SIP (2019), vol. 8, e14, page 1 of 11 © The Authors, 2019.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.
doi:10.1017/ATSIP.2019.7

ORIGINAL PAPER

Combining acoustic signals and medical records to improve pathological voice classification

SHIH-HAU FANG[1]CHI-TE WANG,[1,2,3] JI-YING CHEN,[1,3] YU TSAO[4] AND FENG-CHUAN LIN[2,3]

This study proposes two multimodal frameworks to classify pathological voice samples by combining acoustic signals and medical records. In the first framework, acoustic signals are transformed into static supervectors via Gaussian mixture models; then, a deep neural network (DNN) combines the supervectors with the medical record and classifies the voice signals. In the second framework, both acoustic features and medical data are processed through first-stage DNNs individually; then, a second-stage DNN combines the outputs of the first-stage DNNs and performs classification. Voice samples were recorded in a specific voice clinic of a tertiary teaching hospital, including three common categories of vocal diseases, i.e. glottic neoplasm, phonotraumatic lesions, and vocal paralysis. Experimental results demonstrated that the proposed framework yields significant accuracy and unweighted average recall (UAR) improvements of 2.02–10.32% and 2.48–17.31%, respectively, compared with systems that use only acoustic signals or medical records. The proposed algorithm also provides higher accuracy and UAR than traditional feature-based and model-based combination methods.

Keywords: Pathological voice, Diseases classification, Acoustic signal, Medical record, Artificial intelligence

Received 17 January 2019; Revised 1 May 2019

I. INTRODUCTION

Deep learning technology has shown excellent performance in a wide variety of practical applications (e.g. energy [1, 2], aviation [3, 4], software [5, 6], traffic [7–10], etc). Biomedical engineering [11] combines the knowledge of science and techniques of engineering to solve clinical problems in medicine. A common task in biomedical engineering is to classify and predict the presence of diseases in the human body through biological images, sounds, or patient provided information (e.g. alcohol or tobacco consumption, medical history, and symptoms). Although previous studies had already accomplished the detection of disease of abnormal status using one of the above-mentioned biomedical features, using two or more categories of features had rarely been attempted before. To our knowledge, this is the first study to classify voice disorders based on acoustic signals and medical history, which brings great advancements to both modeling techniques and clinical practicability.

From a health science perspective, the pathological status of the human voice can substantially reduce the quality of life and occupational performance, which results in considerable costs for both the patient and the society [12, 13]. Current standards recommend the use of laryngeal endoscopy for the accurate diagnoses of voice disorders [13], which requires well-trained specialists and expensive equipment. In places without sufficient medical resources, and for patients lack of adequate insurance coverage, correct diagnosis and subsequent treatment may be delayed [14]. A previous study had also noticed that even among professional vocalists, reluctance to seek medical intervention is frequent [15].

To mitigate these problems and lowering the barriers, noninvasive screening methods have been proposed for clinical applications [16]. Because of laryngeal disorders, particularly those originating from the membranous vocal folds, almost always result in the change of voice quality, an automatic recognition framework was developed to detect the presence of vocal diseases based on features extracting from acoustic signals [14].

Voice disorders are one of the most common medical diseases in modern society, especially for patients with occupational voice demand. Common etiologies include neoplasm (e.g., squamous cell carcinoma), phonotraumatic lesions (e.g., vocal polyps and cysts), and neurogenic dysfunction (e.g., unilateral vocal palsy); which can substantially reduce an individuals quality of life [17]). In recent decades, automatic detection of voice pathologies gathered

[1] Department of Electrical Engineering, Yuan Ze University, and MOST Joint Research Center for AI Technology and All Vista Healthcare Innovation Center, Taoyuan, Taiwan
[2] Department of Otolaryngology Head and Neck Surgery, Far Eastern Memorial Hospital, New Taipei City Taiwan
[3] Department of Special Education, University of Taipei, Taipei, Taiwan
[4] Research Center for Information Technology Innovation, Academia Sinica, Taipei Taiwan

Corresponding author:
Yu Tsao,
Email: yu.tsao@citi.sinica.edu.tw

a lot of academic interest, using various machine-learning-based classifiers such as support vector machine [18–22], Gaussian mixture model (GMM) [23–26], convolutional neural network [27–29], and long short-term memory [30, 31]. Some approaches take advantage of automatic speech recognition for acoustical analysis and assessment of pathological speech [32]. Previous studies had already demonstrated the potential to detect pathological voice samples [33–35]. Our recent study used a deep neural network (DNN) approach and achieved the highest accuracy of 99.32% [14] among current literature for a well-established MEEI database [36, 37]. Another study from our research group had also pointed out the potential advantage of using patient-provided information to differentiate several categories of voice disorders [38].

Despite great success in the detection of abnormal voice signals using various machine learning algorithms and signal processing techniques, a further classification of pathological voice samples into several different etiologies has seldom been attempted before. Epidemiological studies clearly demonstrated risk factors and specific symptoms for different categories of voice disorders [39]. Personal habitual behaviors may also contribute to the development of voice disorder, e.g., laryngeal neoplasms typically resulted from chronic exposure to tobacco and alcohol [40]. In this study, we integrate a more comprehensive data set including demographics, medical history, clinical symptoms, and acoustic signals from dysphonic patients to examine if multimodal learning can be applied to classify common voice disorders. In earlier works, multimodal learning has been used to combine voice signals with other modalities for speech recognition [41, 42], speech emotion recognition [43, 44], and speech enhancement [45] tasks. Experimental results have confirmed the effectiveness of incorporating the information from additional modalities to improve the performance in target objectives.

This study proposes two multimodal learning frameworks to classify common voice disorders. In the first framework, called hybrid GMM and DNN (HGD), acoustic signals are first converted into a one-dimensional static supervector via a GMM. Then, a DNN fuses supervectors with medical record data and performs classification. The second framework, referred to as two-stage DNN (TSD), performs classification in a two-stage manner. In the first-stage, acoustic signals and medical records are processed by DNNs individually, and each of the two DNNs generates a three-dimensional output vector indicating the possibility of three target voice disorders. The second-stage DNN fuses the output vectors of the first-stage DNNs and generate final probabilities of disease prediction. Experimental results show that the proposed TSD framework outperforms systems that rely solely on acoustic signals and medical records, respectively, with notable accuracy improvements of 10.32% (from 76.94 to 87.26%) and 5.70% (from 81.56 to 87.26%), and UAR improvements of 17.34% (from 64.25 to 81.59%) and 7.94% (from 73.65 to 81.59%). Model-based combination is an effective scheme for diverse-source learning. In contrast with feature-based combination, the key principle is to train a model to characterize each modality individually. Then a fusion module is used to combine model outputs to perform disease diagnoses. To the best of our knowledge, this is the first study combining both acoustic signals and patient-provided information in the task of computerized classification of voice disorders.

Table 1. FEMH data description.

	Number		Mean age (*y*)		Age range (*y*)		Standard deviation	
	♂	♀	♂	♀	♂	♀	♂	♀
Neoplasm	84	15	57.63	59.93	27–86	36–87	13.44	14.18
Phonotrauma	97	269	43.92	39.19	21–77	20–75	12.95	10.94
Vocal Palsy	76	48	59.91	55.69	28–87	24–84	14.10	14.55

Abbreviations: ♂, male; ♀, female.

Table 2. Phonotrauma data description.

	Phonotrauma		
	Nodules	Polyps	Cysts
♂	11	69	17
♀	121	118	30

Abbreviations: ♂, male; ♀, female.

II. PATHOLOGICAL VOICE CLASSIFICATION FRAMEWORKS

A) Study subjects

1) Acoustic signals

Pathological voice samples were collected from a voice clinic in a tertiary teaching hospital (Far Eastern Memorial Hospital, FEMH, Taiwan). The database includes 589 samples of three common voice disorders, including phonotraumatic diseases (i.e. vocal nodules, polyps, and cysts), glottic neoplasm, and unilateral vocal paralysis (Tables 1 and 2). The clinical diagnosis of voice disorders is based on videolaryngostroboscopic examination [13]. All patients received voice recordings of a sustained vowel /a:/ at a comfortable level of loudness, with a microphone-to-mouth distance of approximately 15—20 cms, using a unidirectional microphone and a digital amplifier (CSL model 4150B, Kay Pentax). The sampling rate was 44100 Hz with a 16-bit resolution, saved in the uncompressed .wav format.

2) Medical record

Besides voice recordings, study subjects also completed a detailed questionnaire in Table 3 about demographic features (e.g., age and sex), duration (years and months), and onset of dysphonia, dysphonic symptoms, occupations (e.g., teacher, stage performer, and business), and occupational vocal demands (using a 4-point Likert scale with the following anchors: always, frequent, occasional, and minimal). Cigarette smoking was classified as active, past, and never. Alcohol consumption was classified as never, occasionally,

Table 3. Medical records of demographics and symptoms feature retrieved from the FEMH database.

Items	Respond	Coding	Neoplasm ($n = 99$)	Phonotrauma ($n = 366$)	Vocal palsy ($n = 124$)
Husky voice	No/yes	0/1	9/90	9/357	20/104
Narrow pitch range	No/yes	0/1	63/36	155/211	86/38
Decreased volume	No/yes	0/1	64/35	239/127	47/77
Fatigue	No/yes	0/1	75/24	141/225	72/52
Dysphonia	No/yes	0/1	68/31	203/163	78/46
Dryness	No/yes	0/1	59/40	150/216	77/47
Lumping	No/yes	0/1	70/29	234/132	100/24
Heartburn	No/yes	0/1	91/8	348/18	120/4
Night meal	No/yes	0/1	91/8	318/48	119/5
Choking	No/yes	0/1	89/10	329/37	64/60
Eyes dryness	No/yes	0/1	88/11	289/77	112/12
Postnasal drip	No/yes	0/1	86/13	264/102	111/13
Diabetes	No/yes	0/1	92/7	362/4	117/7
Hypertension	No/yes	0/1	70/29	346/20	108/16
Coronary artery disease	No/yes	0/1	95/4	363/3	114/10
Head and neck cancer	No/yes	0/1	96/3	364/2	106/18
Head injury	No/yes	0/1	99/0	365/1	122/2
Cerebral vascular accident	No/yes	0/1	99/0	366/0	123/1
Smoking	Never/past/active	0/1/2	24/20/55	278/25/63	83/32/9
Alcohol Drinking	Never/past/active	0/1/2	58/1/40	234/1/131	108/1/15
Drinking frequency	Not/occasionally/ weekly/daily	0/1/2/3	56/29/12/2	235/108/ 17/6	110/12/0/2
Noise at work	Not/a little/noisy	0/1/2	57/25/17	111/155/100	91/25/8
VAS	Worst to best	0–10	3.02±1.98	2.74±1.54	2.34±1.65
Maximal phonation time (MPT)	e.g. 10 s	e.g. 10	10.16±6.23	8.93±4.39	4.96±4.35
Voice handicap index - 10	Sum of 10-item voice handicap index	0–40	21.23±9.99	22.85±7.81	28.35±9.46
Reflux symptom index	Sum of reflux symptom index	0–45	11.08±7.42	12.87±7.06	15.52±10.07
Onset of dysphonia	Missing	0	1	5	6
	Sudden	1	22	76	60
	Gradually	2	50	196	24
	On and off	3	20	78	5
	Since childhood	4	0	4	1
	Other	5	6	7	28
Diurnal patterns	Missing	0	3	0	3
	Worse in the morning	1	7	85	7
	Worse in the afternoon	2	18	87	19
	Similar all day	3	40	86	67
	Fluctuating	4	31	108	28
Occupational vocal demand	Missing	0	3	1	5
	Always	1	25	240	33
	Frequent	2	24	94	42
	Occasional	3	26	22	25
	Minimal	4	21	9	19

weekly, and daily. Patients were instructed to self-rate their voice quality using a visual analog scale (VAS) with scores ranging from 0 (worst) to 10 (best) and fill out questionnaires with 10-item voice handicap index (VHI-10) and reflux symptom index (RSI) [46–48].

B) Feature extraction

1) Acoustic signals

The following six steps must be performed to derive 13-coefficient MFCCs from acoustic signals: pre-emphasis, windowing, fast Fourier transform, Mel scale filter bank, nonlinear transformation, and discrete cosine transform. MFCCs frames were extracted from a window length of 16-millisecond and captured 8-millisecond overlap for time shift.

2) Medical record

To simplify the input parameters, we encode each item into digit numbers. For example, binary data (i.e. yes/no) is recorded as 1/0. In ordinal data such as tobacco and alcohol consumption, we encode it as 0/1/2 (never/past/active) or 0/1/2/3 (never/occasionally/weekly/daily), respectively. Coding and definition of all the 34 input variables are presented in Table 3.

C) Typical combination methods

1) Feature-based combination

Feature-based combination is an intuitive approach to learning from diverse information sources. The basic principle is to concatenate heterogeneous features directly to form a new higher dimensional feature, and then a classifier is trained to perform classification with concatenated features as input. If we consider medical records as an individual dynamic feature, the dimension of the combined feature is $(L + d)$, including d acoustic features and L medical record features. Moreover, if features of acoustic signals contain N frames, then medical records will be duplicated N times and used in all frames. In accordance with the concept of feature-based combination, we establish a one-stage DNN (OSD) system. In OSD, acoustic features and medical records are represented by a sequence of MFCC+delta vectors and digit numbers, respectively. We then derive the concatenated feature by combining the MFCC+delta vectors and the numerical digit. Finally, a DNN is used in OSD to perform classification with concatenated features.

2) MODEL-BASED COMBINATION

In this paper, the established model-based combination system is referred to as DNN with Linear Combination (DLC). In DLC, acoustic signals and medical records are processed by DNNs individually. A linear combination function is then used as the fusion module to linearly combine the outputs of the two DNNs. The weights of the linear combination function are estimated based on the training data to maximize classification accuracy.

D) Enhanced feature-based combination algorithm

Although the implementation of feature-based combination is straightforward, the key drawback of OSD is not fully considering the dynamic properties of different information sources. In fact, acoustic waves are rather dynamic compared with the medical record, creating difficulties in model learning. To overcome this limitation, this study proposes an enhanced feature-based combination algorithm, called HGD. In HGD, the acoustic signals are first modeled by a GMM, and then the means of the GMM are concatenated to form a supervector for feature combination instead of using MFCC+delta in OSD.

The basic principle for the GMM-based supervector is to represent a sequence of acoustic features with arbitrary length as a static long vector [49]. This technique is a standard method and has been validated for speaker recognition tasks [50]. GMM-based supervector extraction involves two steps. First, one must train the GMM-universal background model (UBM) using a dedicated data set. The training process in this step is performed in an unsupervised manner and does not require classification labels. Then, each utterance is used to adapt the GMM-UBM to generate an utterance-specific GMM. Finally, a supervector is formed by concatenating the mean vectors into a higher-dimensional vector; for instance, by stacking d-dimensional mean vectors of a M-component adapted GMM into a $M * d$-dimensional GMM-based supervector [51, 52].

Figure 1 shows the architecture of the proposed HGD framework. In this framework, a GMM-UBM is first trained based on the entire data set. Then, acoustic features for each utterance were used to adjust the mean parameters of the GMM-UBM. The adjusted mean parameters are concatenated to form a supervector. The size of a supervector is determined by the number of GMM-UBM. Then, a DNN fuses the supervector and medical records to perform classification.

E) Enhanced model-based combination algorithm

To further improve the performance of DLC, we proposed an enhanced model-based combination algorithm referred to as two-stage DNN (TSD). In contrast with DLC, another deep learning model is used as the fusion module to combine the outputs of the DNNs in order to characterize the joint effects of separate modalities, namely, acoustic signals and medical records in this study, more accurately. The architecture of the TSD framework is shown in Fig. 2. It can be divided into two stages. In the first stage, two DNNs (referred to as first-stage DNNs) are used to process acoustic signals and medical records individually. Each of the two first-stage DNNs generates a three-dimensional output vector indicating the probability of three target voice disorders. Unlike DLC, in which the fusion is a linear combination, the fusion mechanism in TSD is an alternative DNN (termed second-stage DNN). Note that the inputs of the second-stage DNN are acoustic signals (26 dimensions), medical records (34 dimensions), and the outputs of the two first-stage DNNs with six-dimensional vectors (3+3 dimensions). The second-stage DNN fuses outputs of the first-stage DNNs, mean of acoustic signal, and medical records, and then performs classification. Additionally, the architecture of all DNNs contains 300 neurons with three hidden layers and having sigmoid function for the activation function.

III. EXPERIMENTS AND RESULTS

A) Experimental setup

This study focuses on three typical voice disorders including phonotraumatic lesions (i.e. vocal nodules, polyps, and cysts), glottic neoplasm, and unilateral vocal paralysis. The voice samples are recorded by asking the patients to pronounce a sustained vowel (/a:/) for at least 3 s. During the experimental processes, we randomly divided each class of voice disorders into training (80%) and testing (20%) sets. The performance was verified through five-fold cross-validation. In addition to the 13 MFCCs, we added 13 delta features (the first derivative features of MFCCs) to form 26 dimensional MFCCs+delta feature vectors. The delta feature is obtained from the frames of MFCC over time. Because MFCC is the static cepstral features, adding dynamic information is widely used in many recognition tasks [9, 53–55]. The cepstral variance normalization was then applied to the MFCC(N)+delta feature vectors in such a manner that the normalized feature vectors have zero mean and unit variance.

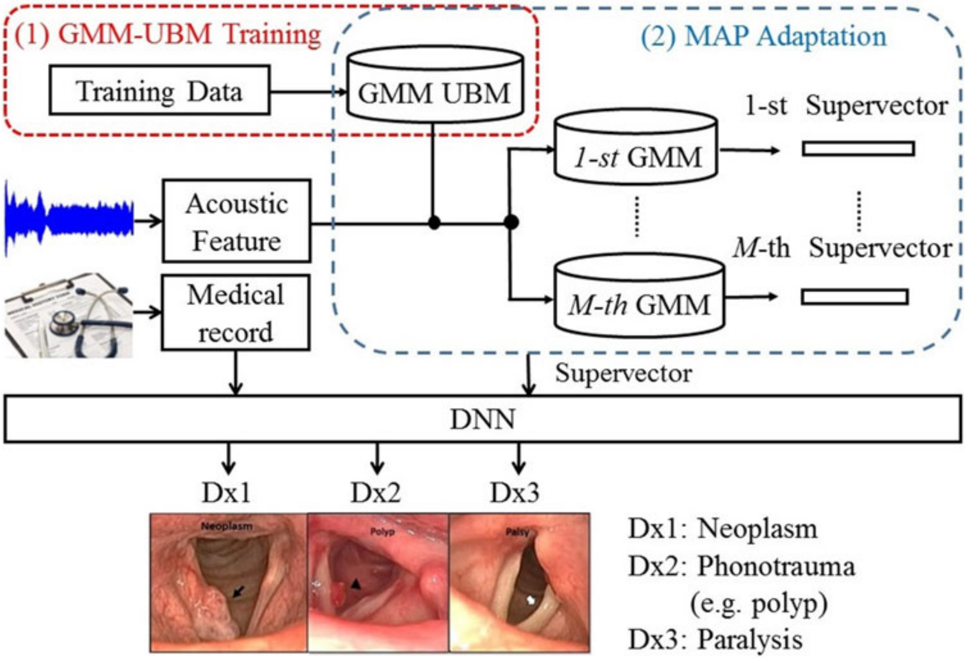
(1) GMM-UBM Training
Training Data
GMM UBM
(2) MAP Adaptation
1-st Supervector
1-st GMM
M-th Supervector
M-th GMM
Acoustic Feature
Medical record
Supervector
DNN
Dx1
Dx2
Dx3
Neoplasm
Polyp
Palsy
Dx1: Neoplasm
Dx2: Phonotrauma (e.g. polyp)
Dx3: Paralysis

Fig. 1. The block diagram of the proposed HGD framework.

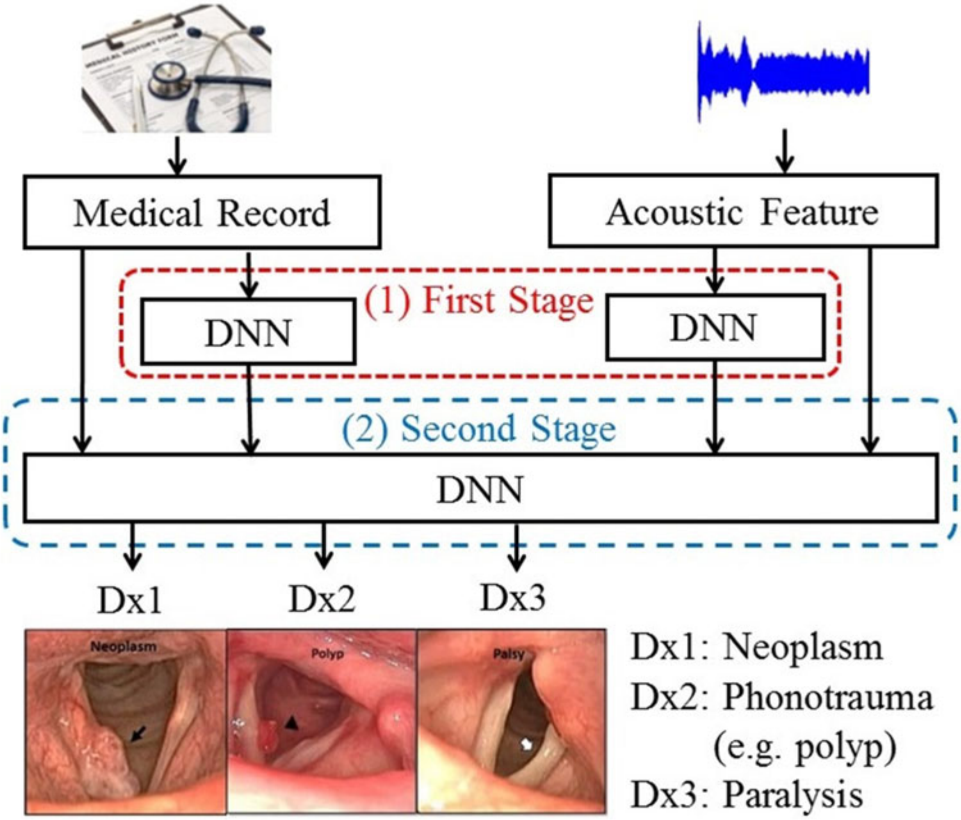
Medical Record
Acoustic Feature
(1) First Stage
DNN
DNN
(2) Second Stage
DNN
Dx1
Dx2
Dx3
Neoplasm
Polyp
Palsy
Dx1: Neoplasm
Dx2: Phonotrauma (e.g. polyp)
Dx3: Paralysis

Fig. 2. The block diagram of the proposed TSD framework.

In this study, we used three performance indexes: overall accuracy (ACC), sensitivity, and UAR. These indexes were widely employed in the classification tasks. First, ACC is the value of the difference between prediction and truth in equation (1), where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

$$ACC = 100\% \times \frac{TN + TP}{TN + TP + FN + FP}. \tag{1}$$

Second, as shown in equations (2)(4), the sensitivity for each disorder is also a common index for the classification task,

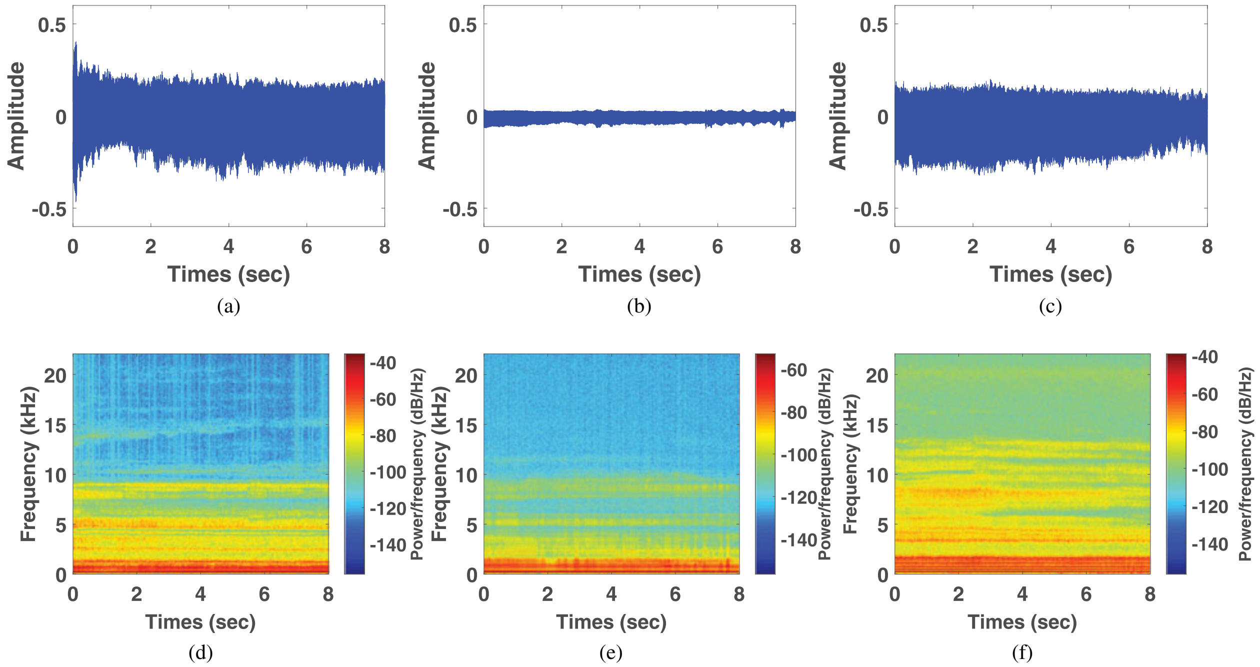

Fig. 3. Waveforms from voice samples of neoplasm (a), vocal palsy (b), and phonotrauma (c). Wide band spectrograms in voice samples of neoplasm (d), vocal palsy (e), and phonotrauma (f).

where neo, pho, and pal stand, respectively, for neoplasm, phonotrauma, and vocal palsy; unneo, unpho, and unpal denote non-neoplasm, non-phonotrauma, and non-vocal palsy, respectively.

$$SN_{neo} = 100\% \times \frac{TP_{neo}}{TP_{neo} + FN_{unneo}}, \quad (2)$$

$$SN_{pho} = 100\% \times \frac{TP_{pho}}{TP_{pho} + FN_{unpho}}, \quad (3)$$

$$SN_{pal} = 100\% \times \frac{TP_{pal}}{TP_{pal} + FN_{unpal}}. \quad (4)$$

Finally, UAR is an alternative index considering unbalanced data, as shown in equation (5), where K denotes the number of classes ($K = 3$ in this study).

$$UAR = 100\% \times \frac{SN_{neo} + SN_{pho} + SN_{pal}}{K}. \quad (5)$$

B) Experimental results

Figure 3 shows acoustic waves and spectrogram plots of glottic neoplasm, phonotrauma, and vocal palsy voice samples. In the waveform plots, all samples had irregular diversification in each period. Because of the loss of normal muscle tone, voice samples from cases of vocal palsy usually demonstrate much lower amplitude of volume (3b) than others, with weak spectrum in high-frequency harmonics (3e). Glottic neoplasm may invade the deeper structures of the vocal folds, such as vocal ligaments of muscles. In contrast, phonotraumatic lesions are usually confined to subepithelial spaces, as they do not violate the original tissue structure. Accordingly, we observed that the harmonic structure of phonotraumatic samples is generally preserved (3f), while cases of glottic neoplasm showed distorted harmonics (3d). These differences are more prevalent in higher frequency domain. However, owing to the limitations from inter-subject variability, it remains difficult to categorize the pathological voice simply by time-domain waves and frequency-domain signals.

Additionally, we also plot the feature distributions to better visualize the results via principle component analysis (PCA). PCA is a widely used approach in which a linear transformation is designed to compress the information among features into the relatively lower dimensions [55–59]. Figures 4 and 5 show the distribution of the first and second principal component among all data sets for acoustic signals and medical record, respectively. Figures 4(b) and 4(c) are the joint distribution in two-dimension principal component space of the first and second principal component for acoustic signals MFCC, respectively. The input data set for PCA is the 26-dimensional MFCCs of all training data. The x- and y-axis in Fig. 4 represent the first and the second principle comments (PC), respectively. The unit of PC is the same as that in MFCC because each PC is, in fact, a linear combination of all MFCCs. By reconstructing the covariance matrix of all training data, the linear weights is determined by the corresponding eigenvalue, which quantifies the information contributed by the corresponding MFCC component. These figures show that the features of three categories almost overlap in the scatter plot, while few neoplasm frames can be distinguished. Figures 4(a) and 4(d) plot the histogram of individual components and show a similar trend. That is, three kinds of voice disorders are overlapped in most area, making the classification difficult while using only two components. We follow the same procedure to plot the histogram of individual components and

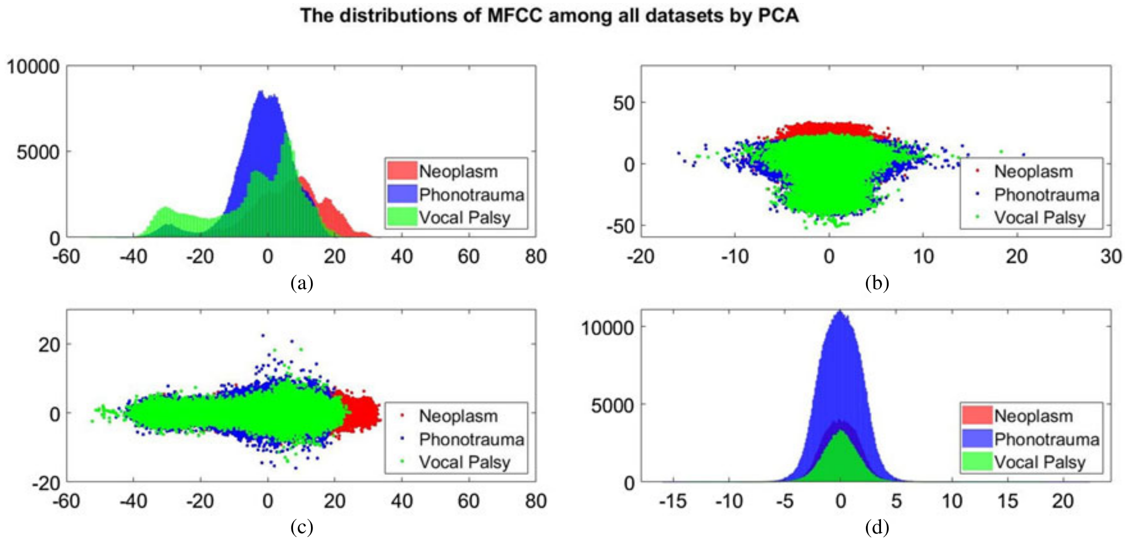

Fig. 4. Distribution of the first and second principal component for acoustic signals MFCC. (a) and (d) Are the histogram of individual components, while (b) and (c) are the joint distribution in two-dimension principal component space (the first and second principal components are placed at different axes in (b) and (c)).

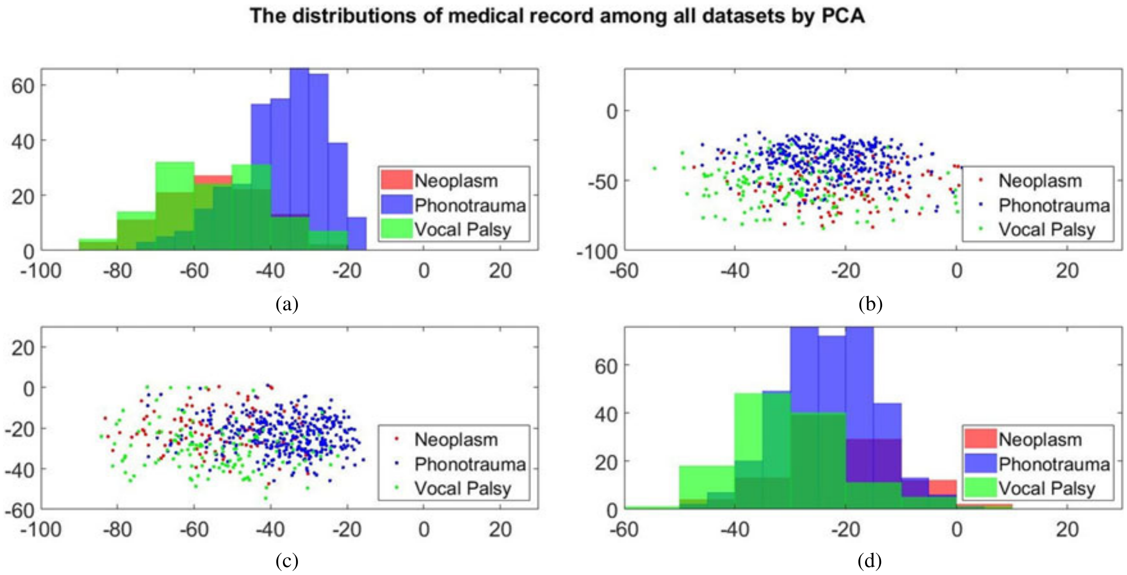

Fig. 5. Distribution of the first and second principal component for medical record. (a) and (d) Are the histogram of individual components, while (b) and (c) are the joint distribution in two-dimension principal component space (the first and second principal components are placed at different axes in (b) and (c)).

the joint distribution for medical record, as shown in Fig. 5. Note that the difference between Figs 5 and 4 is that the MFCC feature is frame-based whereas the medical record is individual-based, thus the samples are unequal. More importantly, results show that the medical record seems to has better distinguish ability among three voice disorders. From Fig. 5(b) and 5(c), the overlapped region between vocal palsy and phonotrauma is smaller than Fig. 4. Accordingly, these additional plots supported our idea of including medical records as an adjuvant features to differentiate three categories of vocal disorders.

Table 4 reports three performance indexes of the OSD, DLC, HGD, and TSD systems. Performance indexes for systems using only acoustic signals or medical records are also listed as the baseline for comparison. From the table, we first note that the system using medical records outperforms the one using acoustic signals, with an accuracy improvement of 4.62% (from 76.94 to 81.56%) and a UAR improvement of 9.40% (from 64.25 to 73.65%), and the difference was statistical significant ($p<0.001$) (Tables 5 and 6). The results confirm that the medical records can contribute more useful information as compared with acoustic signals when classifying multiple voice disorders.

Furthermore, the results in Table 4 demonstrate that DLC, a typical multimodal learning system, outperforms systems using acoustic signals and medical records alone. More specifically, DLC achieved significant accuracy improvements of 6.64% (from 76.94 to 83.58%) and UAR improvements of 9.83% (from 64.25 to 74.08%), compared with systems using acoustic signals ($p<0.001$, Tables 5 and 6). We also noticed that DLC (model-based combination) outperforms OSD (feature-based combination) in

Table 4. Performance comparison.

	Neoplasm	Phonotrauma	Vocal Palsy		
		Sensitivity (Recall)			
	(%)	(%)	(%)	Accuracy (%)	UAR (%)
Acoustic signals	63.00±17.89	95.36±4.39	34.40±20.12	76.94±6.71	64.25±11.04
Medical record	59.00±11.40	91.54±3.67	70.40±2.19	81.56±1.25	73.65±3.49
OSD	53.00±14.83	88.78±4.36	64.00±12.65	77.48±2.38	68.59±5.03
DLC	65.00±17.68	96.44±2.47	60.80±10.35	83.58±3.42	74.08±7.90
HGD	72.00±16.05	94.00±2.49	62.40±4.56	83.58±3.00	76.13±5.43
TSD	79.00±14.75	95.36±3.03	70.40±10.43	87.26±2.23	81.59±5.94

Table 5. P-value for accuracy (ACC).

ACC (P-value)	Acoustic signals	Medical record	OSD	DLC	HGD	TSD
Acoustic signals	–	**2.493×10^{-5}**	0.051	**4.032×10^{-8}**	**1.772×10^{-7}**	**9.619×10^{-14}**
Medical record	–	–	0.008	0.028	0.031	**1.713×10^{-8}**
OSD	–	–	–	**1.455×10^{-5}**	**4.205×10^{-5}**	**7.774×10^{-12}**
DLC	–	–	–	–	0.790	**7.099×10^{-5}**
HGD	–	–	–	–	–	1.294×10^{-3}

Table 6. P-value for unweighted average recall (UAR).

UAR (P-value)	Acoustic signals	Medical record	OSD	DLC	HGD	TSD
Acoustic signals	–	**6.047×10^{-7}**	**2.058×10^{-4}**	**2.669×10^{-6}**	**2.723×10^{-8}**	**1.113×10^{-12}**
Medical record	–	–	0.100	0.571	0.272	**2.053×10^{-4}**
OSD	–	–	–	0.252	9.626×10^{-3}	**1.140×10^{-6}**
DLC	–	–	–	–	0.0984	**2.197×10^{-5}**
HGD	–	–	–	–	–	8.727×10^{-3}

this task. When comparing to OSD, DLC yields a higher accuracy of 6.10% (from 77.48 to 83.58%) (p<0.001, Table 5).

The results suggest that a direct combination of a dynamic signal source (acoustic feature sequence) and static information (medical records) may be problematic in the feature domain, and thus, the performance does not necessarily improve compared with learning from an individual source. In this case, the model-based combination is a more suitable approach. The same trends have been reported in previous multimodal learning works [45].

Table 4 also shows that the proposed HGD framework outperforms OSD in regard to accuracy (p<0.001, Table 5). Notably, when comparing with OSD and DLC, TSD achieved significant improvements of accuracy and UAR (p<0.001, Tables 5 and 6). Note that OSD uses a simple feature concatenation scheme and DLC uses a simple model combination scheme to learn information from multiple modalities. On the other hand, the proposed TSD processes acoustic features and medical records using the first-stage DNNs, whose outputs are fed to the second-stage DNN. Specifically, the TSD jointly optimizes DNNs for feature processing and for the fusion module for classification and is thus able to achieve higher performance than OSD (optimizing a DNN for the fusion module) and DLC (optimizing DNNs for feature processing).

Finally, we note that TSD outperforms HGD, leading to an accuracy improvement of 3.68% (from 83.58 to 87.26%) and a UAR improvement of 5.46% (from 76.13 to 81.59%) with a borderline statistical significance (Tables 5 and 6). Note that HGD adopts a GMM-based supervector for a more suitable combination of the information in acoustic signals and medical data, and thus achieves better performance than OSD and DLC. However, the GMM-based supervector is an average representation of the whole acoustic feature sequence. On the contrary, the TSD that jointly optimizes first-stage DNNs (for feature processing) and the second-stage DNN (for fusion) in order to generate optimal classification results representing a better approach for this pathological voice detection task.

From the results of Table 4, we can observe an interesting diversity of interaction between disease categories and classification models. The four multimodal learning systems and the systems using acoustic signals and medical records alone perform similarly in phonotrauma, in which all the sensitivities are around 90%. The results suggest that when individual classifiers can already yield satisfactory performance, the multimodal learning generate only marginal performance improvements. However, for glottic neoplasm, although systems using only acoustic signals and medical records do not perform well, the proposed HGD and TSD can yield notable improvements of performance by combining acoustic signals and medical records. The TSD approach achieves best performance among all results reported in Table 4, confirming the advantages of the two-stage learning architecture, which jointly considers feature processing and multimodality fusion in a unified framework.

IV. CONCLUSION

This paper proposes two multimodal learning frameworks, namely HGD and TSD, to efficiently exploit the complementary advantages of acoustic signals and medical records. The HGD framework transforms dynamic acoustic waveforms into a static supervector via a GMM; the supervector is then combined with the medical records to form the input vector for the DNN to perform classification. The TSD framework has a two-stage DNN architecture to jointly optimize the feature processing and the fusion module. Experimental results from 589 samples of glottic neoplasm, phonotraumatic lesions, and vocal paralysis demonstrated that the proposed multimodal learning frameworks outperform systems using simply acoustic signals or medical records for classifying voice disorders, and improves the accuracy and UAR by 2.02–10.32% and 2.48–17.31%, respectively. The proposed frameworks also provide higher accuracy and UAR than typical feature-based and model-based combination methods.

In the future, we plan to deploy the proposed multimodal learning frameworks to detect and predict voice disorders in real clinical scenarios. A potential implementation would be via internet and cloud computation. In such environments, acoustic signals may be distorted by environmental noises, quality of recording devices, and channel mismatches. Furthermore, patients provided information may not be as complete as those gathered from medical facilities. More robust and effective refinements of the proposed multimodal learning frameworks are required to predict diverse categories of voice disorders with scarce training data and low-computational costs.

ACKNOWLEDGEMENTS

The study protocol was approved from the Research Ethics Review Committee of Far Eastern Memorial Hospital, Taipei, Taiwan (No. 105139-E). This study is supported by research grants from the Ministry of Science and Technology, Taipei, Taiwan (MOST 106-2314-B-418-003, 107-2314-B-418-008, and 108-2634-F-155-001).

REFERENCES

[1] Kelly, J.; Knottenbelt, W.: Neural nilm: deep neural networks applied to energy disaggregation, in *Proc. of the 2nd ACM Int. Conf. on Embedded Systems for Energy-Efficient Built Environments*, pp. 55–64, ACM, 2015.

[2] Mocanu, E.; Nguyen, P.H.; Gibescu, M.; Kling, W.L.: Deep learning for estimating building energy consumption. *Sustainable Energy, Grids and Networks*, **6** (2016), 91–99.

[3] Akçay, S.; Kundegorski, M.E.; Devereux, M.; Breckon, T.P.: Transfer learning using convolutional neural networks for object classification within x-ray baggage security imagery. IEEE, 2016.

[4] Zheng, Y.-J.; Sheng, W.-G.; Sun, X.-M.; Chen, S.-Y.: Airline passenger profiling based on fuzzy deep machine learning. *IEEE Trans. Neural Netw. Learn. Syst.*, **28** (12) (2017), 2911–2923.

[5] Shi, S.; Wang, Q.; Xu, P.; Chu, X.: Benchmarking state-of-the-art deep learning software tools, in *Cloud Computing and Big Data (CCBD), 2016 7th IEEE Int. Conf.*, pp. 99–104, IEEE, 2016.

[6] Bahrampour, S.; Ramakrishnan, N.; Schott, L.; Shah, M.: Comparative study of deep learning software frameworks. *arXiv preprint arXiv:1511.06435*, 2015.

[7] Huang, W.; Song, G.; Hong, H.; Xie, K.: Deep architecture for traffic flow prediction: deep belief networks with multitask learning. *IEEE Trans. Intell. Transp. Syst.*, **15** (5) (2014), 2191–2201.

[8] Lv, Y.; Duan, Y.; Kang, W.; Li, Z.; Wang, F.-Y. *et al.*: Traffic flow prediction with big data: a deep learning approach. *IEEE Trans. Intell. Transp. Syst.*, **16** (2) (2015), 865–873.

[9] Fang, S.-H.; Fei, Y.-X.; Xu, Z.; Tsao, Y.: Learning transportation modes from smartphone sensors based on deep neural network. *IEEE Sensors J.*, **17** (18) (2017), 6111–6118.

[10] Fang, S.-H. *et al.*: Transportation modes classification using sensors on smartphones. *Sensors*, **16** (8) (2016), 1324.

[11] Mamoshina, P.; Vieira, A.; Putin, E.; Zhavoronkov, A.: Applications of deep learning in biomedicine. *Mol. Pharm.*, **13** (5) (2016), 1445–1454.

[12] Titze, I.: Workshop on acoustic voice analysis: summary statement. National center for voice and speech. 1995.

[13] Stachler, R.J. *et al.*: Clinical practice guideline: hoarseness (dysphonia)(update). *Otolaryngol. Head. Neck. Surg.*, **158** (2018), S1–S42.

[14] Fang, S.-H. *et al.*: Detection of pathological voice using cepstrum vectors: a deep learning approach. *J. Voice.*, 2018.

[15] Gilman, M.; Merati, A.L.; Klein, A.M.; Hapner, E.R.; Johns, M.M.: Performer's attitudes toward seeking health care for voice issues: understanding the barriers. *J. Voice*, **23** (2) (2009), 225–228.

[16] Vaziri, G.; Almasganj, F.; Behroozmand, R.: Pathological assessment of patients' speech signals using nonlinear dynamical analysis. *Comput. Biol. Med.*, **40** (1) (2010), 54–63.

[17] Cohen, S.M.; Dupont, W.D.; Courey, M.S.: Quality-of-life impact of non-neoplastic voice disorders: a meta-analysis. *Ann. Oto. Rhinol. Laryn.*, **115** (2006), 128–134.

[18] Arjmandi, M.K.; Pooyan, M.: An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine. *Biomed. Signal. Process. Control.*, **7** (2012), 3–19.

[19] Markaki, M.; Stylianou, Y.: Using modulation spectra for voice pathology detection and classification, in *2009 Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, September (2009), 2514–2517.

[20] Hammami, I.; Salhi, L.; Labidi, S.: Pathological voices detection using support vector machine, in *2016 2nd Int. Conf. on Advanced Technologies for Signal and Image Processing (ATSIP)*, March (2016), 662–666.

[21] Verde, L.; Pietro, G.D.; Sannino, G.: Voice disorder identification by using machine learning techniques. *IEEE Access.*, **6** (2018), 16246–16255.

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

[23] Arias-Londoño, J.D.; Godino-Llorente, J.I.; Sáenz-Lechón, N.; Osma-Ruiz, V.; Castellanos-Domínguez, G.: Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients. *IEEE Trans. Biomed. Eng.*, **58** (2011), 370–379.

[24] Ali, Z.; Alsulaiman, M.; Muhammad, G.; Elamvazuthi, I.; Mesallam, T.A.: Vocal fold disorder detection based on continuous speech by using mfcc and gmm, in *2013 7th IEEE GCC Conf. and Exhibition (GCC)*, November (2013), 292–297.

[25] Fezari, M.; Amara, F., I.M. El-Emary: Acoustic analysis for detection of voice disorders using adaptive features and classifiers, in *Int. Conf. on Circuits, Systems and Control, ISBN*, January (2014), 978–1.

[26] Wang, J.; Jo, C.: Vocal folds disorder detection using pattern recognition methods, in *2007 29th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, August (2007), 3253–3256.

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

[28] Wu, H.; Soraghan, J.; Lowit, A., G. Di, Caterina: Convolutional neural networks for pathological voice detection, in *40th Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, 2018.

[29] Alhussein, M.; Muhammad, G.: Voice pathology detection using deep learning on mobile healthcare framework. *IEEE Access.*, **6** (2018), 41034–41041.

[30] Gupta, V.: Voice disorder detection using long short term memory (lstm) model. *arXiv preprint arXiv:1812.01779*, 2018.

[31] Hsu, Y.-T.; Zhu, Z.; Wang, C.-T.; Fang, S.-H.; Rudzicz, F.; Tsao, Y.: Robustness against the channel effect in pathological voice detection. *Machine Learning for Health (ML4H) Workshop at NeurIPS*, 2018.

[32] Lee, T. *et al.*: Automatic speech recognition for acoustical analysis and assessment of cantonese pathological voice and speech, in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE Int. Conf.*, pp. 6475–6479, IEEE, 2016.

[33] Dibazar, A.A.; Narayanan, S.; Berger, T.W.: Feature analysis for automatic detection of pathological speech, in *Proc. of the Second Joint 24th Annual Conf.and the Annual Fall Meeting of the Biomedical Engineering Society][Engineering in Medicine and Biology*, vol. 1, pp. 182–183, IEEE, 2002.

[34] Dibazar, A.A.; Narayanan, S.: A system for automatic detection of pathological speech, in *Conference Signals, Systems, and Computers, Asilomar, CA*, 2002.

[35] Henríquez, P.; Alonso, J.B.; Ferrer, M.A.; Travieso, C.M., J.I. Godino-Llorente, F. Díaz-de María: Characterization of healthy and pathological voice through measures based on nonlinear dynamics. *IEEE Trans. Audio, Speech, Language Process.*, **17** (6) (2009), 1186–1195.

[36] Zhang, Y.; Jiang, J.J.: Acoustic analyses of sustained and running voices from patients with laryngeal pathologies. *J. Voice*, **22** (1) (2008), 1–9.

[37] Moran, R.J.; Reilly, R.B.; de Chazal, P.; Lacy, P.D.: Telephony-based voice pathology assessment using automated speech analysis. *IEEE Trans. Biomed. Eng.*, **53** (3) (2006), 468–477.

[38] Tsui, S.-Y.; Tsao, Y.; Lin, C.-W.; Fang, S.-H.; Lin, F.-C.; Wang, C.-T.: Demographic and symptomatic features of voice disorders and their potential application in classification using machine learning algorithms. *Folia Phoniatrica et Logopaedica*, **70** (3–4) (2018), 174–182.

[39] Stemple, J.C.; Roy, N.; Klaben, B.K.: Clinical voice pathology: Theory; management. Plural Publishing, The United States of America, 2014.

[40] Hashibe, M. *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 Epidemiol. Prev. Biomar.*, **18** (2) (2009), 541–550.

[41] Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y.: Multimodal deep learning, in *Proc. of the 28th Int. Conf. on machine learning (ICML-11)*, 2011, 689–696.

[42] Mroueh, Y.; Marcheret, E.; Goel, V.: Deep multimodal learning for audio-visual speech recognition, in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE Int. Conf.*, pp. 2130–2134, IEEE, 2015.

[43] Hsiao, S.-W.; Sun, H.-C.; Hsieh, M.-C.; Tsai, M.-H.; Tsao, Y.; Lee, C.-C.: Toward automating oral presentation scoring during principal certification program using audio-video low-level behavior profiles. *IEEE Trans. Affect. Comput.*, 2017.

[44] Wu, C.-H.; Lin, J.-C.; Wei, W.-L.: Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA Trans. Signal Inf. Process.*, **3**, (2014), e12

[45] Hou, J.-C.; Wang, S.-S.; Lai, Y.-H.; Tsao, Y.; Chang, H.-W.; Wang, H.-M.: Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Trans. Emerging Topics in Computational Intelligence*, **2** (2) (2018), 117–128.

[46] Aaltonen, L.-M. *et al.*: Voice quality after treatment of early vocal cord cancer: a randomized trial comparing laser surgery with radiation therapy. *Int. J. Radiation Oncology* Biology* Physics*, **90** (2) (2014), 255–260.

[47] Hsu, Y.-C.; Lin, F.-C.; Wang, C.-T.: Optimization of the minimal clinically important difference of the mandarin chinese version of 10-item voice handicap index. *J. Taiwan Otolaryngology-Head and Neck Surgery*, **52** (1) (2017), 8–14.

[48] Belafsky, P.C.; Postma, G.N.; Koufman, J.A.: Validity and reliability of the reflux symptom index (rsi). *J. voice*, **16** (2) (2002), 274–277.

[49] Bocklet, T.; Haderlein, T.; Hönig, F.; Rosanowski, F.; Nöth, E.: Evaluation and assessment of speech intelligibility on pathologic voices based upon acoustic speaker models, in *Proc. of the 3rd Advanced Voice Function Assessment Int. Workshop*, pp. 89–92, Citeseer, 2009.

[50] Reynolds, D.A.; Quatieri, T.F.; Dunn, R.B.: Speaker verification using adapted Gaussian mixture models. *Digit. Signal processing*, **10** (1–3) (2000), 19–41.

[51] Campbell, W.M.; Sturim, D.E.; Reynolds, D.A.: Support vector machines using gmm supervectors for speaker verification. *IEEE Signal Process. Lett.*, **13** (5) (2006), 308–311.

[52] Kinnunen, T.; Li, H.: An overview of text-independent speaker recognition: From features to supervectors. *Speech Commun.*, **52** (1) (2010), 12–40.

[53] Kumar, K.; Kim, C.; Stern, R.M.: Delta-spectral cepstral coefficients for robust speech recognition, in *2011 IEEE int. conf. on acoustics, speech and signal processing (ICASSP)*, pp. 4784–4787, IEEE, 2011.

[54] Ahmad, K.S.; Thosar, A.S.; Nirmal, J.H.; Pande, V.S.: A unique approach in text independent speaker recognition using mfcc feature sets and probabilistic neural network, in *2015 Eighth Int. Conf. on Advances in Pattern Recognition (ICAPR)*, pp. 1–6, IEEE, 2015.

[55] Fang, S.-H.; Chuang, C.-C.; Wang, C.: Attack-resistant wireless localization using an inclusive disjunction model. *IEEE Trans. Commun.*, **60** (5) (2012), 1209–1214.

[56] Fang, S.-H.; Wang, C.-H.: A novel fused positioning feature for handling heterogeneous hardware problem. *IEEE Trans. Commun.*, **63** (7) (2015), 2713–2723.

[57] Karamizadeh, S.; Abdullah, S.M.; Manaf, A.A.; Zamani, M.; Hooman, A.: An overview of principal component analysis. *J. Signal Info. Process.*, **4** (03) (2013), 173.

[58] Boualleg, A.; Bencheriet, C.; Tebbikh, H.: Automatic face recognition using neural network-pca, in *2006 2nd Int. Conf. on Information & Communication Technologies*, vol. 1, pp. 1920–1925, IEEE, 2006.

[59] Meng, J.; Yang, Y.: Symmetrical two-dimensional pca with image measures in face recognition. *Int. J. Adv. Robot. Syst.*, **9** (6) (2012), 238.

Shih-Hau Fang (M'07-SM'13) is a Full Professor in the Department of Electrical Engineering, Yuan Ze University (YZU), and MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan. He received a B.S. from National Chiao Tung University in 1999, an M.S. and a Ph.D. from National Taiwan University, Taiwan, in 2001 and 2009, respectively, all in communication engineering. From 2001 to 2007, he was a software architect at Chung-Hwa Telecom Ltd. and joined YZU in 2009. Prof. Fang has received several awards for his research work, including the Young Scholar Research Award (YZU, 2012), Project for Excellent Junior Research Investigators (MOST, 2013), Outstanding Young Electrical Engineer Award (the Chinese Institute of Electrical Engineering, 2017), Outstanding Research Award (YZU, 2018), Best Synergy Award (Far Eastern Group, 2018), and Y.Z Outstanding Professor (Y.Z. Hsu Science and Technology Memorial Foundation, 2019). His team won the third place of IEEE BigMM (Multimedia Big Data) HTC Challenge in 2016, and the third place of IPIN (Indoor Positioning and Indoor Navigation) in 2017. He is currently technical advisor to HyXen and PTCom Technology Company Ltd., an Associate Editor for IEICE Trans. on Information and Systems, and serves as YZU President's Special Assistant. Prof. Fang's research interests include artificial intelligence, mobile computing, machine learning, and signal processing. He is a senior member of IEEE.

Chi-Te Wang received his MD degree from the National Taiwan University, Taipei, Taiwan, in 2003. After resident training from 2003 to 2008, he joined Far Eastern Memorial Hospital as an attending physician. He received PhD degree from the Institute of Epidemiology and Preventive Medicine at National Taiwan University in 2014. He received the PhD degree from the Institute of Epidemiology and Preventive Medicine at National Taiwan University in 2014. During his professional carrier, he visited Mount Sinai Hospital (NYC, 2009), Mayo Clinic (Arizona, 2012), Isshiki voice center (Kyoto, 2015), UC Davis voice and swallow center (Sacramento, 2018), and UCSF voice and swallow center (San Francisco, 2018) for continual exposure on the expertise practice. He is a corresponding member of the American Laryngological Society and member of councils on the Taiwan Otolaryngological Society and Taiwan Voice Society. He has a wide clinical and academic interest on different fields, including phonosurgery, automatic detection and classification of voice disorders, real time monitoring of phonation, and telepractice of voice therapy. He is the winner of Society for Promotion of International Oto-Rhino-Laryngology (SPIO) Award on 2015, and Best Synergy Award of Far Eastern Group on 2018.

Ji-Ying Chen received the B.S. and M.S. degrees in electrical engineering from Yuan Ze University, Taoyuan, Taiwan, in 2017 and 2019, respectively. His research interests include artificial intelligence, machine learning, and signal processing.

Yu Tsao received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1999 and 2001, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2008. From 2009 to 2011, he was a Researcher with the National Institute of Information and Communications Technology, Tokyo, Japan, where he was involved in research and product development in automatic speech recognition for multilingual speech-to-speech translation. He is currently an Associate Research Fellow with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. His research interests include speech and speaker recognition, acoustic and language modeling, audio-coding, and biosignal processing. He was the recipient of Academia Sinica Career Development Award in 2017 and National Innovation Award in 2018.

Feng-Chuan Lin received the MD degree from the University of Taipei, Taipei, Taiwan, in 2009. He became a speech therapist after graduation. He joined the Department of Otolaryngology, Head and Neck Surgery at Far Eastern Memorial Hospital from 2015. He is currently the president of New Taipei City Speech-Language Pathologists Union. His clinical and academic interests include voice therapy, speech training, and swallowing rehabilitation.

APPENDIX

In order to provide statistically objective results, we repeated the experiments for five rounds and obtained a total of 25 validation data sets and examined the performance between different models using Student's *t*-tests. Bonferroni method was applied to adjust the significant level of *p* value, as illustrated below:

0.05 (original significant level of p value)/30 (numbers of repeated statistical tests, Tables 5 and 6) = 0.00167

For easier interpretation, we defined $p<0.001$ as the adjusted significance level.