A Multi-stage Diagnosis Strategy based on Transfer Learning for a Class of Rare Laryngeal Movement Disorders

Yu Yao[a,c], Maria Powell[b], Jules White[c], Jian Feng[a], Quchen Fu[c], Peng Zhang[c], Douglas C. Schmidt[c]

[a]*the College of Information Science and Engineering, Northeastern University, Wenhua Road 3-11, Shenyang, 110819, Liaoning, P. R. China.*
[b]*Vanderbilt University Medical Center, 1211 Medical Center Drive, Nashville, 37232, TN, USA*
[c]*the Department of Electrical Engineering and Computer Science, Vanderbilt University,, 2301 Vanderbilt Place, Nashville, 37235, TN, USA*

Abstract

Background: Laryngeal movement disorders (LMDs) is a class of rare neurological disorders, including three sub-disorders: essential tremor of voice (ETV) and abductor and adductor spasmodic dysphonia (ABSD and ADSD). The debilitating socioemotional and socioeconomic consequences of LMDs motivate researchers to develop automated methods to diagnose these disorders early. Deep learning-based methods have been recently introduced to perform voice disorder detection, but differentiating the three sub-disorders of LMDs remains hard due to limited available databases. Moreover, unlike other voice disorders (such as structural disorders) that can be diagnosed accurately by visualizing the larynx, LMDs are most identified from auditory cues in recording signals. Although prior work has demonstrated that sustained phonations can help detect pathological voice from healthy voice, an open research question is which sub-disorder benefits most from diagnosis based on sustained phonations.

Method: A Transfer learning strategy is developed for LMD diagnosis with limited data. First, an extra vocally healthy database from the International Dialects of English Archive (IDEA) is employed to pre-train a convolutional autoencoder. The proportion of the pre-trained encoder to transfer is then explored, and its impact on LMD diagnosis is evaluated, yielding a two-stage transfer model. Finally, to verify the different effects of applying sustained

phonations on diagnosing the three sub-disorders and to boost the final diagnostic performance, a third stage is designed following the initial two stages, where information of pathological sustained phonations is embedded into the model.
Results: The analysis of clinician-labeled pathological data obtained from the Vanderbilt University Medical Center (VUMC) shows that the proposed multi-stage transfer learning strategy can produce (1) accuracy of 65.3% on classifying normal and other three sub-disorders all at once, (2) accuracy of 85.3% in differentiating normal, ABSD, and ETV, and (3) accuracy of 77.7% for normal, ADSD and ETV. Moreover, diagnosing ETV shows sensitivity to sustained phonations within the current database.

Keywords: Laryngeal Movement Disorders, Autoencoder, Transfer Learning, Limited data, Sustained Phonation

1. Introduction

The challenges of diagnosing laryngeal movement disorders. Laryngeal movement disorders (LMDs), including adductor spasmodic dysphonia (ADSD), abductor spasmodic dysphonia (ABSD) and essential tremor of voice (ETV) are neurogenic disorders affecting the larynx. LMDs are considered rare, with estimated incidence 1:100,000 in North America [1]. These disorders produce involuntary spasms of the vocal folds during voiced (ADSD) or unvoiced sounds (ABSD), and/or rhythmic oscillations of the larynx resulting in tremulous variations in pitch and loudness (ETV). Although LMDs are not life threatening, they can significantly impact patient quality of life [2]. For example, patients report heavy psycho-social burdens for themselves and their families as a result of their decline in communicative function [3].

Currently, there is no definitive medical test to confirm diagnosis of spasmodic dysphonia [4, 5]. The diagnosic process typically involves consultation with multiple providers, and the encounter-to-diagnosis lag time is almost 4.5 years [6]. Delay in diagnosis is driven by a lack of objective assessments to accurately and efficiently screen and diagnose ADSD and ABSD [4]. Unlike other voice disorders that rely on visualization of the vocal folds for accurate diagnosis [7, 8], ADSD and ABSD are most effectively identified from auditory cues in the acoustic signal [9]. Diagnosis based on perceptual analysis requires special clinical expertise in identifying the salient acoustic qualities that differentiate these disorders from each other and/or other functional voice disorders [10, 11].

For individuals with ETV, laryngeal oscillations associated with the disorder can be appreciated via laryngoscopy; however, severe ETV can present auditorily as voice breaks on voiced sounds. Further, ETV may present in isolation, or concomitantly with ADSD and/or ABSD. Thus, it is important that acoustic-based diagnostic tools can differentiate between these disorders to help differentiate mixed phenotypes from pure ADSD or ABSD. Due to the challenges in differentially diagnosing these disorders and the high reliance on the acoustic signal for obtaining diagnosis, acoustic-based machine learning diagnostic tools are a promising avenue of research.

Applying transfer learning to improve LMD diagnosis. The process of detecting pathological voice from healthy voice is a binary classification problem referred to as ”voice disorder detection”. At the next level, the process of diagnosing different types of voice disorders is a multi-class classification problem referred to as ”voice disorder classification”. To date, most voice disorders multi-class classification tasks have focused on discriminating between broader diagnostic categories (e.g. benign lesions vs malignancy vs neurological disorders, etc.). Deep learning models have demonstrated their effectiveness and superiority in addressing these types of voice disorder detection and classification problems [12, 13]. The focus of this paper is to differentially diagnose subtypes within the single diagnostic category of LMDs, specifically pure ADSD, ABSD, and ETV.

LMDs are rare among the general population, unlike the disorders mentioned above that have sufficient data to use for training deep learning models. A limited amount of data is thus available for LMDs, which impedes the development of automated deep learning diagnostic models in this domain. Such difficulty on obtaining a diagnosis further limits the amount of data available to train automated diagnostic techniques based on deep learning.

To address this gap, this paper focuses on applying automated transfer learning methods to diagnose LMDs. Transfer learning is widely employed to handle limited data problem in computer vision and speech recognition fields [14, 15]. This method first pre-trains a deep learning model with a amount of source data related to the target task and then fine-tunes the pre-trained model with the limited target data.

We pre-trained a backbone network using a vocally healthy database from the International Dialects of English Archive (IDEA) and then applied this pre-trained network to develop two transfer approaches that

1. Use the reconstructed error directly, versus
2. Just combine an encoder with an additional classifier.

We then evaluated the performance of each approach on an LMDs database contributed by Vanderbilt University Medical Center (VUMC). To investigate the impact of the sustained phonations on diagnosing the three subdisorders of LMDs, the data of pathological sustained phonations were separately used to fine-tune the model. Finally, a multi-stage transfer learning strategy was developed and evaluated empirically by diagnosing ABSD, ADSD and ETV from the VUMC database.

Paper organization. The remainder of this paper is organized as follows: Section 2 provides an overview of voice disorder detection techniques and compares/contrasts this related work with our transfer learning approach; Section 3 describes databases, data pre-processing, and our transfer learning strategy for diagnosing LMDs; Section 4 analyzes empirical results and discusses their implications for research on LMD diagnosis; and Section 5 presents concluding remarks and future work.

2. Background and Comparison with Related Work

This section presents an overview of prior work on voice disorder detection and compares and contrasts this related work with our transfer learning approach.

2.1. Voice Disorder Detection and Classification Techniques

Automatic voice disorder detection is well studied in recent years. Due to the relatively ease, low cost, and non-invasive nature of data collection, many acoustic features (such as jitter, shimmer, pitch, harmonic-to-noise-ratio etc. [16, 17]) have been devised to detect pathological voice. For example, Harar et al. [18] extracted these conventional features from sustained phonations and successfully detected pathological voice.

Unfortunately, the clinical utility of these measures is diminished by the time-intensive nature of manual extraction. Moreover, the diagnostic relevance of any given measure may be limited to specific genders or specific speech tasks. To address these limitations, features originating from speech recognition and natural language processing have been introduced.

For example, Javanmardi et al. [19] and Chen et al. [20] investigated the impact of Mel-frequency cepstral coefficients (MFCC) on detecting pathological voice from the health. Likewise, Akbari and Arjmandi [21] showed the effectiveness of linear prediction cepstral coefficients (LPCC). Moreover, Hammami et al. [22] used features obtained by discrete wavelet transform (DWT) combined with a support vector machine (SVM) to achieve good performance on the Saarbrücken voice database (SVD) [23]. Beside investigating the discriminative features for voice disorders, various classifiers (such as SVM [24], random forest [25], and the Gaussian mixture model (GMM) [26]) were also developed to process these features to support the final diagnosis.

Recently, deep learning has attracted much interest in the field of vocal disordered. For example, Fang et al. [12] demonstrated how deep learning models are superior to conventional methods with sustained vowel /a/. Convolutional neural networks (CNNs) are explored and compared on the SVD dataset in [27, 28]. Likewise, Zhang et al. [29] employed residual network to improve the robustness of the voice pathology detection system. Moreover, recurrent neural networks (RNNs) coupled with amplitude and phase spectrogram were designed for rating pathological voice [30].

Voice disorder classification is also an important task worthy of study because classification results can help clinicians further locate the causes of the disorders and prescribe the appropriate treatment. For example, Al-Nasheri et al. [31] classified cyst, polyp, and paralysis via a one-versus-rest way, using correlation functions based on different frequency bands. Likewise, by using sustained phonations and running speech data, Ali et al. [32] employed GMM to differentiate ADSD, keratosis, vocal fold nodules, vocal fold polyp, and paralysis, which demonstrated that running speech is more suitable for screening purposes in the context of daily communication. Likewise, Geng et al. [13] fine-tuned the pre-trained 18-layer residual network (ResNet18) with the Mel-spectrogram of voice signal and electroglottography (EGG). Similarly, Ding et al. [33] proposed a deep connected attention ResNet to classify four different types of disorders with different causes.

Much of the prior work distinguished voice disorders of different causes. For example, Geng et al. [13] classified spasmodic dysphonia (SD) and nodules etc., where SD is caused by the problem of nervous system, while nodules is caused by physical changes. In addition, compared with structural disorders, such as cyst, nodules, and laryngitis that were widely studied in the existing works, neurological voice disorders have received relatively little attention. The reason may stem from the fact that current work was mainly conducted based on public databases, such as SVD and the Massachusetts Eye and Ear Infirmary (MEEI) database [34], where the data of structural disorders are well established. However, for neurological voice disorders, especially LMDs, the available data is still limited due to long periods of diagnosis[1] and their rarity in prevalence.

2.2. Open Research Question: Transfer Learning for Voice Disorder Classification

Inspired by the prior work described above, this paper analyzes an open research question: *can voice/speech data from healthy subjects of different genders and ages be used as source data to pre-train a deep learning model that accurately captures the basic universe of features that help classify ADSD, ABSD, and ETV via transfer learning.*

Transfer learning has been applied to achieve voice disorder detection Guedes et al. [35]. For example, VGGish was adopted and pre-trained with AudioSet [36], where the embedding layer was used to extract characteristics matrices. A new classifier was then developed based on the embeddings of the limited target data.

Likewise, Autoencoder (AE) [37, 38] is widely used in transfer learning approaches and can be trained only with "normal" data, which is appropriate for our open research question setforth above. It also has the relatively simpler structure and more straightforward training process when compared with other more complicated and larger networks. We therefore selected AE as the backbone network for transfer learning in our study.

In addition to the structure of the model, this paper also considers the types of data to employ when diagnosing pathological voice in a clinical setting. Prior work [32, 39] has shown the promise of using running speech for voice disorder detection and classification. Supported by this work and thanks to VUMC, the first three sentences of Rainbow Passage from subjects with LMDs were collected and used as running speech in the database.

Likewise, the importance of sustained vowels for voice disorder detection is well demonstrated in recent few years [40]. It is still an open research question, however, which LMD sub-disorder (*i.e.*, ABSD, ADSD and/or ETV) benefits most from applying sustained phonations in the diagnosis process. As discussed in Section 4, we investigated this research question by designing a separate transfer learning stage for sustained phonations and analyzing the model's performance when diagnosing different sub-disorders.

In summary, the key contributions of this paper are:

[1] An estimated average of 4.4 year delay is incurred from initial clinician evaluation to final diagnosis of LMDs [6].

1. Diagnosing a class of specific neurogenic LMDs that differs from prior work on voice disorders stemming from various causes.
2. Creating a multi-stage transfer learning strategy to address the problem of limited available LMD data, where the questions of data representations, what to transfer, and how to transfer are analyzed empirically to determined the most effective transfer approach.
3. Investigating the effects of applying sustained phonations to diagnose the three LMD sub-disorders (*i.e.*, ABSD, ADSD and ETV).

3. Materials and Method for LMDs diagnosis

This section describes the databases, data pre-processing, and transfer learning strategy we applied to diagnose LMDs.

3.1. Database Selection

This study used the voice recordings of 134 (36 male and 98 female) patient encounters across all three LMD sub-disorders (*i.e.*, ADSD (n=44), ABSD (n=45), and ETV (n=45)) collected at the Vanderbilt University Medical Center (VUMC).[2] As part of the standard of care, subjects seen at the Vanderbilt Voice Center provide standardized voice/speech recordings that include the first three sentences of the Rainbow Passage [41] and short phrases, counting, and sustained phonation, such as /a/ and /i/. Patients record their voice/speech in a quiet clinic room using an omnidirectional lapel microphone with a 44.1kHz sampling rate.

Voice/speech recordings of vocally-healthy speakers were obtained from the International Dialect of English Archives (IDEA), which is a repository of voice recordings from native and non-native English speakers from 44 countries and 23 US states. These voice/speech recordings include both the Rainbow Passage [41] and spontaneous speech, which are similar to the VUMC recordings. In our study, the recordings of 145 (61 male and 84 female) native English speakers were used as a control to built the normal database, with 74 speakers from the US and 71 speakers from other countries.[3]

[2] These recordings were collected under IRB # 181191, approved 07/26/2018 by the Medical Institutional Review Board at Vanderbilt University Medical Center in Nashville, Tennessee.

[3] Although the recording environments of the VUMC and IDEA were not identical, each recording was captured in a quiet room with a microphone having a relatively flat frequency response.

3.2. Data Pre-processing

The original recordings were resampled at frequency of 22kHz in the data pre-processing step. For vocally-healthy recordings, 130 out of 145 speakers were then randomly selected for pre-training, which we refer to as “PreData” (Pre-training data). The remaining 15 speakers (8 from the US and 7 from other countries), along with 134 recordings of LMDs, were next used to evaluate our transfer learning strategy performance. We refer to these 149 recordings as “FTData” (fine-tuning data).

Exploring the effects of sustained phonation on diagnosing LMDs is one of our research goals. The sustained phonation data from encounters with LMDs patients was therefore used to create an independent dataset by replicating sustained phonation of each patient encounter up to 25 seconds. We refer to this dataset as “SusData” (sustained phonation data).

The fixed training and testing sets for FTData and SusData were not specifically split since we employed k-fold cross validation in this study due to its limited data size. The subjects/encounters in the dataset were thus divided into k smaller sets of approximately equal size. For each unique set, the model was evaluated on it, but trained with the remaining $k-1$ sets. The final performance of the model was evaluated by summarizing its performance on each fold. Each subject/encounter was assigned to an individual set and stayed in that set for the duration of the procedure to ensure each subject/encounter was used in the testing set one time and used to train the model $k-1$ times.

We next pre-processed the raw recordings into the following two formats provided as input for the backbone network:

- **Time-series waveform**, where the raw recordings were split into several three second segments (66,150 data points) and a single segment was treated as a sample with the format of time-series waveform. The choice of three seconds corresponds to clinician heuristics [42] that this duration contains sufficient information to diagnose LMDs.[4]

- **Mel-frequency spectrogram format**, which is obtained based on the segments of the time-series waveform format. This approach divides a segment into short-time frames by Hann windows with 50% overlap between consecutive frames. We applied this step to avoid losing the

[4] Future work will investigate the optimal segment length for diagnosis.

frequency contours of the signal over time when performing a Fourier transform across the entire segment.

The time-series waveform segment was converted to the time-frequency domain format by applying a Fourier transform on each frame, which is also called Short-Time-Fourier-Transform (SFTF). The obtained spectrum was then computed with a filter bank on a Mel-scale to extract frequency bands, with the goal of mimicking the human ear perception of sound. The relationship between frequency (f) and Mel-scale (m) can be approximated by the following equation:

$$m(f) = 2595log_{10}(1 + f/700) \tag{1}$$

Moreover, the filter bank can be modeled by

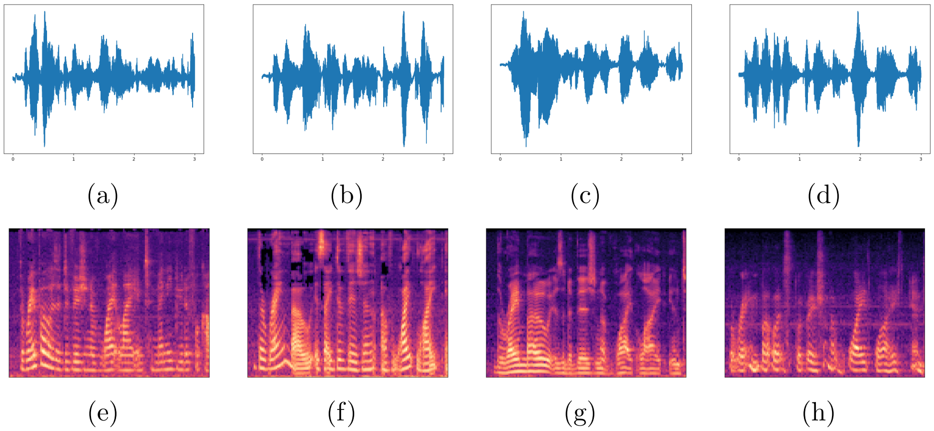
$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \le k < f(m) \\ 1, & k = f(m) \\ \dfrac{f(m+1) - k)}{f(m+1) - f(m)}, & f(m) \le k < f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{2}$$

where $f(m)$ is the center frequency of the mth filter and $H_m(\cdot)$ is the response. Each filter in the filter bank is triangular, with a response of one at the center frequency and decreasing linearly towards zero until it reaches the center frequencies of the two adjacent filters.

Our study set the frame size to 46ms and the number of Mel-scale filters to 128, which transformed the segment into 128 × 128 Mel-spectrogram. Figure 1 depicts the recordings of normal, ADSD, ABSD, and ETV in each of the two data formats for the diagnosis of (a, e) normal, (b, f) ADSD, (c, g) ABSD and (d, h) ETV. The upper row and the lower row are the first format (time-series waveform) and the second format (Mel-spectorgram) of the recordings, respectively.

3.3. Pre-training the Convolutional Autoencoder

The backbone network of our transfer learning strategy is a designed Convolutional Autoencoder (CAE), which consists of an encoder and a decoder. The encoder's architecture stacks alternating convolutional and pooling layers, as follows:



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 1: Visualizing the Rainbow Passage Segments.

- The convolutional layers have the structure of $k_1 \times k_2@c$, where $k_{1,2}$ is the kernel size and c is the number of filters.
- The pooling layers have the pooling size of $p_1 \times p_2$.

The decoder has a similar architecture, but it removes the pooling layers and uses alternating upsampling and convolutional layers.

We introduced noise into the inputs before feeding them into the model to avoid copying input to output in the model without learning features. In particular, Gaussian noise, along with salt and pepper noise, was added into first and second formats of inputs, respectively. The CAE applies these corrupted samples while training to recover the original samples.

Figure 2 visualizes the elements in each of the CAE layers. In this figure the light green, dark green and blue layers denote convolutional layers of $3 \times 3@16$, $3 \times 3@1$ and $3 \times 3@32$, respectively. The orange layer is 2×2 maxpooling layer, and the gray layer is 2×2 upsampling layer. The yellow

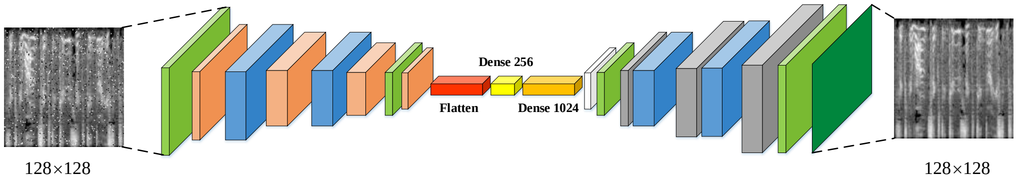


Figure 2: Our Convolutional Autoencoder Architecture.

and gold layers are dense layers of 256 and 1024 neurons, respectively. The red layer is the flatten output of the previous adjacent layer, while the white layer with the size of $8 \times 8@16$ is the re-shape of the previous dense layer.

Figure 2 also shows the use of Mel-spectrogram as the input, where the kernel size is $k_1 = k_2 = 3$ and the pooling size is $p_1 = p_2 = 2$. When the time-series waveform is used as the input, the CAE has the same architecture with Figure 2. However, the kernel and pooling sizes are updated as $k_1 = 1, k_2 = 3$ and $p_1 = 1, p_2 = 2$, respectively.

3.4. Transfer Learning Strategy

Our multi-stage transfer learning strategy is described below, focusing on what to transfer and how to transfer for LMD diagnosis.

3.4.1. Deciding What to Transfer

Based on the pre-trained CAE, we considered two directions to achieve the final diagnosis. The first direction uses both the CAE encoder and decoder, whereas the second direction uses just the encoder.

(a) Based on entire CAE.

This approach stems from work on voice pathology detection. For these types of anomaly detection problems, reconstruction error (RE) is often adopted to measure the extent of a new sample deviates from the distribution of the normal samples. The precondition of this approach is that the distribution of abnormal samples is different from normal samples.

For multi-class classification problems, we formulated the following hypotheses:

- Samples from ABSD, ADSD, and ETV have different deviations in the embedding space when compared with normal samples given that the CAE was pre-trained on just vocally-healthy samples.
- This difference can be measured by RE, so therefore the three LMD sub-disorders can be diagnosed accurately by RE.

(b) Based on just encoder of the CAE. In this context, the decoder in the pre-trained CAE at the top of the encoder was replaced with a fully connected layer (FCL) [43] to give the final diagnosis. The number of the neurons of the FCL was set to the number of health conditions to diagnose. For convenience, the convolutional layers and output dense layer in the encoder were set to $Convi$ $(i = 1, 2, 3, 4)$ and Enc.

This operation, however, is not new in transfer learning-related work. For an encoder trained with PreData, it is an open question what proportion of the encoder is trainable in the transfer process. Our study on diagnosing LMDs thus tested three different transfer approaches according to the different trainable proportions of the encoder, including:

(a) only the *Enc* was fine-tuned,

(b) *Conv*4 (the last convolutional layer), as well as *Enc* were fine-tuned,

(c) *Conv*3, *Conv*4 and *Enc* were fine-tuned.

For approach (a) the majority of the encoder is fixed, so the knowledge learned on PreData significantly impacts the diagnosis results. In contrast, with more trainable layers, approaches (b) and (c) mix more LMD information into the model. As a consequence, we expected these latter two approaches would perform diagnosis more effectively than approach (a).

3.4.2. Deciding How to Transfer

For pathological voice detection and diagnosis, existing work related to transfer learning usually adopts a one-stage strategy, *i.e.*, the pre-trained encoder is fixed and the classifier is just trained with the target data. This strategy may not capture the features of LMDs to some extent, however, so in the previous step we explore the best proportion of a trained encoder to fine-tune, which is denoted as p_1^*.

We therefore applied a two-stage transfer learning strategy. In the first stage, the whole encoder is fixed and only the FCL classifier is trained with FTData. This stage can be viewed as the initialization of classifier parameters. The second stage, in contrast, is designed to adapt the LMD characteristics, such as pitch and rhythm of the voice with LMDs. The proportion of p_1^* of the encoder, as well as the classifier initialized from the first stage, are then fine-tuned with FTData.

In addition, an aim of this study is to verify the effects of sustained phonation on diagnosing the three LMD sub-disorders. Sustained phonation, such as /a/ and /i/, has been demonstrated play an important role in detecting pathological voice [19, 44] based on using sustained phonation, which is a promising approach to further improve the model's diagnosis performance on LMDs. We therefore designed a third stage following the first two stages, where the model obtained by the second stage is further refined on SusData with the trainable proportion of p_2.

In this study we set the pool size to $p_2 < p_1^*$ because we expect the information from sustained phonation and running speech will be mixed rather than replacing each other. The multi-stage transfer learning strategy for diagnosing LMDs we designed is shown in Figure 3. Stage 1 is designed primarily to properly initialize parameters of the FCL classifier. In Stage 2, the model is fine-tuned with the best proportion of p_1^* to make the model suitable for LMDs, while holding the capability of capturing general characteristics of voice/speech. Stage 3 is designed to boost the model's performance and explore the effect of sustained phonation on diagnosing the three LMD subdisorders.

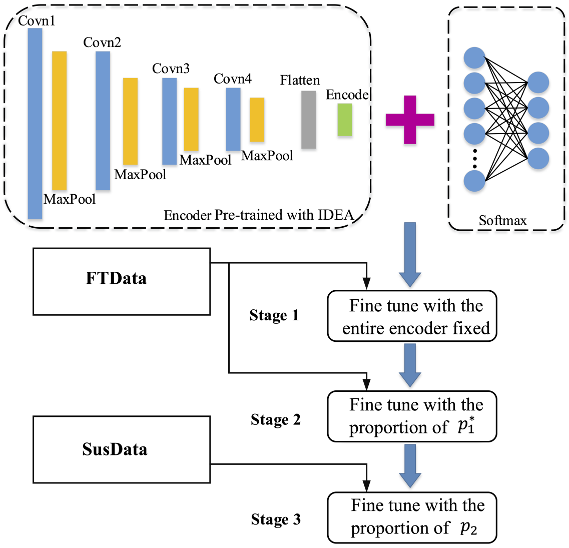

Covn1
Covn2
Covn3
Covn4
Flatten
Encode
MaxPool
MaxPool
MaxPool
MaxPool
Encoder Pre-trained with IDEA
Softmax
FTData
SusData
Stage 1
Stage 2
Stage 3
Fine tune with the entire encoder fixed
Fine tune with the proportion of p_1^*
Fine tune with the proportion of p_2

Figure 3: Our Three-stage Transfer Learning Strategy.

For convenience, the models from different stages in the rest of the paper are denoted as Stage-I, Stage-II and Stage-III, respectively.

4. Results from Experiments

This section presents and analyzes the results of experiments for diagnosing LMDs, focusing on the data formats, transfer strategies, and effects of sustained phonation.

4.1. Summary of the Metrics Applied

The following are the metrics we used to evaluate our models and transfer learning strategies:

- ***Sample-wise accuracy***. All recordings in our experiments were cut into several segments with three second length. Sample-wise accuracy is thus defined as the percentage of correctly classified segments out of all segments.
- ***Subject-wise accuracy***. To get subject-wise prediction, the predicted categories of all segments belonging to the same subject were aggregated and the percentage of the subject in each category was calculated, where the category with the highest percentage was defined as the final diagnosis of the subject. The subject-wise accuracy was measured as the percentage of correctly diagnosed subjects out of all subjects.
- ***Sensitivity (SN score)***, which is defined as TP/(TP + FN)=(Number of true positive assessment)/(Number of all positive assessment).
- ***Specificity (SP score)***, which is defined as TN/(TN + FP)=(Number of true negative assessment)/(Number of all negative assessment) [45].

The results were obtained by the average of five-fold cross validation.

4.2. Testing Data Formats for Diagnosing LMDs

The performance of our CAE with different data formats is described below.

4.2.1. Experiment Description

As described in Section 3.2, we trained two CAEs with the data in formats of time-series waveform and Mel-spectrogram, respectively. Both models employed the loss function of Mean Square Error (MSE) and an Adam optimizer with initial learning rate of 0.001. The training process was halted after 10 epochs without a reduction in the loss value.

Losses in the training process were visualized to compare the convergence of the two models with different data formats. The distributions of samples in the feature spaces (Dense256 in Figure 2) from two model were visualized in 2D figures using the TSNE [46] method.

4.2.2. Raw Results

The loss of the model trained on the Mel-spectrogram format dramatically dropped in the first 15 epochs and converged to a small value after 150 epochs, as shown in Figure 4. In contrast, this figure shows the loss of the model trained on the time-series waveform decreased in the first five epochs and then remained at around 0.015 until the training process stopped.

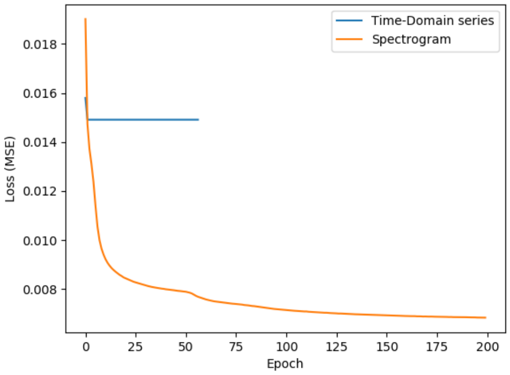
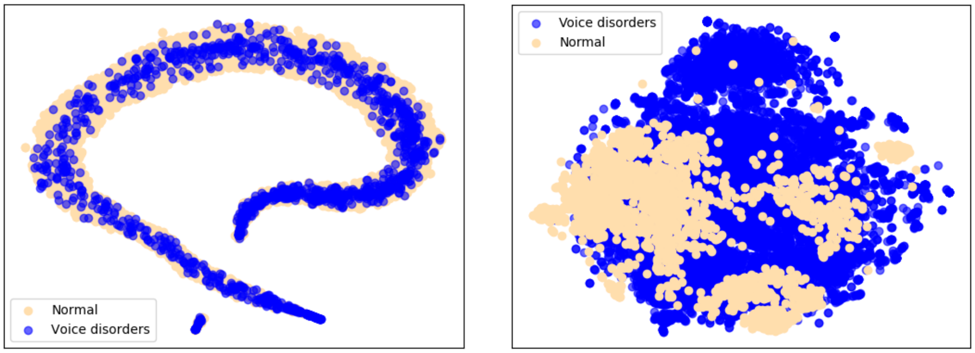


Figure 4: Losses of Two Models in the Training Process.

The visualized feature spaces of two model are shown in Figure 5. For the model trained on the time-series waveform (Figure 5(a)), samples from



(a)

(b)

Figure 5: The Distribution of FTData in the Feature Space from (a) the Model Trained on a Time-series Waveform and (b) the Model Trained on Mel-spectrogram Images

normal and LMDs had roughly the same distribution in the feature space. For the model trained on the Mel-spectrogram (Figure 5(b)), normal and LMDs samples were also overlapped in the feature space. The normal voice samples, however, were concentrated on the left and lower region in Figure 5(b).

4.2.3. Analysis of Results

As shown in Figure 4, approximately 50 epochs were needed train the model using the time-series waveform data format. The loss did not drop noticeably from the initial value (around 0.016), however, and stayed at 0.015 until training was halted. Based on these results we infer that the training failed.

Likewise, Figure 5(a) demonstrates poor training results to some extent. The characteristics of LMDs and normal samples theoretically different, so these differences should be revealed in the feature space because the model was just trained with normal samples. Training probably failed because the time-series waveform with 66,150 points is too complicated to reconstruct given the computational resources, the structure of autoencoder, and the limited amount of available data.

In contrast, better performance was obtained using the Mel-spectrogram data. The loss value steadily decreased from high to low and the distributions of normal and LMDs samples were not exactly the same in the feature space, though there was some overlap among them. Mel-spectrogram therefore shows potential to better distinguish normal and LMDs in higher dimensional spaces.

Our results also show that given the current computational resources and depth of the neural network, the data format of Mel-spectrogram is more suitable and efficient for deep learning methods.

4.3. Testing If RE Alone can Diagnose LMDs

The results of the first direction in Section 3.4.1 are presented and analyzed below, *i.e.*, using RE alone to diagnose ABSD, ADSD, and ETV.

4.3.1. Experiment Description

Based on the results in section 4.2, Mel-spectrogram was applied to train the CAE. The MSE loss function and Adam optimizer with an initial learning rate of 0.001 were also employed. The training process was stopped after 10 epochs without an improvement in loss. Meanwhile, the reconstruction errors (REs) of FTData were calculated to verify if RE can distinguish different diagnostic categories.

4.3.2. Raw Results

Figure 6 shows that REs of normal samples are concentrated within range (0.011, 0.016), with just a few outliers distributed around the top and bottom boundaries of the box. Conversely, REs of samples from LMDs showed more outliers and higher values of outliers (the highest reached 0.022 from ETV). Likewise, the medians of REs of ETV, ADSD, and ABSD were markedly different from those of normal samples.

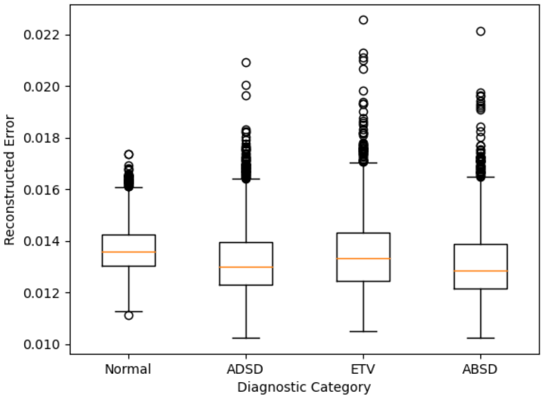


Figure 6: The Distribution of Reconstructed Error of FTData.

4.3.3. Analysis of Results

The few outliers and compact shape of the box in Figure 6 shows that the trained CAE can reconstruct normal samples well. Conversely, the number of outliers and high values of REs for disordered samples indicate the CAE cannot reconstruct those disordered voice samples effectively. These results are expected because normal and LMDs samples have some differences and our model was only trained with normal data.

In addition, REs for ETV samples showed outliers with the highest value and the median that deviated from that of normal and other disordered voice samples, which means it may be easier to distinguish ETV. In contrast, REs for ADSD and ABSD showed similar box shape and median, which indicate that these two categories may be misdiagnosed. Nevertheless, for the REs of the majority samples, including normal and LMDs, the lower and upper quartiles ranged within 0.12 and 0.17. These results are consistent with the results shown in Figure 5(b), which demonstrate that REs alone cannot diagnose the three LMD sub-disorders, though there are differences between normal and LMDs samples.

4.4. Testing the Proportion of the Encoder to Transfer for LMDs

The results of the model with different transferred proportions are described below.

4.4.1. Experiment Description

As discussed in Section 3.4.1, we considered three transfer approaches according to different trainable layers in just the pre-trained encoder. In contrast to training the CAE, the model fine-tuning process in this experiment set the loss function to cross entropy and the process was halted after 20 epochs without a reduction in the loss value. Both sample-wise and subject-wise accuracy were employed to test the performance of three approaches.

4.4.2. Raw Results

Table 1 compares the results of applying the three transfer approaches on FTData. Fine-tuning the *Enc* layer alone performed the worst according to either sample-wise or subject-wise accuracy. Fine-tuning *Enc*+$Covn4$ layers and *Enc*+$Covn3, 4$ layers had the same performance based on sample-wise, achieving the accuracy of 52.26%. Conversely, the results from subject-wise showed that fine-tuning *Enc*+$Covn4$ performed best, obtaining the highest accuracy of 56.32%.

Table 1: Sample-wise and Subject-wise Accuracy of Three Transfer Approaches

Approach	Sample-wise (%)	Subject-wise(%)
Enc	50.12	51.66
Enc+$Covn4$	52.26	56.32
Enc+$Covn3, 4$	52.26	54.97

4.4.3. Analysis of Results

Table 1 shows that fine-tuning too few or too many layers yielded poor performance. Just fine-tuning the *Enc* layer caused the model lose the ability to capture general features of LMD data. Though adjusting *Enc*+$Covn3, 4$ theoretically can make the model more appropriate to LMD samples, a larger amount of data is then required. Adjusting $Covn3, 4$ is therefore not the best way to diagnose voice disorders with a limited amount of data. I our study, therefore, the best proportion p_1^* is to train Enc+$Covn4$.

4.5. Testing the Multi-stage Transfer Strategy and Effects of Sustained Phonation for LMDs

We now present the core results of the proposed strategy. Based on the previous results discussed in Section 4.4, the transfer strategy was determined first and then the impact of sustained phonation on LMD diagnosis was discussed.

4.5.1. Experiment Description

As shown in Figure 3, in the first two stages, the model was fine-tuned with FTData. The entire encoder was fixed in stage 1, while the encoder was fine-tuned through Enc+$Covn4$ layers in stage 2. In stage 3, the Enc layer of the model from stage 2 was further fine-tuned on SusData, where p_2 (training Enc) $< p_1^*$ (training Enc+$Covn4$). Cross entropy was employed as the loss function and the fine-tuning process was stopped after 20 epochs without an improvement in loss. We tested our three-stage transfer strategy on the following tasks:

T_1) differentiate Normal, ADSD, ABSD, and ETV all at once;

T_2) differentiate Normal, ADSD, and ETV;

T_3) differentiate Normal, ABSD, and ETV.

4.5.2. Raw Results

Table 2 shows the subject-wise ACC, SN, and SP when diagnosing four health conditions by each stage of the designated transfer strategy. The results shown in this table reveal that performance improved with each stage sequence, achieving an overall accuracy of 65.3% by Stage-III compared with that of 50.8% in stage 1. For ETV identification, the SN score increased from 46.7% in stage 1 to 66.7% in stage 3. Likewise, the SP score also increased from 73.1 to 81.9%.

Table 2 also shows the SN and SP for ADSD and ABSD identification exhibited tendencies similar to the ETV identification. However, the SN scores for Stage-III are relatively low, *i.e.*, 45.6% and 48.9% for ADSD and ABSD, respectively. With three stages, all models performed well on normal

Table 2: SN and SP for Each Diagnostic Category at Each Stage in One-versus-all Way and Overall Subject-wise Accuracy (ACC).

Stage #	Overall	SN(%)				SP(%)			
	ACC(%)	Normal	ADSD	ETV	ABSD	Normal	ADSD	ETV	ABSD
Stage-I	50.8	100.0	18.6	46.7	37.8	97.0	79.1	73.1	67.4
Stage-II	59.9	86.7	43.9	62.2	46.7	98.5	79.1	73.1	84.7
Stage-III	65.3	100.0	45.6	66.7	48.9	99.2	79.4	81.9	81.4

identification, where the highest scores of SN=100% and SP=99.2% were obtained in Stage 3.

The results obtained by our transfer learning strategy on T_2 are shown in Table 3. The overall accuracy achieved the highest score (77.7%) at Stage 3 by fine-tuning the model on disordered running speech and sustained phonation data. With the three-stage transfer strategy, ADSD identification benefited most, whereas the SN score improved by 19.3% from a relatively low score of 42.8% in Stage 1 to 62.1% in Stage 3. Likewise, the model's SN and SP scores for diagnosing ETV climbed to 71.1% and 72.1% at Stage 3, respectively, indicating less mis-classification compared with 66.7% and 64.2% at Stage 2.

Table 3: SN and SP for T_2 at Each Stage in One-versus-all way and Overall Subject-wise Accuracy (Acc).

Stage #	Overall	SN(%)			SP(%)		
	ACC(%)	Normal	ADSD	ETV	Normal	ADSD	ETV
Stage-I	65.4	86.7	42.8	66.7	94.4	75.0	62.4
Stage-II	69.9	86.7	56.4	66.7	100.0	75.0	64.2
Stage-III	77.7	100.0	62.1	71.1	100.0	78.3	72.1

Table 3 also shows that no disordered subject was classified as normal at Stage 2 and 3 (SP=100%). However, normal subjects were often diagnosed as disordered by Stage-II (SN=86.7%). After going through all stages, the model perfectly diagnosed normal samples.

Figure 7 shows the performance of our three-stage transfer strategy on T_3 at each stage. This task had similar performance to T_1, achieving the highest

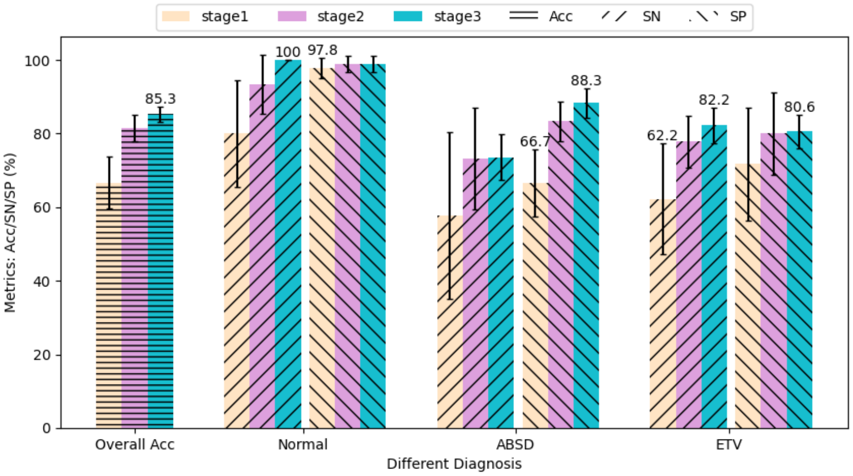


Figure 7: SN and SP for T_3 in One-versus-All Way, and Overall Subject-Wise Accuracy.

overall accuracy of 85.3% after going through all the stages. Differentiating ABSD and ETV benefited most from our method, where the SN score for ETV identification improved by 20.0% from 62.2% at stage 1 to 82.2% at Stage 3 and. Likewise, the SP score for ABSD identification improved by 21.6% from 66.7% at stage 1 to 88.3% at Stage 3.

Figure 7 also shows that few subjects with LMDs are classified as normal for all stages (SP $\geq$ 97.8%). Moreover, the model became more stable with smaller deviations for all the diagnoses with this stage sequence.

4.5.3. Analysis of Results

Our experiments showed that metric scores were relatively low when attempting to diagnose four health conditions all at once, as shown in Table 2. We thus inferred that the different sub-disorder features may negative effect each other in the feature space. Our approach performed better on T_2 and T_3 based on the results in Table 3 and Fig 7, where the SN scores for ADSD and ABSD identification obviously improved. In contrast, the corresponding scores were lower than 50% in Table 2.

Normal identification achieved good SN and SP scores when diagnosing four health conditions all at once. This result is consistent with results shown in Figure 6, where the median of reconstructed error for normal samples deviated significantly from those of the three LMD sub-disorders. Likewise, based on Table 3 and Fig 7, the models from all stages performed worse when diagnosing disorders (ADSD/ETV or ABSD/ETV) compared with diagnosing normal voices.

These results further underscore that interference occurs among features of LMD sub-disorders in the diagnosis process. On the one hand, these results indicate that our current encoder has room to improve to diagnose LMDs. However, the good results for normal voices is reasonable because the encoder was derived from CAE that was just pre-trained with normal data.

Table 3 and Fig 7 suggest FTData and SusData play different roles for differentiating the three LMD sub-disorders. For ETV identification in T_2, Stage-II and Stage-I got the same SN. In contrast, both SN and SP scores increased to a new level after embedding the information of sustained phonation into the model.

Moreover, T_3 Stage-III improved significantly from Stage-II when diagnosing ETV compared with diagnosing ABSD, which just 0.3% gained on SN. This performance improvement indicates that

- ETV symptoms are more sensitive to sustained phonation and
- Running speech may mask the symptoms.

In addition, according to Table 3, Stage-III also gained the capacity of diagnosing ADSD. We found that after fine-tuning the model on SusData, more ADSD subjects with severity of severe were classified correctly, which improved the SN score. This finding is consistent with medical views [47] that vowels carry information related to severe ADSD. Running speech and sustained phonation are clearly both important for diagnosing LMDs and provide different information regarding laryngeal function.

4.6. Comparison with Other Methods on the VUMC Dataset

A comparison of results among the our transfer learning strategy and other approaches is presented below.

4.6.1. Experiment Description

The approaches we compared with our transfer learning strategy include the widely used MFCC feature [48] and classical classifiers, such as support vector machine (SVM) [24] and random forest (RF) [25]. The details of the approaches we compared are as follows:

(a) SVM with RBF kernel fed by features obtained by Encoder of the pre-trained CAE, which is denoted as Encodes_SVM,

(b) RF with depth of 10 fed by features obtained by Encoder of the pre-trained CAE, which is denoted as Encodes_RF,

(c) SVM with RBF kernel fed by MFCC, which is denoted as MFCC_SVM,

(d) RF with depth of 10 fed by MFCC, which is denoted as MFCC_RF.

As described in Section 1, MFCC is a conventional feature that is effective for diagnosing pathological voice. The MFCC was obtained via the cut segment (time-series waveform). According to Fraile et al. [49], the number of MFCC dimension between 15 to 20 is better for pathological voice diagnosis, so we set the MFCC dimension to 20 in this study.

4.6.2. Raw Results

Table 4 shows the comparison results on T_2. Our transfer learning strategy performed best, achieving an overall accuracy of 77.7%. In contrast, Encode_RF performed worst among the compared methods, *i.e.*, only 60.3% of all subjects were diagnosed correctly.

Table 4: Comparison Results on T_2 for the Proposed Method and Four Baselines

Methods	Overall	SN(%)			SP(%)		
	ACC(%)	Normal	ADSD	ETV	Normal	ADSD	ETV
Stage-III	77.7	100.0	62.1	71.1	100.0	78.3	72.1
Encode_RF	60.3	73.3	54.2	53.3	96.7	65.0	64.1
Encode_SVM	67.8	86.7	54.4	62.2	95.6	71.6	69.4
MFCC_RF	67.7	80.0	43.1	80.0	97.8	85.0	55.8
MFCC_SVM	72.3	100.0	48.1	68.9	97.6	76.7	64.8

Table 4 also shows that among the four compared methods, MFCC_SVM got the highest overall accuracy of 72.3%, and almost perfectly diagnosed normal with SN=100% and SP=97.6%. It was also more sensitive to ETV (SN=68.9%) when compared with ADSD (SN=48.1%), which differed by 20.8%. Encode_SVM and MFCC_RF obtained similar accuracy of 67.8% and 67.7%, respectively. However, MFCC_RF showed the lowest SN score of 43.1% but the highest SP score of 85% for ADSD identification. It also showed the highest SN score of 80%, but the lowest SP score of 55.8% for ETV identification.

Comparison results on T_3 showed were similar to the T_2 results, as shown in Table 5. All metrics for our Stage-III strategy were superior to the four compared methods, especially for the overall accuracy, which reached 85.3%. Likewise, MFCC_SVM achieved good diagnostic performance, whereas Encode_RF exhibited the worst performance, with overall accuracy of just 62.2%.

Table 5: Comparison Results on T_3 for Our Transfer Learning Strategy and Four Baselines

Methods	Overall	SN(%)			SP(%)		
	ACC(%)	Normal	ABSD	ETV	Normal	ABSD	ETV
Stage-III	85.3	100.0	73.6	82.2	98.9	88.3	80.6
Encode_RF	62.2	80.0	57.8	48.9	95.6	60.0	71.7
Encode_SVM	70.4	93.3	55.6	62.2	96.7	70.0	71.7
MFCC_RF	70.4	73.3	60.0	77.8	98.9	80.0	68.3
MFCC_SVM	78.5	86.7	71.1	77.8	96.7	83.3	80.0

Table 5 also showed that Encode_SVM and MFCC_RF had the same overall accuracy of 70.4%, though MFCC_RF tended to improperly diagnose normal subjects in the disorder categories, *i.e.*, its SN of 73.3% for diagnosing normal subjects was the lowest among all compared methods. In addition, MFCC_RF was sensitive to ETV (SN=77.8%), while it obtained relatively low specificity of 68.3%. However, it showed the opposite performance for ABSD identification, with a low SN of 60% but high SP of 80%.

4.6.3. Analysis of Results

Based on results in Table 4 and Table 5, we conclude that our transfer learning strategy outperformed four compared conventional methods, due largely to differences in the feature extraction process. For conventional Encode-based methods, the encoders were based on the pre-trained CAE, which was fixed and not fine-tuned on LMD data. It therefore had limited ability to capture features of disorders, resulting in poor diagnostic performance. Conversely, the encoder of Stage-III in our three-stage transfer learning strategy could capture salient features of LMDs from running speech and sustained phonation data.

For MFCC-based methods, MFCC provides a *shadow feature* that cannot carry as rich distinguished information compared with the *deep feature* obtained by our fine-tuned deep network. These MFCC-based methods did achieve higher overall accuracy than conventional Encode-based methods. However, they were *biased* models according to metrics of each category.

For example, MFCC-based methods were sensitive to ETV identification on T_2, but their corresponding low SP scores showed the models tended to predict samples to ETV. Given that their SN scores for ADSD identification failed to reach 50%, it is clear that these models struggled to distinguish ADSD and ETV. Based on Table 5, however, MFCC-based methods achieved acceptable results differentiating ABSD and ETV, which means MFCC carries distinguished information but the information it holds is insufficient, yielding imbalanced performance of models trained with it.

Despite feature extraction, SVM-related methods were clearly superior to RF-related methods. SVM is a binary-classification method. When it was applied to our experiments several sub-models were established following an one-versus-all approach, *i.e.*, it can emphasize different features of sub-disorders at multiple times. In contrast, though RF held the depth of 10, it was a single model that addressed the multi-classification problem, where the features of LMDs were only used once.

5. Concluding Remarks

This paper presented the results of experiments that addressed the following open research questions related to applying deep autocoders to diagnose laryngeal movement disorders (LMDs), such as as adductor spasmodic dysphonia (ADSD), essential tremor of voice (ETV), and abductor spasmodic dysphonia (ABSD):

1. How can the limited amount of LMDs data be overcome?
2. What data representation is most efficient for deep learning in diagnosing LMDs?
3. What transfer learning strategies best improve diagnostic performance?
4. Do traditional machine learning techniques, such as support vector machines (SVM) and random forest (RF), outperform completely neural networks (*i.e.*, combining the autoencoder combined with FCL and transfer learning)?

The following is a summary of lessons learned from the research reported in this paper:

- Using a deep model with a transfer learning on public normal recording databases is a promising way to overcome the limited amount of data in this space, thereby generalizing the model and enabling more effective diagnosis.

- Given the computational resources and the limited available LMD data in this study, Mel-spectrogram is a more efficient and suitable data format for training deep learning models compared with the time-series waveform because it has a smaller size and represents rich frequencies better.

- Reconstructed error (RE) obtained by our pre-trained autoencoder helped diagnose vocally-healthy voice and LMDs, but could not diagnose different LMD sub-disorders alone.

- Sustained phonation and running speech play different roles in diagnosis. In particular, sustained phonation impacted ETV identification and was related to severe ADSD, which is consistent with clinical perspectives [47].

- SVM and RF combined with traditional Mel-frequency cepstral coefficients (MFCC) outperformed those combined with a stacked autoencoder according to overall diagnosis accuracy. However, our completely neural network structure exhibited much better performance for differentiating LMDs.

Based on the results from our experiments, our future work will focus on the following topics:

- The structure of our Convolutional Autoencoder (CAE) has room for improvement since it cannot handle three LMD sub-disorders all at once. Designing a more appropriate structure to obtain better results is a focus of our future work.

- Our current study just diagnosed ETV, ADSD, and ABSD, but did not take into account subjects with mixed sub-disorders. Dealing with this practical and complex problem is also a focus of our future work.

- Recognizing the severity of each diagnostic category is an important issue that comes after diagnosis and also something we are addressing in future work.

6. Acknowledgments

For healthy-vocally speakers, recordings and associated text files are copyright and used by special arrangement with the International Dialects of English Archive at https://www.dialectsarchive.com.

References

[1] E. C. Konkiewitz, I. Trender-Gerhard, C. Kamm, T. Warner, Y. Ben-Shlomo, T. Gasser, B. Conrad, A. Ceballos-Baumann, Service-based survey of dystonia in munich, Neuroepidemiology, 21 (2002) 202–206.

[2] S. Hegde, S. Shetty, S. Rai, T. Dodderi, A survey on machine learning approaches for automatic detection of voice disorders, Journal of Voice, 33 (2019) 947–e11.

[3] D. Isetti, T. Meyer, Workplace productivity and voice disorders: A cognitive interviewing study on presenteeism in individuals with spasmodic dysphonia, Journal of Voice, 28 (2014) 700–710.

[4] C. L. Ludlow, C. H. Adler, G. S. Berke, S. A. Bielamowicz, A. Blitzer, S. B. Bressman, M. Hallett, H. Jinnah, U. Juergens, S. B. Martin, et al., Research priorities in spasmodic dysphonia, Otolaryngology—Head and Neck Surgery, 139 (2008) 495–505.

[5] A. D. Hillel, Classifying and diagnosing laryngeal dystonia—are we artists or are we scientists?, JAMA Otolaryngology–Head & Neck Surgery, 144 (2018) 666–667.

[6] F. X. Creighton, E. Hapner, A. Klein, A. Rosen, H. A. Jinnah, M. M. Johns, Diagnostic delays in spasmodic dysphonia: a call for clinician education, Journal of Voice, 29 (2015) 592–594.

[7] V. Uloza, A. Vegiene, V. Saferis, Correlation between the quantitative video laryngostroboscopic measurements and parameters of multidimensional voice assessment, Biomedical Signal Processing and Control, 17 (2015) 3–10.

[8] J. Lin, E. S. Walsted, V. Backer, J. H. Hull, D. S. Elson, Quantification and analysis of laryngeal closure from endoscopic videos, IEEE Transactions on Biomedical Engineering, 66 (2019) 1127–1136.

[9] P. Daraei, C. R. Villari, A. D. Rubin, A. T. Hillel, E. R. Hapner, A. M. Klein, M. M. Johns, The role of laryngoscopy in the diagnosis of spasmodic dysphonia, JAMA Otolaryngology–Head & Neck Surgery, 140 (2014) 228–232.

[10] D. M. Johnson, E. R. Hapner, A. M. Klein, M. Pethan, M. M. Johns III, Validation of a telephone screening tool for spasmodic dysphonia and vocal fold tremor, Journal of Voice, 28 (2014) 711–715.

[11] J. M. Barkmeier, J. L. Case, Differential diagnosis of adductor-type spasmodic dysphonia, vocal tremor, and muscle tension dysphonia, Current Opinion in Otolaryngology & Head and Neck Surgery, 8 (2000) 174–179.

[12] S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, C.-T. Wang, Detection of pathological voice using cepstrum vectors: A deep learning approach, Journal of Voice, 33 (2019) 634–641.

[13] L. Geng, Y. Liang, H. Shan, Z. Xiao, W. Wang, M. Wei, Pathological voice detection and classification based on multimodal transmission network, Journal of Voice, (2022).

[14] S. Chakraborty, B. Uzkent, K. Ayush, K. Tanmay, E. Sheehan, S. Ermon, Efficient conditional pre-training for transfer learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,, 2022, pp. 4241–4250.

[15] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu, et al., Transfer learning from speaker verification to multispeaker text-to-speech synthesis, Advances in neural information processing systems, 31 (2018).

[16] H. Kim, J. Jeon, Y. J. Han, Y. Joo, J. Lee, S. Lee, S. Im, Convolutional neural network classifies pathological voice change in laryngeal cancer with high accuracy, Journal of Clinical Medicine, 9 (2020) 3415.

[17] J. P. Teixeira, A. Gonçalves, Algorithm for jitter and shimmer measurement in pathologic voices, Procedia Computer Science, 100 (2016) 271–279.

[18] P. Harar, Z. Galaz, J. B. Alonso-Hernandez, J. Mekyska, R. Burget, Z. Smekal, Towards robust voice pathology detection: Investigation

of supervised deep learning, gradient boosting, and anomaly detection approaches across four databases, Neural Computing and Applications, 32 (2020) 15747–15757.

[19] F. Javanmardi, S. R. Kadiri, M. Kodali, P. Alku, et al., Comparing 1-dimensional and 2-dimensional spectral feature representations in voice pathology detection using machine learning and deep learning classifiers, in: Interspeech,, International Speech Communication Association, 2022.

[20] Z. Chen, P. Zhu, W. Qiu, J. Guo, Y. Li, Deep learning in automatic detection of dysphonia: Comparing acoustic features and developing a generalizable framework, International Journal of Language & Communication Disorders, 58 (2023) 279–294.

[21] A. Akbari, M. K. Arjmandi, Employing linear prediction residual signal of wavelet sub-bands in automatic detection of laryngeal pathology, Biomedical Signal Processing and Control, 18 (2015) 293–302.

[22] I. Hammami, L. Salhi, S. Labidi, Voice pathologies classification and detection using emd-dwt analysis based on higher order statistic features, Irbm, 41 (2020) 161–171.

[23] B. Woldert-Jokisz, Saarbruecken voice database, Institut für Phonetik, Universität des Saarlandes, (2007). URL: `https://stimmdatenbank.coli.uni-saarland.de/help_en.php4`.

[24] G. Muhammad, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat, K. H. Malki, A. Al-Nasheri, M. A. Bencherif, Voice pathology detection using interlaced derivative pattern on glottal source excitation, Biomedical signal processing and control, 31 (2017) 156–164.

[25] D. Hemmerling, A. Skalski, J. Gajda, Voice data mining for laryngeal pathology assessment, Computers in biology and medicine, 69 (2016) 270–276.

[26] Z. Ali, G. Muhammad, M. F. Alhamid, An automatic health monitoring system for patients suffering from voice complications in smart cities, IEEE Access, 5 (2017) 3900–3908.

[27] H. Wu, J. Soraghan, A. Lowit, G. Di Caterina, A deep learning method for pathological voice detection using convolutional deep belief networks, Interspeech 2018, (2018).

[28] G. Muhammad, M. Alhussein, Convergence of artificial intelligence and internet of things in smart healthcare: a case study of voice pathology detection, Ieee Access, 9 (2021) 89198–89209.

[29] J. Zhang, J. Liss, S. Jayasuriya, V. Berisha, Robust vocal quality feature embeddings for dysphonic voice detection, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31 (2023) 1348–1359.

[30] S. Hidaka, Y. Lee, K. Wakamiya, T. Nakagawa, T. Kaburagi, Automatic estimation of pathological voice quality based on recurrent neural network using amplitude and phase spectrogram., in: INTERSPEECH,, 2020, pp. 3880–3884.

[31] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat, K. H. Malki, M. A. Bencherif, An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification, Journal of Voice, 31 (2017) 113–e9.

[32] Z. Ali, I. Elamvazuthi, M. Alsulaiman, G. Muhammad, Automatic voice pathology detection with running speech by using estimation of auditory spectrum and cepstral coefficients based on the all-pole model, Journal of voice, 30 (2016) 757–e7.

[33] H. Ding, Z. Gu, P. Dai, Z. Zhou, L. Wang, X. Wu, Deep connected attention (dca) resnet for robust voice pathology detection and classification, Biomedical Signal Processing and Control, 70 (2021) 102973.

[34] Massachusetts eye ear infirmary voice speech lab., Disordered Voice Database Model 4337 (Ver. 1.03). Boston, MA: Kay Elemetrics Corp; (1994).

[35] V. Guedes, F. Teixeira, A. Oliveira, J. Fernandes, L. Silva, A. Junior, J. P. Teixeira, Transfer learning with audioset to voice pathologies identification in continuous speech, Procedia Computer Science, 164 (2019) 662–669.

[36] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, M. Ritter, Audio set: An ontology and human-labeled dataset for audio events, in: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP),, IEEE, 2017, pp. 776–780.

[37] S. Sardari, B. Nakisa, M. N. Rastgoo, P. Eklund, Audio based depression detection using convolutional autoencoder, Expert Systems with Applications, 189 (2022) 116076.

[38] H. Huang, X. Hu, Y. Zhao, M. Makkie, Q. Dong, S. Zhao, L. Guo, T. Liu, Modeling task fmri data via deep convolutional autoencoder, IEEE transactions on medical imaging, 37 (2017) 1551–1561.

[39] T. Haderlein, C. Moers, B. Möbius, E. Nöth, Automatic rating of hoarseness by text-based cepstral and prosodic evaluation, in: Text, Speech and Dialogue: 15th International Conference, TSD 2012, Brno, Czech Republic, September 3-7, 2012. Proceedings 15,, Springer, 2012, pp. 573–580.

[40] M. Alhussein, G. Muhammad, Automatic voice pathology monitoring using parallel deep models for smart healthcare, Ieee Access, 7 (2019) 46474–46479.

[41] G. Fairbanks, The rainbow passage, Voice and articulation drillbook, 2 (1960).

[42] M. E. Powell, M. Rodriguez Cancio, D. Young, W. Nock, B. Abdelmessih, A. Zeller, I. Perez Morales, P. Zhang, C. G. Garrett, D. Schmidt, et al., Decoding phonation with artificial intelligence (dep ai): proof of concept, Laryngoscope Investigative Otolaryngology, 4 (2019) 328–334.

[43] S. S. Basha, S. R. Dubey, V. Pulabaigari, S. Mukherjee, Impact of fully connected layers on performance of convolutional neural networks for image classification, Neurocomputing, 378 (2020) 112–119.

[44] J. R. Orozco-Arroyave, E. A. Belalcazar-Bolanos, J. D. Arias-Londoño, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, K. Daqrouq, F. Hönig, E. Nöth, Characterization methods for the detection of multiple voice disorders: neurological, functional, and laryngeal diseases, IEEE journal of biomedical and health informatics, 19 (2015) 1820–1828.

[45] W. Zhu, N. Zeng, N. Wang, et al., Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas implementations, NESUG proceedings: health care and life sciences, Baltimore, Maryland, 19 (2010) 67.

[46] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research, 9 (2008) 2579–2605.

[47] A. Suppa, F. Asci, G. Saggio, L. Marsili, D. Casali, Z. Zarezadeh, G. Ruoppolo, A. Berardelli, G. Costantini, Voice analysis in adductor spasmodic dysphonia: Objective diagnosis and response to botulinum toxin, Parkinsonism & Related Disorders, (2020).

[48] J. Andén, S. Mallat, Deep scattering spectrum, IEEE Transactions on Signal Processing, 62 (2014) 4114–4128.

[49] R. Fraile, N. Saenz-Lechon, J. Godino-Llorente, V. Osma-Ruiz, C. Fredouille, 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, 61 (2009) 146–152.