Detection of Pathological Voice Using Cepstrum Vectors: A Deep Learning Approach

***Shih-Hau Fang, †Yu Tsao, *Min-Jing Hsiao, *Ji-Ying Chen, ‡Ying-Hui Lai, §Feng-Chuan Lin, and §,¶,**Chi-Te Wang,** **Taoyuan and †‡§¶**Taipei, Taiwan*

Summary: Objectives. Computerized detection of voice disorders has attracted considerable academic and clinical interest in the hope of providing an effective screening method for voice diseases before endoscopic confirmation. This study proposes a deep-learning-based approach to detect pathological voice and examines its performance and utility compared with other automatic classification algorithms.

Methods. This study retrospectively collected 60 normal voice samples and 402 pathological voice samples of 8 common clinical voice disorders in a voice clinic of a tertiary teaching hospital. We extracted Mel frequency cepstral coefficients from 3-second samples of a sustained vowel. The performances of three machine learning algorithms, namely, deep neural network (DNN), support vector machine, and Gaussian mixture model, were evaluated based on a five-fold cross-validation. Collective cases from the voice disorder database of MEEI (Massachusetts Eye and Ear Infirmary) were used to verify the performance of the classification mechanisms.

Results. The experimental results demonstrated that DNN outperforms Gaussian mixture model and support vector machine. Its accuracy in detecting voice pathologies reached 94.26% and 90.52% in male and female subjects, based on three representative Mel frequency cepstral coefficient features. When applied to the MEEI database for validation, the DNN also achieved a higher accuracy (99.32%) than the other two classification algorithms.

Conclusions. By stacking several layers of neurons with optimized weights, the proposed DNN algorithm can fully utilize the acoustic features and efficiently differentiate between normal and pathological voice samples. Based on this pilot study, future research may proceed to explore more application of DNN from laboratory and clinical perspectives.

Key Words: Nodule–Polyp–Neoplasm–Spasmodic dysphonia–Sulcus.

INTRODUCTION

From a health science perspective, the pathological status of the human voice can substantially reduce the quality of life and occupational performance,[1] which results in considerable costs for both the patient and the society. Common pathologies of impaired vocal function include structural lesions (eg, vocal fold nodules, polyps, and cysts), neoplasms, and neurogenic disorders (eg, vocal fold paralysis and adductor spasmodic dysphonia).[2] Current standards recommend the use of laryngeal endoscopy for the accurate diagnoses of voice disorders,[3] which requires well-trained specialists and expensive equipment. In places without sufficient medical resources, and for patients without adequate insurance coverage, correct diagnosis and subsequent treatment may be delayed.

To mitigate these problems, noninvasive screening methods have been proposed for clinical applications.[4] Because laryngeal disorders, particularly those originating from the membranous vocal folds, almost always result in the change of voice quality, an automatic speech recognition framework was developed to differentiate between normal and pathological voices. A previous study reported an overall accuracy of 93.4% using a linear discriminant analysis to detect continuous speech samples from the voice disorder database of Massachusetts Eye and Ear Infirmary (MEEI).[5,6] Later studies using Mel frequency cepstral coefficients (MFCCs) and Gaussian mixture model (GMM) further improved the classification accuracy to 94.1%.[7] Another scheme combining the modified MFCCs and hidden Markov model has been proposed by Costa et al.[8] Meanwhile, other work applied neural networks to classify MFCC features[9] and showed that the performance can be improved by differentiating the speaker's gender.[10] Combining MFCCs and other acoustic features, Arias-Londono et al developed a multistaged classifier and effectively increased the accuracy to 98%.[11,12] Aside from the MFCC features, wavelet transform and the modulated spectrum were also applied for the tasks of abnormal voice detection using various algorithms with good accuracy.[13–15]

Despite the success of the abovementioned studies, automatic expert systems for voice disorders have not popularly used as initially envisioned because people are still reluctant to rely on a machine to receive a diagnosis. In recent years, the concept of deep learning has attracted considerable attention. Deep neural network (DNN) approaches have been extensively implemented to model data in numerous applications.[16–19] With increasing daily exposure to machine learning and big data empowered health services of the general population,[20–22] it might be a good timing to reinvestigate the potential of computerized classification systems of voice disorders. To the best of our knowledge, no study has attempted to use DNN for the detection of pathological voice samples before. Accordingly, we conducted this pilot study with the following objectives: (1) to propose a

Accepted for publication February 6, 2018.

From the *Department of Electric Engineering, Yuan Ze University, Taoyuan, Taiwan; †Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan; ‡Institute of Biomedical Engineering, National Yang-Ming University, Taipei, Taiwan; §Department of Otolaryngology Head and Neck Surgery, Far Eastern Memorial Hospital, Taipei, Taiwan; ¶Department of Special Education, University of Taipei, Taipei, Taiwan; and the **Department of Otolaryngology Head and Neck Surgery, National Taiwan University College of Medicine, Taipei, Taiwan.

Address correspondence and reprint requests to Chi-Te Wang, Department of Otolaryngology Head and Neck Surgery, Far Eastern Memorial Hospital, 21, Section 2, Nan-Ya South Road, Pan Chiao District, New Taipei City, Taiwan. E-mail: drwangct@gmail.com

Journal of Voice, Vol. ■■, No. ■■, pp. ■■-■■
0892-1997

https://doi.org/10.1016/j.jvoice.2018.02.003

DNN-based system for detecting features extracted from voice samples, (2) to examine the performance of DNN in differentiating between normal and pathological voice samples, and (3) to validate the accuracy of the DNN using the widely applied voice disorder database from MEEI.

MATERIALS AND METHODS

Study subjects

Voice samples were obtained from a voice clinic in a tertiary teaching hospital (Far Eastern Memorial Hospital, FEMH), which included 60 normal voice samples and 402 samples of common voice disorders, including vocal nodules, polyps, and cysts; glottic neoplasm; vocal atrophy; laryngeal dystonia (ie, spasmodic dysphonia and tremor); unilateral vocal paralysis; and sulcus vocalis (Tables 1 and 2). Voice samples of a 3-second sustained vowel sound /a:/ were recorded at a comfortable level of loudness, with a microphone-to-mouth distance of approximately 15–20 cm, using a high-quality microphone (Model: SM58, SHURE, IL),[23] with a digital amplifier (Model: X2u, SHURE) under a background noise level between 40 and 45 dBA. The sampling rate was 44,100 Hz with a 16-bit resolution, and data were saved in an uncompressed .wav format.

Feature extraction from MFCCs

Derived through pre-emphasis, windowing, fast Fourier transform, Mel filtering, nonlinear transformation, and discrete cosine transform, MFCCs have been widely used in acoustic research.[24] For example, MFCC and MFCC + delta features were selected for voice disorder detection,[7,10,11] and the normalized version was selected for performance comparison.[25,26] To capture these MFCC features, first, the raw waveform was divided into N frames (or segments), represented by the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ (Figure 1). A total of N frames were then transformed into N MFCC vectors, representing the acoustic features. Next, for the second feature, we calculated the trajectories of the MFCCs over time (delta MFCCs) and appended them to the original MFCCs. Finally, we normalized the MFCCs such that all of the coefficients had zero mean and unit variance, and appended the delta MFCCs to the normalized MFCCs to form the third feature vectors. The details of MFCCs and their variations were outlined in a previous publication.[27]

The experimental setups for the acoustic signal processing and feature extraction procedures are described later. The first feature, made up of 13-dimension MFCCs, was extracted from a 16-millisecond windowed signal using an 8-millisecond frameshift. A window length of 16 milliseconds is used to capture the fast dynamic acoustic waves, whereas the 8-millisecond frameshift enables smoothness between frames. Similar settings were applied in many previous studies.[28,29] The next feature, MFCC + delta, was created by appending 13 velocity features to the original 13-dimension MFCCs and thus had 26 dimensions. The third feature, denoted by MFCC(N) + delta for convenience, has the same dimensions as that of MFCC + delta. The only difference is that the former normalized all MFCC coefficients with zero mean and unit variance.

DNN

A DNN model comprises multiple hidden layers to form a complex mapping function between the inputs and outputs. Previous studies have verified that a DNN model can provide a satisfactory performance in speech enhancement.[30] In DNN, the relationship between the input, $\boldsymbol{x}$, and the output of the first hidden layer, $\boldsymbol{h}_1$, is described as

$$\boldsymbol{h}_1 = f(\boldsymbol{W}_1\boldsymbol{x}) + \boldsymbol{b}_1, \quad (1)$$

where $\boldsymbol{W}_1$ and $\boldsymbol{b}_1$ are the weight matrix and bias vector, respectively, and $f(.)$ is the activation function. In this study, we use the sigmoid function for the activation function, namely, $f(z) = [1 + \exp(-z)]^{(-1)}$, based on the better performance among different activation functions (Appendix 1 of the Supplementary material). The relationship between the current hidden layer and the next hidden layer can be expressed as

$$\boldsymbol{h}_{i+1} = f(\boldsymbol{W}_{i+1}\boldsymbol{h}_i + \boldsymbol{b}_{i+1}), i = 1, 2, \ldots, L-1, \quad (2)$$

TABLE 1.
Demographics of the 462 Normal and Pathological Voice Samples

	Number		Mean Age (y)		Age Range (y)		Standard Deviation	
	M	F	M	F	M	F	M	F
Normal	16	44	30.7	30.1	23–37	22–47	3.93	5.79
Pathological	189	213	56.1	44.2	20–87	20–87	15.9	14.9

Abbreviations: M, male; F, female.

TABLE 2.
Disease Categories of the 402 Pathological Voice Samples

	Nodules	Polyp	Cyst	Neoplasm	Atrophy	Dystonia	Vocal Palsy	Sulcus
M	1	18	17	43	39	2	41	28
F	51	33	34	5	16	17	26	31

Abbreviations: M, male; F, female.

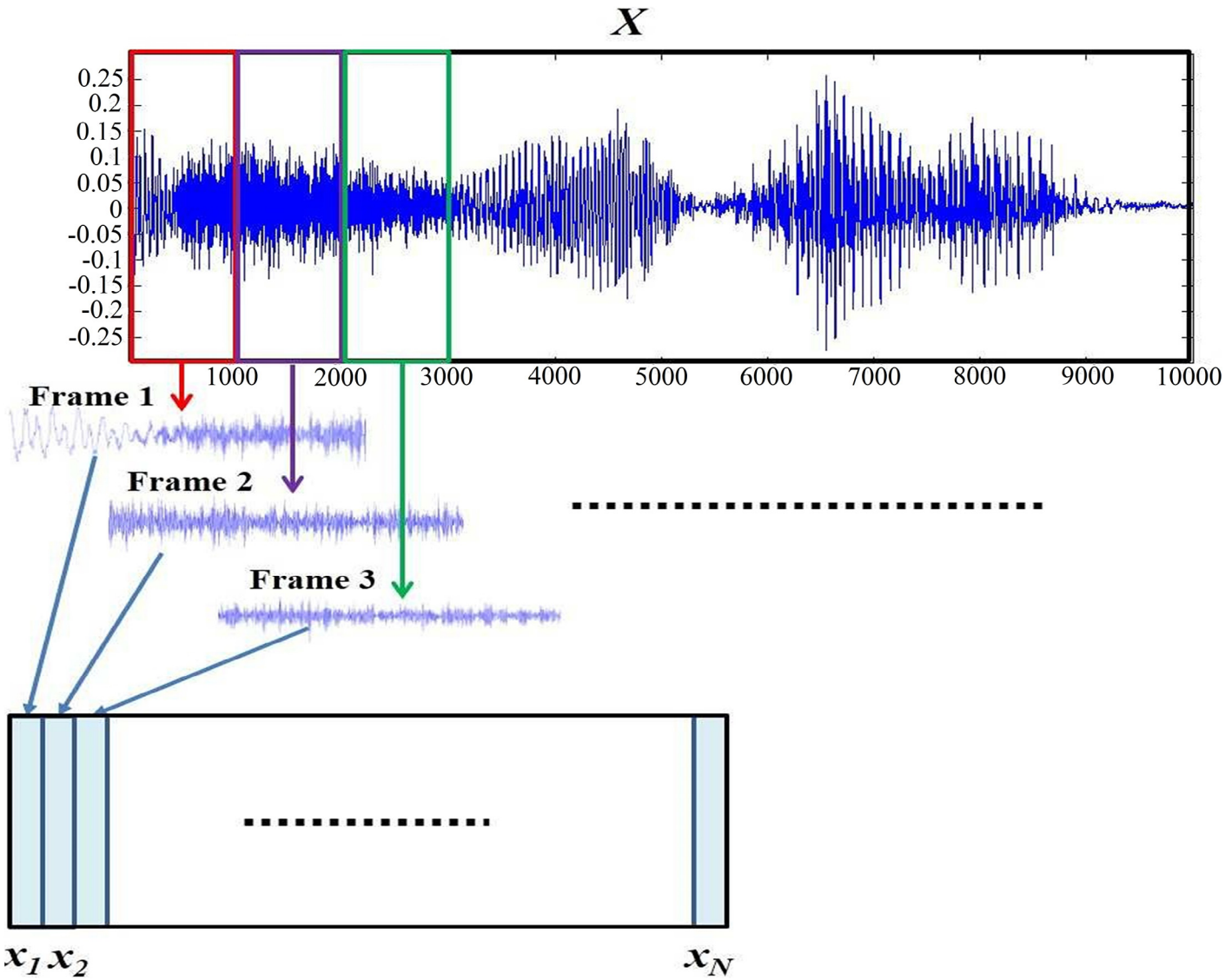

FIGURE 1. Illustration of the acoustic waveform segmentation procedure.

where L is the DNN hidden layer number. Finally, another function, $g(.)$, is adopted on the output layer to form the output vector $\hat{y}$. Thus, we have

$$\hat{y} = g(\boldsymbol{h}_L). \tag{3}$$

For classification tasks, the softmax function is usually adopted for $g(.)$. To compute the parameters in the DNN model with the training samples $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_N]$ and the corresponding labels $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_i, \ldots, \boldsymbol{y}_N]$, where N is the total number of training samples, we formulate an objective function:

$$O\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}; \boldsymbol{X}, \theta\right) = \frac{-1}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J}[y_{i,j}\log\hat{y}_{i,j}], \tag{4}$$

where $\theta = \{\boldsymbol{W}_l, \boldsymbol{b}_l, l = 1, 2, L\}$ is the DNN parameter set, and $\hat{\mathrm{Y}} = [\hat{y}_1, \hat{y}_i, \ldots, \hat{y}_N]$ is the DNN output ($\hat{y}_i$ is the ith DNN output given input $\boldsymbol{x}_i$); $\boldsymbol{y}_{i,j}$ and $\hat{y}_{i,j}$ denote the jth element of $\boldsymbol{y}_i$ and $\hat{y}_i$, respectively. The parameter is then estimated by

$$\theta^* = \arg\min_{\theta} O(\boldsymbol{y}, \hat{\boldsymbol{y}}; \boldsymbol{X}, \theta), \tag{5}$$

where the standard back-propagation algorithm is applied to compute θ^* in Eq. (5).

Experimental setup

We examined 1–16 Gaussian mixtures for the GMM and tested the performance of different kernel functions for the SVM, including linear, quadratic, and Gaussian functions. The DNN was structured into multiple hidden layers with varying neuron numbers in each layer. The best combination of hidden layers and number of neurons was determined based on the experimental results. The threshold of the ratio of the pathological feature vectors was investigated from 0.1 to 0.9, with a 0.1 increment. The performance was evaluated through a fivefold cross-validation. We utilized general accuracy, which is widely used for detection tasks, as the main performance metric.

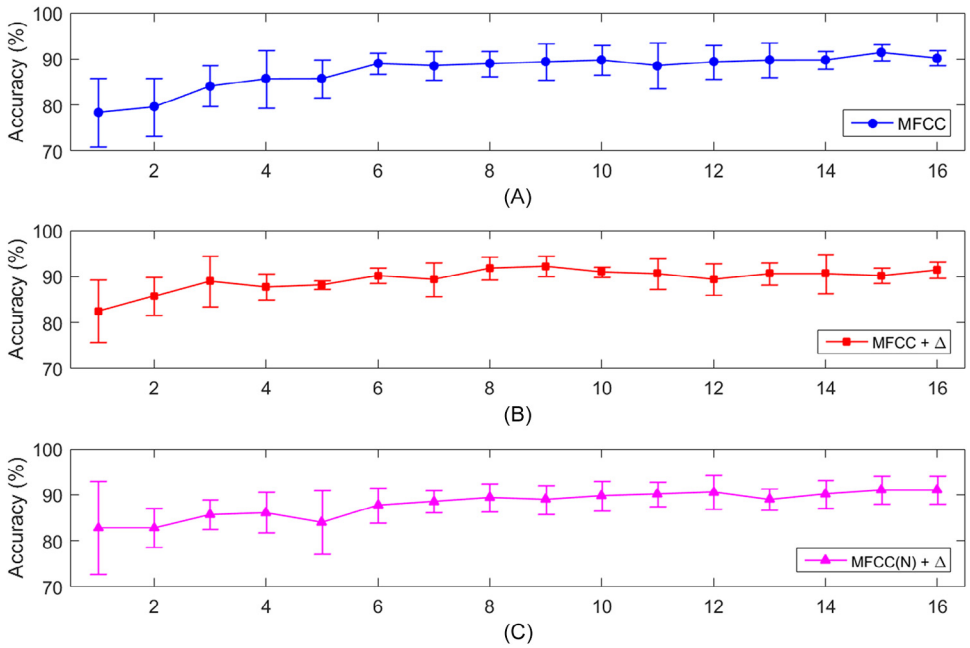

FIGURE 3. Accuracy of different numbers of mixtures in GMM with **(A)** MFCC, **(B)** MFCC + delta, and **(C)** MFCC (N) + delta features.

RESULTS AND DISCUSSION

Spectrogram and acoustic waveforms

Figure 2 shows the waveform and spectrogram plots of normal and pathological voice samples. From the waveform plots, the pathological voice sample (B) showed irregular and wider variations of amplitude compared with the normal voice sample (A). Meanwhile, from the spectrogram plots, the normal voice sample (C) presented clearer harmonic structures, especially in the low-frequency areas. In contrast, the pathological voice sample (D) showed blurred harmonic structures and contained noise-like components in the high-frequency region.

Optimal setting for SVM, GMM, and DNN

We performed multiple experiments to determine the optimal kernel functions for SVM. The Gaussian radial basis kernel function was chosen because of its highest accuracy for female voice samples with the lowest variation of accuracy among male samples (Appendix 2 of the Supplementary material). We also compared the accuracies of GMM using different numbers of Gaussian mixtures with three MFCC features. The results show that the accuracy increases as the number of mixtures increases from 1 to 6, whereas the performance becomes saturated when the number of Gaussian mixtures ranges from 8 to 16 (Figure 3). Accordingly, we used eight Gaussian mixtures as a representative GMM model in the following study. We also examine the performance among different DNN structures (ie, hidden layers and number of neurons); the results showed that the best performance was achieved when using 3 hidden layers with 300 neurons in each layer (Appendix 3 of the Supplementary material). Finally, we investigated the ratio of the pathological feature vectors to determine the adequate cut-off value.

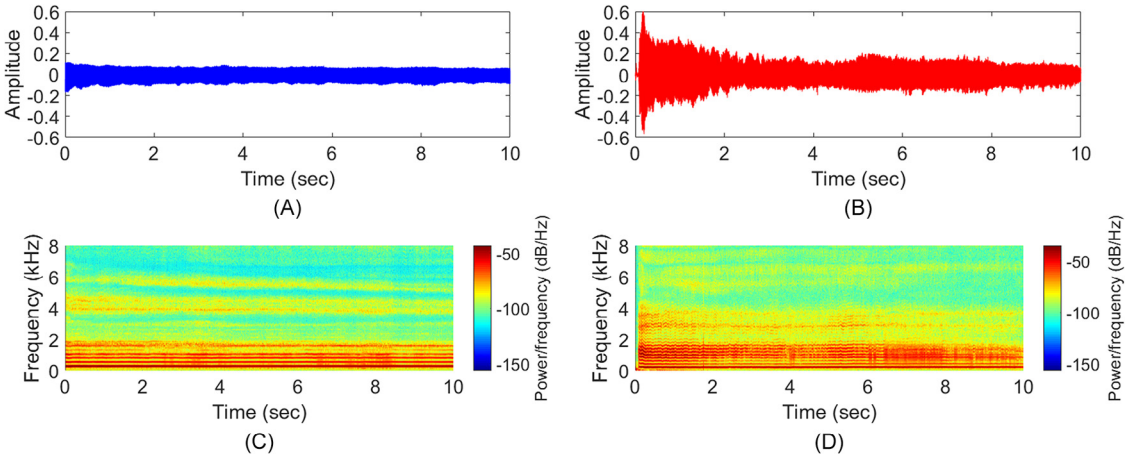

FIGURE 2. Waveform of a normal voice sample **(A)** and a pathological voice sample **(B)**. Wide band spectrogram in the corresponding normal voice sample **(C)** and pathological voice sample **(D)**.

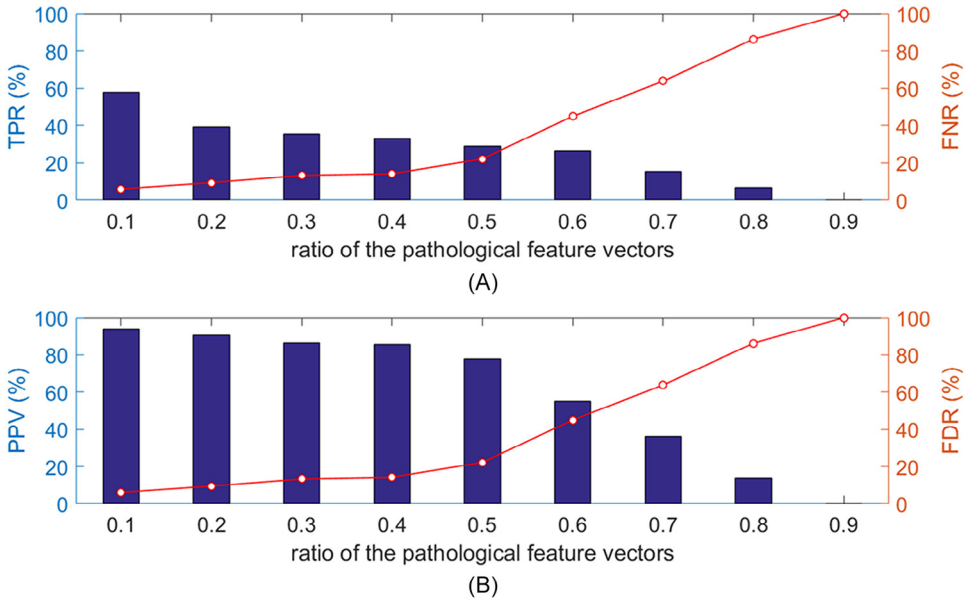

FIGURE 4. **(A)** True-positive rate (TPR), false-negative rate (FNR); **(B)** positive predictive value (PPV) and false detection rate (FDR) among different ratios of pathological feature vectors on DNN.

Experimental results indicated that 0.5 could be a good balance point between the true-positive rate, false-negative rate, positive predictive value, and false detection rate (Figure 4).

Accuracy of DNN in comparison with SVM and GMM

Table 3 compares the performance of the three algorithms and three features among the voice samples of the male subjects from FEMH. The table indicates that the DNN coupled with the MFCC(N) + delta feature provides higher accuracy (94.26%), with a lower standard variation (2.25%) than those of the other two classification algorithms (GMM and SVM) and the other two features (MFCC and MFCC + delta). Data on the performance of the DNN among female subjects are provided in Table 4, which also demonstrates that the DNN and MFCC(N) + delta feature achieve the highest accuracy for detecting pathological voice samples. However, the overall accuracy among the female voice samples is lower than that among the male samples. Similar to our results, a previous study by Fraile et al also reported a higher accuracy for pathological voice detection in men than in women.[10] The authors proposed that such discrepancy might be explained by the wider distribution of the values of MFCC features in women compared with the narrower distribution for men.[10]

Compared with a previous study using an artificial neural network containing one layer of hidden node (accuracy: men: 90.95%; women: 86.50%),[10] DNN model with multiple layers of hidden nodes further increases the accuracy rates by around 4% (men: 94.26%; women: 90.52%, Tables 3 and 4). Accordingly, our results indicate that DNN achieves the highest performance for both female and male subjects, confirming the capability of DNN for detecting pathological voice samples. Moreover, the velocity (delta) features and normalization are useful for improving the performance, particularly for the GMM and DNN methods, in the FEMH database of voice disorders.

Validation of DNN performance using MEEI voice disorder database

To validate the aforementioned results, which indicates that DNN outperforms SVM and GMM in detecting pathological voice samples, we applied a common voice disorder database from MEEI under the same experimental setting. We retrieved 53 normal and 173 pathological samples from the MEEI database,[5] identical to a previously published study.[7] Results in Table 5 showed that DNN provides greater accuracy and a lower standard deviation than SVM and GMM, indicating the same tendency as the previous results obtained using the FEMH data (Tables 3 and 4). Similarly, compared with previous studies using neural networks with a single hidden layer to detect pathological voice samples from MEEI database,[15,31] this study utilized DNN with three hidden layers and exhibited a better performance, further confirming the advantages of the proposed DNN-based approach. Although the detailed settings for extracting the MFCC features and numbers of GMM mixtures were not identical with the previous study by Godino-Llorente et al,[7] this study also showed that dynamic delta features of MFCCs do not enhance the capability of the MEEI model in the detection of voice disorders (Table 5). Such a concordance may be due to the fact that the MFCC produces sufficient discriminative information when voice samples are recorded in a well-controlled environment. Accordingly, the appended delta trajectory may be redundant and result in learning confusions. In contrast, under circumstances in which the voice samples are recorded in suboptimal settings, adding temporal derivatives (delta feature) might be helpful to increase the robustness of performance (Tables 3 and 4).

TABLE 3.
Classification Accuracies of Three Classification Algorithms and Three MFCC Features Among Male Subjects

	SVM	GMM	DNN
	Accuracy ± Standard Deviation	Accuracy ± Standard Deviation	Accuracy ± Standard Deviation
MFCC	92.24 ± 2.66%	89.00 ± 1.79%	93.86 ± 2.05%
MFCC + delta	92.24 ± 2.66%	91.02 ± 3.38%	93.86 ± 2.05%
MFCC(N) + delta	93.04 ± 2.74%	90.24 ± 4.18%	94.26 ± 2.25%

TABLE 4.
Classification Accuracies of Three Classification Algorithms and Three MFCC Features Among Female Subjects

	SVM	GMM	DNN
	Accuracy ± Standard Deviation	Accuracy ± Standard Deviation	Accuracy ± Standard Deviation
MFCC	85.18 ± 0.72%	83.56 ± 2.12%	86.14 ± 1.43%
MFCC + delta	85.18 ± 0.72%	86.12 ± 4.35%	87.74 ± 1.43%
MFCC(N) + delta	87.40 ± 1.92%	90.20 ± 3.83%	90.52 ± 2.00%

TABLE 5.
Detection of Pathological Voice Samples in the MEEI Voice Disorder Database

	SVM	GMM	DNN
	Accuracy ± Standard Deviation	Accuracy ± Standard Deviation	Accuracy ± Standard Deviation
MFCC	98.28 ± 2.36%	98.26 ± 1.80%	99.14 ± 1.92%
MFCC + delta	93.04 ± 2.74%	90.24 ± 4.18%	94.26 ± 2.25%
MFCC(N) + delta	87.40 ± 1.92%	90.20 ± 3.83%	90.52 ± 2.00%

Limitations

This study found that the classification accuracy using DNN model is higher for the MEEI data than for the FEMH data. A similar result was also reported in a previous study, in which the MEEI database revealed a much higher accuracy of pathological voice detection than another database,[12] probably owing to differences in the recording environment (eg, background noise levels and recording instruments). For example, the frequency response of the microphone used in this study was not ideally flat with a peak around the 2–4 kHz region, which may have affected the recognition results. Nevertheless, as voice samples used for training and testing were recorded using the same microphone, the experimental results shall remain credible.[32] Future study using a better microphone (with an ideally flat frequency response) might result in higher recognition accuracy than the current study. Besides, larger numbers of pathological samples in FEMH database might introduce higher variability, resulting in lower accuracy for automatic detection. Another possible explanation is that the ratio between normal and pathological voice samples is different between the two databases. Further study shall gather more voice samples from normal subjects to provide a wider base of feature identification for machine learning algorithms. In addition, other techniques of feature extraction, for example, wavelet transformation[15] and spectra modulation,[12] can be combined with DNN for a better performance and wider application on different voice databases.

CONCLUSION

This study proposed an enhanced pathological voice detection system that combines a DNN classifier and normalized MFCC features. The results based on both the FEMH and the MEEI databases indicated that DNN outperforms traditional GMM and SVM in improving detection accuracy based on three representative features. Successful aspects of the current offline model (recording followed by analysis) can be implemented in the future development of an online model (simultaneous recording and analysis) for telepractice, in which the voice samples may be screened in real time through modern cloud-based computation (Figure 5).

Acknowledgments

This study was supported by research grants from the Ministry of Science and Technology (MOST 105-2221-E-155-013-MY3, 106-2314-B-418-003 and 107-2634-F-155-001). The authors thank Prof. Chii-Wann Lin, PhD, and Mr. Sheng-Yang

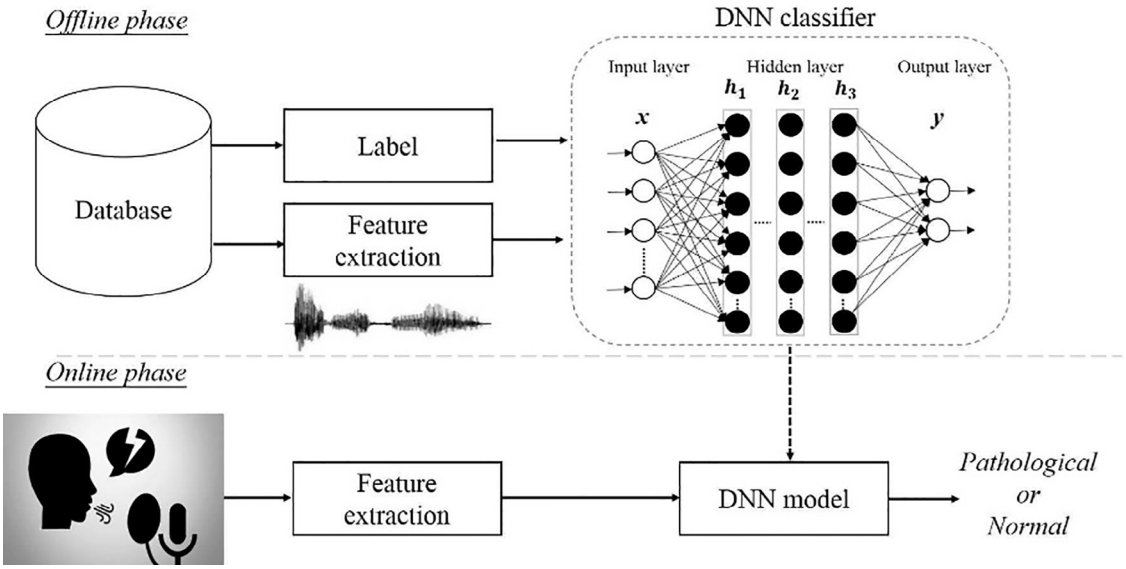

FIGURE 5. Online and offline models of the proposed pathological voice detection system.

Tsui, BSc, for their help in the analysis of acoustic data and neural network modeling.

SUPPLEMENTARY DATA

Supplementary data related to this article can be found online at doi:10.1016/j.jvoice.2018.02.003.

REFERENCES

1. Titze IR. Workshop on acoustic voice analysis: Summary statement. National Center for Voice and Speech; 1995.
2. Stemple JC, Roy N, Klaben BK. *Clinical Voice Pathology Theory and Management*. San Diego: Plural Publishing; 2014.
3. Schwartz SR, Cohen SM, Dailey SH, et al. Clinical practice guideline: hoarseness (dysphonia). *Otolaryngol Head Neck Surg*. 2009;141:S1–S31.
4. Vaziri G, Almasganj F, Behroozmand R. Pathological assessment of patients' speech signals using nonlinear dynamical analysis. *Comput Biol Med*. 2010;40:54–63.
5. Elemetrics K. Disordered voice database 1.03ed, 1994.
6. Umapathy K, Krishnan S, Parsa V, et al. Discrimination of pathological voices using a time-frequency approach. *IEEE Trans Biomed Eng*. 2005;52:421–430.
7. Godino-Llorente JI, Gomez-Vilda P, Blanco-Velasco M. Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters. *IEEE Trans Biomed Eng*. 2006;53:1943–1953.
8. Costa SC, Neto BGA, Fechine JM. Pathological voice discrimination using cepstral analysis, vector quantization and hidden Markov models. In: *IEEE International Conference on Bioinformatics and Bioengineering*. Athens, Greece; 2008:1–5.
9. Salhi L, Mourad T, Cherif A. Voice disorders identification using multilayer neural network. *Int Arab J Inf Technol*. 2010;7:177–185.
10. Fraile R, Saenz-Lechon N, Godino-Llorente JI, et al. Automatic detection of laryngeal pathologies in records of sustained vowels by means of Mel-frequency cepstral coefficient parameters and differentiation of patients by sex. *Folia Phoniatr Logop*. 2009;61:146–152.
11. Arias-Londono JD, Godino-Llorente JI, Saenz-Lechon N, et al. Automatic detection of pathological voices using complexity measures, noise parameters, and Mel-cepstral coefficients. *IEEE Trans Biomed Eng*. 2011;58:370–379.
12. Arias-Londono JD, Godino-Llorente JI, Markaki M, et al. On combining information from modulation spectra and Mel-frequency cepstral coefficients for automatic detection of pathological voices. *Logoped Phoniatr Vocol*. 2011;36:60–69.
13. Markaki M, Stylianou Y. Voice pathology detection and discrimination based on modulation spectral features. *IEEE Trans Audio, Speech, Language Proc*. 2011;19:1938–1948.
14. Muhammad G, Mesallam TA, Malki KH, et al. Multidirectional regression (MDR)-based features for automatic voice disorder detection. *J Voice*. 2012;26:e819–e827.
15. Arjmandi MK, 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*. 2012;7:3–19.
16. Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Proc Mag*. 2012;29:82–97.
17. Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*. 2016;529:484–489.
18. Fang SH, Fei YX, Xu ZZ, et al. Learning transportation modes from smartphone sensors based on deep neural network. *IEEE Sens J*. 2017;17:6111–6118.
19. Li B, Tsao Y, Sim KC. An investigation of spectral restoration algorithms for deep neural networks based noise robust speech recognition. INTERSPEECH; 2013:3002–3006.
20. Tawalbeh LA, Mehmood R, Benkhlifa E, et al. Mobile cloud computing model and big data analysis for healthcare applications. *IEEE Access*. 2016;4:6171–6180.
21. Sahoo PK, Mohapatra SK, Wu SL. Analyzing healthcare big data with prediction for future health condition. *IEEE Access*. 2017;99:1.
22. Ma Y, Wang Y, Yang J, et al. Big health application system based on health internet of things and big data. *IEEE Access*. 2016;PP:1.
23. Fu S, Theodoros DG, Ward EC. Delivery of intensive voice therapy for vocal fold nodules via telepractice: a pilot feasibility and efficacy study. *J Voice*. 2015;29:696–706.
24. Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans Acoust*. 1980;28:357–366.
25. Hamawaki S, Funasawa S, Katto J, et al. Feature Analysis and Normalization Approach for Robust Content-Based Music Retrieval to Encoded Audio with Different Bit Rates. In: Huet B, Smeaton A, Mayer-Patel K, et al., eds. *Advances in Multimedia Modeling: 15th International Multimedia Modeling Conference, MMM 2009, Sophia-Antipolis, France, January 7–9, 2009. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009:298–309.

26. Boril H, Hansen JHL. Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments. *IEEE Trans Audio, Speech Lang Proc*. 2010;18:1379–1393.
27. Zhang D, Gatica-Perez D, Bengio S, et al. Semisupervised adapted HMMS for unusual event detection. *IEEE Comput Soc Conf Comput Vis Pattern Recognit*. 2005;1:611–618.
28. Chan CP, Wong YW, Tan L, et al. Two-dimensional multi-resolution analysis of speech signals and its application to speech recognition 1999. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, Vol. 401. ICASSP99 (Cat. No.99CH36258). Phoenix, AZ, USA; 1999:405–408.
29. Dahmani M, Guerti M. Vocal folds pathologies classification using Naïve Bayes Networks Systems and Control (ICSC), 2017:p. 426–432.
30. Lu X, Tsao Y, Matsuda S, et al. Speech enhancement based on deep denoising autoencoder, 2013:436–440.
31. Godino-Llorente JI, Gomez-Vilda P. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. *IEEE Trans Biomed Eng*. 2004;51:380–384.
32. Li J, Deng L, Gong Y, et al. An overview of noise-robust automatic speech recognition. *IEEE/ACM Trans Audio, Speech Lang Proc*. 2014;22:745–777.