DYSPHONIA DETECTION BASED ON MODULATION SPECTRAL FEATURES AND CEPSTRAL COEFFICIENTS

M. Markaki, Y. Stylianou

Multimedia Informatics Lab, CSD, UoC, Greece

J.D. Arias-Londoño, J.I. Godino-Llorente

ICS EUIT de Telecomunicación, Spain

ABSTRACT

In this paper, we combine modulation spectral features with mel-frequency cepstral coefficients for automatic detection of dysphonia. For classification purposes, dimensions of the original modulation spectra are reduced using higher order singular value decomposition (HOSVD). Most relevant features are selected based on their mutual information to discrimination between normophonic and dysphonic speakers made by experts. Features that highly correlate with voice alterations are associated then with a support vector machine (SVM) classifier to provide an automatic decision. Recognition experiments using two different databases suggest that the system provides complementary information to the standard mel-cepstral features.

***Index Terms*—** pathologic voice detection, modulation spectrum, feature normalization, mutual information, SVD

1. INTRODUCTION

Objective voice quality assessment has been introduced to assist the perceptual evaluation of dysphonic voice quality used by the clinicians. Many studies in voice function assessment try to identify descriptive parameters for acoustic phenomena that highly correlate with pathological voice qualities. Acoustic measures that highly correlate with voice alterations can be associated then with a classification system to provide an automatic decision.

Organic pathologies modify the morphology of vocal folds resulting in abnormal vibration patterns and increased turbulent airflow at the level of the glottis [1]. The perceived voice abnormality is assumed to originate at the vocal source rather than resulting from abnormalities in the vocal tract configuration. Hence, many studies have focused on parameters such as pitch perturbation quotient (PPQ), jitter, shimmer, harmonics to noise ratio, etc. [2, 3, 4]. Perturbations at the glottal level will also affect the spectral properties of the recorded speech signal. There are both parametric and non parametric approaches for identifying the abnormal glottal activity based on analysis of speech signals. The parametric approaches are based on the source filter theory for the speech production and on the assumptions made for the glottal signal [5]. The non parametric approaches are based on magnitude spectrum of speech. Mel frequency cepstral coefficients (MFFC) - representing the vocal tract resonances - have been sucessfully used in voice pathology detection [6, 7]. Other non parametric approaches include time-frequency representations [8], and amplitude-modulation [9] or modulation spectral features [10].

Dysphonic voices are characterized by frequency-band dependent, time-varying amplitude fluctuations [9]. Similar to amplitude-modulation features, modulation spectra [11] can capture a class of source mechanism characteristics related to voice qualities. In this paper we pursue previous work in which we built an automatic dysphonia recognition and classification system based on modulation spectral representations [10]. Specifically, we investigate the complementary information that *normalized* modulation spectral features provide to MFCC for pathological voice detection in two different databases.

The paper is organized as follows: In Section 2 we briefly review modulation spectral features and their normalization, as well as the method of dimensionality reduction and feature selection we use. Section 3 describes the experiments we conducted using the same features and classifiers on the two databases. Finally in Section 4 we summarize our approach and discuss next steps.

2. MODULATION SPECTRA

The most common modulation frequency analysis framework [11] for a discrete signal $x(n)$, initially employs a short-time Fourier transform (STFT) $X_k(m)$

$$\begin{aligned} X_k(m) &= \sum_{n=-\infty}^{\infty} h(mM-n)x(n)W_K^{kn}, \\ k &= 0, \dots, K-1, \end{aligned} \tag{1}$$

where $W_K = e^{-j(2\pi/K)}$ and $h(n)$ is the acoustic frequency analysis window with a hop size of M samples (m denotes time). Mel scale filtering can be employed at this stage in order to reduce the number of frequency bands.

Subband envelope detection proceeds by taking the magnitude $|X_k(m)|$ of the subband. The distribution of envelope amplitudes of voiced speech has a strong exponential component. Hence we use a log transformation of the amplitude values $|X_k(m)|$ and subtract their mean log amplitude :

$$\hat{X}_k(m) = \log|X_k(m)| - \overline{\log|X_k(m)|} \quad (2)$$

where $\{\bar{\cdot}\}$ denotes the average operator over m.

Frequency analysis of subband envelopes with Fourier transform is performed next:

$$\begin{aligned} X_l(k,i) &= \sum_{m=-\infty}^{\infty} g(lL-m)|X_k(m)|W_I^{im}, \quad (3) \\ i &= 0,\ldots,I-1, \end{aligned}$$

where $g(m)$ is the modulation frequency analysis window and L the corresponding hop size (in samples); k and i are referred to as the "Fourier" (or acoustic) and "modulation" frequency, respectively. Tapered windows $h(n)$ and $g(m)$ are used to reduce the side lobes of both frequency estimates.

A modulation spectrogram representation then, displays modulation spectral energy $|X_l(k,i)|$ (magnitude of the subband envelope spectra) in the joint acoustic/modulation frequency plane. In order to enable cross-database portability of the classification system, feature subband normalization has been employed according to [12] (further details can be found in [12]). We normalize every acoustic frequency subband with the marginal of the modulation frequency representation:

$$X_{l,sub}(k,i) = \frac{X_l(k,i)}{\sum_i X_l(k,i)} \quad (4)$$

In previous work [12] it was shown that this subband normalization is important when there is a mismatch between training and testing conditions, or in other words, when the detection system is employed in real (testing) conditions.

2.1. Dimensionality Reduction and Feature Selection

Assuming a frame-by-frame analysis of speech, modulation spectra produce 3-D features (or tensors). We used a generalization of SVD to tensors referred to as Higher Order SVD (HOSVD) [13] to reduce dimensions in acoustic and modulation frequency subspaces separately. HOSVD enables the decomposition of tensor $\mathcal{D}$ to its $n-$mode singular vectors (or, principal components). Ordering of these $n-$mode singular values implies that the "energy" of tensor $\mathcal{D}$ is concentrated in the singular vectors with the lowest indices. Each singular matrix containing the $n-$mode singular vectors, can be truncated then by setting a predetermined threshold so as to retain only the desired number of principal axes in each mode. Projection of modulation spectral features on the principal axes with the higher energy in each subspace results in a compact set of features with minimum redundancy.

We further select features which are more relevant to the given classification task using mutual information (MI). That is, relevance is defined as the mutual information $I(x_j;c)$ between feature x_j and class c. *Maximal relevance* (MaxRel) feature selection criterion simply selects the features most relevant to the target class c [14]. Through a sequential search, which does not require estimation of multivariate densities, the top m features in the descent ordering of $I(x_j;c)$ are then selected.

3. AUTOMATIC DYSPHONIA RECOGNITION

We devised an automatic system to categorize speech as either pathological or normal. We will show that normalized modulation spectra-based features have good discrimination power in classifying dysphonic from normophonic voices in a cross-database experiment, while they provide complementary information to mel-cepstral coefficients. Therefore, combination of these two feature sets improves the classification performance.

3.1. Data and Methods

The first dysphonic voice corpus we used was the Kay Voice Disorders Database [15], which contains recordings of sustained vowels (/a/) and is commercially available. We will refer to this database as MEEI. A subset of 173 pathological and 53 normal speakers were selected according to [8], with similar age and sex distributions. The second database was recorded by Universidad Politécnica de Madrid, and it is referred to as Príncipe de Asturias (PdA) Hospital in Alcalá de Henares of Madrid database [16]. Similar to MEEI, PdA contains recordings of sustained vowels (/a/) and was developed for voice function assessment purposes. For the following experiments, the voices of 200 dysphonic subjects (74 men and 126 women, aged 11 to 76) affected by nodules, polyps, oedema, etc, as well as 199 normal subjects (87 men and 112 women, aged 16 to 70) were used. All the tests were conducted on signals sampled at 25 kHz. A 4-fold stratified cross-validation scheme - repeated 4 times - produced 16 different groupings of the voices, each using $\sim 75\%$ of the utterances for training and $\sim 25\%$ for testing. For the cross-database evaluation, we used PdA for training and MEEI for testing or vice-versa (in order to simulate the situation of completed unseen, to the classification system, data).

In each case, modulation spectra were computed in a frame-by-frame basis using long windows in time (250 ms) which were shifted by 50ms. We used Mel scale filtering with 53 bands while the size of the Fourier transform for the time-domain transformation was set to 257 (up to π). Therefore, each modulation spectrum consisted of $I_1 = 53$ acoustic frequencies and $I_2 = 257$ modulation frequencies, resulting therefore in an 53×257 image per frame. The normalized modulation spectra computed in each frame were stacked to produce a third order tensor $\mathcal{D} \in R^{I_1 \times I_2 \times I_3}$, where I_3 is the number of frames in the training dataset. Applying the High Order SVD algorithm described previously, the near-optimal projections or principal axes (PCs) of features were detected among those contributing more than 0.1% to the "energy" of $\mathcal{D}$. For MEEI, we detected 44 PCs in the acoustic frequency and 29 PCs in the modulation frequency subspace. This resulted in a reduced space of $44 \times 29 = 1276$ features. For PdA, the corresponding reduced space had dimensions of $53 \times 36 = 1908$. Next, the features which were more correlated to the voice pathology detection task were selected for each database, using the Maximal Relevance criterion (MaxRel) . For details about the application of the MaxRel criterion on this task please refer to [12].

To extract MFCC features, each utterance was first run through the standard mel-cepstrum filterbank (using 12 filters) at a 25-ms frame interval. The cepstrum was computed and channel compensation techniques were applied according to [7]. In order to combine MFCC with mRMS features, the mean and variance of the 12 MFCC features over 10 frames were extracted, every 2 frames (a 50 ms shift). Delta features were not included since the improvement over MFCC features alone was not found to be statistically significant in [7].

The features were then fed as input to a support vector machine (SVM) classifier with a radial basis function kernel [17]. Detection-error tradeoff (DET) curves and the equal error rate (EER) were used to compare the performance of different systems on MEEI and PdA.

3.2. Results

DET curve results for standard mel-cepstrum, mRMS and the concatenated feature vector (including both MFCC and mRMS features) are plotted in Figure 1 for MEEI and in Figure 2 for PdA. The top m mRMS features were selected for each database using 4-fold cross validation. The optimum detector based on mRMS features alone was obtained by considering the $m = 125$ most relevant features for both MEEI and PdA. As shown, the equal error rate (EER) - the point where the false alarm probability equals the miss probability - of mel-cepstrum alone is 8.47% on MEEI and 22.86% on PdA, with mRMS features yielding 6.29% on MEEI and 17.67% on PdA, and the concatenated vector resulting in 3.63% on MEEI and 12.15% on PdA (Table 1).

In the cross-database experiments, when training is performed on the $m = 125$ most relevant features of PdA and testing on the same mRMS features for MEEI, the EER for MFCC is 28.24%, for mRMS is 24.40% and for the concatenated features 16.87% (see Figure 3 and Table 1). When training is performed on the $m = 125$ most relevant features of MEEI and testing on the same number of mRMS features for PdA, the performance of the system significantly deteriorates. We had to consider the top $m = 450$ most relevant features - relevance estimated on MEEI - in order to capture dysphonia in PdA. In that case, the EER of mRMS is 26.07%, of MFCCs is 30.97% and of concatenated features 21.86%. Table 1 summarizes the classification scores for the different conducted experiments. The last two rows of the Table provide information for the cross-database experiment where PdA-MEEI means training on PdA and testing on MEEI and vice versa for MEEI-PdA. In brackets we note the number of the mRMS features used in each experiment.

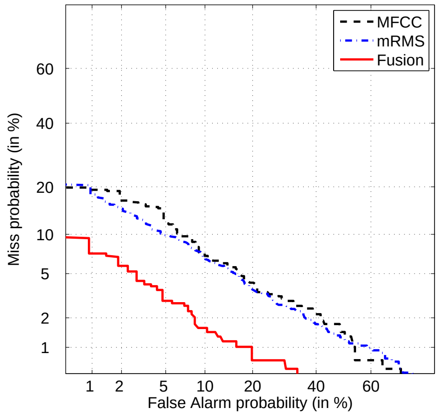

Fig. 1. Performance of MFCC and mRMS features in MEEI.

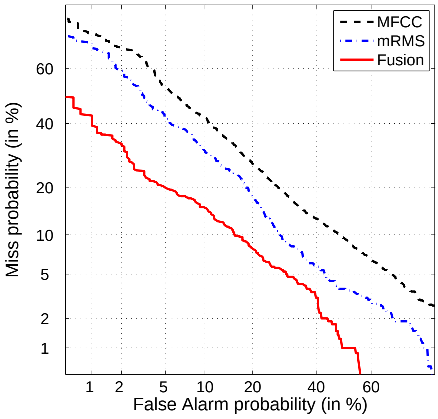

Fig. 2. Performance of MFCC and mRMS features in PdA.

Table 1. Equal Error Rate (EER) in % for mRMS features, MFCC and both of them in MEEI and PdA.

	MFCC	mRMS	Fusion
MEEI	8.47	6.29 (125)	3.63
PdA	22.86	17.67 (125)	12.15
PdA-MEEI	28.24	24.40 (125)	16.87
MEEI-PdA	30.97	26.07 (450)	21.86

4. DISCUSSION

Pathological voice is characterized by an increase of the vocal folds mass, a subsequent lack of closure or an elasticity change of the vocal folds and surrounding tissue [7]. Dys-

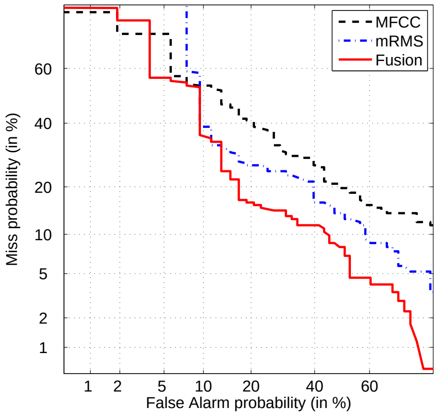

Fig. 3. Performance of mRMS features, MFCC and their fusion when training is performed in PdA and testing in MEEI.

phonia recognition experiments on MEEI and PdA confirmed that modulation spectral features provide complementary information to MFCC. The low bands of the MFCC reflect alterations related with the mucosal waveform due to an increase of mass whereas the noisy components induced by lack of closure are modeled by the higher bands [7]. Modulation spectra on the other hand capture the amplitude envelope fluctuations evident on sustained vowel phonations [9].

Regarding cross-database experiments, features selected from PdA alone were more successful in capturing class specific information in MEEI than vice versa. A potential reason for this is that some of the normal speakers in MEEI database were recorded at different sites and over possibly different channels than the pathological subjects [9]. This makes the MEEI an easy database for classification tests. This is not the case with PdA, where the same recording conditions were used for normal and dysphonic speakers. It follows then, that it is better to train the classifier on PdA than on MEEI.

We have simply concatenated the mean and variance of MFCC over the same segments that mRMS were estimated from; the concatenated feature vector was given as input to the SVM classifier. A better strategy, would be to combine different classifier schemes for every feature set. We ran additional experiments with MFCC and GMM classifier, as well as mRMS and GMM classifier on the same datasets for normal/pathological distinction. Configuration of MFCC with GMM classifier (the system described in [7]) was better than using MFCC with SVM - still, in all experiments MFCC plus GMM produced inferior results to the fusion of features combined with SVM. On the other hand, mRMS plus SVM configuration clearly superseded mRMS plus GMM. The reason is the large number of mRMS features and the corresponding quadratic increase of the number of parameters of GMM classifier. In the future, therefore, we will explore the fusion of classifiers at the decision level and not the fusion at the feature level.

5. REFERENCES

[1] R.J. Baken, *Clinical measurement of speech and voice*, College Hill Press, Boston, 1987.

[2] S.B. Davis, "Computer evaluation of laryngeal pathology based on inverse filtering of speech," SCRL Monograph Number 13, 1976.

[3] R.A. Prosek, A.A. Montgomery, B.E. Walden, and D.B. Hawkins, "An evaluation of residue features as correlates of voice disorders," *Journal of Communication Disorders*, vol. 20, pp. 105–117, 1987.

[4] V. Parsa and D.G. Jamieson, "Identification of pathological voices using glottal noise measures," *J. Speech, Language, Hearing Res.*, vol. 43(2), pp. 469–485, 2000.

[5] A. Askenfelt and B. Hammarberg, "Speech waveform perturbation analysis revisited," *Speech Transmission Lab. - Quartely Progress and Status Report*, vol. 22(4), pp. 49–68, 1981.

[6] A.A.Dibazar, T.W.Berger, and S.S.Narayanan, "Pathological voice asessment," in *IEEE, 28th Eng. in Med. and Biol. Soc.* IEEE, 2006, pp. 1669–1673.

[7] J.I. Godino-Llorente, P. Gómez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on GMMs and short-term cepstral parameters," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 10, pp. 1943–1953, 2006.

[8] K. Umapathy, S. Krishnan, V. Parsa, and D.G. Jamieson, "Discrimination of pathological voices using time-frequency approach," *IEEE Trans. Biomedical Engineering*, vol. 52, no. 3, pp. 421–430, 2005.

[9] N. Malyska, T.F. Quatieri, and D. Sturim, "Automatic dysphonia recognition using biologically inspired amplitude-modulation features," in *Proc. ICASSP*, 2005, pp. 873–876.

[10] M. Markaki and Y. Stylianou, "Using modulation spectra for voice pathology detection and classification," in *Proc. EMBS*, Minessota, U.S.A., 2009.

[11] L. Atlas and S.A. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Applied Signal Processing*, vol. 7, pp. 668–675, 2003.

[12] M. Markaki and Y. Stylianou, "Normalized modulation spectral features for cross-database voice pathology detection," in *Proc. Interspeech*, Brighton, U.K., 2009.

[13] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 21, pp. 1253–1278, 2000.

[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226–1238, 2005.

[15] Massachusetts Eye and Ear Infirmary, "Elemetrics Disordered Voice Database (Version 1.03)," Voice and Speech Lab, Boston, MA, Oct. 1994, Kay Elemetrics Corp.

[16] J.I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, I. Cobeta-Marco, R. González-Herranz, and C. Ramírez-Calvo, "Acoustic analysis of voice using WPCVox: a comparative study with multi dimensional voice program," *European Archives of Otolaryngology*, vol. 265(4), pp. 465–476, 2008.

[17] T. Joachims, *Advances in Kernel Methods - Support Vector Learning*, chapter Making large-scale SVM Learning Practical, MIT-Press, 1999.