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 (MFCC) - representing the vocal tract resonances - have been successfully 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) \( X_k(m) \)

\[
X_k(m) = \sum_{n=-\infty}^{\infty} h(mM-n)x(n)W_K^n, \quad (1)
\]

\[
k = 0, \ldots, K - 1,
\]

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
modulation spectral energy 
reduce the side lobes of both frequency estimates.

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

\[
X_l(k, i) = \sum_{m=-\infty}^{\infty} g(lL - m)|X_k(m)|W_{im}^m,
\]

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 sub-
band envelope spectrum) in the joint acoustic/modulation fre-
quency 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 sub-
band with the marginal of the modulation frequency represen-
tation:

\[
X_{l,sub}(k, i) = \frac{X_l(k, i)}{\sum_i X_l(k, i)}
\]

In previous work [12] it was shown that this subband nor-
malization 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 general-
ization 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 de-
composition of tensor \( D \) to its \( n \)-mode singular vectors (or,
principal components). Ordering of these \( n \)-mode singular
values implies that the “energy” of tensor \( D \) is concentrated
in the singular vectors with the lowest indices. Each singular
matrix containing the \( n \)-mode singular vectors, can be trunc-
ated then by setting a predetermined threshold so as to retain
only the desired number of principal axes in each mode. Pro-
jection 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)

3. Automatic Dysphonia Recognition

We devised an automatic system to categorize speech as ei-
ther 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, com-
bination 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 sus-
tained 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 re-
ferrred to as Príncipe de Asturias (PdA) Hospital in Alcal´ a de
Henares of Madrid database [16]. Similar to MEEI, PdA con-
tains 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 con-
ducted 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 \( \pi \)). There-
fore, each modulation spectrum consisted of \( I_1 = 53 \) acoustic
frequencies and \( I_2 = 257 \) modulation frequencies, resulting
therefore in an \( 53 \times 257 \) image per frame. The normalized
modulation spectra computed in each frame were stacked to
produce a third order tensor \( D \in R^{I_1 \times I_2 \times I_3} \), 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

values \(|X_k(m)|\) and subtract their mean log amplitude:

\[
\hat{X}_k(m) = \log |X_k(m)| - \log |\bar{X}_k(m)|
\]

where \([\cdot]\) denotes the average operator over \( m \).

3. AUTOMATIC DYSPHONIA RECOGNITION

feature selection criterion simply selects the features most rel-
evant 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.
among those contributing more than 0.1% to the “energy” of \( D \). 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].

In the cross-database experiments, when training is performed on the \( m = 125 \) most relevant features of PdA and testing on the same number of features for MEEI, 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.

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

<table>
<thead>
<tr>
<th></th>
<th>MFCC</th>
<th>mRMS</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEEI</td>
<td>8.47</td>
<td>6.29</td>
<td>3.63</td>
</tr>
<tr>
<td>PdA</td>
<td>22.86</td>
<td>17.67</td>
<td>12.15</td>
</tr>
<tr>
<td>PdA-MEEI</td>
<td>28.24</td>
<td>24.40</td>
<td>16.87</td>
</tr>
<tr>
<td>MEEI-PdA</td>
<td>30.97</td>
<td>26.07</td>
<td>21.86</td>
</tr>
</tbody>
</table>

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-
phonation 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 surpassed 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