with neoplasm using the proposed classification system, his/her appointment can be brought forward to reduce the waiting time. On the other hand, if the self-screening result shows healthy, the user can avoid the risk of infection while traveling to the hospital and the waste of medical resources, especially during epidemics.

In recent decades, several non-invasive screening methods have been proposed, and their potential to identify samples of pathological voices has been demonstrated [8], [9], [10], [11]. Furthermore, pathological voices always accompany changes in the voice quality [12], [13], [14], [15]. Thus, in previous research, acoustic features, such as Mel frequency cepstral coefficients (MFCCs) [16], [17], [18], glottal features [19], [20], and gammatone spectral latitude (GTSL) [21], were used as inputs of classic machine learning (ML) classifiers, including Gaussian mixture models (GMM) [22], support vector machine (SVM) [23], [24], [25], [26], [27], [28], [29], and k-nearest neighbors (KNN) [30], [31]. On the other hand, various neural network (NN)-based models also verified the reliability of deep learning (DL) [32], [33], [34], [35], [36], [37], [38]. Automatic speech recognition systems were used to assess voice disorders as well [39], [40]. Besides, some studies have added auxiliary inputs, for example, medical records [41] and the GRBAS

scale [42], to help classify pathological voices. Based on the promising performance under ideal conditions, Hsu et al. [43] further addressed the channel effect due to hardware variation; Fan et al. [44] and Jinyang et al. [45] investigated the sample imbalance between voice disorders. With the development of the Internet of Things (IoT), IoT and cloud technology has also been applied to voice pathology monitoring [46], [47], [48], [49]. In addition to the above research, the FEMH Challenge [50], an international competition held by the IEEE Big Data conference in Seattle in 2018, provided a common dataset and evaluation metrics for the development of voice disorder classification systems [51], [52], [53], [54], [55], [56]. The dataset was published by the Far Eastern Memorial Hospital (FEMH), Taiwan.

Nevertheless, current NN-based solutions are not optimal for practical applications due to two main challenges. First, to achieve state-of-the-art performance, large and deep model structures of neural networks are typically designed. However, the limited memory and computational resources of embedded systems provide little room for models to increase the number of parameters. Second, the domain mismatch between the standardized training data and testing data acquired from real-world scenarios substantially degrades the accuracy.

Because a large model requires a huge memory capacity and computational resources, there are typically three common solutions to achieve real-time processing on embedded systems: quantization techniques, knowledge distillation, and factorized convolutional neural networks (CNNs). Quantization is a technique that replaces the arithmetic of 32-bit floating points with that of integers [57] or powers of two [58], allowing for a much lower latency on commonly available hardware. Moreover, owing to the limited number of quantized values, representations with even lower bits can be applied to further reduce the usage of memory. An alternative solution, knowledge distillation is the process of transferring knowledge from a large network, particularly for an ensemble of models, to a small one [59]. The outputs generated by the cumbersome but well-trained network act as additional labels for the distilled network. By imitating the behavior of a large network, a distilled network can achieve better performance than using only true labels. Finally, the standard CNN can be regarded as a combination of spatial convolution in each channel (also called *intra-channel convolution*) and linear projection across channels simultaneously [60]. The factorized CNN was developed to rearrange the spatial convolution [61] or address these two parts separately [62] to reduce memory and computation.

Another issue in the real-world scenarios is the *domain mismatch*. Although joining abundant labeled data from different environments is likely to improve the generalizability of models, it is not feasible to prepare rather diverse and out-of-clinic data in the biomedical area due to extreme time consumption. Additionally, labeling such data requires a strong professional background, which further increases the difficulty. As a result, data collected in the laboratory or clinics are the few (and sometimes the only) labeled data available. Our intention is to focus on unsupervised domain adaptation, which requires no labeled data from the real-world domain but labeled data from the clinic domain during the training process. In general, labeled data defined as the *source domain* have one probability distribution, while unlabeled data, which we intend to adapt to, called the *target domain*, have another. There are a few methods for generalizing a model to an unseen target domain, including the Generative Adversarial Network (GAN)-based and discrepancy-based methods. In the GAN-based method, a *generator* generates plausible target domain data with labels from the given source data [63], where the plausibility is governed by a *discriminator*. Subsequently, the labeled data from the source domain together with the generated target data are utilized for the main task training. Thus, information from both the source and the target domains is revealed to the main task model. Another discrepancy-based method intends to learn extracted features by minimizing the gap between the probability distributions of the source and the target domains so that a well-trained model can be directly applied to the target domain to fit our purpose. To derive such domain-invariant features, predefined statistics [64], [65], [66] or domain classifier [67], [68] are introduced to assess the discrepancy of probability distributions between the domains.

In this study, we propose a new voice disorder classification system customized for embedded devices, which adapts to daily noisy environments simultaneously. The proposed model consists of factorized CNNs to obtain compact architecture and is augmented with a domain adversarial training (DAT) module during training to equip it with the ability to operate in noisy environments. The results showed that the unweighted average recall (UAR) in the noisy real-world domain improved by 13%, while that in the clinic domain remained at 80% with only slight degradation. In addition, the numbers of parameters and Multiply–Accumulate Operations (MACs) were significantly reduced by 73.9% and 77.0%, respectively.

The remainder of this paper is organized as follows: In Section II, the related works, including MobileNet and DAT, are reviewed. Section III introduces the proposed robust voice disorder classification system. The experimental results and an ablation study are presented in Section IV. Section V claims our plans for the future work. Finally, the conclusions are presented in Section VI.

II. Related Works

A. Model Compression

To achieve a high penetration rate of self-diagnosis at home, classification systems will confront the limits of memory and computational resources on embedded devices. Thus, in addition to accuracy, the model size and computational cost are also prior considerations.

In speech signal processing, filter-based conversions, such as Mel-spectrograms and MFCCs, are typically applied to temporal signals. Therefore, CNN-based models can effectively extract the local pathological characteristics from the (2D image-like) converted inputs. From this perspective, the factorized CNNs are suitable for reducing the difficulties.

An Efficient Residual Factorized Network (ERFNet) [61] rearranges spatial convolutions to achieve lower resource usage. It factorizes a 3×3 convolution into the union of perpendicular 3×1 and 1×3 convolutions, and residual connections are used

to improve the training efficiency while retaining remarkable accuracy under the constrained scenario.

Another factorized method, separable convolution, proposed in MobileNet [62], is aimed at mobile and embedded applications. A *separable* convolutional layer factorizes a standard convolutional layer into a depth-wise convolutional layer and a point-wise convolutional layer. A standard convolution deals with the spatial convolution in each channel and the linear transformation across channels simultaneously, whereas a separable convolution splits this operation into two stages. Specifically, in the first stage, the depth-wise convolution applies a single filter to each input channel, and in the second stage, the point-wise convolution (simply a 1×1 convolution) then performs a linear projection on the previous depth-wise convolution outputs. This separate consideration of the relationship in each channel and the relationship between channels drastically reduces the computation and model size. Two variants, MobileNetV2 [69] and MobileNetV3 [70], were proposed to further improve the accuracy and reduce the latency on the successful base of MobileNet.

B. Unsupervised Domain Adaptation

Compared to undisturbed clinics or studios where pathological voices are recorded, background noise is inevitable in daily life where our system is aimed for application. Moreover, there are very few annotated data in out-of-clinic scenarios available, since it is hard to perform standard and unified experiments for the general public without the assistance of experienced specialists. Thus, dealing with domain mismatches between the labeled source data and unlabeled target data poses a challenge. In general, the distributions of these two domains are expected to be similar but not exactly coincident. In fact, they are required to be "similar" by nature due to the same learning task. However, slight differences are inevitable between the ages or genders of the subjects, the environments where the data are generated, etc. The existence of these differences causes performance degradation, especially in data-driven neural networks. Our purpose is to rectify the data deviation. Because the GAN-based method [63], which aims to generate target domain data with labels, requires a large amount of training data, it is not favorable for each situation. Therefore, the discrepancy-based method is more feasible.

A classification model consists of two parts: a feature extractor and a label predictor. The feature extractor is designed to extract useful information from the input; subsequently, the label predictor utilizes the extracted features for classification. Several domain adaptation techniques typically rely on a feature extractor deriving features invariant across domains, *e.g.,* ignoring the background noise or the difference between recording devices [43], so that a model can generalize on the target domain while preserving a low risk of misclassification on the source domain [71]. If the extracted features are perplexing across domains at all times, those features are considered *domain-invariant* in this study. Based on this idea, statistical techniques or an NN-based domain classifier are introduced to assess the domain invariability of the extracted features. In the former, Maximum Mean Discrepancy (MMD) [64], [72] and Optimal Transport (OT) [65] serve as loss functions to calculate the distance between the probability distributions of the extracted features across domains to be minimized together with classification losses. In the latter, the extracted features are adversarially trained to perplex the domain classifier and remain high prediction accuracy, where DAT is one of the most popular algorithms augmented with an NN-based domain classifier.

To fool the domain classifier, DAT instructs the feature extractor to update in the opposite direction of minimizing the domain classification loss. For this purpose, the study of DAT introduced a novel gradient reversal layer (GRL) glued by two functions R and $\widetilde{R}$ at different stages, requiring no parameters such that:

$$\begin{aligned} R(\mathbf{z}) &= \mathbf{z} && \text{(forward propagation)} \\ \widetilde{R}(\mathbf{z}) &= -\mathbf{z} \;\Leftrightarrow\; \nabla_{\mathbf{z}}\widetilde{R} = -\mathbf{I} && \text{(backward propagation)} \end{aligned} \quad (1)$$

where $\mathbf{z}$ and $\mathbf{I}$ denote the input and identity matrix, respectively. It should be noted that in a general layer with forward function $\mathbf{z} \mapsto f(\mathbf{z})$, the backward propagation naturally has the derivative $\mathbf{z} \mapsto \nabla_{\mathbf{z}} f$ from the same f. The GRL deliberately splits the forward and backward function into two to achieve the designated purpose. As such, the GRL acts as an identity function during the forward propagation, but multiplies the gradient by -1 during back propagation. Owing to the GRL, the DAT algorithm can be implemented on any existing ML package with little effort.

III. Methodology

A. Proposed Method

We propose a system for voice disorder classification consisting of separable convolutional layers equipped with a DAT architecture. Our backbone model replaces the standard deep CNN-based convolutional layers with separable convolutional layers. The standard CNN-based model referenced ensures the performance and the efficiency with reduced computation. The experiments in Section IV verify that the performance of the proposed method is comparable to that of the referenced standard CNN-based method.

We note that the associated training process of DAT is given by the min-max algorithm:

$$\arg\max_{\theta_d} \min_{\theta_f, \theta_y} \; \mathbb{E}_{x \sim D_s}[L_y] + \mathbb{E}_{x \sim D_s \times D_t}[-\lambda L_d] \quad (2)$$

$$L_y = -\log \mathrm{P}(\mathbf{y}|\mathbf{x}, \theta_{\mathrm{f}}, \theta_{\mathrm{y}}) \quad (3)$$

$$L_d = -\log \mathrm{P}(\mathbf{d}|\mathbf{x}, \theta_{\mathrm{f}}, \theta_{\mathrm{d}}) \quad (4)$$

where a data point {*input, label, domain*} is denoted by $\{\mathbf{x}, \mathbf{y}, \mathbf{d}\}$, and θ_f, θ_y, θ_d denote the parameters of the feature extractor, label predictor and domain classifier, respectively. D_s and D_t denote the source domain and the target domain data distribution; L_y represents the cross-entropy loss of the labels, and L_d is that of the domains, where $\lambda \geq 0$ is the coefficient that regularizes the loss in (2). To improve the performance of the main task, the classification of pathological voices in this work, the feature extractor and the label predictor shall jointly minimize L_y. Due to the unsupervised domain adaptation scheme, L_y can only be computed on the source domain data. On the other hand, the domain classifier minimizes L_d to enhance

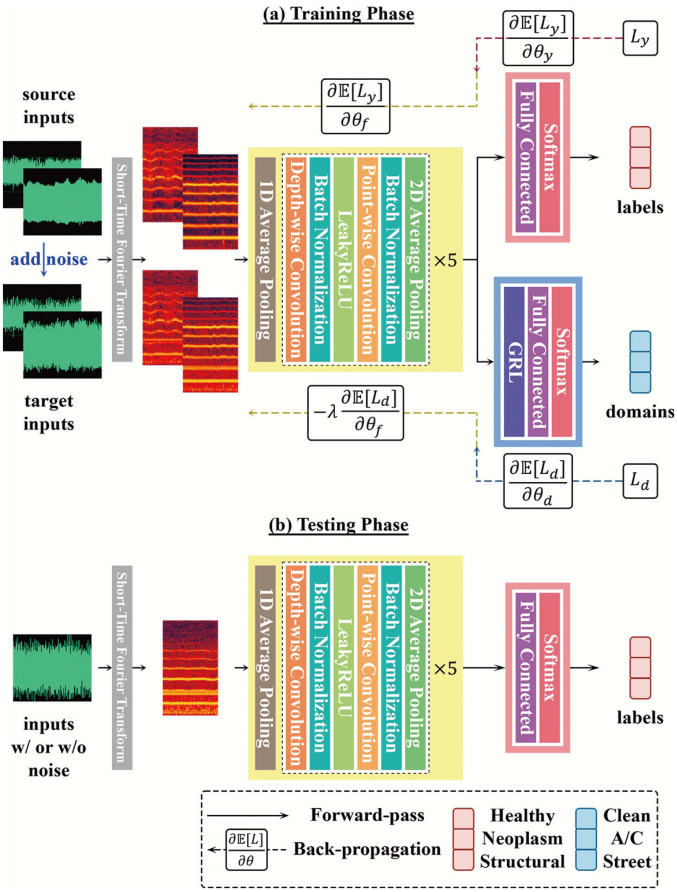

Fig. 1. Feature extractor consists of 5 separable convolutional blocks; The disease predictor and domain classifier are simply combinations of a fully connected layer and a softmax layer.

the ability to discriminate domains, whereas the feature extractor maximizes L_d to obtain the opposite gradients. For simplicity, we substitute $-L_d$ for L_d in (2) and invert minimization and maximization operations correspondingly.

To realize the adversarial min-max (2), we first formulate the updates of parameters θ_f, θ_y and θ_d via the gradient descent [73] as follows:

$$\theta_f \longleftarrow \theta_f - \alpha\left(\frac{\partial \mathbb{E}[L_y]}{\partial \theta_f} - \lambda \frac{\partial \mathbb{E}[L_d]}{\partial \theta_f}\right) \tag{5}$$

$$\theta_y \longleftarrow \theta_y - \alpha \frac{\partial \mathbb{E}[L_y]}{\partial \theta_y} \tag{6}$$

$$\theta_d \longleftarrow \theta_d - \alpha\,\lambda \frac{\partial \mathbb{E}[L_d]}{\partial \theta_d} \tag{7}$$

where α is the learning rate. The flowchart of the training and testing phases is shown in Fig. 1. Thus, the training phase in Fig. 1(a) illustrates the gradient descent process. The updated formula in (5) and (7) are observed to have opposite signs with respect to the gradient of L_d to conform to the GRL [67].

To provide more details, we will elaborate on the flowchart shown in Fig. 1. In this study, the source domain is defined as the clean recording environment like a clinic, whereas the target domain is the environment with noises at home. During the training phase, the source domain utterances recorded in the clinic are augmented with noises of *air conditions (A/Cs)* and *streets* to synthesize the target domain data and then both are transformed into Log Power Spectrums (LPSs). The details of the data preprocessing are clarified in the next section. Next, the lightweight backbone model learns to identify utterances of health, neoplasm, and benign structural diseases from both clean and noisy environments through the DAT technique. However, real-world utterances with or without noises can be directly used to diagnose diseases through the trained model during the testing phase. Compared to the high UAR in the source domain, our system only suffers slight degradation of the UAR in the target domain.

B. Model Architecture

The proposed voice disorder classification system is shown in Fig. 1. The feature extractor consists of separable convolutional layers receiving inputs of size (127, 251) with the first and second dimensions denoting the *frame length* and the *frequency basis* of LPSs, respectively. First, the 1D average-pooling layer (with a kernel size of 2 and stride of 2) reduces the input size along the dimension of frequency to be (127, 126), leaving the frame length unchanged. Subsequently, the downsampled inputs are extended with a dimension of channels to be (127, 126, 1). After the first 1D average-pooling layer, five identically separable convolutional blocks follow, each of which comprises a depth-wise convolutional layer (with akernel size of (3, 3, C_{i}) and stride of 1), a point-wise convolutional layer (with a kernel size of (1, 1, C_{i}, 16) and stride of 1), and a 2D average-pooling layer (with a kernel size 2 and stride of 2). Here, C_{i} is set to 1 in the first block and 16 in the rest. Batch normalization [74] is applied after all convolutional layers (depth-wise and point-wise), while LeakyReLUs [75] (with a negative slope of 0.2) are inserted after the depth-wise convolutional layers only. The final output of the feature extractor, viewed as a 1D vector of dimension 256, is regarded as the extracted feature carrying domain-invariant information to pass on to the next disease predictor and the domain classifier. The disease predictor is a fully connected layer of matrix size (256, k) with a softmax layer concatenated right after to predict the probability between the k distinct diseases, in our case, $k = 3$ (health, neoplasm, and benign structural diseases). The domain classifier is similar to the disease predictor, as a three-class classification task, only with the augmentation of a GRL ahead. It classifies the domains into clean, A/C, and street. Specifically, *the two types of noises, A/C and street, are annotated separately for the domain classifier*. The ADAM [76] optimizer with a learning rate of 0.001 and the regularization coefficient $\lambda = 0.5$ are used throughout the experiments unless otherwise specified.

C. Memory Usage and Computational Cost

First, we derive the capability of the separable convolutional layer. Consider a general 3D input $\mathbf{I} \in \mathbb{R}^{W_i \times H_i \times C_i}$, an output $\mathbf{O} \in \mathbb{R}^{W_o \times H_o \times C_o}$ and a corresponding convolution kernel $\mathbf{K} \in \mathbb{R}^{W_k \times H_k \times C_i \times C_o}$ in a standard convolutional layer, where W_i, H_i and C_i denote the width, height and number of channels of

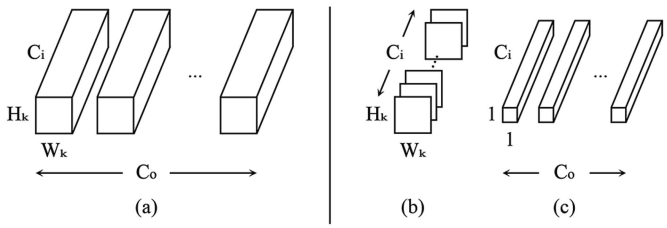

Fig. 2. (a) Standard convolutional filters are factorized into (b) depth-wise convolutional filters for intra-channel convolution and (c) point-wise convolutional filters for cross-channel projection.

an input feature respectively; similarly W_o, H_o, C_o, W_k, H_k denote those of an output feature and a convolution kernel. In the following, a filter is defined by a group of kernel parameters decomposed along the dimension of output channels. Therefore, $\mathbf{K}$ is regarded as C_o filters of size $W_k \times H_k \times C_i$.

Because a standard convolutional layer is parameterized by its convolution kernel $\mathbf{K}$ (Fig. 2(a)), we can directly derive the number of parameters in a single layer:

$$W_k \cdot H_k \cdot C_i \cdot C_o \tag{8}$$

However, due to only intra-channel convolutions, a convolution kernel $\widehat{\mathbf{K}}$ (Fig. 2(b)) of the depth-wise convolutional layer comprised of C_i filters with a size of $W_k \times H_k$. Moreover, the number of channels for the output features remains C_i. Therefore, a point-wise convolutional layer combining the information between channels and mapping to the desired shape is essential. Since we focus on the convolutions across channels, the width and height of a point-wise convolution kernel $\dot{\mathbf{K}}$ (Fig. 2(c)) are set to 1, forming a 1×1 convolutional layer. In total, the number of parameters of a separable convolutional layer is:

$$W_k \cdot H_k \cdot C_i + 1 \cdot 1 \cdot C_i \cdot C_o \tag{9}$$

By (8) and (9), the reduction ratio of the model size is:

$$\frac{W_k \cdot H_k \cdot C_i + 1 \cdot 1 \cdot C_i \cdot C_o}{W_k \cdot H_k \cdot C_i \cdot C_o} = \frac{1}{C_o} + \frac{1}{W_k \cdot H_k} \tag{10}$$

Next, we discuss the reduction in computation. The computational cost is dominated by multiplications of floating points, so we analyze the number of multiplications in convolutional layers. Because each element in the output feature is the dot product of the specific filter and part of the input feature, the number of multiplications is the filter size multiplied by the output size. The point-wise convolutions merely affect the dimension of channels, and hence the output feature of a depth-wise convolutional layer sizes $W_o \times H_o \times C_i$ after intra-channel convolutions. The following are computational costs for each type of convolution layer:

$$\begin{aligned} \text{Standard:} \quad & (W_k \cdot H_k \cdot C_i) \cdot (W_o \cdot H_o \cdot C_o) \\ \text{Depth- wise:} \quad & (W_k \cdot H_k) \cdot (W_o \cdot H_o \cdot C_i) \\ \text{Point- wise:} \quad & (1 \cdot 1 \cdot C_i) \cdot (W_o \cdot H_o \cdot C_o) \end{aligned} \tag{11}$$

The computational reduction ratio is then:

$$\frac{(W_k \cdot H_k) \cdot (W_o \cdot H_o \cdot C_i) + (1 \cdot 1 \cdot C_i) \cdot (W_o \cdot H_o \cdot C_o)}{(W_k \cdot H_k \cdot C_i) \cdot (W_o \cdot H_o \cdot C_o)}$$

$$= \frac{1}{C_o} + \frac{1}{W_k \cdot H_k} \tag{12}$$

On the other hand, one important reason to use DAT as the domain adaptation method in our system is that it does not increase any memory load or computational cost in the testing phase. As shown in the comparison of Fig. 1(a) and (b). During the training phase, the feature extractor collaborates with the domain classifier to jointly learn the domain-invariant features. However, in the testing phase, the disease predictor utilizes the well-trained features from both domains to diagnose diseases without the interference of the domain classifier. Thus, the number of parameters remains unchanged regardless of whether the DAT technique is used.

Therefore, (10) and (12) reveal that the entire memory usage and the computational cost can both be significantly reduced by over 73.9% in our design. Additional experimental details are provided in Section IV-B.

IV. Experiments

A. Dataset and Preprocessing

The voice samples were collected from the Far Eastern Memorial Hospital (FEMH) using a unidirectional microphone and a digital amplifier (CSL model 4150B, Kay Pentax). All participants were asked to sustain the vowel /:a/ for at least 2 seconds during recordings. The sampling rate was 44.1 kHz with a 16-bit resolution and the data were saved in an uncompressed wave format.

A total of 523 voice samples were recorded in a voice clinic as the source domain, containing 108 healthy voices, 112 glottic neoplasms and 303 benign structural diseases (*i.e.,* vocal nodules, polyps and cysts). Another 30 voice samples, 10 of each category, were synthesized with various noises of A/Cs and streets as the target domain data, where the labels were not provided during the training phase, for unsupervised domain adaptation. In this study, a 10-times 5-fold cross-validation approach was applied to validate the proposed system [77]. In each 5-fold cross-validation, the 523 speech samples were randomly divided into five equal partitions and each partition served as the testing fold used for evaluation in turns. This approach can reduce the bias (resulting from the environment) in evaluating the system. In Section IV-B, all scores reported were the averages of the 10×5 testing folds. The significance of the performance between different approaches was statistically measured on the 50 testing folds using independent t-test at 95% ($p < 0.05$). The voice samples in the testing fold were inferred under both the source domain in the clinic and the target domain corrupted by noises of A/Cs and streets to assess the effect of our system. That is, the target domain data are a corrupted versions of the source domain data during the testing phase. The approving institution of this study is Far Eastern Memorial Hospital, under the IRB/ethics board protocol number: 109063-E, and the date of approval was May 10, 2020.

In the target domain, we considered the two types of noises (A/C and street) at the same time. For the 30 voice samples used for unsupervised domain adaptation during training, half of the

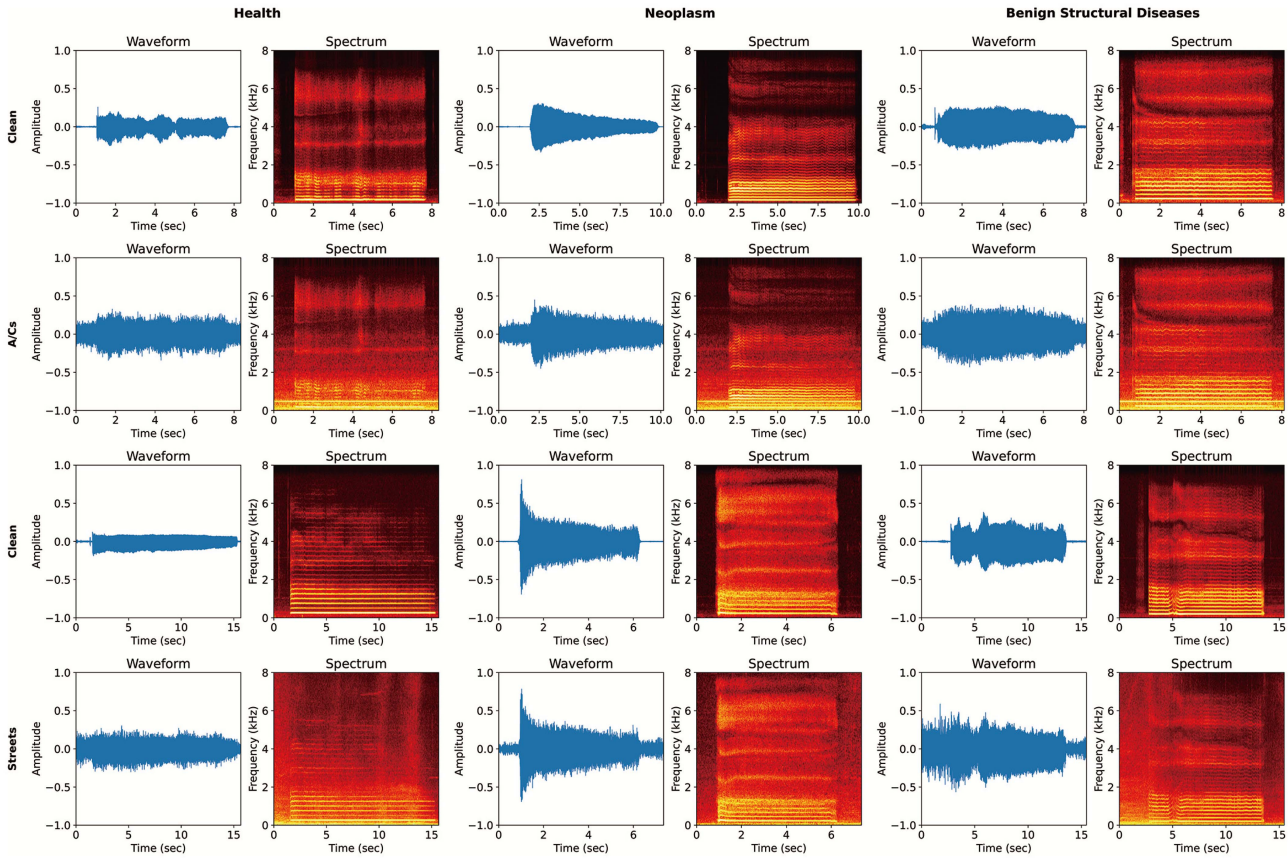

Fig. 3. Waveform and spectrogram plots of health, neoplasm, and benign structural diseases speech samples (the vowel /:a/ sound); the first and third rows list the plots of clean utterances, and the second and fourth rows list the plots of noisy utterances corrupted by noises of A/Cs and streets, respectively.

samples were corrupted with noises of A/Cs; the remaining 15 samples were corrupted with noises of streets. During the testing phase, each testing fold (including 104-105 voice samples) was corrupted using the same rule. Half of the samples (about 50) were corrupted with noises of A/Cs; another half (about 50) were corrupted with noises of streets. The magnitudes of noises were totally distinct between the training and testing phases, with signal-to-noise ratios (SNRs) of 0, 5 and 10 dB for training and 3, 6 and 9 dB for testing. Here, it should be emphasized that we computed the UAR in the target domain over all 104-105 voice samples. In other words, the target data UAR is the average score of the two types of noises.

Prior to training, the raw waves were first down-sampled from 44.1 kHz to 16 kHz and subsequently converted into LPS features using a Hamming filter, with a 31.25 ms window size and half of the window size as the frame shift. The LPS features were normalized by the standard score before fed to the models. During training, random segments of 127 frames (2 seconds) from the normalized LPS features were chosen as inputs for every epoch to increase the training variety of the models. During testing, the first 2 seconds of the normalized LPS features were fixed as the inputs.

Fig 3 visualizes the speech utterances (the vowel /:a/ sound) involved in the experiments. In the first and third rows of Fig. 3, the left, center, and right columns show the paired waveform and spectrogram plots of health, neoplasm, and benign structural diseases voice signals recorded under a clean condition, respectively. In the second and fourth rows of Fig. 3, the left, center, and right columns, demonstrate the paired waveform and spectrogram plots of health, neoplasm, and benign structural diseases voice signals under two noisy conditions (A/Cs and streets), respectively.

First, by comparing the plots of clean utterances, we can observe that health, neoplasm, and benign structural diseases sounds exhibit very different waveform-domain and time-frequency properties. Accordingly, we believe that a deep learning model can effectively classify these three types of sounds. Next, by comparing the plots of clean utterances and noisy utterances, we can clearly note that the noises of A/Cs corrupted the detailed structures of the voice signals, especially in regions below 400 Hz. In addition to the low frequencies, the noises of streets also influenced the details of high frequencies. From the noisy waveform and spectrogram plots in Fig. 3, we can infer that voice disorder classification is more challenging since the key structural details of voice signals have been considerably covered by the noise signals. Moreover, the obvious differences occurred not only between the clean and noisy plots, but also between the plots of the distinct noise types. This is the reason why the domain classifier was designed to identify the A/C and street domains separately in the proposed system.

TABLE I
COMPARISON OF SEPCONV-DAT WITH ITS VARIANTS FOR DOMAIN ADAPTATION

Model	Source Domain				Target Domain			
	Health	Neoplasm	Structural	**UAR**	Health	Neoplasm	Structural	**UAR**
StdConv	0.92	0.85	0.82	0.87*	0.48	0.79	0.61	0.63*
SepConv	0.88	0.85	0.80	0.85*	0.39	0.78	0.60	0.59*
SepConv-tgt	0.46	0.38	0.71	0.52*	0.64	0.69	0.56	0.63*
SepConv-ft	0.78	0.70	0.67	0.72*	0.74	0.84	0.60	0.72
SepConv-jnt	0.85	0.79	0.74	0.79	0.69	0.74	0.65	0.69*
SepConv-mmd	0.87	0.76	0.73	0.79	0.66	0.84	0.61	0.70*
SepConv-dat	0.88	0.79	0.72	0.80	0.70	0.81	0.64	0.72

*Indicates a significant difference (p-value < 0.05) between SepConv-dat and other variants.

TABLE II
USAGE OF RESOURCES FOR REPLACING STANDARD CONVOLUTIONS WITH SEPARABLE CONVOLUTIONS

Model	# Parameters ($\times 10^3$)	# MACs ($\times 10^6$)
StdConv	10.29	15.82
SepConv	2.69	3.64

B. Results

Table I lists the overall performances of the various baselines and the proposed method. First, we verified the effectiveness of the separable convolutional layers, **SepConv**. **SepConv** is similar to the architecture mentioned in Section III-B with the domain classifier removed and only source domain data were used during training. **StdConv**, on the other hand, replaces the separable convolutional layers in **SepConv** with standard layers such that the arguments of the input channels, output channels, kernel size, etc. are identical in these two baselines. Table I shows that the degradation is insignificant in **SepConv** with UARs reduced by only 4% in the source domain and 2% in the target domain when compared to **StdConv**. With almost no dropping performance in the UARs, **SepConv** significantly reduced the model size and computational cost by 73.9% in the number of parameters and 77.0% in the number of MACs, as presented in Table II. Because a MAC is the basic arithmetic unit of operation a model performs, counting the number of MACs in one forward-pass prediction of one input datum, which is independent of the hardware and platforms used, is one of the most common and fair approaches for comparing the computational cost. Otherwise, The computation time may vary when different computing hardware is used. In the following domain adaptation experiments, **SepConv** was the basis for comparison. Besides, the two baseline scores also showed that the noises tend to cause the models to misjudge the noisy inputs as neoplasms without domain adaptation.

Our proposed system, based on **SepConv** with 30 target domain samples provided in the DAT, is denoted by **SepConv-dat**. Three other "supervised" variants (**SepConv-tgt**, **SepConv-ft** and **SepConv-jnt**) and one "unsupervised" variant (**SepConv-mmd**) were constructed for systematic comparison, with the architecture fixed as **SepConv** yet the training strategies slightly altered as follows:

- **SepConv-tgt**: The model was trained *only on the 30 target domain samples with labels.* In turn, the target domain was an exposed domain to **SepConv-tgt**, yet the source domain became unseen.
- **SepConv-ft**: This variation used a pretrained **SepConv** as an initial state and was then fine-tuned using the 30 target samples with labels.
- **SepConv-jnt**: **SepConv** was trained from scratch with labeled data from both domains jointly.
- **SepConv-mmd**: MMD served as an unsupervised variant using statistical-based domain adaptation.

It is observed that **SepConv-dat** outperforms all baselines in the target data UAR, particularly the other systems **SepConv-tgt**, **SepConv-ft**, and **SepConv-jnt** with extra labels of the target data provided. Compared to **SepConv**, the UAR increased by 13% (from 59% to 72%) in the target data, with only slight degradation in the source domain, maintaining 80%.

In **SepConv-tgt**, there is no doubt that the target data UAR is better than that of the source data, due to the exposed target domain. However, the UAR of **SepConv-tgt** in the exposed (target) domain was not comparable with that of **SepConv** in the exposed (source) domain, with a reduction of up to 22%. Moreover, the UAR in the exposed (target) domain was even worse than that of the other compared systems in the unseen (target) domain. The poor performance reflects the impact of the small dataset.

The fine-tuning of **SepConv-ft** successfully improved the UAR in the target domain by 13%, but the degradation in the source domain was also obvious. The average UAR of **SepConv-ft** is almost equal to that of **SepConv**, which means that we simply obtained a trade-off between the two domains. The generalizability of model was not fully achieved by fine-tuning the target domain.

Intuitively, **SepConv-jnt** should achieve the best performance with sufficient data in both domains. However, owing to the

extremely small amount of target data under the proposed scenario, a severe imbalance of the two domains confines the improvement of generalizability. Even the target data become distractions that degrade the source domain score. Therefore, for the scenario with severe data imbalance that we intend to overcome, unsupervised domain adaptation algorithms are more suitable than supervised ones.

Compared to **SepConv-jnt**, **SepConv-mmd** reduces the effect of data imbalance by computing the distance between the means of the extracted features across domains. However, obtaining statistical values that can reflect the entire target domain through only 30 target data points is almost impossible. Finally, the total scores of **SepConv-mmd** were still worse than our proposed **SepConv-dat**. Consequently, **SepConv-dat** is the best method among these variants, which yields the largest improvement in the target domain by overcoming data imbalance due to the extremely deficient target data. Meanwhile, the high performance maintained in the source domain validates the generalizability.

Furthermore, we can observe that **SepConv-dat** is significantly different between the baselines and variants in most UARs, except for the target data UAR of **SepConv-ft**; the source data UAR of **SepConv-jnt** and **SepConv-mmd**. From the results, we first note that **SepConv-ft** specializes in the performance of the target domain after fine-tuning, and our system still achieves a comparable and similar target data UAR. Second, the labels of the source domains are exposed to **SepConv-dat**, **SepConv-mmd**, and **SepConv-jnt**, thus, they achieve a rather stable source data UAR. In conclusion, these three scores without significant differences happen to be the strengths of the corresponding variants. This confirms the power of **SepConv-dat**.

C. Ablation: Impact of Domain Classifier Via λ

We conducted an ablation study to learn how the regularization coefficient λ in (5) and (7) affect the domain adaptation performance. Owing to the unstable training procedure of the adversarial min-max method, the statistics of 200 models with random initialized weights were considered for each setup in the ablation experiment. Fig. 4 only shows the results of a specific fold, but the other folds are similar.

The coefficient λ was tuned from 0.01 to 10 to understand the effect. The results in Fig. 4 indicate that when $\lambda \geq 5$, the UAR performances in both domains broke down promptly. It was due to the gradient of L_d being over-amplified by the large λ, such that the overall update was guided away from the direction of minimizing L_y to affect the diagnosis prediction. The detailed process of increasing λ from 1 to 5 exhibited in Fig. 5 corroborates our inference: With the increment of λ, the classification gradually got worse.

On the other hand, the observed fact that when $\lambda \to 0$, the less domain-invariant the features are meets our intuition. The coefficient is somewhere between $[0, 5)$ to best construct the domain-invariant features. In this study, $\lambda \leq 0.5$ was observed to yield converging UARs in the source data, so that $\lambda = 0.5$ was eventually chosen to balance the accuracy and the domain invariance.

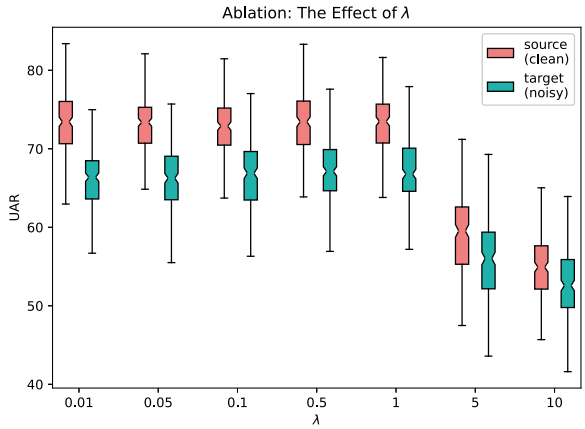

Fig. 4. Box plots of different λ. The source domain is the clean data collected in the clinics, and the target domain is the data corrupted by the noises of A/Cs and streets.

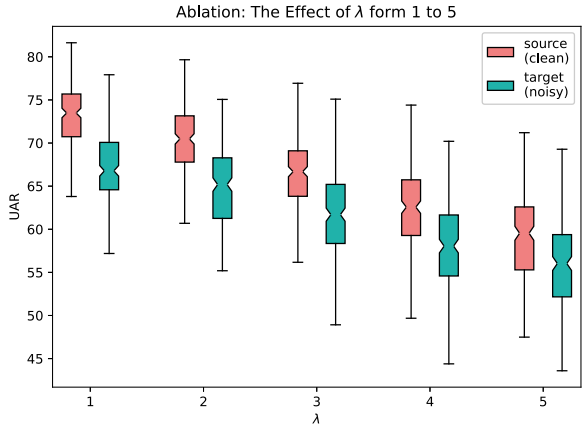

Fig. 5. Detailed process of λ increasing from 1 to 5.

D. Ablation: Reduction of Performance in Source Domain

To create the domain-invariant features, DAT guides the feature extractor to adjust in the opposite direction of minimizing the domain classifier loss L_d. Nevertheless, this approach inevitably diverges from the direction of minimizing the label predictor loss L_y, leading to a reduction in the UAR in the source domain (clean environment).

In this section, we propose a potential solution, a two-stage inference system, to alleviate the reduction of performance in the source domain. We introduced a noise detector that can identify whether the input voice samples are corrupted by noise or noise-free. In the first stage, the noise detector categorizes input voice samples as either corrupted by noise or noise-free. In the second stage, if the input voice samples are detected as corrupted by noise, we adopt the proposed **SepCNN-dat**, whereas if they are identified as noise-free, we employ the original **SepConv**.

In our experiment, we utilized a noise detector formed by CNN with the output layer having two dimensions. The presented results are the average of 10×5 testing folds. The noise detector showed an impressive classification accuracy of 91% in the first stage of the inference process. Overall, the two-stage

TABLE III
THE PERFORMANCE OF THE PROPOSED TWO-STAGE INFERENCE

Model	Source UAR	Target UAR
SepConv	0.85	0.59
SepConv-dat	0.80	0.72
Two-Stage	0.84	0.71

inference system achieved UARs of 0.84 and 0.71 in the source and target domains, respectively, which is very close to the best result of the single model (0.85 for SepConv and 0.71 for SepConv-dat). The results of the two-stage inference system are provided in Table III.

E. Visualization: Domain Invariance of Extracted Features

In addition to the significant progress of the UAR in the target domain shown in Table I, the visualization of the distributions of features extracted from the feature extractor further proves the effectiveness of the proposed system. Because the testing samples in the target domain are the same as those in the source domain except for the corruption by the noises, each source-target pair of extracted features should be close if high domain-invariant features are extracted. In Fig. 6, the t-SNE is used to visualize the distributions of the extracted features. Fig. 6(a) is the t-SNE of **SepConv** for a specified fold, but the other folds are similar. When investigating each category in Fig. 6(a), these two distributions are different and have no correlation. This explains the low accuracy of the target domain data for **SepConv**. However, in Fig. 6(b), the t-SNE of **SepConv-dat** exhibits that most samples in the source domain and the target domain are in pairs. Whether the samples are corrupted with the A/C noises or the street noises, the proposed system could map them to corresponding clean features successfully.

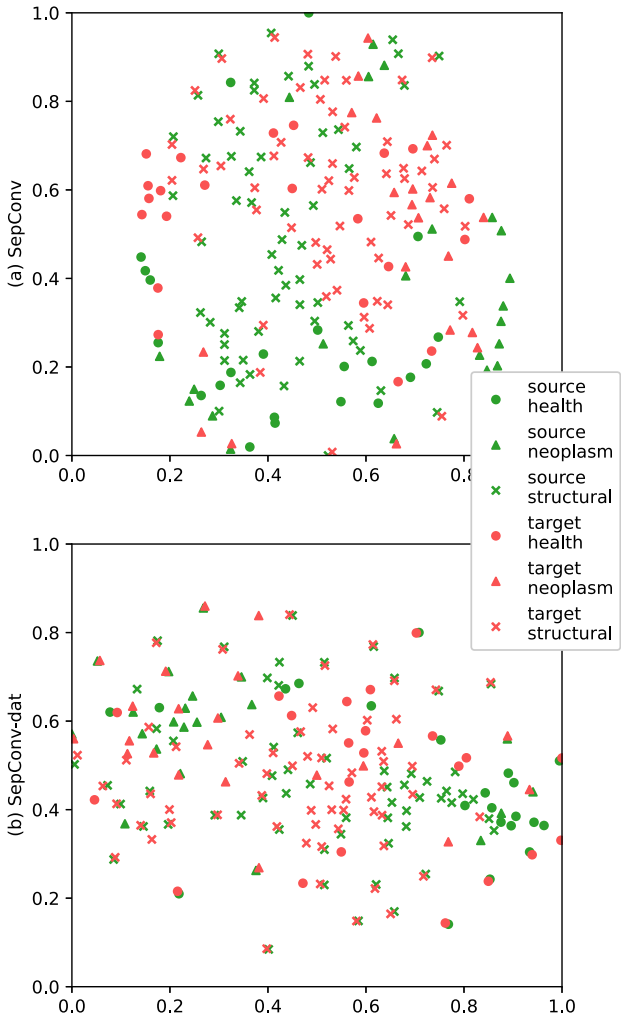

Fig. 6. t-SNE plots of the latent features extracted by the feature extractor in (a) the **SepConv** and (b) the **SepConv-dat**. The data in the source domain are marked in green, whereas the data in the target domain are marked in red.

V. FUTURE WORKS

We consider our future work from two perspectives. First, in the aspect of the clinic, we are planning to perform both internal and external validations of our proposed system. The internal validation will verify the proposed system on pathological data collected within Far Eastern Memorial Hospital (FEMH), the approval institution of this study. Conversely, the external validation will be conducted in partnership with other hospitals to test the robustness of our system in recording environments that are unseen in this study. In fact, We have already begun internal validation under noisy conditions, and interim results are included in the Appendix.

Second, we are also devoted to technological improvement. Although this study investigated a more practical real-world application and achieved significant progress, the functionality of adapting to hardware mismatch should be incorporated into the proposed system, especially if implementing our approach with IoT technology or evolving it to personalized healthcare. Therefore, our next step in technological improvement will introduce two domain classifiers, one for background noises and another for recording devices. The interaction effects of the two min-max objective functions make the training procedure more challenging. However, this integration allows our system to be more applicable in practice.

VI. CONCLUSION

In the past decade, the automatic detection and classification of pathological voices have achieved outstanding performance with the advancement of machine learning methods. Nevertheless, two main challenges arise in practical applications: (1) state-of-the-art models often require increasing memory load and computational cost, whereas resources are rather limited for embedded systems; and (2) the domain mismatch between the training and real-world data significantly degrades the classification performance. To overcome these difficulties, we utilized separable convolutional layers and a DAT module to build a compressed and domain-robust system. Seven experiments were conducted and their results were compared. The effect of λ was also discussed. Therefore, We proposed an unsupervised

TABLE IV
EXPERIMENTAL RESULTS OF **SEPCONV** AND **SEPCONV-DAT** TESTED WITH REAL-WORLD DATA AND SYNTHETIC TARGET DATA

Model	Real-World Domain			Synthetic Target Domain		
	Health	Structural	**UAR**	Health	Structural	**UAR**
SepConv	0.30	0.45	0.38	0.39	0.60	0.50
SepConv-dat	0.60	0.63	0.62	0.70	0.64	0.67

domain adaptation system that is jointly trained by using sufficient labeled data in the source domain and a small amount of unlabeled data in the target domain. The results showed that the UAR in the noisy real-world domain improved by 13%, and that in the clinic domain remained at 80% with only slight degradation. Moreover, the numbers of parameters and MACs were significantly reduced by 73.9% and 77.0%, respectively.

It is concluded that our proposed system efficiently reduces computational and memory usage, and effectively eliminates the domain mismatch.

APPENDIX

We gathered voice data from 21 participants, with 10 having healthy voices and 11 having benign structural diseases (unfortunately, we were unable to obtain voice samples from patients with neoplasms due to limited revision time). All of the voice samples we collected were corrupted by air conditioning noise in the real world.

Table IV shows the real-world experimental results of **SepConv** and **SepConv-dat**. It should be noted that we employed the same **SepConv** and **SepConv-dat** as in Section IV-B and evaluated their performance on two distinct test sets: real-world (noisy) data and synthetic target data. As neoplasm data was not available, the UARs are the average over the recall score of healthy and benign structural diseases. From the table, it can be observed that the prediction results of both **SepConv** and **SepConv-dat** are affected when the input is changed from synthetic data to real-world data due to acoustic mismatches. Notably, SepConv-dat, which employs DAT techniques, still exhibits better performance than **SepConv**. This validates that despite being trained with synthetic noisy data, **SepConv-dat** still provides notable performance improvements against noise in a real-world scenario.

It is also noteworthy that **SepConv-dat** exhibits only a 5% reduction in UAR, while the original **SepConv** displays a decrease of 12% in UAR. If we regard the real-world data as an unseen domain during the training phase, **SepConv-dat** again provides better generalization ability when encountering unknown domains.

REFERENCES

[1] J. K. Laguaite, "Adult voice screening," *J. Speech Hear. Disord.*, vol. 37, no. 2, pp. 147–151, 1972.

[2] D. E. Morley, "A ten-year survey of speech disorders among university students," *J. Speech Hear. Disord.*, vol. 17, no. 1, pp. 25–31, 1952.

[3] L. O. Ramig and K. Verdolini, "Treatment efficacy: Voice disorders," *J. Speech, Lang., Hear. Res.*, vol. 41, no. 1, pp. S101–S116, 1998.

[4] N. Roy et al., "Voice disorders in the general population: Prevalence, risk factors, and occupational impact," *Laryngoscope*, vol. 115, no. 11, pp. 1988–1995, 2005.

[5] S. M. Cohen, "Self-reported impact of dysphonia in a primary care population: An epidemiological study," *Laryngoscope*, vol. 120, no. 10, pp. 2022–2032, 2010.

[6] S. R. Best and C. Fakhry, "The prevalence, diagnosis, and management of voice disorders in a National Ambulatory Medical Care Survey (NAMCS) cohort," *Laryngoscope*, vol. 121, no. 1, pp. 150–157, 2011.

[7] S. M. Cohen et al., "Prevalence and causes of dysphonia in a large treatment-seeking population," *Laryngoscope*, vol. 122, no. 2, pp. 343–348, 2012.

[8] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Proc. IEEE 2nd Joint 24th Annu. Conf. Annu. Fall Meeting Biomed. Eng. Soc. Eng. in Med. and Biol.*, 2002, pp. 182–183.

[9] P. Henríquez et al., "Characterization of healthy and pathological voice through measures based on nonlinear dynamics," *IEEE Trans. audio, speech, Lang. Process.*, vol. 17, no. 6, pp. 1186–1195, Aug. 2009.

[10] G. Vaziri, F. Almasganj, and R. Behroozmand, "Pathological assessment of patients' speech signals using nonlinear dynamical analysis," *Comput. Biol. Med.*, vol. 40, no. 1, pp. 54–63, 2010.

[11] J. M. Miramont, M. A. Colominas, and G. Schlotthauer, "Emulating perceptual evaluation of voice using scattering transform based features," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1892–1901, Aug. 2022.

[12] M. Cooke, O. Scharenborg, and B. T. Meyer, "The time course of adaptation to distorted speech," *J. Acoustical Soc. Amer.*, vol. 151, no. 4, pp. 2636–2646, 2022. [Online]. Available: https://doi.org/10.1121/10.0010235

[13] M. Illa et al., "Pathological voice adaptation with autoencoder-based voice conversion," in *Proc. 11th ISCA Speech Synthesis Workshop*, 2021, pp. 19–24.

[14] B. M. Halpern et al., "An objective evaluation framework for pathological speech synthesis," in *Proc. Speech Commun.; 14th ITG Conf.*, 2021, pp. 1–5.

[15] J. Unger et al., "A noninvasive procedure for early-stage discrimination of malignant and precancerous vocal fold lesions based on laryngeal dynamics analysis early-stage detection of malignant vocal fold lesions," *Cancer Res.*, vol. 75, no. 1, pp. 31–39, 2015.

[16] R. Fraile et al., "Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficient parameters and differentiation of patients by sex," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 3, pp. 146–152, 2009.

[17] S. C. Costa, B. G. A. Neto, and J. M. Fechine, "Pathological voice discrimination using cepstral analysis, vector quantization and hidden Markov models," in *Proc. IEEE 8th Int. Conf. Bioinform. Bioeng.*, 2008, pp. 1–5.

[18] P. Harar et al., "Towards robust voice pathology detection," *Neural Comput. Appl.*, vol. 32, no. 20, pp. 15747–15757, 2020.

[19] K. Umapathy et al., "Discrimination of pathological voices using a time-frequency approach," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 3, pp. 421–430, Mar. 2005.

[20] M. Pützer and W. Wokurek, "Electroglottographic and acoustic parametrization of phonatory quality provide voice profiles of pathological speakers," *J. Voice*, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199721001211

[21] C. Zhou et al., "Gammatone spectral latitude features extraction for pathological voice detection and classification," *Appl. Acoust.*, vol. 185, 2022, Art. no. 108417.

[22] J. I. Godino-Llorente, P. Gomez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 10, pp. 1943–1953, Oct. 2006.

[23] M. K. Arjmandi and M. Pooyan, "An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine," *Biomed. Signal Process. Control*, vol. 7, no. 1, pp. 3–19, 2012.

[24] M. Markaki and Y. Stylianou, "Voice pathology detection and discrimination based on modulation spectral features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 1938–1948, Sep. 2011.

[25] I. Hammami, L. Salhi, and S. Labidi, "Pathological voices detection using support vector machine," in *Proc. IEEE 2nd Int. Conf. Adv. Technol. Signal Image Process.*, 2016, pp. 662–666.

[26] L. Verde, G. De Pietro, and G. Sannino, "Voice disorder identification by using machine learning techniques," *IEEE Access*, vol. 6, pp. 16246–16255, 2018.

[27] M. Pishgar et al., "Pathological voice classification using mel-cepstrum vectors and support vector machine," in *Proc. IEEE Int. Conf. Big Data*, pp. 5267–5271, 2018.

[28] J. D. Arias-Londoño et al., "On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices," *Logopedics Phoniatrics Vocol.*, vol. 36, no. 2, pp. 60–69, 2011.

[29] J. D. Arias-Londoño et al., "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.

[30] M. Dahmani and M. Guerti, "Recurrence quantification analysis of glottal signal as non linear tool for pathological voice assessment and classification," *Int. Arab J. Inf. Technol.*, vol. 17, no. 6, pp. 857–866, 2020.

[31] A. Basalamah et al., "A highly accurate dysphonia detection system using linear discriminant analysis," *Comput. Syst. Sci. Eng.*, vol. 44, no. 3, pp. 1921–1938, 2023.

[32] H. Wu et al., "Convolutional neural networks for pathological voice detection," in *Proc. IEEE 40th Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2018, pp. 1–4.

[33] V. Gupta, "Voice disorder detection using long short term memory (LSTM) model," 2018, *arXiv:1812.01779*.

[34] S.-H. Fang et al., "Detection of pathological voice using cepstrum vectors: A deep learning approach," *J. Voice*, vol. 33, no. 5, pp. 634–641, 2019.

[35] C.-H. Hung et al., "Using SincNet for learning pathological voice disorders," *Sensors*, vol. 22, no. 17, 2022, Art. no. 6634.

[36] J.-Y. Lee, "Experimental evaluation of deep learning methods for an intelligent pathological voice detection system using the Saarbruecken voice database," *Appl. Sci.*, vol. 11, no. 15, 2021, Art. no. 7149.

[37] W. Ariyanti et al., "Ensemble and multimodal learning for pathological voice classification," *IEEE Sens. Lett.*, vol. 5, no. 7, pp. 1–4, Jul. 2021.

[38] K. G. Dávid Sztahó and T. M. Gábriel, "Deep learning solution for pathological voice detection using LSTM-based autoencoder hybrid with multi-task learning," in *Proc. 14th Int. Joint Conf. Biomed. Eng. Syst. Technol.*, 2021, pp. 135–141.

[39] T. Lee et al., "Automatic speech recognition for acoustical analysis and assessment of cantonese pathological voice and speech," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, 2016, pp. 6475–6479.

[40] Y. Liu et al., "Acoustical assessment of voice disorder with continuous speech using ASR posterior features," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 6, pp. 1047–1059, Jun. 2019.

[41] S.-H. Fang et al., "Combining acoustic signals and medical records to improve pathological voice classification," *APSIPA Trans. Signal Inf. Process.*, vol. 8, 2019, Art. no. e14.

[42] T. Kojima et al., "Objective assessment of pathological voice using artificial intelligence based on the GRBAS scale," *J. Voice*, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199721004094

[43] Y.-T. Hsu et al., "Robustness against the channel effect in pathological voice detection," *NIPS Mach. Learn. Health Workshop*, 2018. [Online]. Available: https://arxiv.org/abs/1811.10376

[44] Z. Fan et al., "Class-imbalanced voice pathology detection and classification using fuzzy cluster oversampling method," *Appl. Sci.*, vol. 11, no. 8, 2021, Art. no. 3450.

[45] Q. Jinyang et al., "Pathological voice feature generation using generative adversarial network," in *Proc. IEEE Int. Conf. Sens., Meas. Data Analytics Era Artif. Intell.*, 2021, pp. 1–6.

[46] G. Muhammad et al., "Smart health solution integrating IoT and cloud: A case study of voice pathology monitoring," *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 69–73, Jan. 2017.

[47] G. Muhammad et al., "Edge computing with cloud for voice disorder assessment and treatment," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 60–65, Apr. 2018.

[48] M. S. Hossain, G. Muhammad, and A. Alamri, "Smart healthcare monitoring: A voice pathology detection paradigm for smart cities," *Multimedia Syst.*, vol. 25, pp. 565–575, 2019.

[49] S. U. Amin et al., "Cognitive smart healthcare for pathology detection and monitoring," *IEEE Access*, vol. 7, pp. 10745–10753, 2019.

[50] A. Ramalingam, S. Kedari, and C. Vuppalapati, "IEEE FEMH voice data challenge 2018," in *Proc. IEEE Int. Conf. Big Data*, 2018, pp. 5272–5276.

[51] M. Pham, J. Lin, and Y. Zhang, "Diagnosing voice disorder with machine learning," in *Proc. IEEE Int. Conf. Big Data*, 2018, pp. 5263–5266.

[52] T. Grzywalski et al., "Parameterization of sequence of MFCCs for DNN-based voice disorder detection," in *Proc. IEEE Int. Conf. Big Data*, 2018, pp. 5247–5251.

[53] C. Bhat and S. K. Kopparapu, "FEMH voice data challenge: Voice disorder detection and classification using acoustic descriptors," in *Proc. IEEE Int. Conf. Big Data*, 2018, pp. 5233–5237.

[54] K. Degila, R. Errattahi, and A. El Hannani, "The UCD system for the 2018 FEMH voice data challenge," in *Proc. IEEE Int. Conf. Big Data*, 2018, pp. 5242–5246.

[55] J. D. Arias-Londoño et al., "Byovoz automatic voice condition analysis system for the 2018 FEMH challenge," in *Proc. IEEE Int. Conf. Big Data*, 2018, pp. 5228–5232.

[56] K. A. Islam, D. Perez, and J. Li, "A transfer learning approach for the 2018 FEMH voice data challenge," in *Proc. IEEE Int. Conf. Big Data*, 2018, pp. 5252–5257.

[57] B. Jacob et al., "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2704–2713.

[58] Y.-C. Lin et al., "SEOFP-NET: Compression and acceleration of deep neural networks for speech enhancement using sign-exponent-only floating-points," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 30, pp. 1016–1031, 2021.

[59] G. Hinton et al., "Distilling the knowledge in a neural network," *NIPS Deep Learn. Representation Learn. Workshop*, 2015. [Online]. Available: https://arxiv.org/abs/1503.02531

[60] M. Wang, B. Liu, and H. Foroosh, "Factorized convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 545–553.

[61] E. Romera et al., "ERFNET: Efficient residual factorized ConvNet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, Jan. 2017.

[62] A. G. Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861* .

[63] K. Bousmalis et al., "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3722–3731.

[64] E. Tzeng et al., "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*.

[65] H.-Y. Lin et al., "Unsupervised noise adaptive speech enhancement by discriminator-constrained optimal transport," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 19935–19946 .

[66] H. Hu et al., "A variational Bayesian approach to learning latent variables for acoustic knowledge transfer," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 4041–4045.

[67] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.

[68] P. O. Pinheiro, "Unsupervised domain adaptation with similarity learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8004–8013.

[69] M. Sandler et al., "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[70] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.

[71] I. Redko et al., "A survey on domain adaptation theory: Learning bounds and theoretical guarantees," 2020, *arXiv:2004.11829*.

[72] Y. Zhang et al., "Pathological voice detection using joint subsapce transfer learning," *Appl. Sci.*, vol. 12, no. 16, 2022, Art. no. 8129.

[73] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*.

[74] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[75] B. Xu et al., "Empirical evaluation of rectified activations in convolutional network," *ICML Deep Learn. Workshop*, 2015. [Online]. Available: https://arxiv.org/abs/1505.00853

[76] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR Conf. Track*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[77] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognit.*, vol. 48, no. 9, pp. 2839–2846, 2015.