Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Pathological Voice Assessment

Alireza A. Dibazar, Theodore W. Berger, and Shrikanth S. Narayanan,

Abstract — While there are number of guidelines and methods used in practice, there is no standard universally agreed upon system for assessment of pathological voices. Pathological voices are primarily labeled based on the perceptual judgments of specialists, a process that may result in different label(s) being assigned to a given voice sample. This paper focuses on the recognition of five specific pathologies. The main goal is to compare two different classification methods. The first method considers single label classification by assigning a new label (single label) to the ensembles to which they most likely belong. The second method employs all labels originally assigned to the voice samples. Our results show that the pathological voice assessment performance in the second method is improved with respect to the first method.

I. INTRODUCTION

Analysis of voice signal is usually performed by the extraction of acoustic parameters using digital signal processing techniques. These parameters are analyzed to determine the particular characteristic of the voice. In the domain of pathological voice assessment, several methods have been proposed in the literatures many of which relying on the calculation of signal statistics reflecting cycle-to-cycle variation of the time domain voice parameters [1]. Highest, lowest, average, and standard deviation of fundamental frequency are the basic features which have generally been employed. Evaluation of the period to period variability of the pitch period (Jitter) and its statistics have also been utilized in the analysis and assessment of pathological voice by many researchers. Amplitude perturbation [2], voice break analysis [3], subharmonic analysis [4], and noise related analysis [5] have all been investigated for measuring voice quality. As discussed in a number of previous papers on vocal fold pathology [6-7], detection of vocal fold pathology typically considers the excitation signal. Therefore, research in this area has investigated glottal inverse filtering schemes to estimate the source signal from the speech.

A. A. Dibazar is with the Neural Dynamic Laboratory, Biomedical Engineering Department, University of Southern California, DRB 140, Los Angeles, CA 90089-1111, USA (phone: 213-740-9359; fax: 213-821-2368; e-mail: dibazar@usc.edu).

T. W. Berger is the Professor of Biomedical Engineering department; and director of Neural Dynamic Laboratory, University of Southern California, DRB 140, Los Angeles, CA 90089-1111, USA (phone: 213-740-9360; fax: 213-821-2368; e-mail: berger@bmsrs.usc.edu).

S. S. Narayanan is the Professor of Electrical Engineering, Linguistics and Computer Science; director of Speech Analysis & Interpretation Laboratory (SAIL), University of Southern California, 3740 McClilntock Avenue, Room EEB 430 Los Angeles, CA 90089-2564, USA (phone: 213-740-6432; Fax: 213-740-4651; e-mail: shri@sail.usc.edu)

Although these temporal perturbation measures are useful in many circumstances, and measures based on voice perturbation have become widely available through commercial voice analysis systems, however perturbation measures may not be consistently measured. The measurement of these perturbations is limited to variations in either fundamental frequency or peak amplitude of the glottal wave. Moreover, these features present quantities for specific characteristics of the voice signal over long period of the time i.e., static features. Even though static features (long term measures) can be measured more reliably, dynamic features (short-term measures) are informative about acoustic correlates of perceptual dimensions of voice quality which are useful for differential diagnosis.

One of the key properties that make dynamic features useful is that it considers changes in temporal structure of the excitation signal. Static classifiers remove all temporal dependency and therefore dynamic pattern classifiers are needed to handle explicit temporal dependencies in the pathological voices.

We have previously shown that short-term Mel frequency cepstral coefficients – MFCCs – features together with fundamental frequency, both of which are not computationally intensive to measure, can be reliably employed for large scale, rapid assessment of normal and pathological voices [8]. In order to show the effectiveness, consistency, and reliability of the system we focused on the assessment of wide variety of voice pathologies reflecting vocal fold and vocal tract disorders. The system was tested with recordings of the sustained vowel /a/ from a comprehensive database recorded by the Massachusetts Ear and Eye Infirmary [9].

In this paper we extend the preliminary work in [8] in two ways. Here the assessment of five specific pathologies is considered by formulating a multi-class recognition problem. The same above mentioned features and phoneme are used for the assessment of five different pathologies: anterior-posterior (AP) squeezing, hyper-function, ventricular compression, paralysis, and gastric reflux. First, we assume that classes of pathologies are mutually exclusive meaning that a given speech signal token has only one recognition tag attached to. Second, we examine a different scenario, wherein the classes are, by definition, not mutually exclusive. Such a problem arises in multiple pathology recognition where pathological ensembles have more than one labels. The Maximum A Posterior – MAP – estimation is used for recognition of multi-label pathological classes.

The outline of rest of the paper is as follows: in section II, the method of this study is explained. Section III describes experiments and results and conclusion of this work is presented in section IV.

II. METHOD

In our implementation, two requirements were imposed. First, the features had to be efficient in terms of measurement cost and time. Second, both the vocal tract and excitation source information had to be included. The MFCC features were obtained by a standard short-term speech analysis, along with frame-level pitch, to form the feature vectors. Then, set of Hidden Markov Model – HMM – classifiers were applied for the assessment of feature vectors.

A. Features

Twelve Mel frequency cepstral coefficients using 25 msec Hamming window frame were extracted. The employed filters were triangular and equally spaced along the entire Mel-scale.

Period-to-period pitch variation is a classic method of evaluating voice pathologies [10]. However the irregularities of the disordered voice make the pitch extraction algorithms to be inaccurate. The method of pitch extraction which has been employed by multi-dimensional voice program – MDVP – reportedly is reliable in the presence of pathologies [11]. In this study, similar to MDVP, the following method of pitch detection algorithm was used.

1. Autocorrelation-based fixed frame fundamental frequency estimation based on short-term autocorrelation analysis with hard threshold *sgn* function. The signals are low-pass filtered at 1800Hz in order to eliminate higher harmonics of F0.
2. F0 verification: Autocorrelation based adaptive frame for F0 verification (pitch-synchronous) to suppress the influence of sub-harmonic components.
3. Pitch-synchronous Momentum fundamental period (τ_0) extraction made on the original signal
4. Three point linear interpolation

This approach reduces voiced/unvoiced, harmonic/sub harmonic, and other pitch extraction errors in disordered voice signal. Fig. 1 shows an example of the estimated fundamental frequency for healthy and pathological subjects.

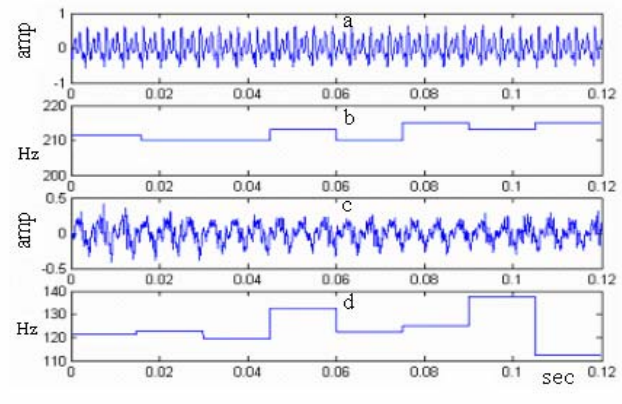

Fig. 1: a) Microphone output signal (time domain) of normal voice (AXH1NAL) b) variation of the pitch; MEAN=212.29 and STD=2.12 c) Microphone output signal (time domain) of pathological voice subject (AMC23AN) and d) variation of the pitch; MEAN=124.64 and STD=7.64 e)

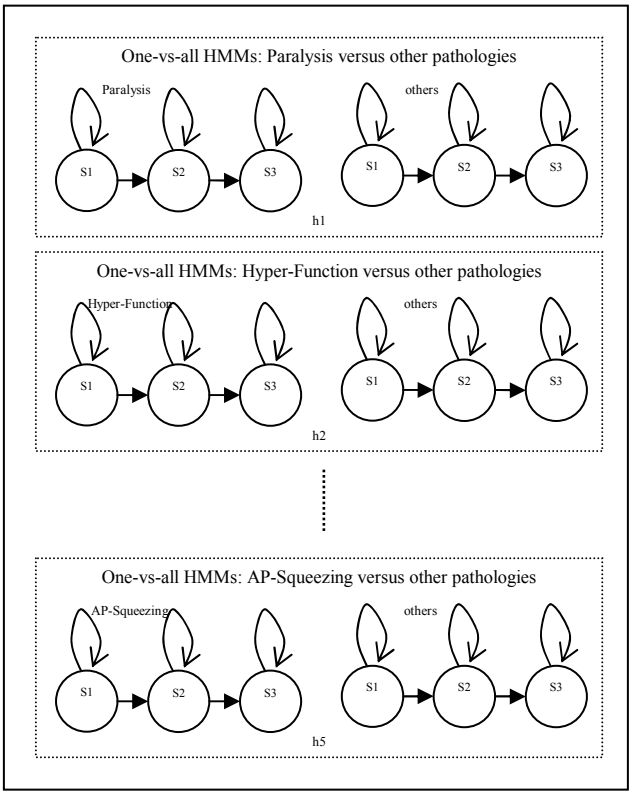

Fig. 2: Schematics of classifier set for multi-label classification

B. HMM Classifier

Let's assume that each of utterances can be represented by a sequence of feature vectors O, which are the MFCCs and associated pitch frequency. Pathological speech recognition then can be regarded as computing:

$$arg\text{MAX}\{P(Pathol_i \mid O)\} \quad (1)$$

where, $Pathol_i$ = *{normal, pathology}*. In practice, if a parametric production model such as the Markov model is assumed, then computing the joint probabilities, which are necessary for solving (1), can be replaced by estimating the Markov model parameters.

The employed HMMs have the following specifications. Three-states, 3-mixtures, left to right HMMs were used based on 14 features (13 MFCCs + fundamental frequency). The EM algorithm was used to train the HMMs. In all of the experiments, the expectation maximization – EM – algorithm iterated seven times for convergence.

For multi-label classification sets of one-vs-all classifiers were used. Fig. 2 shows the schematics of these classifiers. The method of adapting and test of these models will be discussed in section IV.3. In general, the training examples of each class are used more than once; using each example as a positive example of each of the classes to which it belongs. Each sets of classifiers outputs a score which is used for the evaluation.

III. EXPERIMENTS AND RESULTS

In the current paper, the recognition of voice disorders is performed based on vowel /a/. The reason for this has fairly explained in the work by Vieira [12]. In summary, in the production of short vowels, the poor control of respiratory system is not significant; hence, vowels phonated in a sustained fashion with comfortable levels of pitch and loudness are interesting and useful from clinical point of view. The result of this study shows that there is consistency between electro glottal graph – EGG – parameters and acoustic signal features of sustained vowel /a/. According to the report, this is because of the larger and sharper peaks of time domain acoustic signal of /a/ with respect to the other vowels.

The database employed in this study has been developed by Massachusetts Eye and Ear Infirmary Voice and Speech Laboratory (MEEI4337). It contains voice samples of 710 subjects. Included are sustained phonations of speech samples from patients with a wide variety of organic, neuralgic, traumatic, and psychogenic voice disorders, as well as 53 normal subjects. Fig. 3 shows the distribution of different pathologies, based on gender, in the employed database. The mean and standard deviation of age of normal and pathological subjects were 36.00/8.36 and 48.06/20.64 respectively.

There could be four possible strategies for training multi-label pathological classes. The first strategy is very straightforward; the data with multiple labels considered as a new class and build a model for that. One important problem of this method is the number of possible classes in the classification of five pathologies may rise up to 32 classes, rendering the training procedure to be expensive in terms of training time and complexity. With the above mentioned number of classes the database will be spars. The second possible method is labeling the multi-label data with the one class to which the data most likely belonged. This is done by reassigning new labels to the data by a pathological voice specialist. We will examine this case in this paper. The third method would be simply to ignore the multi-label data while training the classifiers. However, this method can not be taken into our consideration because the majority of data has multi-label and dropping them from training makes the remaining data to be too sparse. Therefore the resulting models will be unusable. The last approach is to use the multi-label data more than once while training; using each example as a positive example of each of the classes to which it belongs.

A. Multi-label classifier

We train one-vs-all HMM classifiers for each of the pathological classes (Fig. 2). The training samples of each class are used as a positive example of each of the classes to which it belongs. This means that some training samples may be used more than once. Then, Maximum A Posterior (MAP) classification is performed. The models are built for each base class which avoids sparseness. The test examples are labeled with the class corresponding to the HMM that

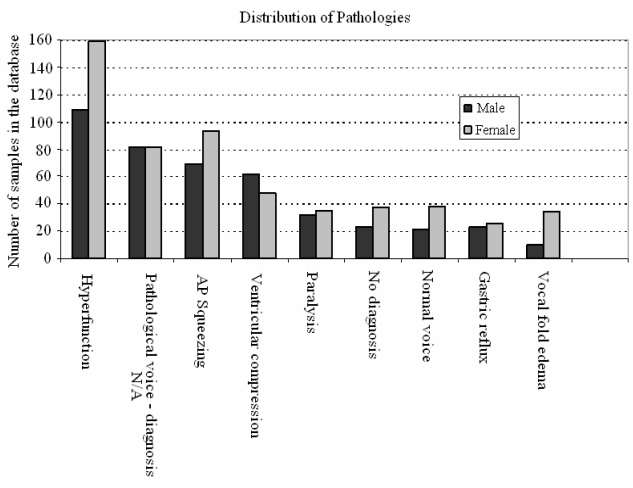

Fig. 3: Distribution of seven major voice pathologies in the database

outputs the maximum scores, even if multiple scores are positive. It is also possible that for some test samples, none of HMM scores are positive. Following is the generalized one-vs-all method for multi-label classification:

Test samples are labeled by all of the classes corresponding to positive scores. If there is no positive score, labeling is done based on the sample with maximum corresponding score. The classification is done by measuring the closeness of top scores (regardless whether they are positive or negative) using MAP. The method can be formulated as follows:

Given a test sample, t there are two HMM scores h_1, and h_2, for corresponding pathological classes of c_1, and c_2. The question is should we label t with only c_1 or c_1 and c_2? This question can be answered by using MAP:

$$C = \arg\max_i P(E_i | dist) \tag{2}$$

$$C = \arg\max_i P(dist | E_i) \cdot P(E_i) \tag{3}$$

where E_i is the event which labels t with the corresponding single class c_i, and $dist$ is the difference between output scores of HMMs, i.e., $h_1 - h_2$ (if $h_1 > h_2$). The probability $P(E_i)$ and joint probability distribution $P(dist | E_i)$ are estimated through training.

B. Classification results

Two sets of experiments were undertaken. First a new label for the multi-label data was assigned with the one class to which the data most likely belonged. This was carefully done by a specialist. Table 1 shows the coincidence of pathological voice in the employed database and Table 2 shows the distribution of data after assigning new labels for the data of Table 1. The HMMs were adapted for each class of pathology and the results of Table 3 were obtained.

Second, the method of section III.B was employed for training and test of multi-label classes. The joint

TABLE 1
COINCIDENCE OF PATHOLOGICAL VOICE IN MEEII DATABASE; THE DIAGONAL ELEMENTS ARE THE NUMBER OF INDIVIDUALS WHICH HAVE ONLY ONE LABEL, THE NON-DIAGONAL ELEMENTS ARE THE COINCIDENCE BETWEEN TWO DIFFERENT CLASSES

	hyper function	A-P squeezing	ventricular compression	paralysis
hyper function	38	152	100	45
A-P squeezing	152	27	94	30
ventricular compression	100	94	31	29
paralysis	45	30	29	25

TABLE 2
DISTRIBUTION OF PATHOLOGIES WITH NEW SINGLE LABELS

Hyper function	205
A-P squeezing	153
ventricular compression	122
paralysis	70
gastric reflux	45

TABLE 3
CORRECT CLASSIFICATION WITH NEW LABELS (SINGLE LABEL): CONFUSION MATRIX

	hyper function %	A-P squeezing %	ventricular compression %	paralysis %	gastric reflux %
hyper function	61.3	18.11	11.25	6.84	2.50
A-P squeezing		63.21	14.54	2.17	1.97
Ventri. comp.			68.75	3.33	2.13
paralysis				75.12	12.54
gastric reflux					65.33

probabilities of $P(dist|E_i)$ were estimated by Gaussian mixtures. The overlap between two classes of gastric reflux and paralysis were dropped because of sparseness of data.

The evaluation of multi-label classification is different than single label. The results of multi-label could be partly correct, fully correct, or fully incorrect. The evaluation method for such a problem was borrowed from studies of Sebastiani and Boutell [13-14]. Suppose that Y_x is the set of actual labels assigned to the samples by specialist and P_x is the set of label predicted by classifiers. Let H_x^c for any $c \in P_x$ and $c \in Y_x$, 0, otherwise. Similarly, let $\hat{Y}_x^c = 1$ for any $c \in Y_x$, 0, otherwise. Moreover, let $\hat{P}_x^c = 1$ if $c \in P_x$, 0, otherwise. Then base-class recall and precision on data set, D, are defined as follows:

$$Score(C) = \frac{\sum_{x \in D} H_x^c}{\sum_{x \in D} \hat{Y}_x^c} \quad (4)$$

$$\Pr ecision(C) = \frac{\sum_{x \in D} H_x^c}{\sum_{x \in D} \hat{P}_x^c} \quad (5)$$

where $Score(C)$ and $\Pr ecision(C)$ are the correct classification rate and predicted correct classification rate for class c respectively. The results of multi-label classification for the above mentioned pathologies are shown in Table 4. This table is based on equation 4 for base classes. The results show that for all base classes the performance has been improved with respect to the Table 3. A-P squeezing with %6.36 and paralysis with %1.03 has maximum and minimum improvements respectively. The average overall performance of Table 4 is 6.35% better than average overall performance of the results of Table 3. We hypothesis the above mentioned training method makes the pathological models to be rich with respect to the single label models.

Table 5 contains the classification scores and precisions of single label and multi-label classification. Based on the results of this table the precession of average multi-label is slightly higher than precession of average single label classification.

TABLE 4
CORRECT CLASSIFICATION WITH MULTIPLE LABELS: CONFUSION MATRIX

	hyper function %	A-P squeezing %	ventricular compression %	paralysis %	gastric reflux %
hyper function	66.42	14.05	10.58	6.23	2.72
A-P squeezing		69.57	10.41	2.85	3.12
Ventri. comp.			72.46	3.00	3.55
paralysis				76.09	11.83
gastric reflux					70.83

IV. CONCLUSION

In this paper the use of standard speech features which are easy to measure and that can be robustly used for assessment of normal versus pathological voice was considered. HMM classifiers applied for short-term features. The results indicated that short term features with dynamic classifiers can be used in the classification of pathological voices.

TABLE 5
PRECISION OF BASE CLASSES

	single label		multi-label	
	Score %	Precision %	Score %	Precision %
hyper function	61.30	85.11	66.42	86.35
A-P squeezing	63.21	80.30	69.57	79.98
ventricular compression	68.75	79.02	72.46	79.91
paralysis	75.12	79.33	76.09	76.85
gastric reflux	65.33	80.84	70.83	84.15

In addition, the classification of five specific pathological voices was explored. Two scenarios were considered. First it was assumed that pathological classes are mutually exclusive. Therefore a new label was assigned to the samples with the one to which the sample most likely belongs. Second, we presented a method for assessment of a practical issue in the domain of biomedical signal processing. Recall that sometimes clinical vital signals have more than one recognition tag which makes standard pattern classification methods to fail. To overcome this difficulty, the modified one-vs-all HMM classifiers were employed. The training samples with multiple labels were used as a positive sample of each class. Labels were assigned to the test samples based on their closeness to the base class. New joint probabilities were introduced to measure the closeness, based on MAP estimation. The classification scores and precession of Table 5 justified that using samples with multi-labels as a positive sample of each class, is more efficient which makes the models to be richer therefore enhances the average overall performance.

In this article we demonstrated assessment of five pathological voices however sparseness of data prohibits the extension of the work for other pathologies. It has been planned for future work to investigate the performance of presented method in the assessment of other pathologies. Moreover, the generalization of the proposed technique in the recognition of pathologies from continuous speech will be focused in the future work. Preliminary researches indicated that this approach could be used with continuous speech, such as telephone conversations, while maintaining the performance.

ACKNOWLEDGMENT

The authors would like to thank Dr. Eilnaz A. Azari, for her help for reassigning new labels for pathological voices (single label) and Dr. Sageev George for his programming help. We also would like to express our appreciation for all members of Neural Dynamic Laboratory and Speech Analysis & Interpretation Laboratory (SAIL) for their invaluable collaborations.

REFERENCES

[1] R. J. Baken, and R. Orlikoff, Clinical measurement of speech and voice, 2nd Edition, Singular Publishing Group (2000).

[2] S. Bielamowicz, J. Kreiman, B. R. Gerratt, M. S. Dauer, G. S. Berke, “Comparison of voice analysis systems for perturbation measurement,” Journal of Speech and Hearing Research, 39, 254–266 (1996).

[3] L. Eskenazi, D. G. Childers, D. M. Hicks,“Acoustic correlates of vocal quality,” Journal of Speech and Hearing Research, 33, 298–306 (1990).

[4] D. Michaelis, M. Fröhlich, H. W. Strube, “Selection and combination of acoustic features for the description of pathologic voices,” Journal of Acoustical Society of America, 103, 1628–1638 (1998).

[5] V. Parsa, D. G. Jamieson, “Identification of pathological voices using glottal noise measures,” Journal of Speech, Language, and Hearing Research, Vol. 43, 469–485 (2000).

[6] D. E. Veeneman, S. L. BeMent, “Automatic glottal inverse filtering from electroglottographic signals,” IEEE trans. on Acoustic Speech and Signal Processing, Vol. ASSP 33, no. 2, pp. 369-377 (1985).

[7] D. Y. Wong, J. D. Markel, A. H. Gray, “Least square glottal inverse filtering from the acoustic speech waveform,” IEEE trans. on Acoustic Speech and Signal Processing, Vol. ASSP 27, no. 4, 350-355 (1979).

[8] A. A. Dibazar, S. Narayanan, T. W. Berger, “Feature analysis for automatic detection of pathological speech,” Proc. of IEEE EMBS meeting (2002).

[9] “Disorder Dtabase Model 4337,” Massachusetts Eye and Ear Infirmary Voice and Speech Lab, Boston, MA, (2002).

[10] S. Iwata, “Periodicities of pitch perturbation in normal and pathologic larynges,” Laryngoscope, vol. 82, pp. 87-96, (1972).

[11] R. D. Kent, H. K. Vorperian, J. R. Duffy, “Reliability of the multi-dimensional voice program for the analysis of samples of subjects with dysarthria,” American journal of speech-language pathology, vol. 8, pp.129-136 (1999).

[12] M. N. Vieira, F. R. Mclnnes, M. A. Jack, “On the influence of laryngeal pathologies on acoustic and electroglottographic jitter measures,” JASA vol. 111 (2002).

[13] F. Sebastiani, “Machine learning in automated text categorization,” ACM Compu. Surveys 34 (1), pp 1–47 (2002).

[14] M. R. Boutell, J. Luo, X. Shen, C. M. Brown,“Multi-label scene classification,” Pattern Recognition 37, pp 1757-1771 (2004).