IEEE Sensors Letters Volume 2(3) (2017)
Article Subject (see Index Terms below and write one here) _____

Ensemble and Multimodal Learning for Pathological Voice Classification

Whenty Ariyanti[1], Tassadaq Hussain[2], Jia-Ching Wang[1], Chi-Tei Wang[3], Shih-Hau Fang[4], Yu Tsao[2]

[1] *National Central University*
[2] *Academia Sinic*
[3] *Far Eastern Memorial Hospital*
[4] *Yuan Ze University*
** Senior Member, IEEE*
*** Fellow, IEEE*

Received 1 Nov 2016, revised 25 Nov 2016, accepted 30 Nov 2016, published 5 Dec 2016, current version 15 Dec 2016. (Dates will be inserted by IEEE; "published" is the date the accepted preprint is posted on IEEE Xplore®; "current version" is the date the typeset version is posted on Xplore®).

Abstract—Voice disorders are one of the most common medical diseases in modern society, especially for those who with occupational voice demand. In this paper, we investigate a stacked ensemble learning method to classify pathological voice disorders by combining acoustic signals and medical records. In the proposed ensemble learning framework, stacked support vector machines (SVMs) form a set of weak classifiers, and a deep neural network (DNN) acts a meta leaner. Acoustic features and medical records are combined to attain better classification performance based on the high complexity of meta learner. Results showed that the proposed approach significantly outperforms individual SVM and DNN classifiers, and showed a performance improvement over the two-stage-DNN based fusion classifier. The proposed approach achieved 89.83% accuracy and 85.84% unweighted average recall in a three-disorder classification task, confirming the effectiveness of the ensemble learning for pathological voice classification.

Index Terms—pathological voice, acoustic signal, ensemble learning, binary classification

I. INTRODUCTION

Acoustic sensing enables a variety of applications and health monitoring is one of the possibility [1, 2, 3]. Voice disorders are one of the most common medical disease in modern society. The health condition of the human voice can significantly affect social life of an individual. Diagnosing the early-stage voice disorders is critical. The standard to detect voice disorders is generally by a meticulous laryngeal examination procedure called "laryngeal endoscopy" [1, 4]. A major problem performing this procedure is that well-trained experts are needed because laryngeal endoscopy requires physical intrusion in the patient's body and sufficient medical knowledge. This standard procedure makes people, who lives in remote-area, difficult to have convenient diagnose because of limited access to medical services.

To mitigate the problem and lower the examination costs, non-invasive screening methods have been proposed [5]. A non-intrusive screening approach to detect voice disorder analyzes the patient's voice directly. Clinically, voice is considered as multi-dimensional measurement tool and can be used to identify the reason, severity, and prognosis of a disease as well as to develop a therapeutic program for the cure [6]. Mean airflow rate (MAFR) can be reduced because of abnormalities in the respiratory system. These phenomena can be further diagnosed to identify the clinical reason behind it [6], e.g.: vocal fold paralysis causes high MAFR whereas specific dysphonia results lower MAFR [6]. An expert can estimate the deviation from the optimum, which can be useful for therapy modeling through patients phonating the vowels /a:/, /i:/ and /u:/ [6]. Although this method is non-intrusive, it still requires an experienced operator to give us the diagnosis. In contrast, automated pathological voice diagnosis can be used as an inexpensive pre-screening tool to identify the disease remotely. [4, 5]. Some study has proposed an automatic detection of voice disorders based on the analysis of the acoustic quality of the voice signal patients to provide a better system for the diagnosis. This method is attractive since it enables an early screening for voice disorders in settings where optimal medical expertise and or facilities are not required.

A machine learning model using voice signal can be an effective way to identify the pathological voices [4, 7]. Numerous algorithms for effective pattern recognition have been developed in recent years. Support vector machines (SVM) [8] and AdaBoost are two well-known examples. When the task is binary classification, SVM can provide satisfactory performance [9]. While there are multiple classes at the output, more complex models are needed to reach good performance. For example, in [10], a multiclass version of AdaBoost was proposed for speech recognition. A more common approach is to construct the multiclass classifier by combining the outputs of multiple binary classifiers [9]. Typically, the combination is done via a simple nearest-neighbor rule, which finds the class that is closest to the given input and the final output.

In this paper, we proposed a stacked ensemble learning approach to perform pathological voice detection. Specifically, we extract features from voice samples provided by Fast Eastern Memorial Hospital (FEMH). The dataset contains voice samples of patients

Corresponding author: F. A. Author (f.author@psu.edu). If some authors contributed equally, write here, "F. A. Author and S. B. Author contributed equally."

IEEE Sensors Letters discourages courtesy authorship; please use the Acknowledgment section to thank your colleagues for routine contributions.
Digital Object Identifier: 10.1109/LSEN.XXXX.XXXXXXX (inserted by IEEE).

Table 3: *Performance Comparison*

Method	Sensitivity (Recall)			Accuracy (%)	UAR (%)
	Neoplasm %	Phono trauma %	Vocal Palsy %		
Medical Record (SVM)	21.05 %	94.59 %	60.0 %	75.42 %	58.55 %
Acoustic Features (SVM)	57.89 %	91.89 %	36.0 %	74.57 %	61.93 %
One-Stage DNN	53.00 %	88.76 %	64.00 %	77.48 %	68.59 %
Two-stage DNN	79.00%	95.36%	70.40%	87.26 %	81.59%
Stacking Ensemble	**84.21 %**	**97.30 %**	**76.00 %**	**89.83 %**	**85.84 %**

phonating the vowel [a:] for 3-second sustained three common voice disorders including neoplasm, phono trauma and vocal palsy. From this dataset, we use stacking ensemble learning to perform classification of the three classes. The stacking ensemble model is based on multimodality, namely acoustic waveforms and medical records. We first use a stack of SVM models, each performing binary classification. Then, the outputs of these SVMs are then fused by a deep neural network (DNN) to generate the final output. Experimental results show that the stacking method of ensemble learning showed a performance improvement about 2.57% (test accuracy) over the recent Two Stage DNN (TSD) based multiclass classifier. The proposed stacking method also shows significantly improved accuracy, as compared to individual learning algorithms. The proposed approach achieved 89.83% for accuracy and 85.84% for UAR, confirming that the ensemble learning can be utilized in various similar scenarios in the field of biomedical recognition tasks. The remaining part of this paper is organized as follows. Section II presents the proposed method for pathological voice detection. Section III introduce our experiments and results, while Sections IV describes the conclusion and future work.

II. Proposed Method

In this work, we have proposed stacking ensemble method to classify pathological voice disorder by combining acoustic signals and medical records. In this section, we first introduce the acoustic and medical record data. Then, we present the feature extraction methods. Finally, we present the stacking ensemble classifier used in this study.

A. Data Description

In this section, we present the acoustic signal and medical records used in this study.

Pathological voice samples were obtained from a voice clinic in a tertiary teaching hospital (Fast Eastern Memorial Hospital, FEMH, Taiwan), which included 589 samples of three common voice disorders, including glottis neoplasm, phono traumatic disease (i.e. vocal nodules, polyps, cysts), and unilateral vocal paralysis (Tables 1 and 2). Voice samples of a three-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 (CSL model 4150B, Kay Pentax). The sampling rate was 44.100 Hz with a 16-bit resolution and data were saved in .wav format.

The dataset also contained of medical records, including age, gender, jobs, habits and symptoms, when the voice is the worst, how did it happen, whether experienced previous surgery or not, gastroesophageal reflux, voice questionnaire, etc. Each person has 33 dimensions of medical records and produce a matrix of size 1*33.

Table 1: *FEMH Data Description.*

	Number		Total
	♂	♀	
Neoplasm	84	15	99
Phono trauma	97	269	366
Vocal Palsy	76	48	124

Abbreviations: ♂, male; ♀, female.

Table 2: *Phono trauma data description.*

	Phono trauma		
	Nodules	Polyps	Cysts
♂	11	69	17
♀	121	118	30

Abbreviations: ♂, male; ♀, female

B. Feature Extraction

In this study we used the Mel-frequency Cepstral Coefficients (MFCCs) as the acoustic features, which have been widely used by the most speech processing system. MFCCs uses a Mel-scale that is based on human hearing (i.e., humans are more sensitive to low frequencies and can distinguish even small change occurring at low frequencies) [11, 12, 13]. The following step must be performed to derive 13-coefficient MFCCs from acoustic signals: pre-emphasis, windowing, fast Fourier transform, Mel scale filter bank, non-linear transformation, and discrete cosine transform. In this study, the MFCCs frames were extracted from a window length of 16-millisecond and captured 8-millisecond overlap for time shift. Thereafter, we compute six summary statistics (the mean, the median, the maximum, the minimum, the skewness and the standard deviation) over all frames to represent the speech through a fixed length supervector. All the values are finally scaled to have a mean of 0 and a standard deviation of 1 using scikit-learn Standard Scaler [14] for normalization.

For medical record, we encode each item into digit number to simplify the input parameter. For example, binary data (i.e. yes/no) is recorded as 1/0 or 0/1/2/3 (never/occasionally/ weekly/daily) in ordinal data such as tobacco and alcohol consumption, respectively.

In this paper we established model-based combination. A linear combination function is used as the fusion module to linearly combine

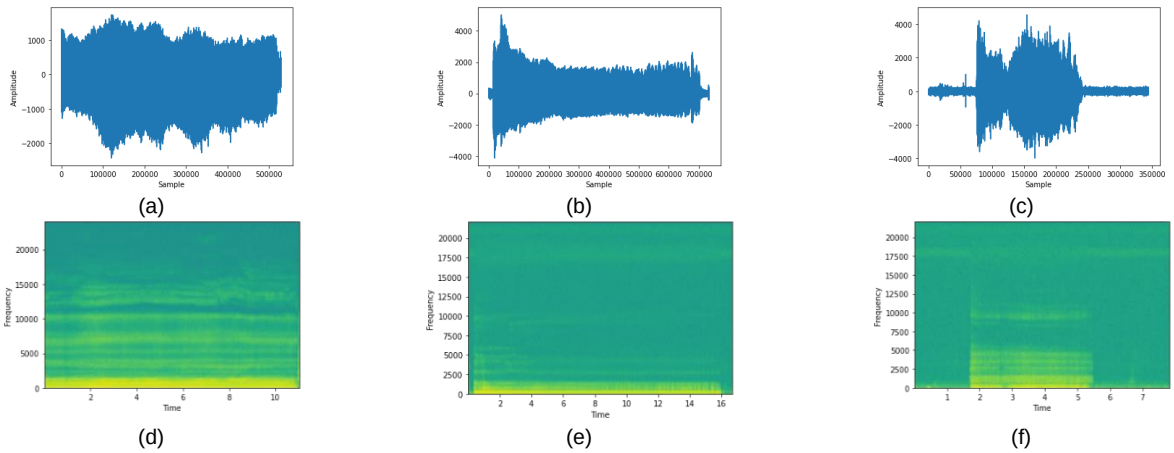

Figure 1: *Waveforms from voice samples of neoplasm (a), phono trauma (b) and vocal palsy (c). Wide band spectrograms in voice samples of neoplasm (d), phono trauma (e), and vocal palsy (f)*

the output of the two SVMs as a base-learner. In this step, acoustic feature and medical records are processed by SVMs individually.

C. Classifier Design

We firstly design and test a binary classifier and a multiclass classifier using the most widely used SVN models. The conventional SVM is used as a baseline for comparison.

To design the stacking ensemble method, we combine the binary classifier (SVM) as a base learner and DNN as a meta learner. The binary classifier is designed to classify acoustic signals and medical records individually as shown in Figure 2.

SVM is one of the most popular machine learning techniques. SVM is known as a soft margin classification and it classifies class by finding the optimal hyperplane distinguish two data. An important consideration in using SVM is the use of kernel trick. In this paper we apply Gaussian radial basis function (RBF), which is the most widely used kernel trick.

Since SVM is a binary classifier, we stack several SVMs to design a multiclass classifier. Typically, there are one-versus-all (OAA) and one-versus-one (OAO) method. In this study we apply OAO approach since it can give better performance. If the number of the given classes is M, in this case, we should consider combination cases of $M \times (M-1)/2$ classifiers. The OAO approach is more complicated.

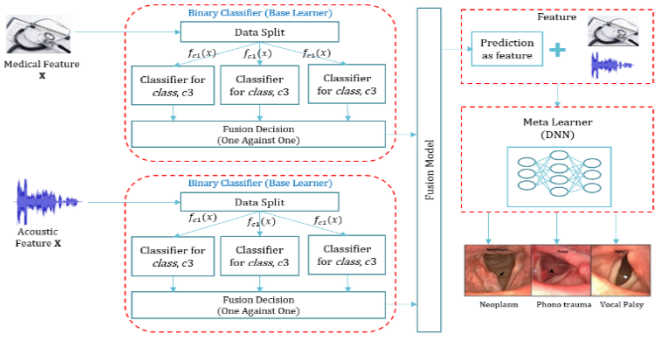

Figure 2: *Ensemble and Multimodal Learning Framework*

DNN is a neural network with multiple hidden layer between the input layer and the output layer. Different results may ne produced when different numbers of layers and the number of nodes in each layer are specified.

A stacking method is a useful tool for combining classifier with different structure. By combining the prediction results obtained from several predictors, we can get better recognition results. This learning method is called ensemble learning. The idea of stacking ensemble learning is that even though each predictor is a weak classifier, a combination of many ensembles may become a strong predictor by combine them by training a meta model to output predictions based on the multiple prediction returned by these weak models.

The training phase in stacking is to train a new prediction model on top of the last layer that aggregates predictions. The new predictor on the last layer called meta-learner. This technique involves the information of linear combinations of different predictors to give overall improved result. It composed of two phases. In the first phase, we use different models such as SVM to learn and perform the first stage prediction. Here, the data partitioning techniques is used to achieve this goal. In the second phase, the whole set of training data is then used to train the meta learning in order to give the final output.

III. Experiments and Results

A. Experimental Setup

In this study, we have performed our empirical experiment using three typical voice disorders including phono traumatic lesions (i.e. vocal nodules, polyps, and cysts), glottic neoplasm, and unilateral vocal paralysis. During the experimental process, we split the training dataset into a train and validation set. We used cross-validation for verification. 75% of total samples were used for the training set, 12.5% were used for the validation set and the remaining 12.5% were used for test set. Through SVM is binary classifier, the system adopted OAO. In this approach each class compared to each other class. We built binary classifier to discriminate between each pair of classes individually for acoustic signals and medical records, while discarding the rest of classes. The labels of data set used to train and test the classifiers should be associated or the same. For example, the output labeling of the binary classifier is “0” or “1”.

First we split the training data into two folds because predictions on data that have been used for the training on the weak learners are not relevant for the training on the meta-learner. We use the predictions of base learner as a feature to train on meta-learner. So, we can produce relevant predictions for each observation of our dataset and then train our meta-learner on all these predictions. Then

we choose weak learners and fit them to data of the first fold to make predictions for observations in the second fold. Afterwards, we fit the meta-learner on the second fold using predictions made by the weak learners as input.To evaluate the performance, we used three performance indexes: overall accuracy (ACC), sensitivity, and Unweighted Average Recall (UAR). These indexes were widely employed in the classification tasks. The accuracy is the ratio of the correctly labeled subjects to the prediction as shown in equation (1).

$$ACC = 100\% \times \frac{TN + TP}{TN + TP + FN + FP} \quad (1)$$

TP, *TN*, *FP*, and *FN* denote true positive, true negative, false positive, false negative, as shown in equation (2, 3, 4). Sensitivity or Recall refers to our model's ability to correctly classify, where neo, pho, pal stand respectively for neoplasm, phono trauma, and vocal palsy; unneo, unpho, and unpal stands for non-neoplasm, non-phono trauma, and on-vocal palsy respectively.

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

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

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

UAR is the average of the recall or sensitivity of each voice disorder class when it is considered as the positive class as shown in equation (5) where *K* denotes the number of classes. In this study $K = 3$.

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

B. Experimental Results

The performance of the proposed stacking ensemble method to classify pathological voice disorder is compared with other reported systems. Table 3 demonstrate the results comparison of some previous studies conducted by combining acoustic signals and medical feature. Results show that the proposed stacking ensemble method outperforms traditional approaches, including the latest two stage DNN. The results show that our stacking ensemble approach provided 89.83% for accuracy, and 85.84% in UAR, demonstrating the best performance as compared with other classification systems.

IV. CONCLUSION

In this study, we have proposed a novel and effective methodology of ensemble learning to classify pathological voice disorder by fusing acoustic signals and medical records. First, we trained an SVM binary classifier for each acoustic signals and medical record to distinguish the three classes. Then, we use the outputs from the base-learner as a new feature and fed it into DNN model as a meta-learner to get the final prediction. Experimental results show that the stacking method of ensemble learning showed a performance improvement about 2.57% (test accuracy) over the recent Two Stage DNN (TSD) based multiclass classifier.

The proposed stacking method also shows significantly improved accuracy, as compared to individual learning algorithms. The proposed approach achieved 89.83% for accuracy and 85.84% for UAR, confirming that the ensemble learning can be utilized in various similar scenarios in the field of biomedical recognition tasks.

REFERENCES

[1] [1] S.-H. Fang., C-T. Wang., J-Y. Chen., Y. Tsao., F-C. Lin., "Combining acoustic signals and medical records to improve pathological voice classification," in *APSIPA Transaction on Signal and Information Processing*, 2019.

[2] J. Zhou and R. N. Miles, "Directional Sound Detection by Sensing Acoustic Flow," in *IEEE Sensors Letters*, vol. 2, no. 2, pp. 1-4, Art no. 1501204, 2018.

[3] G. Baldini, I. Amerini and C. Gentile, "Microphone Identification Using Convolutional Neural Networks," in *IEEE Sensors Letters*, vol. 3, no. 7, pp. 1-4, Art no. 6001504, 2019.

[4] S. R. Schwartz., S. M. Cohen., S. H. Dailey., R. M. Rosenfeld., E. S. Deutsch., M. B. Gillespie., E. Granieri., E. R. Hapner., C. E., Kimball., H. J. Krouse *et al*., "Clinical practice guideline: hoarness (dysphonia)," in *Otolaryngology-Head and Neck Surgery*, vol. 141, pp.1-31, 2009.

[5] Vaziri. G., Almasganj. F., Behroozmand. R., "Pathological assessment of patients speech signals using nonlinear dynamical analysis," in *Computers in Biology and Medicine*, vol.40(1), pp.128-134, 2006.

[6] S. R. Savithri., "Clinical voice evaluation," http://docplayer,.net/53758736-Clinical-voice-evaluation.html, (Date last accessed March 20, 2020)

[7] 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.

[8] R. Herbrich, "Learning Kernel Classifiers: Theory and Algorithm," in *MIT Press*, 2002

[9] E. Allwein., R. Schapire., Y. Singer., "Reducing multiclass to binary: a unifying approach for margin classifiers," in *Journal of Machine Learning Research*, pp. 113-141, 2000.

[10] G. Zweig., "Boosting Gaussian mixture in an LVCSR system," in *Acoustic, Speech and Signal Processing (ICASSP)*, pp. 1527-30, 2000.

[11] J.I. Godino-Liorente., P. Gomez Vilda., and M. Blanco-Velasco., "Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameter," in *IEEE Transactions on Biomedical Engineering*, vol.53, no.10, pp.1943-1953.

[12] Li. H., Kinnuen T., "An overview of text-independent speaker recognition: from features to supervectors," in *Speech Communication*, pp.12-40.

[13] 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," in Journal of Voice, pp.634-641, 2019.

[14] D. Zhang., D. Gatcia-Perez., S. Bengio and I. McCowan., "Semi-supervised adapted HMMs for unusual event detection," in *IEEE Comp Society Conference*, vol.1, pp.611-618, 2005.

[15] Dzeroski. S., Zenko. B., "Is combining classifiers with stacking better than selecting the best one?" in *Machine Learning*, pp.255-273, 2004.