Learning Strategies for Voice Disorder Detection

Conference Paper · January 2019

DOI: 10.1109/ICOSC.2019.8665504

2 authors:

Hongzhao Guan
Georgia Institute of Technology
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Alexander Lerch
Georgia Institute of Technology
75 PUBLICATIONS **553** CITATIONS
SEE PROFILE

Learning Strategies for Voice Disorder Detection

Hongzhao Guan
Center for Music Technology
Georgia Institute of Technology
Atlanta, Georgia 30332
Email: hguan7@gatech.edu

Alexander Lerch
Center for Music Technology
Georgia Institute of Technology
Atlanta, Georgia 30332
Email: alexander.lerch@gatech.edu

***Abstract*—Voice disorder is a health issue that is frequently encountered, however, many patients either cannot afford to visit a professional doctor or neglect to take good care of their voice. In order to give a patient a preliminary diagnosis without using professional medical devices, previous research has shown that the detection of voice disorders can be carried out by utilizing machine learning and acoustic features extracted from voice recordings. Considering the increasing popularity of deep learning and feature learning, this study explores the possibilities of using these methods to assign voice recordings into one of the two classes—*Normal* and *Pathological*. While the results show the general viability of deep learning and feature learning for the automatic recognition of voice disorder, they also demonstrate the shortcomings of the existing datasets for this task such as insufficient dataset size and lack of generality.**

I. Introduction

Examples of voice disorders include vocal fold nodules, polyps, and cysts could be caused by overusing or misusing your voice. Among people who use their voice professionally, voice disorders are a frequently observed problem [1]. There exists evidence from a 1990s survey that 69 percent of singers and 41 percent of non-singers report a vocal disability over a time period of 12 months [1]. Another survey shows that singers and teachers might be forced to end their career earlier if they do not take care of their voice in time, and 39 percent of the polled population of this study is not willing to seek medical attention for their voice [2].

The diagnosis of voice disorders requires professional skills and medical equipment providing a close-up view of the vocal folds. The motivation of this study stems from the desire to provide people about a preliminary voice diagnosis and inform them with potential issues regarding their vocal health. As opposed to many professional diagnosis devices, this study focuses on voice disorder diagnosis using audio recordings. The goal is to assess the patients' vocal health condition from voice recordings by assigning them into one of the two classes—*Normal* and *Pathological*. In order to do so, this study explores both feature learning and deep learning methods in the context of voice disorder detection. A detailed analysis of intra dataset evaluation vs. inter dataset evaluation enables us to point out current data challenges in this field.

The main contributions of this paper are:

(1) the evaluation of feature learning and deep learning to voice disorder detection,

(2) the usefulness of data augmentation when applying feature learning and deep learning to voice disorder detection, and

(3) a systematic study of shortcomings of commonly used datasets on the task of voice disorder detection.

The remainder of this paper is structured as follows: Section II presents the related work in voice disorder detection and the significance of cross-dataset evaluation on machine learning tasks. In Sec. III, the proposed method is described. The experiment setup and results are discussed in Sect. IV and Sect. V. The conclusion and future research directions can be found in the final Sect. VI.

II. Related Work

Computational systems to discriminate between normal and pathological voices have been proposed as early as the 1980s [3]. Many approaches proposed in the literature using supervised learning algorithms and custom-designed acoustic features [4]–[6]. For example, Wallen and Hansen propose to use features such as pitch perturbation and amplitude perturbation and a neural network classifier [6]. Alonso et. al also propose to use neural network classifiers but with a different group of features such as bispectrum and chaos [4], [5]. Among all of these approaches, many of them use support vector machine (SVM) as a classifier [7]–[11]. On the other hand, Mel-Frequency Cepstral Coefficients (MFCCs) are commonly chosen as features to train different types of classifiers [11]–[15].

In the past few years, deep learning has received considerable attention due to its superior performances on various machine learning tasks including voice disorder detection. Fang et al. propose a system that passing MFCCs into a multilayer Deep Neural Network (DNN) with a sigmoid activation function and a softmax layer as the output layer [16]. Nevertheless, one existing challenge of using deep learning is the models' requirements on the amount of training data. For this problem, Muhammad et al. propose a system that uses transfer learning and adopts CaffeNet [17]. CaffeNet is a Convolutional Neural Network (CNN) which is powerful on image classifications [18], and input representations such as mel-spectrogram and octave-spectrogram can be treated as image representations of an audio signal [17]. In order to perform voice disorder detection using CaffeNet, the softmax layer in CaffeNet is replaced by a new softmax layer that has two neurons.

Feature learning differs from the supervised approaches mentioned above in that it attempts to automatically learn the most representative features from data as opposed to using features carefully designed by experts. It has been been successfully applied to speech-related tasks such as speech recognition [19] and speech emotion recognition [20]. However, feature learning has not been explored in the context of voice disorder detection.

It is worth pointing out that the vast majority of publications on this topic utilizes only one dataset, the Massachusetts Eye and Ear Infirmary (MEEI) dataset [21]. Considering the distribution of the features such as gender and age, some works only use a subset of the MEEI database to evaluate their systems. For example, Parsa and Jamieson employed a subset that contains 53 normal files and 173 pathological files [22], and this subset is repeatedly used in later research [8], [10], [16], [23]. Godino et al. use another subset consisting of 53 normal files and 82 pathological files with the pathological files are randomly selected from the whole database [14].

Very high classification scores on the MEEI dataset are reported in the literature. The system proposed by Fang et al., for example, achieves a 99.32% classification rate on a specific subset [16], while Al-nasheri et al. present their system with a 99.81% classification rate on the same subset [8]. Furthermore, Godino et al. report a best accuracy of around 95% when experimenting on their subset [14]. There is one publication by Dibazar et al. that reports results for the complete MEEI dataset as high as 98.3%; in this case, however, the publication does not provide sufficient details on system or methodology [12]. However, as pointed out by Torralba and Efros that every dataset has its own bias [24], so that assumption might be true for the MEEI dataset as well.

III. Systems

The flowchart that describes the proposed systems is shown in Fig. 1. Pre-processing is performed to extract equal-length snippets from the dataset recordings. Data augmentation is applied to four specific experiments that are discussed in Sect. IV. Then all voice recordings are framed and transformed into three different input representations. After the input representations are obtained, four different machine learning approaches are investigated: (1) support vector machine (SVM), (2) CNN, (3) CNN followed by SVM, and (4) autoencoder (AE) followed by SVM.

A. Pre-processing and Data Augmentation

Silence at the start and end of the audio recordings is removed, and all files are down-sampled to $16\,\mathrm{kHz}$ afterwards. Next, each file is segmented into multiple $500\,\mathrm{ms}$ long snippets, with a $400\,\mathrm{ms}$ overlap of subsequent snippets. It is worth mentioning that one recording from the MEEI dataset is shorter than $0.5\,\mathrm{s}$; therefore, it is discarded and leaves the MEEI dataset with 655 pathological audio files before framing.

To investigate the impact of the amount of training data on the results, dataset augmentation is implemented for both

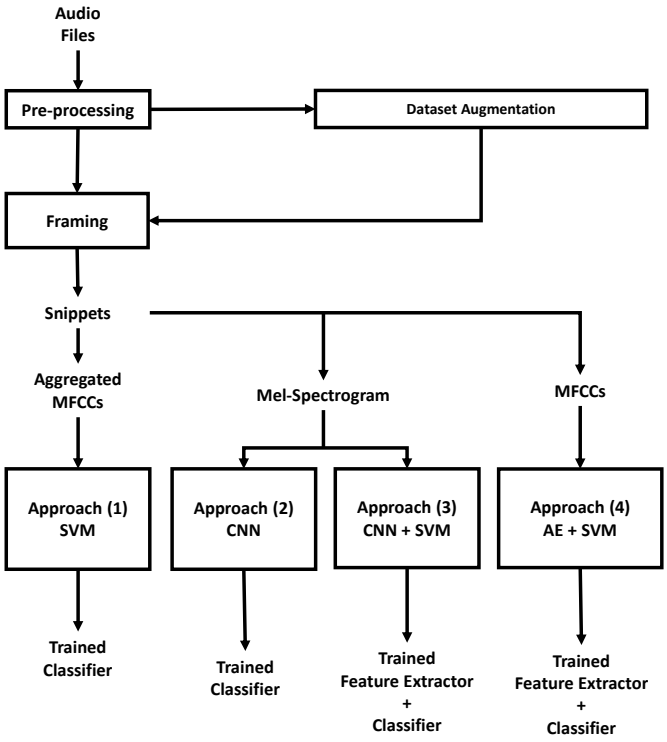

Fig. 1. Flowchart of the system overview.

datasets before framing them into snippets. In the augmentation step, each pre-processed file is pitch-shifted by 8 half-semitones up and 8 half-semitones down using a professional grade commercial pitch-shifting solution.[1] The goal of this data augmentation is to generate more training data and ensure pitch independence of the trained models. Using a commercial, de-facto standard, pitch shifting engine allows us to make the assumption of a reasonable realistic, i.e., artifact-free and natural-sounding pitch-shifted dataset extension.

B. Input Representation

In order to account for different requirements of different machine learning approaches, the audio data is firstly converted into two widely used input representations, a feature-based representation with MFCCs and a mel-spectrogram.

Each snippet is divided into multiple blocks for a Short-time Fourier transform (STFT) with a block size of 512 and a hop size of 128 samples, respectively. Then, for each block, 20 MFCCs are extracted to form a 20×63 input matrix and their mean values and standard deviations over blocks are calculated, resulting in a 40-dimensional feature vector per snippet. A detailed description of MFCCs and audio feature aggregation can be found in [25]. MFCCs and aggregated MFCCs are separately normalized to a range of 0 to 1.

A low-level acoustic representation, mel-spectrogram, is chosen as the input representation for the deep learning

[1] zplane.development: elastique Pro http://www.time-stretching.com Last Access Date: 2018-10-20

models. A mel-frequency spectrogram is computed from the Fourier spectrogram by applying a nonlinear transformation to the frequency axis. The block size and hop size are identical to the vector representation, and the mel-length is fixed at 128, resulting in an input matrix of 128 $\times$ 63 dimension. At last, every mel-spectrogram is normalized to a range of 0 to 1.

C. Baseline System with Support Vector Machines

SVM is a widely-used binary classifier. In audio classification, a SVM with aggregated MFCC input can be considered as a standard baseline. A linear kernel is used and a parameter search is performed on c= $\{0.1, 1, 10, 100\}$ during training.

D. Deep Learning with Convolutional Neural Networks

Convolutional Neural Network is a widely popular deep learning architecture. A general CNN is comprised of one or more convolutional layers and pooling layers followed by one or more fully connected layers [26]. CNNs have been extensively applied to image classification tasks [27], but also have been increasingly popular for audio analysis such us speech recognition [28] and instrument activity detection [29]. Borrowing the idea of passing images into a CNN, a mel-spectrogram can be treated as a single channel image. An input will go through multiple convolutional layers, pooling layers, and fully-connected layers to reach a classification result. A softmax function at the output layer of the CNN can predict the probabilities of each class for a data point, and this data point will be assigned to the class with the highest probability.

Due to the limited amount of data points, the designed model's structure is relatively simple. The CNN architecture used in this study consists of two convolutional layers, a pooling layer, and multiple fully-connected layers. In this CNN, the convolutional layers and pooling layer are designed to be the features extractors, and the fully-connected layers are treated as a classifier. The first 2-dimensional convolutional layer has 9 output channels, and the kernel is chosen to be a 5 $\times$ 5 moving window, followed by a (1,1) stride in both directions. The second convolutional layer has 15 output channels with a 3 $\times$ 3 kernel, and it has the same stride as the previous convolutional layer. On the other hand, the following pooling layers applies 2-dimensional average pooling, and the nine fully-connected layers contain 1024, 512, 256, 128, 64, 32, 16, 8, and 4 neurons.

E. Feature Learning with CNNs

Feature learning is motivated by the idea of learning meaningful features from the data itself, as opposed to relying on experts to design meaningful features with the expert knowledge. Compared to the raw input representation, the learned representation is expected to be more representative.

The CNN described in Sec. III-D is used to extract compressed data from the input representation because the higher layers of a CNN contain highly compressed representations of the input. After the CNN is trained, output from any fully-connected layer can be collected as a vector of learned features, and the decision of taking which layer's output is left to be a hyperparameter for tuning during training session. The learned features are collected and passed to a SVM, and the SVM is trained based on the learned features and ground-truth label targets.

F. Autoencoder

An autoencoder is a neural network whose input vector and output vector have the same dimension, and the goal of training an autoencoder is to minimize the difference between the input and the output [27]. A hidden layer is selected as encoding layer, and its dimension is called encoding dimension, denoted as k_{AE}. The neural network from the input to the encoding layer is treated as an encoder; similarly, the neural network between the encoding layer and the output layer is treated as a decoder.

An input vector for an autoencoder in this study is made by concatenating the columns of the normalized MFCCs. With k_{AE} is fixed at 4, the number of hidden layers is set to nine for both the encoder and the decoder. The hidden layers for the encoder contain 1024, 512, 256, 128, 64, 32, 16, and 8 neurons, and the decoder's structure is symmetric to the encoder's.

After the autoencoder is properly trained, the outputs of the encoding layer or any hidden layers of the encoder can be interpreted as a vector of learned features, and the decision of taking which layer's output is left to be a hyperparameter for tuning during training session. The rest of training processes is similar to the processes explained above: a SVM will be trained on the learned features with the ground truth label targets.

IV. Experimental Setup

A. Datasets

1) MEEI Dataset: A commercial dataset developed by the Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Labs [21] is commonly used for the classification of voice disorders. This dataset contains 709 recordings of sustained phonations of the vowel /ah/. All recordings were collected in a controlled environment. The recorded audio files have sample rate of either $25\,\mathrm{kHz}$ or $50\,\mathrm{kHz}$. The dataset is split into the two classes: (1) *Normal* with 53 samples with a length or approximately $3\,\mathrm{s}$ and (2) *Pathological* with 656 samples of length $1\,\mathrm{s}$ or less. The differences in length are possibly caused by the fact that pathological patients have difficulties in phonating for a long time. Among the pathological recordings, the top five most common voice disorders are: Hyperfunction, A-P squeezing, Ventricular Compression, Paralysis, and Gastric Reflux.

2) UPM Dataset: The UPM dataset was recorded by the Technical University of Madrid (UPM) [30], [31]. It contains 440 recordings of sustained phonations of Spanish vowel /aa/ with a $50\,\mathrm{kHz}$ sampling rate. These 400 recordings consist of 239 normal files with a length 3 seconds and 200 pathological files with lengths between $1\,\mathrm{s}$ and $3\,\mathrm{s}$. Different to the MEEI dataset, the UPM dataset contains another set of voice disorders. The top five most common voice disorders are: Bilateral

Nodule, Bilateral Reinke Edema, Pedicled polyp, Sulcus in Stria, and Epidermoid Cyst.

B. Experiments

In order to investigate the neural networks' performance on this task, we perform two experiments (original vs. augmented training data) on two different datasets (MEEI and UPM), resulting in overall 4 experiments as described below. For each of the experiments, we compare a baseline SVM, a CNN, features from a CNN used with an SVM, and features from an Autoencoder used with an SVM. All experiments use 5-fold cross-validation. It is worth noticing that the systems were originally designed for the MEEI dataset. As mentioned in Sect. IV-A, the MEEI dataset and the UPM dataset contain different types of voice disorders, hence cross-dataset evaluation is not implemented in this study.

The four experiments are listed below in details:

- Experiment 1:
  Focusing on the non-augmented MEEI dataset with a 5-fold cross-validation.
- Experiment 2:
  Focusing on the augmented MEEI dataset with a 5-fold cross-validation.
- Experiment 3:
  Focusing on the non-augmented UPM dataset with a 5-fold cross-validation. The same
- Experiment 4:
  Focuses on the augmented UPM dataset with a 5-fold cross-validation.

These experiments are designed to achieve the following research goals:

- To investigate the relevance of data augmentation in the context of deep neural networks for this task by comparing the results from Exp. 1 and Exp. 2.
- To demonstrate the viability of using deep learning and feature learning on voice disorder detection and exhibit different approaches' capabilities using the results from Exp. 2.
- To show the importance of the amount of training data when detecting different types of voice disorder detection by comparing the results from Exp. 3 and Exp. 4.
- To test the robustness of the proposed approaches when they are conducted on another dataset by comparing the results from Exp. 1 and Exp. 3. This research goal can also be reached by comparing the results from Exp. 2 and Exp. 4.

The hyperparameters of the CNNs and AEs as well as the SVM parametrization are set during the training phase with a validation set. The separation of the training set, validation set, and the test set is based on the original files, that is, all snippets framed from one audio file will be in one of those three sets. When conducting 5-fold cross-validation, twenty percent files are isolated as the testing set, and ten percent of the the rest of files are randomly selected as the validation set.

In Exp. 1 and Exp. 3, training data takes snippets from the non-augmented dataset; however, in Exp. 2 and Exp. 4, it is formed by snippets from the augmented dataset. The validation and test sets, on the other hand, contain only snippets from the non-augmented dataset.

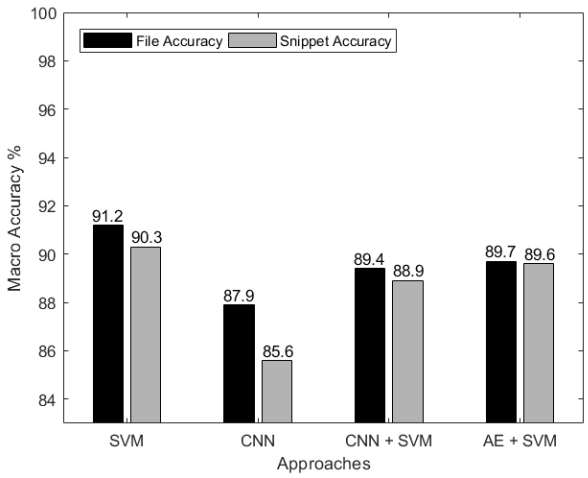

Fig. 2. Result of Experiment 1: non-augmented MEEI dataset

C. Metrics

In the case of k-fold cross-validation, the k confusion matrices are summed to get an overall confusion matrix, then the following two metrics are computed from the overall confusion matrix.

1) Snippet Accuracy: Since the number of data points in the classes is unbalanced, the macro accuracy, i.e., the average accuracy over classes is computed. This is to avoid the class containing more data points dominating the classification results. The following equation defines macro accuracy in case of two classes:

$$C = \frac{1}{2} \times (\frac{TP}{TP+FP} + \frac{TN}{FN+TN})$$

2) File Accuracy: The accuracy can be calculated on the file level using a majority vote per file. If more than 50 percent of the snippets of a file are assigned to a certain class, then the complete file is assumed to belong to that class. Then, macro accuracy is applied again to calculate the average accuracy.

V. Results

Cross-validation is applied on all experiments, and 5 folds are randomly separated before each run; therefore, the accuracy of one specific approach can vary slightly between runs, and an approximate value which is close to most of the results is reported.

A. Experiment 1 with non-augmented MEEI Dataset

Fig. 2 presents the results of different approaches on the non-augmented MEEI dataset. The SVM has the best performance among all approaches. There are two possible explanations for this observation: (1) the learned features

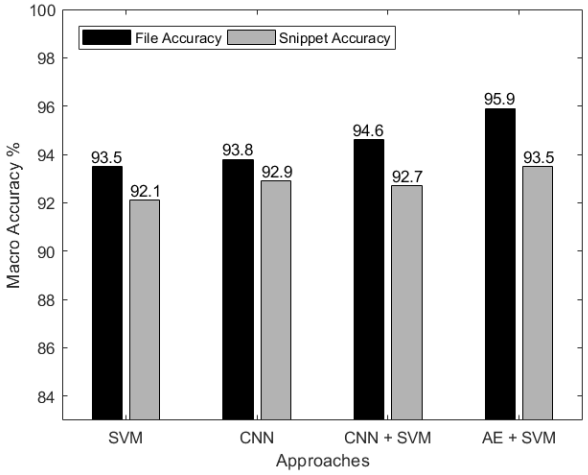

Fig. 3. Result of Experiment 2: augmented MEEI dataset

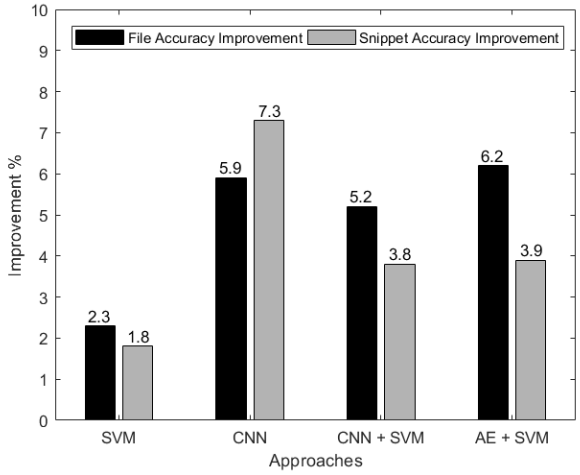

Fig. 4. Improvement on MEEI dataset after data-augmentation

are not superior to the features designed by experts such as MFCCs or (2) the insufficient amount of training data limits the abilities of the neural networks, in other words, the neural networks overfits the training data.

B. Experiment 2 with augmented MEEI Dataset

Fig. 3 shows the results for Exp. 2, using the same procedure as Exp. 1 but with augmented training data.

The first observation is, with an increased amount of training data, as expected, all deep learning and feature learning approaches can improve over the baseline's performance.

Secondly, although all deep learning and feature learning approaches only slightly outperform the baseline, the effects of data augmentation can be clearly observed. Every approach's improvement is shown in Fig. 4. The approaches using deep neural networks have substantial improvements improvements compared to the baseline. These observations prove the fact that the amount of training data can significantly affect the trained models especially when applying deep learning and feature learning.

The next observation from both Exp. 1 and Exp. 2 is, applying majority vote consistently increases the score. Two confusion matrices are provided in Table. I and Table. II, and these confusion matrices are recorded after running the autoencoder and SVM. As there is only one label per file, the majority vote is expected to remove outliers and provide more robust results.

	Normal	Pathological
Normal	96.23 %	3.77%
Pathological	4.43%	95.57%

TABLE I
CONFUSION MATRIX ON FILES, MACRO ACCURACY = 95.9%
AE + SVM APPROACH FROM EXP. 2

https://www.overleaf.com/project/5bbd544239ff686f15c337a5

	Normal	Pathological
Normal	93.79%	6.21%
Pathological	6.85%	93.15%

TABLE II
CONFUSION MATRIX ON SNIPPETS, MACRO ACCURACY = 93.5%
AE + SVM APPROACH FROM EXP. 2

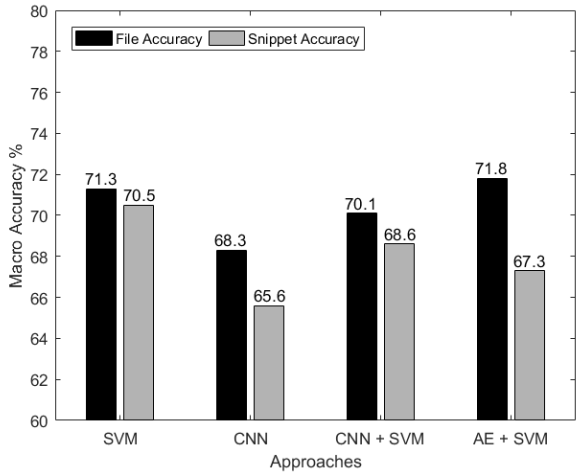

Fig. 5. Result of Experiment 3: non-augmented UPM dataset

C. Experiment 3 with non-augmented UPM Dataset

Different datasets probably have different characteristics. In Exp. 3 and Exp. 4, the proposed system is trained and evaluated on the UPM dataset. A robust system is expected to have similar performances on another dataset.

The result of experiment 3 can be found in Fig. 5. Comparing the results from Exp. 1 and Exp. 3, drastically reduced accuracies for all approaches are detected. Five speculations can be based on the results, and three speculations can be made from the observations on the results. First, since the proposed system architecture was originally designed for the MEEI dataset, it might only perform well on the types of

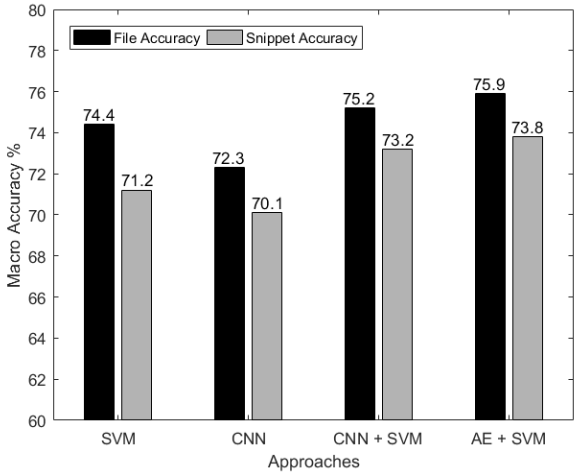

Fig. 6. Result of Experiment 4: augmented UPM dataset

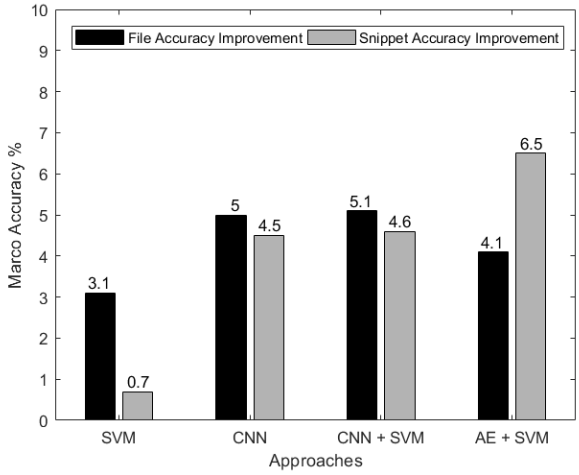

Fig. 7. Improvement on UPM dataset after data-augmentation

voices disorders contained in the MEEI dataset. Second, the different recording circumstances and environments might lead to particular characteristics that are only contained in a particular dataset. Third, the baseline accuracy also decreases; this observation implies that those particular characteristics are captured by low-level acoustic features as well. Another explanation for this phenomenon is that the aggregated MFCCs are more suitable to separate the types of disorders in the MEEI dataset from unimpaired vocals.

D. Experiment 4 with augmented UPM Dataset

The last experiment' results are shown in Fig. 6. Similar to Exp. 3, the accuracies are substantially lower than the results from Exp. 2. The similar three speculations can still be made in under these circumstances.

On the other hand, as shown in Fig. 7, the accuracy improvement is similar to the improvement in Exp. 2. This is an evidence supporting the hypothesis that the amount of training data is a key to implement deep neural network on voice disorder detection.

VI. Conclusion

This paper first proposed several approaches using deep learning and feature learning to classify normal and pathological voice recordings. These approaches, while generating acceptable results, have high dependency on the amount of training data. The experimental results showed that data augmentation is a valid solution to overcome this difficulty; however, a dataset contains more recordings is still more desirable.

This paper also demonstrated the importance of system evaluation using the MEEI dataset and the UPM dataset. Since these two datasets are consisted of different set of voice disorders and the system is originally designed for the MEEI dataset, the performance on the UPM dataset is not as great as expected. Generally speaking, to better test the robustness and the generality of an approach, more and possibly less homogeneous datasets are needed.

There are two general future directions of this work. First, the capabilities of different machine learning approaches, such as transfer learning, could be explored. These approaches might provide novel insights into voice disorder detection. Second, a successful and practically workable voice disorder detector in real-life should be able to categorize unseen audio recordings recorded in various circumstances and environments. Therefore, if there are multiple datasets contain similar types of disorders, cross-dataset evaluation will be an important and necessary research topic on voice disorder detection.

References

[1] D. J. Phyland, J. Oates, and K. M. Greenwood, "Self-reported voice problems among three groups of professional singers," *Journal of Voice*, vol. 13, no. 4, pp. 602–611, 1999.

[2] M. Gilman, A. L. Merati, A. M. Klein, E. R. Hapner, and M. M. Johns, "Performer's attitudes toward seeking health care for voice issues: understanding the barriers," *Journal of Voice*, vol. 23, no. 2, pp. 225–228, 2009.

[3] H. Kasuya, S. Ogawa, K. Mashima, and S. Ebihara, "Normalized noise energy as an acoustic measure to evaluate pathologic voice," *The Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1329–1334, 1986.

[4] J. B. Alonso, J. De Leon, I. Alonso, and M. A. Ferrer, "Automatic detection of pathologies in the voice by hos based parameters," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 275–284, 2001.

[5] J. B. Alonso, F. Díaz-de María, C. M. Travieso, and M. A. Ferrer, "Using nonlinear features for voice disorder detection," in *ISCA tutorial and research workshop (ITRW) on non-linear speech processing*, 2005.

[6] E. J. Wallen and J. H. Hansen, "A screening test for speech pathology assessment using objective quality measures," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2. IEEE, 1996, pp. 776–779.

[7] M. S. Hossain and G. Muhammad, "Healthcare big data voice pathology assessment framework," *IEEE Access*, vol. 4, pp. 7806–7815, 2016.

[8] 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.

[9] 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, 2018.
[10] R. Amami and A. Smiti, "An incremental method combining density clustering and support vector machines for voice pathology detection," *Computers & Electrical Engineering*, vol. 57, pp. 257–265, 2017.
[11] N. Souissi and A. Cherif, "Dimensionality reduction for voice disorders identification system based on mel frequency cepstral coefficients and support vector machine," in *Modelling, Identification and Control (ICMIC), 2015 7th International Conference on*. IEEE, 2015, pp. 1–6.
[12] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint*, vol. 1. IEEE, 2002, pp. 182–183.
[13] C. Maguire, P. d. Chazal, R. B. Reilly, and P. D. Lacy, "Identification of voice pathology using automated speech analysis," in *Third International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2003.
[14] J. I. Godino-Llorente and P. Gomez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 2, pp. 380–384, 2004.
[15] J. C. Saldanha, T. Ananthakrishna, and R. Pinto, "Vocal fold pathology assessment using mel-frequency cepstral coefficients and linear predictive cepstral coefficients features," *Journal of medical imaging and health informatics*, vol. 4, no. 2, pp. 168–173, 2014.
[16] 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*, 2018.
[17] 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.
[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.
[19] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks - studies on speech recognition," in *International Conference on Learning Representations (ICLR)*, 2013.
[20] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.
[21] M. Eye and E. Infirmary, "Voice disorders database, version. 1.03 (cd-rom)," *Lincoln Park, NJ: Kay Elemetrics Corporation*, 1994.
[22] V. Parsa and D. G. Jamieson, "Identification of pathological voices using glottal noise measures," *Journal of speech, language, and hearing research*, vol. 43, no. 2, pp. 469–485, 2000.
[23] ——, "Acoustic discrimination of pathological voice: sustained vowels versus continuous speech," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 2, pp. 327–339, 2001.
[24] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1521–1528.
[25] A. Lerch, *An introduction to audio content analysis: Applications in signal processing and music informatics*. John Wiley & Sons, 2012.
[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[27] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[28] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 4277–4280.
[29] S. Gururani, C. Summers, and A. Lerch, "Instrument activity detection in polyphonic music using deep neural networks," in *International Society for Music Information Retrieval (ISMIR)*, 2018.
[30] J. I. Godino-Llorente, P. Gómez-Vilda, F. Cruz-Roldán, M. Blanco-Velasco, and R. Fraile, "Pathological likelihood index as a measurement of the degree of voice normality and perceived hoarseness," *Journal of Voice*, vol. 24, no. 6, pp. 667–677, 2010.
[31] J. D. Arias-Londoño, J. I. Godino-Llorente, M. Markaki, and Y. Stylianou, "On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices," *Logopedics Phoniatrics Vocology*, vol. 36, no. 2, pp. 60–69, 2011.