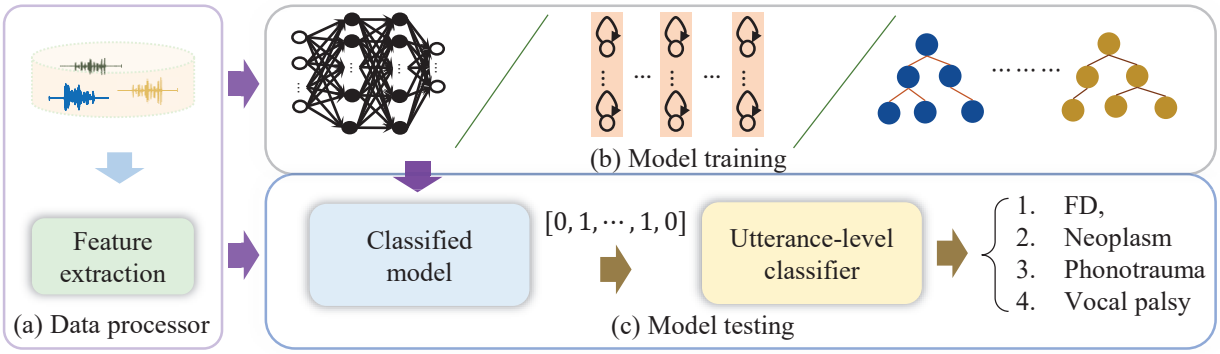
Continuous Speech for Improved Learning Pathological Voice Disorders

Syu-Siang Wang, Chih-Chung Lai, Chi-Te Wang,
Yu Tsao, *Senior Member, IEEE*, Shih-Hau Fang, *Senior Member, IEEE*



The block diagram of the proposed system, which comprises of three stages: (a) data processing, (b) model training, and (c) testing stages.

The graphic-summary figure illustrates the blockdiagram of the proposed pathological voice classification approach using continuous speech. The proposed system is composed of data processing, model training, and online testing stages. From this figure, an acoustic feature is extracted from the input waveform in the feature extraction module. The supervised model is derived accordingly from mapping the relationship between the feature and frame-level classification label in the next training stage. The prediction is then achieved by passing the speech feature through the model followed by the utterance-level classification in the final testing stage.

IEEE Open Journal of Engineering in Medicine and Biology

Continuous Speech for Improved Learning Pathological Voice Disorders

Syu-Siang Wang, Chi-Te Wang, Chih-Chung Lai,
Yu Tsao, *Senior Member, IEEE*, Shih-Hau Fang, *Senior Member, IEEE*

Abstract*—*Goal: **Numerous studies had successfully differentiated normal and abnormal voice samples. Nevertheless, further classification had rarely been attempted. This study proposes a novel approach, using continuous Mandarin speech instead of a single vowel, to classify four common voice disorders (i.e. functional dysphonia, neoplasm, phonotrauma, and vocal palsy).** ***Methods:*** **In the proposed framework, acoustic signals are transformed into mel-frequency cepstral coefficients, and a bi-directional long-short term memory network (BiLSTM) is adopted to model the sequential features. The experiments were conducted on a large-scale database, wherein 1,045 continuous speech were collected by the speech clinic of a hospital from 2012 to 2019.** ***Results:*** **Experimental results demonstrated that the proposed framework yields significant accuracy and unweighted average recall improvements of 78.12–89.27% and 50.92–80.68%, respectively, compared with systems that use a single vowel.** ***Conclusions*****: The results are consistent with other machine learning algorithms, including gated recurrent units, random forest, deep neural networks, and LSTM. The sensitivities for each disorder were also analyzed, and the model capabilities were visualized via principal component analysis. An alternative experiment based on a balanced dataset again confirms the advantages of using continuous speech for learning voice disorders.**

***Index Terms*— Pathological voice, diseases classification, acoustic signal, artificial intelligence.**

***Impact Statement*— Deep learning can detect common voice disorders using continuous speech. Future practice can screen patients who truly need hospital visits and reduce unnecessary medical demands, especially during COVID-19 pandemic.**

I. INTRODUCTION

Voice disorders are one of the most common health complaints, with the lifetime prevalence as high as 30% in the general population [1], [2]. Therefore, in recent decades, automatic detection of voice pathologies gathered much academic interest, and recent works verified the possibility by using the machine-learning-based classifiers, and acoustic signal features [3]–[8].

For example, the works in [9] and [10] extracted vocal-fold-related acoustic features and combined them with a support vector machine for voice pathology detection. The work in [11] performed correlation analyses on the sub-band signal to detect pathological voices. In addition, deep learning and convolutional neural networks were investigated for pathological voice detection [12]–[18]. The recent works applied unsupervised domain adaptation to address the hardware variation [19]. Various acoustic features, including cepstral features [20], [21], vocal jitter [22], and entropy [23] were also investigated in the literature. The IEEE Big Data conference held an international competition in Seattle 2018, called FEMH-Challenge, in which voice pathology detection systems from different research groups worldwide are evaluated empirically on the same dataset, which was published by Far Eastern Memorial Hospital (FEMH), Taiwan [24]. This competition established a systematic evaluation methodology with rigorous metrics for the comparison of voice disorders detection in fair conditions, and over one hundred teams participated in this challenge [21], [22], [24]–[28].

Although numerous published studies had successfully differentiated normal and abnormal voice samples, further classification has rarely been attempted. One possible reason may be the limitation of the single vowel speech signal. Therefore, sustained vowels were investigated for the voice disorder classification, and voice quality measures [13], [29]–[32], [32]. The running speech from both English and Arabic databases was considered as well in [33], and the associated speech features were evaluated in terms of voice-disorder classifiers. Personal habits and behaviors, such as laryngeal tumors caused by long-term use of tobacco and alcohol, were studied in [34]. The previous works combined acoustic signals and medical records to classify three typical voice disorders, including glottic neoplasm, phonotraumatic lesions, and vocal paralysis [35], [36]. Although the results have confirmed the effectiveness of incorporating diverse information, using medical records or questionnaire data still requires an alternative human effort for data collection and a laptop or device for patients typing during the testing phase. These procedures may result in considerable extra costs for both the patient and society. From a health science perspective, people should minimize the contact possibility, especially during the COVID-19 pandemic period.

In general, using the acoustic signal is believed to be the easiest and the most convenient way for the noninvasive screening of voice disorders. Recent studies also demonstrated that reading a text passage can significantly reveal larger ranges of fundamental frequency and sound pressure level (i.e. intensity) [37], [38]. This motivates us to use continuous

Chi-Te Wang is with the Department of Otolaryngology Head and Neck Surgery, Far Eastern Memorial Hospital, Taiwan. Yu Tsao is with the Research Center for Information Technology Innovation, Academia Sinica, Taiwan. Syu-Siang Wang, Chih-Chung Lai, Chi-Te Wang, and Shih-Hau Fang are with the Department of Electrical Engineering, Yuan Ze University, Taiwan (Email: shfang@saturn.yzu.edu.tw). The study protocol was approved by the Research Ethics Review Committee of Far Eatsern Memorial Hospital (FEMH-IRB No. 109063-E), date of approval: May 10th 2020.

speech instead of a single vowel for this task because the multiple syllables may provide richer information to improve the performance. On the other hand, some specialists, such as vocalists, could easily prevent vowel-based disorder recognition systems from working correctly by altering the way of vocalization. Fortunately, altering abdominal vocalization for a long time is difficult during the Chinese syllable transitions, even for vocalist experts. Thus, an alternative advantage of using continuous speech is that it may provide valid detection information resistant to unintentionally abdominal/dantian vocalization from vocalist experts.

This study proposes a novel pathological voice classification approach using continuous speech. To the best of our knowledge, this is the first study using sentence-based Mandarin speech signals to classify voice disorders automatically. In the proposed framework, acoustic signals are transformed into temporal mel-frequency cepstral coefficients (MFCCs) and a bi-directional long-short term memory network (BiLSTM) is adopted to model the sequential features. The experiments were conducted on the large-scale Far Eastern Memorial Hospital (FEMH) voice disorder database, wherein all speech recordings were collected by the speech clinic of FEMH from 2012 to 2019. There is 1,045 continuous speech, each including 7 Chinese sentences, and 1,061 single vowel voice recordings, distributed to functional dysphonia (FD), neoplasm, phonotrauma, and vocal palsy disorders. These four types of vocal diseases are the most common diagnosis for patients with hoarseness [39]. Notably, FD is subjective dysphonia but normal in endoscopic findings. For the classification task, experimental results demonstrated that the proposed framework yields significant accuracy and unweighted average recall (UAR) improvements of 78.12–89.27% and 50.92–80.68%, respectively, compared with systems that use only a single vowel. The results show that the dynamic models with memory architecture successfully extract the robust features from continuous speech, thus improving the effectiveness in the voice disorders classification task.

The results are consistent with other machine learning algorithms, including GRU (Gated Recurrent Unit), RF (Random Forest), DNN (deep neural networks), and LSTM. The sensitivities for each disorder were presented and analyzed. Results show that BiLSTM provides the highest sensitivity scores on FD (86.25%) and vocal palsy (68.00%), respectively, and competitive performance on neoplasm and phonotrauma. The experiments carried out principal component analysis (PCA) to demonstrate the classified capability from testing a model on the continuous-speech corpus. Finally, the experiments were conducted on an alternative public FEMH-Challenge database, which is more balance in the categories with fewer samples. The results again confirm the advantages of the proposed approach using continuous speech.

II. MATERIALS AND METHODS

A. Database Description

We conducted our experiments on the FEMH voice disorder database, wherein all speech recordings were collected by the speech clinic of FEMH from 2012 to 2019. A seven-sentence script that was proposed in [40] was applied to prepare the FEMH database. For each sentence, there are 1,045 voice-disorder recordings. In addition to these continuous speeches, 1,061 */a/*-phone voices were also recorded for providing additional samples. The length of each */a/* sound was about three seconds. The distribution of these voice samples with respect to FD, neoplasm, phonotrauma, and vocal palsy was listed in Table I. All waveforms were recorded at 44,100 Hz sampling rate with a 16-bit integer resolution and using a high-quality microphone (model: SM58, Shure, IL) with a digital amplifier (MDVP, Model 4500, Kay Elemetrics) under a background noise level between 40 and 45 dBA. For each of seven-sentence-speech- and */a/*-sound-corpus, 80% voices were applied to form the training and developing set while the remaining 20% sounds were used to provide the testing set. Notably, there is no overlapped speaker between training and testing sets.

TABLE I. The number of voice-disorder samples for the FEMH database.

	FD	Neoplasm	Phonotrauma	Vocal palsy
continuous speech	100	103	718	124
***/a/* sound**	100	102	735	124

On the other hand, an alternative dataset published for international competition (called FEMH challenge) is employed for further evaluation in the experiments. The statistics for the FEMH-Challenge database are organized in Table II. Comparing Table II with Table I, it can be observed that this database is relatively small but more balance among the three categories. It allows us to provide a fair comparison with existing methods from different perspectives. To be more specific, this database is composed of 150 */a/*-vowel sounds that were pronounced by 150 different patients and classified into neoplasm, phonotrauma, and vocal palsy. In the meanwhile, these speakers were asked to utter the seven-sentence script. The front-end data collection procedure, environment, and separation were identical in these two datasets.

TABLE II. The number of voice-disorder samples for the FEMH-Challenge database.

	Neoplasm	Phonotrauma	Vocal palsy
continuous speech	40	60	50
***/a/* sound**	40	60	50

B. Proposed method

Figure 1 shows the block diagram of the proposed approach, including data processing, model training, and online testing stages. From this figure, an acoustic feature extraction module is applied to extract speech features from the input waveform. The supervised model is performed by mapping the relationship between the feature and frame-level classification label in the following training stage. The prediction is then achieved by passing the speech feature through the model followed by

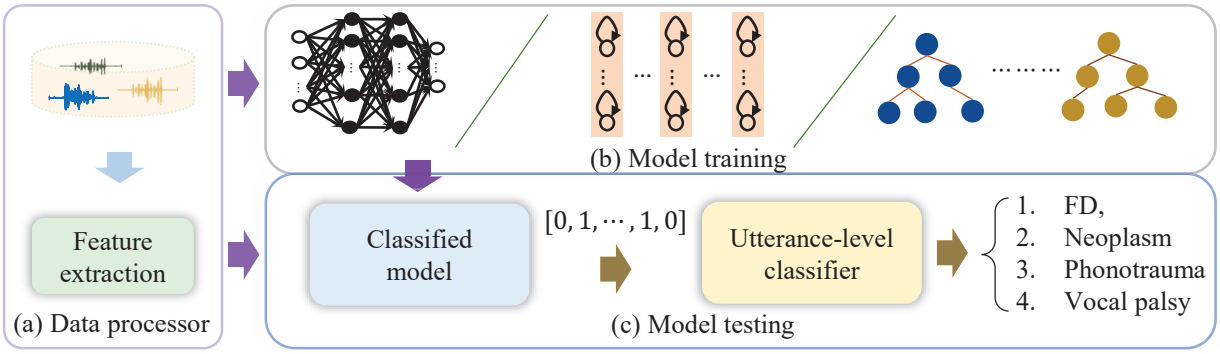


Fig. 1. The block diagram of the proposed system, which comprises of three stages: (a) data processing, (b) model training, and (c) testing stages.

the utterance-level classification in the final testing stage. The Research Ethics Review Committee of Far Eastern Memorial Hospital approved the study protocol (FEMH-IRB No. 109063-E), date of approval: May 10th, 2020. The details of these stages are described in the following subsections.

C. Data Processing Stage

In the data processing stage, voice activity detection and feature extraction operations were utilized for pre-processing the input acoustic waveform.

1) Voice Activity Detection: Voice segments of an utterance were detected by the voice activity detection (VAD) technique, which assumed uncorrelated noises degraded the speech. The Gaussian statistical model was used to model each voice and non-voice component for calculating the likelihood ratio. Those speech present frames were determined thereafter with respect to the predefined threshold. In this study, the threshold was set to 0.9, and the VAD we applied was developed from the VOISEBOX tool[1]. In addition, the prior and posterior signal-to-noise ratio (SNR) for VAD was calculated through the minimum-mean square error short-time spectral amplitude estimator method.

2) Feature Extraction: MFCCs were extracted from those speech segments in the following steps. The hamming-window framing function was applied to split input voice segments into a sequence of frame signals, wherein the frame length and hop size were 32ms and 16ms, respectively. A high-pass filter was applied for each frame to pre-emphasize the input and then decomposed to magnitude spectra through the discrete Fourier transformation. The spectra were filtered through a set of mel filters and then processed by logarithm and power operations to generate 26 filter-bank coefficients. A 13-dimensional static MFCC of a frame was then made-up from this 26-dimensional vector by applying discrete cosine transformation. The MFCC-delta was derived from static MFCC in terms of the delta operator. Finally, the proposed algorithm used 26-dimensional MFCC (MFCC and MFCC-delta) for this task.

D. Model Training Stage

The goal of the paper is to identify the pathological voice from an input continuous speech. Specifically, the model input

[1]http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

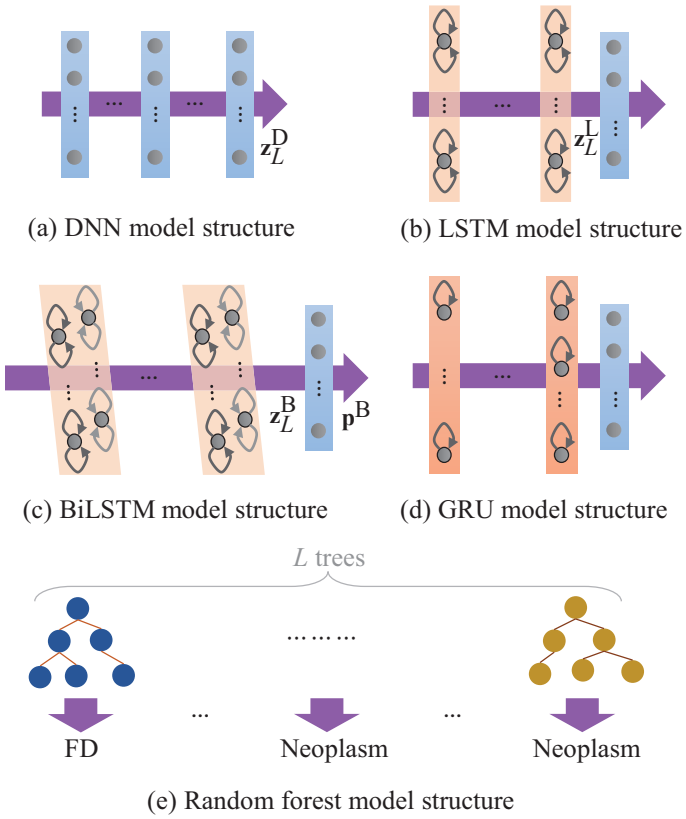


Fig. 2. The (a) DNN, (b) LSTM, (c) BiLSTM, (d) GRU, and (e) RF model structure.

is an MFCC frame, while the output is the frame-wise one-hot label vector. Because the varied length of different recordings results in altered MFCC size, we expect models to achieve two characters: (1) effectively extracted representative features from the size-varied input and (2) handling the contextual information from the input waveform for classifying the voice. Thus, we investigated five learning-based models in the proposed framework, where the model structure is depicted in Fig. 2, and introduced in the following subsections.

1) Deep Neural Network: A DNN model illustrated in Fig. 2 (a) comprising of multiple hidden layers was leveraged to convert a 26-dimensional MFCC frame to the frame-level one-hot label vector. The output of each hidden layer was modified

by using the relu activation function, σ, while the output value of DNN was scaled to the range between 0 and 1 by the softmax function. For an arbitrary ℓth hidden layer, the formulation of input-output ($\mathbf{z}_{\ell}^{\mathfrak{D}}$, $\mathbf{z}_{\ell-1}^{\mathfrak{D}}$) is showed by

$$\mathbf{z}_{\ell}^{\mathfrak{D}} = \sigma_{\ell}\{h_{\ell}(\mathbf{z}_{\ell-1}^{\mathfrak{D}})\}, \tag{1}$$

where $h_{\ell}(\cdot)$ is the linear transformation function. The model parameters were then derived in terms of the cross-entropy cost function in the training process. Because the non-linear transformation capability of a DNN, we believe that the model can be utilized to extract the representative acoustic features from vowel voice and the continuous speech in this classification task.

2) Long Short-Term Memory: The LSTM-classification system depicted in Fig. 2 (b) comprised L LSTM and a feed-forward hidden layers. By passing the nth MFCC frame across all LSTM hidden layers, the contextual feature $\mathbf{z}_{L}^{\mathfrak{L}}[n]$ in the output side of LSTM were derived by jointly considering the previous n MFCC input vectors. $\mathbf{z}_{L}^{\mathfrak{L}}[n]$ was then placed on the input of a feed-forward layer and then provided the predicted likelihood at the output of the system. For performing an LSTM-classification system, the cross-entropy cost function was implemented to calculate the distance between ground truth label vectors and model predictions and minimized thereafter to provide each LSTM and feed-forward layer parameters for the next testing stage.

3) Bi-directional Long Short-term Memory: With respect to an input MFCC feature, a BiLSTM and feed forward blocks were carried out for performing the BiLSTM-classified system, which was showed in Fig. 2 (c). From the figure, the ℓth hidden layer input-output relationship ($\mathbf{z}_{\ell}^{\mathfrak{B}}[n]$, $\mathbf{z}_{\ell+1}^{\mathfrak{B}}[n]$) at the nth frame in the BiLSTM block is formulated as

$$\begin{aligned} \mathbf{z}_{\ell}^{\mathfrak{b}}[n] &= \mathrm{LSTM}_{\ell}^{\mathfrak{b}}\{\mathbf{z}_{\ell}^{\mathfrak{B}}[n]\}, \\ \mathbf{z}_{\ell}^{\mathfrak{e}}[n] &= \mathrm{LSTM}_{\ell}^{\mathfrak{e}}\{\mathbf{z}_{\ell}^{\mathfrak{B}}[n]\}, \ \ell = 1, \cdots, L, \\ \mathbf{z}_{\ell+1}^{\mathfrak{B}}[n] &= \mathbf{z}_{\ell}^{\mathfrak{b}}[n] + \mathbf{z}_{\ell}^{\mathfrak{e}}[n], \end{aligned} \tag{2}$$

where $\mathrm{LSTM}^{\mathfrak{b},\mathfrak{e}}\{\cdot\}$ represents an LSTM cell. From Eq. (2), not only extracting features from the beginning of an MFCC (i.e. $\mathrm{LSTM}_{\ell}^{\mathfrak{b}}\{\cdot\}$), the BLSTM block performed acoustic representations by further considering the speech structure from the very end of an utterance (i.e. $LSTM_{\ell}^{\mathfrak{e}}\{\cdot\}$). Therefore, contextual structures from both ends of MFCC were modeled at the nth output $\mathbf{z}_{\ell+1}^{\mathfrak{B}}[n]$ in the BiLSTM block for the following feed forward layers.

The one-layer feed forward block was used in this study for performing the output prediction $\mathbf{p}^{\mathfrak{B}}[n]$ with respect to $\mathbf{z}_{L+1}^{\mathfrak{B}}[n]$ and is formulated by

$$\mathbf{p}^{\mathfrak{B}}[n] = \mathrm{softmax}\{\mathbf{W}\mathbf{z}_{L+1}^{\mathfrak{B}}[n] + \mathbf{b}\}, \tag{3}$$

where $\mathbf{W}$ is the weight matrix while $\mathbf{b}$ represents the bias vector. We then derived both BiLSTM and feed-forward blocks by leveraging the softmax activation function and the cross-entropy cost function.

4) Gated Recurrent Unit: GRU proposed in [41] was explored in this study to make the recurrent unit capture the long time dependency and extract the robust features from MFCC. Similar to LSTM, the GRU model that was illustrated in Fig. 2 (d) was performed in terms of the gated mechanism to control the information flow through the model. However, fewer gates were used in GRU than those in LSTM, and thereby reducing the size of the model as well as the system latency on processing input data [42]. GRU has been vastly investigated on speech recognition systems. For example, the work in [43] studied the behavior of LSTM and GRU and observed the robustness of GRU in noisy environments. The works in [44] and [45] proposed output-gate projected GRU (OPGRU) that applied an additional transformation matrix on the output GRU. According to these studies, the extracted OPGRU features further improved the recognized accuracy of an acoustic model while improving the decoding speed from the input waveform to the output phone-level sequence.

Inspired by those works on speech recognition that recognized words from input speech, the GRU model was leveraged in this study for identifying the category of disease on the vocal tract through voice. Herein, we followed the work of [41] for constructing the GRU model, which contains L hidden layers and a feed-forward layer followed by the softmax activation function. The system output is the frame-level predicted vectors. The cross-entropy cost function was applied for performing the model training procedure.

5) Random Forest: RF was composed of a series of decision trees as showed in Fig. 2 (e), wherein each tree created a decision rule for the classification. The rule was derived in terms of the training data attributed, whereas the number of leaf nodes for each tree algorithm was determined subsequently. A simple majority vote was then performed across all trees to shrink the predicting variance. In the training stage, m random samples selected from training data were applied to achieve one unpruned classification and regression tree and split afterward at each node to improve classification accuracy. After performing all T trees, an optimized fitting function was applied in the testing stage.

E. Online Testing Stage

In the online testing stage, an MFCC frame was put on the input side of a classification system that generated the predicted vector at the output terminal. The frame-level classification was then derived from the value in the vector that provided the highest class likelihood. All predictions were aggregated in the label predictor in Fig. 1 and determined the final disease category by applying a majority vote.

F. Model Training Setup

Five learning models were investigated for the pathological-voice classification task in this study. In addition, the five-fold learning strategy was applied to evaluate each machine-learning model. For DNN, there are three hidden layers, and the size of each layer is 200. For LSTM, two hidden layers with the dropout rate of 0.2 and 50 cells per layer were used for constructing the LSTM-classification system. Similar settings for performing LSTM model structure were also employed for providing GRU, two 50-cell GRU hidden layers with a 0.2 dropout rate.

Meanwhile, two BiLSTM hidden blocks were used, where each block contains two 50-cell LSTM architectures. The dropout operation was also implemented for each of both LSTMs with a rate of 0.2. All deep-learning models were optimized by using the Adam optimizer with the learning rate of 0.001 for DNN and GRU while that of 0.0005 was used for LSTM and BiLSTM to achieve the optimized performance in the testing stage. As for the RF machine learning model, the bootstrap method is used to sample randomly from training data. In addition, twenty-six trees $T = 26$ were used in RF for optimizing the classified results.

G. Evaluation Metrics

We evaluated the classified performance of the proposed system in terms of the overall accuracy, sensitivity, and unweighted average recall (UAR) metrics. Among these indexes, the score was calculated from the combination of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The accuracy score provided in Eq. (4) was performed to demonstrate the predicted correctness between prediction and truth.

$$\text{Accuracy} = 100\% \times \frac{\text{TN}+\text{TP}}{\text{TN}+\text{TP}+\text{FN}+\text{FP}} \tag{4}$$

The sensitivity value performed in Eq. (5) was calculated for each of four voice-disorder classes ($\mathfrak{D} \in \{\text{FD, neoplasm, phonotrauma, vocal palsy}\}$).

$$\text{Sensitivity}_{\mathfrak{D}} = 100\% \times \frac{\text{TP}_{\mathfrak{D}}}{\text{TP}_{\mathfrak{D}}+\text{FN}_{\mathfrak{D}}}. \tag{5}$$

Then, the final UAR score was obtained by averaging these sensitivity values and is showed in the following equation.

$$\text{UAR} = 100\% \times \frac{\sum_{\mathfrak{D}} \text{Sensitivity}_{\mathfrak{D}}}{K}, \tag{6}$$

where $K = 4$ was used for FEMH while $K = 3$ was set for FEMH-Challenge.

III. Results

A. Sentence Selection from FEMH

The experiments first evaluate the effectiveness of each sentence from the classified performance. In this experiment, we performed a BiLSTM-classified system on the sentence-based FEMH, that is 836 speech-label training pairs, and tested it on the associated testing data. All Chinese characters of each sentence were transferred to Pinyin with tone marks, as shown in Table III. From this table, the second sentence, "lan3 lan3 de5 shuo1 le5 yi4 sheng1 ching3 jin4 lai2", achieves the highest accuracy and UAR scores among seven Chinese sentences. The best score implies that this sentence provides the most balanced pronunciation effort, the articulate structure, and the classified information for both patients and systems. Therefore, this study used the second sentence as the continuous speech for the following evaluation subsections.

B. Evaluations on FEMH Voice Disorder Database

In this subsection, five machine learning classifiers (BiLSTM, GRU, LSTM, DNN, and RF systems) introduced in Section II-D were implemented using FEMH continuous-speech and /*a*/-vowel training corpus to classify four types of vocal disorders. Figures 3 and 4 illustrate the averaged accuracy scores and UAR, respectively, under continuous-speech and /*a*/-vowel testing conditions. Both figures show that continuous speech significantly outperforms the single /*a*/-vowel, except for the UAR using DNN and RF systems. The observation suggests that abundantly fine-structure and vowel-transition speech attributes in the continuous-speech further promote the model capability from identifying the pathological voice

TABLE III. Accuracy and UAR (%) from each sentence based on the BiLSTM disease-classifier system.

Chinese sentence	Accuracy	UAR
wo3 ting1 dao4 you3 ren2 chiao1 men2	82.88	75.39
lan3 lan3 de5 shuo1 le5 yi4 sheng1 ching3 jin4 lai2	**89.27**	**80.68**
men2 kai1 le5 wo3 kan4 dao4 yi4 ge5 nian2 ching1 ren2	84.63	74.09
shou4 chang2 de5 shen1 ti3 ming2 liang4 de5 yan3 jing1	86.23	75.55
hai2 you3 yi4 jhang1 cheng2 ken3 de5 lian3	80.17	74.78
kan4 ta1 lian3 shang4 de5 biao3 ching2 yi3 ji2 yan2 su4 de5 tai4 du4	82.13	67.47
jhen1 siang4 you3 shen2 me5 shih4 ching2 yao4 wo3 bang1 jhu4	81.37	73.75

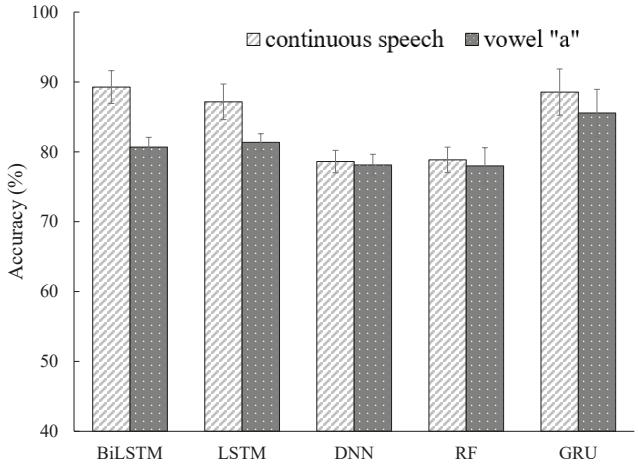


Fig. 3. Accuracy comparison between continuous-speech and /*a*/-phone using FEMH dataset with five machine learning algorithms

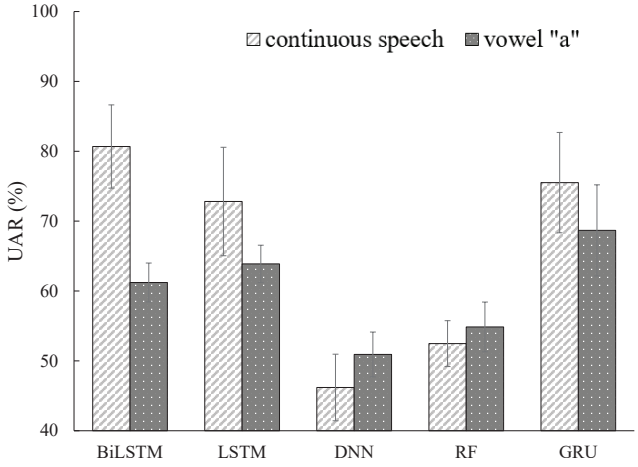


Fig. 4. UAR comparison between continuous-speech and /*a*/-phone using FEMH dataset with five machine learning algorithms

in the classification task. Because DNN and RF are static model structures, extracting information from the continuous speech is difficult. This may explain the exception in Fig. 4. On the other hand, for those evaluations on the continuous-speech database, the classified scores from BiLSTM, GRU, and LSTM systems were better than those from DNN and RF. Three dynamic models perform similarly in the accuracy metrics while BiLSTM achieves the highest UAR. Specifically, BiLSTM achieved the highest accuracy (89.27%) and UAR (80.68%). The results confirm that the dynamic models with memory architecture successfully extract the robust features from continuous speech, thus improving the effectiveness in the voice disorders classification task.

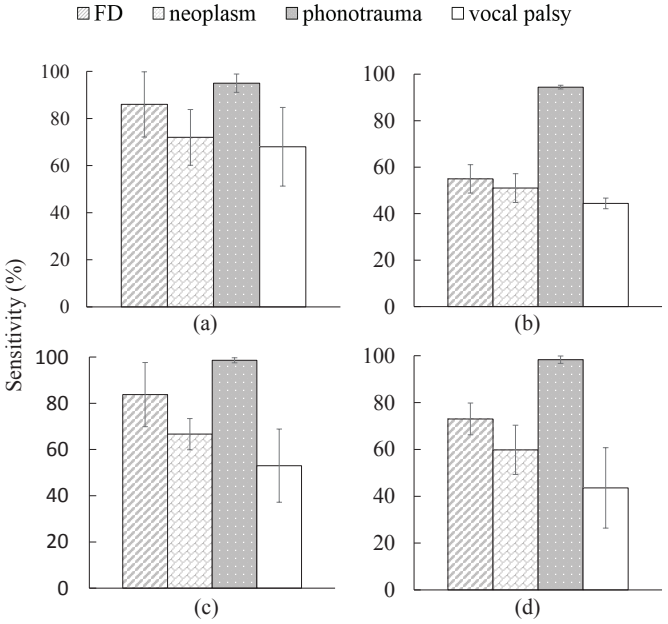


Fig. 5. Sensitivity for each vocal disorders (a) BiLSTM with continuous-speech (b) BiLSTM with single vowel (c) GRU with continuous-speech and (d) GRU with single vowel.

In addition to accuracy and UAR, the sensitivities for each disorder, including FD, neoplasm, phonotrauma, and vocal palsy, were presented and analyzed. Figures 5 (a) and (b) show the sensitivity of continuous-speech and single vowel using BiLSTM models while figures 5 (c) and (d) report that using the GRU models. In the viewpoint of disorder types, when comparing Figs. 5 (a) with (b) and (c) with (d), both BiLSTM and GRU systems on continuous-speech corpus provided better classified results than those on */a/*-vowel database. Again, Fig. 5 show that both BiLSTM and GRU systems using continuous-speech provided better classified results than those on */a/*-vowel ones. Comparing Figs. 5 (a) with (c), BiLSTM provides the highest Sensitivity scores on FD (86.25%) and vocal palsy (68.00%), respectively, and competitive performance on neoplasm and phonotrauma. Comparing Figs. 5 (a)(c) with (b)(d), only phonotrauma shows the comparable performance between continuous-speech and single vowel. The other three disorders were difficult to classify without continuous speech information. The result again confirms the effectiveness of applying continuous-speech corpus to classify voice disorders

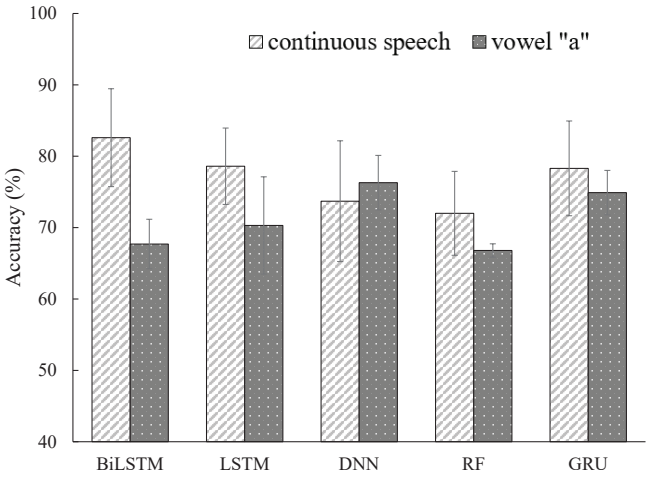


Fig. 6. Accuracy comparison between continuous-speech and */a/*-phone using FEMH-Challenge dataset with five machine learning algorithms

and the ability of BiLSTM to achieve the highest sensitivity. We also noted that scores on phonotrauma are higher than other types of diseases in all subfigures. One possible inference is that the extensive training samples for phonotrauma may result in the model bias issue. Fortunately, the BiLSTM model can effectively shrink the bias phenomena from the extracted speech features in this task.

C. Evaluations on FEMH-Challenge Database

In this subsection, a balanced FEMH-Challenge database comprising neoplasm, phonotrauma, and vocal palsy voice samples was applied for performance evaluation. Tables II and I showed the different samples sizes of FEMH-challenge with the full FEMH dataset. Figures 6 and 7 depicted the accuracy and UAR scores of continuous-speech and */a/*-vowel sound using this balance dataset. Experimental results demonstrated that the proposed framework yields significant accuracy improvements compared with systems that use only a single vowel; the only exception is DNN. The results are consistent with the previous section. Both experiments from a large-scale FEMH or a balance FEMH-Challenge dataset confirm the advantage of the proposed approach using continuous speech. However, Fig. 7 shows that only BiLSTM provides the comparable improvement of UAR as that of accuracy, and the DNN-based approach does not work well for this performance metric. The results imply that the DNN-classified system couldn't effectively leverage the input's contextual information to provide decent output predictions. Meanwhile, the decreased performance of LSTM and DNN may reflect the impact of small dataset size. The training samples are insufficient to train a relatively complex memory structure in LSTM. On the other hand, GRU remains a similar performance because it reduces the number of gates in a neural network. The results also demonstrate that BiLSTM performs well on a small database, achieving the highest accuracy and UAR in both experiments.

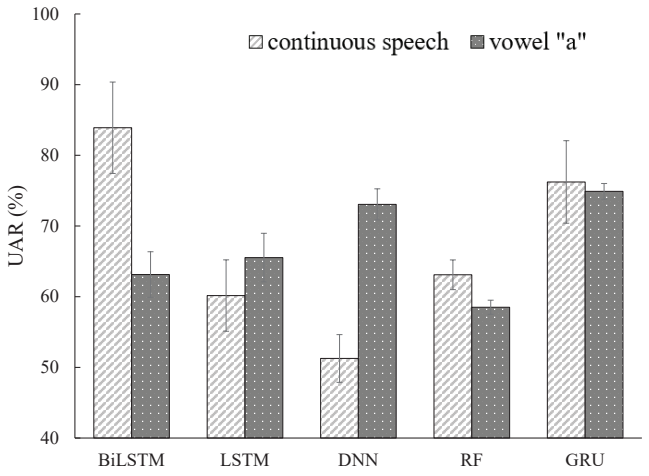


Fig. 7. UAR comparison between continuous-speech and /*a*/-phone using FEMH-Challenge dataset with five machine learning algorithms

IV. Discussion

This section provides discussion from two perspectives, visualization of the disorders classification and the clinical impacts of the proposed technology. First, we carried out PCA to demonstrate system performance in the feature level from testing a model on the continuous-speech corpus. The analyses were achieved in the following steps. Each of the continuous testing utterances in FEMH was converted to MFCC streams. These frame-level MFCCs were used and passed through a classifier for extracting the acoustic features accordingly from the output of the latest hidden layer, i.e. the input of the feed-forward layer, and then averaged afterward to provide the utterance-wise representation. Finally, we collected all extracted acoustic features for PCA. For simplicity, classifier-processed utterance-level features are denoted as "$\mathbf{F}_{c,s}$", wherein the subscript "s" represents a classified system, that is BiLSTM or GRU in this analysis. In addition, the notation "c" denotes that this feature-extraction process was performed on a continuous database. The same procedure introduced above was also applied for extracting utterance-level features on the /*a*/-phone corpus in FEMH from each of BiLSTM and GRU and ultimately represented with the shorthand notation, "$\mathbf{F}_{a,s}$". Thereafter, each feature type ($\mathbf{F}_{c,BiLSTM}$, $\mathbf{F}_{c,GRU}$, $\mathbf{F}_{a,BiLSTM}$ and $\mathbf{F}_{a,GRU}$) labelled as the four vocal-disease classes, FD, neoplasm, phonotrauma and vocal palsy, were then processed by PCA for dimension reduction from 50 to 2 for further visualization. The resulting two-dimensional coefficients were depicted in Fig. 8.

From left to right, the upper row of the figure represents the PCA-processed $\mathbf{F}_{c,BiLSTM}$ and $\mathbf{F}_{a,BiLSTM}$, while the bottom two subfigures illustrated the PCA-processed $\mathbf{F}_{c,GRU}$ and $\mathbf{F}_{a,GRU}$. From the figure, we have the following observations:

- The PCA coefficients of those $\mathbf{F}_{c,BiLSTM}$ and $\mathbf{F}_{c,GRU}$ features reveal more clear boundary among four classes than those of the associated $\mathbf{F}_{a,BiLSTM}$ and $\mathbf{F}_{a,GRU}$, respectively. The result shows the benefit of performing a model on continuous-speech corpus for the followed vocal-disease classifications.

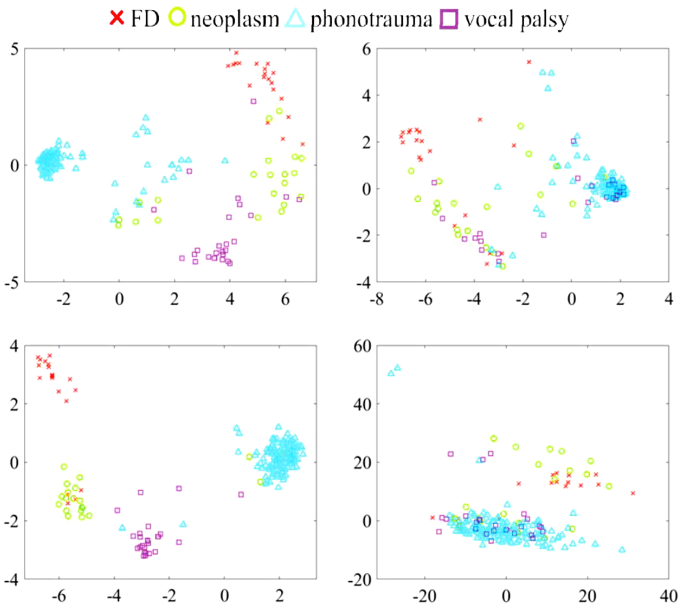


Fig. 8. The scatter plot of PCA coefficients with respect to FD, neoplasm, phonotrauma and vocal palsy clusters. From left to the right, the upper row illustrates PCA-processed $\mathbf{F}_{c,BiLSTM}$ and $\mathbf{F}_{a,BiLSTM}$, while the bottom row shows the result of those of PCA-processed $\mathbf{F}_{c,GRU}$ and $\mathbf{F}_{a,GRU}$ features.

- For all features labeled by phonotrauma, the PCA coefficients show less variety than those of other classes. Also, the small overlap between the phonotrauma cluster and others can be observed from the figure. These observations suggest the decent identified performance in the phonotrauma class.
- Notably, the PCA was performed on utterance-wise acoustic features and thus might not directly reflect the ultimate classification results, which was provided in terms of the frame-level majority vote from the output of classification model in Fig. 1. However, in terms of the cluster mean, we can observe the larger inter-class distance from the PCA coefficients of $\mathbf{F}_{c,BiLSTM}$ than those of $\mathbf{F}_{c,GRU}$. The observation implies that the higher sensitivity of each class from conducting BiLSTM on the continuous-speech testing condition can be obtained than those performed on the GRU system, especially for those of neoplasm and vocal palsy types.

Next, we discuss the social and clinical impacts of the proposed technology. Using the acoustic signal is the easiest and the most convenient way for the noninvasive screening of voice disorders. This motivates us to use continuous speech instead of a single vowel because the multiple syllables may provide richer information to improve the performance. Experimental results show that deep learning algorithms can detect common voice disorders using continuous speech. An alternative advantage of using continuous speech is that it may provide valid detection information resistant to unintentionally abdominal vocalization from vocalist experts. In addition, people should minimize the contact possibility, especially during the COVID-19 pandemic period. With the proposed technology, future practice can screen patients who truly need hospital visits and

reduce unnecessary medical demands, especially during the COVID-19 pandemic.

V. Conclusion

This study proposes a novel pathological voice classification approach using continuous speech. Unlike traditional methods, which rely on the single vowel acoustic signal, continuous speech provides richer information to improve performance. In the proposed framework, acoustic signals are transformed into MFCC features, and BiLSTM is adopted to model the sequential feature vectors. Experimental results demonstrated that the proposed framework yields significant accuracy and UAR improvements of 78.12–89.27% and 50.92–80.68%, respectively, compared with systems that use only a single vowel. The sensitivities for each disorder were analyzed, and the model capabilities were visualized via PCA. An alternative experiment, based on FEMH-Challenge, again confirms the advantages of using continuous speech for learning voice disorders.

References

[1] N. Roy, R. M. Merrill, S. L. Thibeault, R. Parsa, S. D. Gray, and E. M. Smith, "Prevalence of voice disorders in teachers and the general population.," *Journal of speech, language, and hearing research*, vol. 47, no. 2, pp. 281–93, 2004.

[2] D. D. Mehta, J. H. Van Stan, M. Zañartu, M. Ghassemi, J. V. Guttag, V. M. Espinoza, J. P. Cortés, H. A. Cheyne, and R. E. Hillman, "Using ambulatory voice monitoring to investigate common voice disorders: Research update," *Frontiers in bioengineering and biotechnology*, vol. 3, p. 155, 2015.

[3] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE Transactions on biomedical engineering*, vol. 58, no. 2, pp. 370–379, 2010.

[4] G. Muhammad, M. F. Alhamid, M. Alsulaiman, and B. Gupta, "Edge computing with cloud for voice disorder assessment and treatment," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 60–65, 2018.

[5] M. Markaki and Y. Stylianou, "Using modulation spectra for voice pathology detection and classification," in *Proc. EMBC*, pp. 2514–2517, 2009.

[6] I. Hammami, L. Salhi, and S. Labidi, "Pathological voices detection using support vector machine," in *Proc. ATSIP*, pp. 662–666, 2016.

[7] P. Barche, K. Gurugubelli, and A. K. Vuppala, "Towards automatic assessment of voice disorders: A clinical approach.," in *Proc. Interspeech*, pp. 2537–2541, 2020.

[8] P. Barche, K. Gurugubelli, and A. K. Vuppala, "Comparative study of different epoch extraction methods for speech associated with voice disorders," in *Proc. ICASSP*, pp. 6923–6927, 2021.

[9] Z. Ali, M. Alsulaiman, I. Elamvazuthi, G. Muhammad, T. A. Mesallam, M. Farahat, and K. H. Malki, "Voice pathology detection based on the modified voice contour and SVM," *Biologically Inspired Cognitive Architectures*, vol. 15, pp. 10–18, 2016.

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

[11] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, and Z. Ali, "Investigation of voice pathology detection and classification on different frequency regions using correlation functions," *Journal of Voice*, vol. 31, no. 1, pp. 3–15, 2017.

[12] H. Wu, J. Soraghan, A. Lowit, and G. Di-Caterina, "A deep learning method for pathological voice detection using convolutional deep belief networks," in *Proc. Interspeech*, pp. 446–450, 2018.

[13] M. A. Mohammed, K. H. Abdulkareem, S. A. Mostafa, M. K. A. Ghani, M. S. Maashi, B. Garcia-Zapirain, I. Oleagordia, H. Alhakami, and F. T. Al-Dhief, "Voice pathology detection and classification using convolutional neural network model," *Applied Sciences*, vol. 10, no. 11, p. 3723, 2020.

[14] S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, and C.-T. Wang, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *Journal of Voice*, vol. 33, no. 5, pp. 634–641, 2019.

[15] H. Wu, J. Soraghan, A. Lowit, and G. Di Caterina, "Convolutional neural networks for pathological voice detection," in *Proc. EMBC*, pp. 1–4, 2018.

[16] M. Dahmani and M. Guerti, "Glottal signal parameters as features set for neurological voice disorders diagnosis using k-nearest neighbors (KNN)," in *Proc. ICNLSP*, pp. 1–5, 2018.

[17] M. Dahmani and M. Guerti, "Vocal folds pathologies classification using naïve bayes networks," in *Proc. ICSC*, pp. 426–432, 2017.

[18] D. Hemmerling, "Voice pathology distinction using autoassociative neural networks," in *Proc. EUSIPCO*, pp. 1844–1847, 2017.

[19] Y.-T. Hsu, Z. Zhu, C.-T. Wang, S.-H. Fang, F. Rudzicz, and Y. Tsao, "Robustness against the channel effect in pathological voice detection," *arXiv preprint arXiv:1811.10376*, 2018.

[20] M. Pishgar, F. Karim, S. Majumdar, and H. Darabi, "Pathological voice classification using mel-cepstrum vectors and support vector machine," *arXiv preprint arXiv:1812.07729*, 2018.

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

[22] T. Grzywalski, A. Maciaszek, A. Biniakowski, J. Orwat, S. Drgas, M. Piecuch, R. Belluzzo, K. Joachimiak, D. Niemiec, J. Ptaszynski, *et al.*, "Parameterization of sequence of MFCCs for DNN-based voice disorder detection," in *Proc. Big Data*, pp. 5247–5251, 2018.

[23] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, K. H. Malki, T. A. Mesallam, and M. F. Ibrahim, "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions," *IEEE Access*, vol. 6, pp. 6961–6974, 2017.

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

[25] C. Bhat and S. K. Kopparapu, "FEMH Voice Data Challenge: Voice disorder Detection and Classification using Acoustic Descriptors," in *Proc. Big Data*, pp. 5233–5237, 2018.

[26] K. Degila, R. Errattahi, and A. El Hannani, "The UCD System for the 2018 FEMH Voice Data Challenge," in *Proc. Big Data*, pp. 5242–5246, 2018.

[27] J. D. Arias-Londoño, J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "Byovoz automatic voice condition analysis system for the 2018 FEMH challenge," in *Proc. Big Data*, pp. 5228–5232, 2018.

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

[29] J. D. Arias-Londoño, J. A. Gómez-García, and J. I. Godino-Llorente, "Multimodal and multi-output deep learning architectures for the automatic assessment of voice quality using the GRB scale," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 413–422, 2019.

[30] S. Fujimura, T. Kojima, Y. Okanoue, K. Shoji, M. Inoue, and R. Hori, "Discrimination of "hot potato voice" caused by upper airway obstruction utilizing a support vector machine," *The Laryngoscope*, vol. 129, no. 6, pp. 1301–1307, 2019.

[31] H. Cordeiro, C. Meneses, and J. Fonseca, "Continuous speech classification systems for voice pathologies identification," in *Proc. DoCEIS*, pp. 217–224, 2015.

[32] Z. Ali, I. Elamvazuthi, M. Alsulaiman, and 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*, vol. 30, no. 6, pp. 757.e7–757.e19, 2016.

[33] T. A. Mesallam, M. Farahat, K. H. Malki, M. Alsulaiman, Z. Ali, A. Al-Nasheri, and G. Muhammad, "Development of the arabic voice pathology database and its evaluation by using speech features and machine learning algorithms," *Journal of Healthcare Engineering*, vol. 2017, no. 8783751, pp. 1–13, 2017.

[34] M. Hashibe, P. Brennan, S.-c. Chuang, S. Boccia, X. Castellsague, C. Chen, M. P. Curado, L. Dal Maso, A. W. Daudt, E. Fabianova, *et al.*, "Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the international head and neck cancer epidemiology consortium," *Cancer Epidemiology and Prevention Biomarkers*, vol. 18, no. 2, pp. 541–550, 2009.

[35] S. Tsui, Y. Tsao, C. Lin, S. Fang, F. Lin, and C. Wang, "Demographic and symptomatic features of voice disorders and their potential application in classification using machine learning algorithms.," *Folia phoniatrica et logopaedica: official organ of the International Association of Logopedics and Phoniatrics*, vol. 70, no. 3-4, pp. 174–182, 2018.

[36] S.-H. Fang, C.-T. Wang, J.-Y. Chen, Y. Tsao, and F.-C. Lin, "Combining acoustic signals and medical records to improve pathological voice classification," *APSIPA Transactions on Signal and Information Processing*, vol. 8, no. E14, pp. 1–11, 2019.
[37] S. H. Chen, "Sex differences in frequency and intensity in reading and voice range profiles for taiwanese adult speakers," *Folia Phoniatrica et Logopaedica*, vol. 59, no. 1, pp. 1–9, 2007.
[38] Y.-Z. Yen, C.-H. Wu, and R. W. Chan, "A mandarin chinese reading passage for eliciting significant vocal range variations," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 4, pp. 1117–1135, 2021.
[39] J. C. Stemple, N. Roy, and B. K. Klaben, *Clinical voice pathology: Theory and management*. Plural Publishing, 2018.
[40] S. H. Chen, *Phonetograms of normal Taiwanese young adults*. University of Wisconsin–Madison, 1996.
[41] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proc. SSST-8*, pp. 103–111, 2014.
[42] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light gated recurrent units for speech recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 92–102, 2018.
[43] Z. Tang, Y. Shi, D. Wang, Y. Feng, and S. Zhang, "Memory visualization for gated recurrent neural networks in speech recognition," in *Proc. ICASSP*, pp. 2736–2740, 2017.
[44] G. Cheng, P. Zhang, and J. Xu, "Automatic speech recognition system with output-gate projected gated recurrent unit," *IEICE Transactions on Information and Systems*, vol. 102, no. 2, pp. 355–363, 2019.
[45] G. Cheng, D. Povey, L. Huang, J. Xu, S. Khudanpur, and Y. Yan, "Output-gate projected gated recurrent unit for speech recognition.," in *Proc. Interspeech*, pp. 1793–1797, 2018.