A PARAMETRIC APPROACH FOR CLASSIFICATION OF DISTORTIONS IN PATHOLOGICAL VOICES

Amir Hossein Poorjam [1] *, Max A. Little* [2,3]*, Jesper Rindom Jensen* [1] *and Mads Græsbøll Christensen* [1]

[1] Audio Analysis Lab, CREATE, Aalborg University, Aalborg, DK
[2] Engineering and Applied Science, Aston University, Birmingham, UK
[3] Media Lab, MIT, Cambridge, Massachusetts, USA
[1] {ahp, jrj, mgc}@create.aau.dk , [2,3] max.little@aston.ac.uk

ABSTRACT

In biomedical acoustics, distortion in voice signals, commonly present during acquisition and transmission, adversely affects acoustic features extracted from pathological voice. Information on the type of distortion can help in compensating for its effects. This paper proposes a new approach to detecting four major types of commonly encountered distortion in remote analysis of pathological voice, namely background noise, reverberation, clipping and coding. In this approach, by applying factor analysis to Gaussian mixture model mean supervectors, distortions in variable-duration recordings are modeled by fixed-length, low-dimensional channel vectors. Then, linear discriminant analysis (LDA) is used to remove the remaining nuisance effects in the channel vectors. Finally, two different classifiers, namely support vector machines and probabilistic LDA classify the different types of distortion. Experimental results obtained using Parkinson's voices, as an example of pathological voice, show 11.4% relative improvement in performance over systems which directly use acoustic features for distortion classification.

Index Terms— Distortion modeling, channel factors, PLDA, SVM, remote pathological voice analysis.

1. INTRODUCTION

Smartphones, as ubiquitous and inexpensive devices with built-in, high-quality microphones, are recently being considered as tools for remote pathological voice analysis [1, 2]. Compared to voice samples recorded in a sound booth or other acoustically-controlled conditions, recordings from smartphones in everyday environments are subject to many types of linear and nonlinear distortion. The presence of distortion in signals affects acoustic features used for subsequent biomedical applications, which not only diminishes performance of voice pathologists in diagnosing voice disorders, but also degrades the performance of algorithms designed to quantify medical symptoms from the voice [3]. Thus, signal enhancement, or selection of good quality segments of voice recordings seems an essential pre-processing step in remote voice analysis. Information on the type of distortion corrupting a signal can be used to inform the choice of appropriate enhancement algorithms.

Several approaches to detect different types of distortion in voice signals have been proposed, most of them focused on detecting a single and specific type of distortion [4–8]. In [9], we proposed a method to classify four major types of distortion in vowels, namely background noise, reverberation, clipping and coding, directly from mel-frequency cepstral coefficients (MFCCs) extracted from short time frames of speech signals. This approach was motivated by the analysis of MFCC behavior under different distortion conditions. We showed that different types and levels of distortion predictably modify the distribution of MFCCs. This method, however, has two limitations. First, MFCCs are sensitive to any change in signal characteristics. This means that they encode not only distortion in signals, which is beneficial for distortion classification, but also other variability such as speaker, articulation and disorder resulting in conflation of distortion with medical disorder. Second, while the classification is performed at the frame-level, the distortion classification decision is made by majority vote over all frames of a signal, and the computation time increases with increasing signal length.

In this paper, we address these issues by modeling distortion in variable duration recordings with fixed-length, low-dimensional vectors that focus mostly on the type of distortion in signals. Since distortion modifies the signal's characteristics, we consider distortions as a source of channel variability in signals. Now, if we corrupt clean pathological voices with various types and levels of distortion and fit a Gaussian mixture model (GMM) to the acoustic features extracted from each of the recordings, the GMM means convey information about speakers, distortions and disorders. Then, assuming that the GMM means can be decomposed into two components, namely a speaker-dependent component (including information about the speaker), and a channel-dependent component (containing information about channel such as distortion, articulation and disorder), by applying a factor analysis (FA) technique to the GMM means and estimating the channel factors, information about the channel effects can be represented by fixed-length, low-dimensional vectors. Finally, applying linear discriminant analysis projection removes other nuisance factors in the channel vectors and makes them more suitable for distortion classification.

This study focuses on classification of distortions in sustained vowels. Although running speech is known to convey information about the speaker's characteristics [10–14], sustained vowels are the most widely-used signals for pathological voice analysis [15] for two main reasons: first, sustained vowels highlight voice disorders since most dysphonic speakers cannot generate steady, sustained vowel sounds [16], and second, the confounding complexities of articulatory movement during running speech are largely avoided [17]. While there are an infinite number of possible levels, types and combinations of distortions in real scenarios, this study focuses on four major types of distortion commonly present during acquisition or transmission in remote voice analysis, aiming at a simplified approach to detecting the most dominant distortion in voice signals. This would be useful in practical applications where it is important to se-

This work was funded by Independent Research Fund Denmark: DFF 4184-00056

lect an appropriate algorithm to enhance a distorted signal before being inspected by a voice pathologist or being used as an input for algorithms for biomedical acoustics applications.

2. SYSTEM DESCRIPTION

2.1. Problem Formulation

In this distortion classification problem, we are given a set of training data, $\mathcal{S}^{\text{tr}} = \{\boldsymbol{\nu}_j, d_j\}_{j=1}^{J}$, where $\boldsymbol{\nu}_j$ denotes the j^{th} recording and d_j denotes the distortion that has corrupted the signal. The goal is to estimate a classifier function so that for a signal not in the training data, the probability of the estimated output being classified to the correct class is maximized. For contrast, we consider two approaches to solve this problem: purely nonparametric, in which the function is directly approximated from acoustic features [9], and parametric approaches, in which variable-duration signals are first converted to fixed-length vectors by estimating the parameters of a statistical model for acoustic features. Although nonparametric techniques are conceptually simple compared to parametric methods, the computation time can dramatically increase if the signal duration increases.

2.2. Distortion Modeling

The first step in the approach converts variable-duration signals into fixed-dimensional vectors suitable for classification. One possible approach is to fit a GMM to acoustic features calculated from frames of a distorted signal such that the distortion in a signal can be characterized by the parameters of the fitted GMM. A supervector constructed by concatenation of the resulting GMM means can represent a single voice recording. However, due to lack of data, fitting a separate GMM to a short recording cannot be performed reliably, particularly in the case of GMMs with a high number of mixtures. To overcome this problem, instead of fitting a separate GMM to each recording, we first fit a GMM to the acoustic features of a large amount of clean and distorted data. We call this GMM a universal background model (UBM). Then, parametric techniques using *maximum-a-posteriori* adapt the UBM to the characteristics of the recordings. Consider a UBM with the following likelihood function:

$$p(\boldsymbol{\rho}_l|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{c=1}^{C} \pi_c p(\boldsymbol{\rho}_l|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \tag{1}$$

where $\boldsymbol{\rho}_l$ is the acoustic vector of dimension F at frame l, π_c is the mixture weight for the c^{th} mixture component, $p(\boldsymbol{\rho}_l|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ is a Gaussian probability density function with mean $\boldsymbol{\mu}_c$ and covariance matrix $\boldsymbol{\Sigma}_c$, and C is the number of mixture components. The parameters of the UBM are estimated on a large amount of clean and distorted training data. Then, the GMM mean supervector of dimension $CF \times 1$ is constructed for each recording by concatenation of the adapted Gaussian means for that recording.

However, these supervectors are of a high dimensionality resulting in high computational cost, and it is difficult to obtain a reliable model for classification when we have limited data. Moreover, besides distortion-specific characteristics, other nuisance factors such as speaker, disorder and articulation variability confound this adaptation process. To tackle this problem, we try to model different sources of variability by applying a FA technique to the GMM mean supervectors. In this case, if we corrupt the clean recordings of a given speaker by different types and levels of distortion, we can artificially produce channel variability in recordings of that speaker due to distortions and then fit a GMM to each recording of the speaker. Now, we can assume that the GMM mean supervector of the r^{th} recording from the s^{th} speaker can be decomposed as [18]:

$$\boldsymbol{M}_{s,r} = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{U}\boldsymbol{x}_{s,r} + \boldsymbol{D}\boldsymbol{z}_s \tag{2}$$

where $\boldsymbol{m}$ is speaker- and channel-independent supervector obtained by concatenation of UBM means, $\boldsymbol{V}$ (the rectangular matrix of low rank, the "eigenvoice" matrix) defines a speaker subspace of dimension $CF \times R_s$, vector $\boldsymbol{y}_s$ is the speaker factors, $\boldsymbol{U}$ (the rectangular matrix of low rank called the eigenchannel matrix) defines a subspace of dimension $CF \times R_c$ with high channel variability, vectors $\boldsymbol{x}_{s,r}$ are the channel factors which contain channel related information, $\boldsymbol{D}$ is a diagonal matrix of dimension $CF \times CF$ describing any remaining speaker variability not modelled by $\boldsymbol{V}$, and vector $\boldsymbol{z}_s$ contains speaker-specific residual factors. The factors $\boldsymbol{x}_{s,r}$, $\boldsymbol{y}_s$ and $\boldsymbol{z}_s$ are assumed to be independent of each other and have a standard normal prior distribution. To estimate the matrices $\boldsymbol{V}$, $\boldsymbol{U}$ and $\boldsymbol{D}$, we first train $\boldsymbol{V}$, assuming that $\boldsymbol{U}$ and $\boldsymbol{D}$ are zero. Next, given the estimate of $\boldsymbol{V}$ and assuming that $\boldsymbol{D}$ is zero, we estimate the "eigenchannel" matrix $\boldsymbol{U}$. Then, given the estimates of $\boldsymbol{V}$ and $\boldsymbol{U}$, we estimate the residual matrix $\boldsymbol{D}$. Finally, using the estimate of these matrices, the speaker, channel and residual factors are calculated.

In this study, the FA framework is considered as a feature extractor to estimate the channel information in the channel factor, $\boldsymbol{x}_{s,r}$. Therefore, we detail the process of estimating $\boldsymbol{U}$ and $\boldsymbol{x}_{s,r}$ in the next subsection. An efficient procedure for estimating all the other parameters of the FA model can be found in [18].

Estimation of the Channel Factor and the Channel Subspace

The channel factor $\boldsymbol{x}_{s,r}$ is a latent variable containing the channel information which is defined by the expected value of the posterior distribution conditioned to the Baum-Welch statistics (calculated using the UBM) for a given recording. Given a sequence of L acoustic frame vectors $\{\boldsymbol{\rho}_1, \cdots, \boldsymbol{\rho}_l, \cdots, \boldsymbol{\rho}_L\}$, the zero- and first-order statistics for each speaker s, recording r and mixture component c with respect to the UBM are calculated respectively as:

$$N_{s,r,c} = \sum_{l=1}^{L} \gamma_{c,l} \tag{3}$$

$$\mathbf{f}_{s,r,c} = \sum_{l=1}^{L} \gamma_{c,l} \big[\boldsymbol{\rho}_l - (\boldsymbol{m}_c + \boldsymbol{V}_c \boldsymbol{y}_s)\big] \tag{4}$$

where $\gamma_{c,l}$ is the posterior probability of the c^{th} mixture generating the feature vector $\boldsymbol{\rho}_l$, $\boldsymbol{m}_c$ and $\boldsymbol{V}_c$ are respectively the subvector of $\boldsymbol{m}$ and the submatrix of $\boldsymbol{V}$ associated with mixture component c. In (4), the speaker shift, $\boldsymbol{m}_c + \boldsymbol{V}_c\boldsymbol{y}_s$, is weighted and subtracted from the first order statistics to remove the speaker effects.

To train $\boldsymbol{U}$ and $\boldsymbol{x}_{s,r}$, we apply an EM algorithm. In the E-step, using a random initialization of $\boldsymbol{U}$, we first set

$$\boldsymbol{L}_s = \boldsymbol{I} + \boldsymbol{U}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{N}_s \boldsymbol{U}, \tag{5}$$

where ^{T} represents matrix transpose, $\boldsymbol{I}$ is an $F \times F$ identity matrix, $\boldsymbol{\Sigma}$ and $\boldsymbol{N}_s$ are $CF \times CF$ block-diagonal matrices with $\boldsymbol{\Sigma}_c$'s and $(\sum_r N_{s,r,c})\boldsymbol{I}$ as their entries, respectively. Assuming a Gaussian prior distribution, the posterior distribution of the channel factor is also Gaussian $\boldsymbol{x}_{s,r} \sim \mathcal{N}(\boldsymbol{\mu}_{s,r}, \boldsymbol{\Lambda}_{s,r})$ [19]. Let $\mathbf{f}_{s,r}$ be the $CF \times 1$ vector obtained by concatenating $\mathbf{f}_{s,r,c}$. The means and covariances of this distribution for each recording are respectively calculated as:

$$\boldsymbol{\mu}_{s,r} = E[\boldsymbol{x}_{s,r}] = \boldsymbol{L}_s^{-1} \boldsymbol{U}^T \boldsymbol{\Sigma}^{-1} \mathbf{f}_{s,r}, \tag{6}$$

$$\boldsymbol{\Lambda}_{s,r} = E[\boldsymbol{x}_{s,r}\boldsymbol{x}_{s,r}^T] = \boldsymbol{\mu}_{s,r}\boldsymbol{\mu}_{s,r}^T + \boldsymbol{L}_s^{-1}. \tag{7}$$

In the M-step, we set

$$\boldsymbol{\Theta}_c = \sum_s \sum_r N_{s,r,c} \boldsymbol{\Lambda}_{s,r} \quad , \quad c = 1, \cdots, C, \tag{8}$$

$$\boldsymbol{\Psi} = \sum_s \sum_r \mathbf{f}_{s,r} \boldsymbol{\mu}_{s,r}^T, \tag{9}$$

where $\boldsymbol{\Theta}_c$ is a matrix of dimension $R_c \times R_c$ and the dimension of $\boldsymbol{\Psi}$ is $CF \times R_c$. Then, for each mixture component $c = 1, \cdots, C$ and for each $f = 1, \cdots, F$, set $i = (c-1)F + f$, $\boldsymbol{U}$ is updated by solving the equations

$$\boldsymbol{U}_i \boldsymbol{\Theta}_c = \boldsymbol{\Psi}_i, \tag{10}$$

where $\boldsymbol{U}_i$ and $\boldsymbol{\Psi}_i$ are respectively the i^{th} row of $\boldsymbol{U}$ and $\boldsymbol{\Psi}$. The EM algorithm typically converges after tens of iterations [18].

The channel subspace $\boldsymbol{U}$, which contains information about the channel characteristics, is estimated over a large training dataset and is used to extract the posterior mean of the channel factors, $\boldsymbol{\mu}_{s,r}$, for each utterance in the training and test subsets using (6).

2.3. Channel Vector Pre-processing

Although the channel vectors are expected to model only channel effects, particularly the distortions in signals, they contain other variability [20] such as speaker and disorder effects. Therefore, the remaining nuisance effects in channel space should be removed before being passed to the classifier. Our pre-processing steps include centering followed by linear discriminant analysis (LDA) projection. LDA projection is a powerful transformation technique to reduce the dimensionality of multidimensional observations by finding new orthogonal axes that minimize the within-class variance and maximize the between-class variance [21]. In this case, each class includes all the recordings corrupted by a specific type of distortion, and the within-class variance is due to other factors than distortion such as speaker or disorder effects. This supervised transformation technique will enhance the class separability and reduce the dimensionality of the channel vectors at the same time. Centering channel vectors around the global mean of all training vectors, $\bar{\boldsymbol{\mu}}$, followed by applying LDA results in a linearly transformed channel vector as:

$$\boldsymbol{\phi}_{s,r} = \boldsymbol{B}(\boldsymbol{\mu}_{s,r} - \bar{\boldsymbol{\mu}}) \tag{11}$$

where $\boldsymbol{B}$ [21] is the LDA projection matrix estimated from the training data, and then used to transform the test channel vectors.

2.4. Classifiers

The proposed distortion modeling approach is evaluated using two different classifiers, namely the support vector machine (SVM) and probabilistic linear discriminant analysis (PLDA). Introduced by Vapnik and Cortes [22], the SVM is a discriminative classifier which attempts to find the maximum margin separation hyper-plane between two classes of data such that it generalizes well to the test data. Although SVM is a binary classifier, an effective multi-class extension based on a pairwise coupling strategy has also been developed [23]. In this study, we used the LIBSVM toolbox [24] to implement a multiclass SVM with radial basis function kernel.

PLDA [25], originally studied in image processing, is a generative classifier. In the PLDA framework, the channel vector generation process is described in terms of latent variables, specifically a distortion-dependent component and a residual component, by applying a FA. Then, given two channel vectors $\boldsymbol{\phi}_{\text{e}}$ and $\boldsymbol{\phi}_{\text{t}}$, the training and the test channel vectors, the verification score in the PLDA framework is computed as:

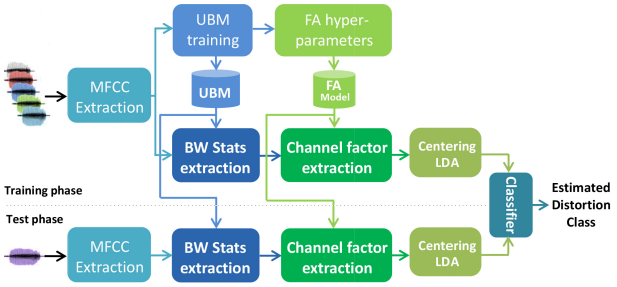

Fig. 1: The block diagram of the proposed distortion classification system in training and testing phases.

$$s^{\text{PLDA}} = \frac{P(\boldsymbol{\phi}_{\text{e}}, \boldsymbol{\phi}_{\text{t}} | H_{\text{s}})}{P(\boldsymbol{\phi}_{\text{e}}, \boldsymbol{\phi}_{\text{t}} | H_{\text{d}})} \tag{12}$$

where H_{s} is the same-distortion hypothesis and implies that both channel vectors, $\boldsymbol{\phi}_{\text{e}}$ and $\boldsymbol{\phi}_{\text{t}}$, originate from the same distortion class, and H_{d} is the different-distortion hypothesis, indicating that channel vectors originate from different distortion classes. Given the Gaussian assumption, a closed form solution for (12) is provided in [26].

2.5. Training and Testing

A block diagram of the proposed system is shown in Fig.1. In the training phase, voice recordings in the training set are mapped to the channel vectors and along with their corresponding distortion labels are used to train the classifier. During the testing phase, the same distortion modeling approach is used to extract a channel vector from a test recording, and the distortion class is predicted using the trained classifier.

3. EXPERIMENTAL SETUP

3.1. Database

The proposed system has been developed and validated using a PD voice database since the vast majority of people with PD exhibit some form of vocal disorder [27]. The database was generated through collaboration between Sage Bionetworks, PatientsLikeMe and Dr. Max Little as part of the Patient Voice Analysis[1] study. The samples are telephone recordings of the sustained vowels /a/ uttered by 750 patients of both genders, sampled at 8 kHz and range from 3 s to 30 s long.

The voice recordings in the database are divided into non-overlapping training and test subsets consisting of 80% and 20% of the speakers, respectively. To create a database for distortion classification, we distorted all recordings by different types and levels of distortion, typically present in the recordings of remote voice analysis, and added them to the database. Specifically, for noise, we used "babble", "white Gaussian" and "office ambiance" noises at 15 dB, 10 dB and 5 dB. For peak clipping, the clipping level was set to 0.3, 0.4, 0.5 and 0.6. Signals were coded using 6.3 kbps, 9.6 kbps and 16 kbps CELP codecs. To provide reverberant signals, recordings were filtered by 8 different real room impulse responses of the AIR database measured with mock-up phone in hand-held and hands-free positions in four realistic indoor environments, namely an office, a lecture room, a corridor and a stairway [28]. Therefore,

[1] Obtained through Synapse ID [syn2321745].

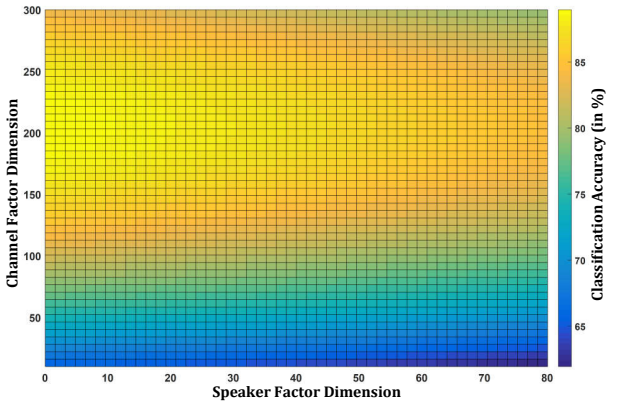

Fig. 2: The classification accuracy of different configuration of the FA model (using GMMs with 256 mixture components).

the extended training and test subsets consist of 3000 and 750 recordings, respectively, which have the same number of recordings per class of distortion.

3.2. Acoustic Features and Distortion Modeling

The proposed approach operates on cepstral features, extracted using a 30 ms Hamming window. For each frame of a signal, 12 MFCCs together with the log energy were calculated along with *delta* and *double-delta* coefficients. They were concatenated to produce a 39-dimensional acoustic feature vector. The variable-duration feature vector sequences are then converted into fixed-dimensional channel vectors using the FA model trained on features from all training utterances. To have a channel space with high distortion variability, it is necessary to have different recordings from the same speakers with different types and levels of distortion. Since it is difficult to collect speech samples from patients under different distortion conditions, we corrupt clean recordings of the speakers by different types and levels of distortion to produce channel variability due to distortions. The number of mixture components and the dimensions of speaker and channel factors were selected using an exhaustive grid search with 5-fold cross-validation (CV) on training data. Fig. 2 shows the classification accuracy as a function of speaker and channel subspace dimensions based on GMMs with 256 mixtures. The plot suggests that the best performance is obtained with a configuration of zero eigenvoice and 210 eigenchannels. Although similar trends have been observed for different numbers of mixture components, using 256 mixtures results in better performance. Setting speaker subspace to zero simplifies (2) to $\boldsymbol{M}_{s,r} = \boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}_{s,r} + \boldsymbol{D}\boldsymbol{z}_s$. Moreover, based on (4), instead of the Gaussian posterior-weighted speaker shift, a speaker-independent shift $\boldsymbol{m}$ is subtracted from the first order statistics for all recordings. Prior to classification, the channel vectors are centered and projected into the LDA subspace. Given 5 classes, the dimensionality of the LDA projected channel vectors is 4.

4. RESULTS

In this study, the MFCC-SVM-based distortion classification system in [9] is considered as the experimental performance baseline. We used 5-fold CV to evaluate the performance of different systems in terms of the number of correctly classified test recordings. The results of the baseline system and the proposed system before and after channel vector pre-processing over all CV repetitions are presented in Table 1 in the form of mean $\pm$ standard deviation (STD). The reported numbers for different classes are the diagonal elements of the confusion matrix for each system and the last column reports the overall classification accuracy.

Table 1: Comparison of the baseline system and the proposed method before and after pre-processing channel vectors using LDA. Results are in the form of mean $\pm$ STD computed using 5-fold CV.

System	Clean	Noisy	Rever.	Clipped	Coded	Overall
Baseline	55±11	97±4	77±4	82±7	85±9	79±3
PLDA	100±0	0±0	0±0	0±0	0±0	20±0
PLDA-LDA	77±4	98±2	86±4	82±2	93±3	87±1
SVM	28±18	33±5	31±16	35±14	68±12	39±4
SVM+LDA	78±3	97±2	87±4	85±2	93±3	88±1

We can observe from these results that although the system is not efficient when unprocessed channel vectors are directly used for classification, applying LDA boosts the classification accuracy by removing the nuisance directions from channel vectors and consequently increasing the class separability in a dimensionality-reduced channel space. The results also show that comparable classification performance can be achieved with either generative or discriminative classifiers on the pre-processed channel vectors. The 11.4% relative improvement in classification accuracy compared to the baseline system shows the effectiveness of the proposed distortion modeling for pathological voices. The lower performance for detecting clean recordings, however, might be due to the fact that the recordings in the PD voice database have already some types of distortion such as noise and reverberation or may have been through one or more codecs since they are collected over the telephone network. This means that some distorted recordings have been presented to the model as "clean" ones during the training phase. We expect better results if an entirely clean pathological database is used. The very good performance in detecting noisy signals is due to the fact that the distribution of MFCCs, from which the channel vectors are extracted, is more affected by additive noise than other types of distortion [9].

The proposed system has two major advantages over the baseline system. First, distortion in variable duration signals is modeled by a fixed-length, low-dimensional vector which is more suitable for classification algorithms. Second, since channel vectors are more robust to small changes in signal characteristics, due to articulatory movements or dysphonias than raw MFCCs, they are more suitable for distortion classification in pathological voices.

5. CONCLUSION

In this study, a new method for classification of four major types of distortion in pathological voices commonly present during acquisition or transmission, namely background noise, reverberation, clipping and coding, has been proposed. This method suggests a new low-dimensional representation of distortion in recordings by applying factor analysis to GMM mean supervectors. We showed that the extracted channel vectors include other variability which can be reduced substantially by applying an LDA technique. Then, SVM and PLDA classifiers were employed to classify the type of distortion in signals. The experimental results over 3750 clean and distorted Parkinson's voices, as an example of pathological voices, show that we can reach 88% overall classification accuracy and improve the performance of the baseline MFCC-SVM-based system by 11.4% which confirms the effectiveness of the proposed method in distortion modeling and classification in pathological voice analysis applications.

6. REFERENCES

[1] S. Arora, V. Venkataraman, A. Zhan, S. Donohue, K. Biglan, E. Dorsey, and M. Little, "Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study," *Parkinsonism Relat. Disord.*, vol. 21, no. 6, pp. 650–653, 2015.

[2] R. J. Moran, R. B. Reilly, P. De Chazal, and P. D. Lacy, "Telephony-based voice pathology assessment using automated speech analysis," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 468–477, 2006.

[3] J. Vasquez-Correa, J. Serra, J. F. Orozco-Arroyave, J.R. Vargas-Bonilla, and E. Noth, "Effect of acoustic conditions on algorithms to detect Parkinson's disease from speech," in *ICASSP*, 2017, pp. 5065–5069.

[4] W. Yuan and B. Xia, "A speech enhancement approach based on noise classification," *Appl. Acoust.*, vol. 96, pp. 11–19, 2015.

[5] K. El-maleh, A. Samouelian, and P. Kabal, "Frame-level noise classification in mobile environments," *ICASSP*, pp. 237–240, 1999.

[6] S. Aleinik and Y. Matveev, "Detection of clipped fragments in speech signals," *Int. J. Electr. Comput. Energ. Electron. Commun. Eng.*, vol. 8, no. 2, pp. 286–292, 2014.

[7] J. Eaton and P. A. Naylor, "Noise-robust detection of peak-clipping in decoded speech," *ICASSP*, pp. 7019–7023, 2014.

[8] J. M. Desmond, L. M. Collins, and C. S. Throckmorton, "Using channel-specific statistical models to detect reverberation in cochlear implant stimuli." *J. Acoust. Soc. Am.*, vol. 134, no. 2, pp. 1112–20, 2013.

[9] A. H. Poorjam, J. R. Jensen, M. A. Little, and M. G. Christensen, "Dominant distortion classification for pre-processing of vowels in remote biomedical voice analysis," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 289–293.

[10] A. H. Poorjam, M. H. Bahari, and H. Van hamme, "A novel approach to speaker weight estimation using a fusion of the i-vector and NFA frameworks," *J. Electr. Syst. Signals*, vol. 3, no. 1, pp. 47–55, 2017.

[11] A. H. Poorjam, M. H. Bahari, V. Vasilakakis, and H. Van hamme, "Height estimation from speech signals using i-vectors and least-squares support vector regression," in *ICTSP*, 2015, pp. 1–5.

[12] A. H. Poorjam, S. Hesaraki, S. Safavi, H. van Hamme, and M. H. Bahari, "Automatic smoker detection from telephone speech signals," in *SPECOM*. Lecture Notes in Computer Science, Vol. 10458, 2017, pp. 200–210.

[13] A. H. Poorjam, M. H. Bahari, and H. Van hamme, "Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals," in *ICCKE*, 2014, pp. 7–12.

[14] A. H. Poorjam, R. Saeidi, T. Kinnunen, and V. Hautam, "Incorporating uncertainty as a quality measure in i-vector based language recognition," in *Odyssey*, Bilbao, 2016, pp. 74–80.

[15] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 59, pp. 1264–1271, 2012.

[16] J. Schoentgen and R. De Guchteneere, "Time series analysis of jitter," *J. Phon.*, vol. 73, pp. 189–201, 1995.

[17] I. Titze, *Principles of voice production*, 2nd ed. Iowa City: National Center for Voice and Speech, 1999.

[18] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[19] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 345–354, 2005.

[20] N. Dehak, "Discriminative and generative approaches for long- and short-term speaker characteristics modeling application to speaker verification." Ph.D. dissertation, Ecole de technologie superieure, 2009.

[21] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, 2nd ed., ser. Springer Series in Statistics. New York, NY: Springer New York, 2009.

[22] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[23] T. F. Wu, C. J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *J. Mach. Learn. Res.*, vol. 5, pp. 975–1005, 2004.

[24] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," National Taiwan University, Tech. Rep., 2016.

[25] S. J. Prince and J. H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in *IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[26] S. J. D. Prince, *Computer vision: models, learning, and inference*, 2012.

[27] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease." *Behav. Neurol.*, vol. 11, no. 3, pp. 131–137, 1998.

[28] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Int. Conf. Digit. Signal Process.*, 2009, pp. 1–5.