MODULATION SPECTRAL FEATURES FOR OBJECTIVE VOICE QUALITY ASSESSMENT: THE BREATHINESS CASE

Maria Markaki¹, Yannis Stylianou¹,²
¹Computer Science Dept, University of Crete, Greece
²Institute of Computer Science, FORTH, Crete, Greece

Abstract: In this paper, we employ normalized modulation spectral features for objective voice quality assessment regarding breathiness. Modulation spectra usually produce a high-dimensionality space. For classification purposes, the size of the original space is reduced using Higher Order Singular Value Decomposition (SVD). Further, we select most relevant features based on the mutual information between the degree of breathiness and the computed features, which leads to an adaptive to the classification task modulation spectral representation. The adaptive modulation spectral features are used as input to a Naive Bayes (NB) classifier. By combining two NB classifiers based on different feature sets a global classification rate of 79\% for breathiness was achieved.

Keywords: Objective voice quality assessment, breathiness, modulation spectrum, mutual information, SVD

I. INTRODUCTION

Objective voice quality assessment has been introduced to assist the perceptual evaluation of dysphonic voice quality used by the clinicians. The most common systems of pathological voice description refer to the degree of “hoarseness” [1]. Hoarseness is perceptually related to the noise generation during phonation. The degree of voice hoarseness can be quantified according to the GRASB (grade, roughness, asthenicity, strain and breathiness) scale proposed by Hirano [1].

The definition of these quantifiable perceptual dimensions (GRASB parameters) is related to a set of descriptive parameters for acoustic phenomena. 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]. Acoustic measures that highly correlate with voice alterations can be associated then with a classification system to provide an automatic decision.

In this work we investigate the correlation of modulation spectral features [7, 8] to the degree of breathiness (B) of pathological voices. Dysphonic voices are characterized by frequency-band dependent, time-varying amplitude fluctuations [5]. Modulation spectral features can capture a class of source mechanism characteristics related to voice qualities (glottal source differences) [5]. Breathiness typically refers to the voice quality related to the audible turbulence generated at the glottal level; this turbulence acts as a noise source to the vocal tract (see [9] and references therein). This paper pursues a previous work in which the authors presented an automatic dysphonia recognition and classification system built on modulation spectral representations [10].

The paper is organized as follows: In Section II we briefly describe the dataset, modulation spectral features and their normalization, and the method of dimensionality reduction and feature selection we use. Specifically, the initial representation is first transformed to a lower-dimensional domain using Higher Order SVD [11]. 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 estimate the relevance of these projections to dysphonic voice characterization based on their mutual information to breathiness class variable. Section III describes the experiments we conducted on breathiness classification using a combination of two naive bayes (NB) classifiers based on different feature sets [13]. Finally in Section IV we summarize features of our approach and discuss next steps.

II. METHOD

A. Dysphonic voice corpus

We used a database provided to us by Universidad Politécnica de Madrid, which is referred to as Príncipe de Asturias (PdA) Hospital in Alcalá de Henares of Madrid database [14]. Similar to MEEI, PdA contains recordings of sustained vowels (/a/) and was developed for voice function assessment purposes. The voices of 201 dysphonic and 209 normal subjects have been classified according to the B parameter (breathiness) of the Hirano’s GRASB
scale. A four-point scoring system is used to rate each subject along the B dimension: 0 denotes no breathiness, 1 means a slightly breathy voice, 2 refers to moderate breathiness, whereas 3 describes a severely breathy voice. For the following experiments, we selected 200 dysphonic subjects (74 men and 126 women, aged 11 to 76) affected by nodules, polyps, oedema, etc., as well as 24 subjects with normal voice (7 men and 17 women, aged 17 to 54). Specifically, we used 26 dysphonic (plus 24 normal) voices with zero breathiness, 50 voices with $B = 1$, 119 with $B = 2$ and 3 voices with $B = 3$. Due to the very small number of subjects with a breathiness rating equal to 3, these were joined with the subjects with a rating 2 breathiness.

### B. Modulation Spectra

The most common modulation frequency analysis framework [8] for a discrete signal $x(n)$, initially employs a short-time Fourier transform (STFT) $X_k(m)$

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

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 - defined as the magnitude $|X_k(m)|$ or square magnitude of the subband - and their frequency analysis with Fourier transform are performed next:

$$X_l(k,i) = \sum_{m=-\infty}^{\infty} g(L-m)|X_k(m)|W_I^m, \quad (2)$$

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 [15].

### C. Normalized modulation spectra

The distribution of envelope amplitudes of voiced speech has a strong exponential component. Hence we calculate modulation spectra using a log transformation of the amplitude values $|X_k(m)|$ and subtracting their mean log amplitude before windowing in (3):

$$\hat{X}_k(m) = \log |X_k(m)| - \log (\langle |X_k(m)| \rangle) \quad (3)$$

where $\langle \cdot \rangle$ denotes the average operator over $m$. This is analogous to the cepstral mean subtraction approach, which is commonly employed to compensate for convolutional noise in the case of MFCC features. Next, we normalize every acoustic frequency subband with the marginal of the modulation frequency representation:

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

Previous work [15] has 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.

### D. Dimensionality reduction and Feature Selection

We used a generalization of SVD to tensors referred to as Higher Order SVD (HOSVD) [11] to reduce dimensions in acoustic and modulation frequency subspaces separately. HOSVD enables the decomposition 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 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 selected features which were more relevant to the given classification task using mutual information (MI). That is, relevance is defined as the mutual information (MI). That is, relevance is defined as the mutual information (MI). That is, relevance is defined as the mutual information ($I(x; c)$) between feature $x$, and class $c$. $Maximal relevance$ (MaxRel) feature selection criterion simply selects the features most relevant to the target class $c$ [12]. Through a sequential search, which does not require estimation of multivariate densities, the top $m$ features in the descent ordering of $I(x_j; c)$ were selected.

Fig. 1 depicts the mutual information of the original normalized modulation spectral features for the classification of the dysphonic phonations of the vowel /AH/ in PdA in 3 scores of breathiness (B0, B1 and B2). Modulations localized lower than $\sim 1600$ Hz on the acoustic frequency axis seem to be more relevant; this is consistent with previous experimental results on pathological voice assessment where frequencies lower than 3000Hz led to an homogeneous discrimination between voices compared with higher frequencies [6].
Accordingly, subject \( x \) will be assigned to class \( \omega_k \) if \( \mu_k(x) \) has the highest value.

### III. RESULTS

Modulation spectra were computed in a frame-by-frame basis using long windows in time (262 ms) which were overlapped by 50%. 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 \( \pi \)). Therefore, 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 modulation spectra computed in each frame were mean subtracted and then they were stacked to produce a third order tensor \( D \in \mathbb{R}^{I_1 \times I_2 \times t} \), where \( I_3 \) is the number of frames in the training dataset. After applying the High Order SVD algorithm, we kept the principal axes (PCs) of features contributing more than 0.1\% to the “energy” of \( D \); i.e., the first 43 PCs in the acoustic frequency and the first 29 PCs in the modulation frequency subspace. This resulted in a reduced space of \( 43 \times 29 = 1247 \) features. Next, the features which were more correlated to the breathiness assessment were selected using the Maximal Relevance criterion (MaxRel). For details about the application of the MaxRel criterion on this task please refer to [10].

Two different feature sets were defined according to the sorted MI values. The first set included the most relevant features when MI estimation also involved voices from (24) normal subjects with zero breathiness. The second feature set was selected using the dataset of dysphonic only voices. We used leave-one-out cross validation to select the top \( m \) features for every NB classifier built on top of each feature set. NB classifier built on top of the \( m = 100 \) most relevant features of the first set was optimum for discriminating class \( B = 0 \). For classes \( B = 1 \) and \( B = 2 \), the optimum NB classifier was obtained by considering the top \( m = 230 \) most relevant features of the second set. By combining the NB classifiers based on these different feature sets [13], a global classification rate of 79.02\% was achieved. Table 1 presents the confusion matrix from the automatic classification of the dysphonic voices into scores of breathiness. This classification has been compared with the original perceptual judgement in the PdA corpus. In Table 2 the performance per breathiness score in terms of correct classification rate is presented. We can observe that the worse performance corresponds to the B0 class which includes 26 dysphonic and 24 normal subjects. However, we note that 21 out the 24 normal speakers have been correctly classified in the B0 class (corresponding to a CCR of 87.50\% for normal only voices). We conclude then that the breathiness of dysphonic only voices has been overestimated in the case of the \( B = 0 \) class.
Table 1: Confusion matrix between scores of breathiness given by the automatic classification system (S-B0, S-B1, S-B2) and the perceptual judgement of phonations (P-B0, P-B1, P-B2).

<table>
<thead>
<tr>
<th></th>
<th>P-B0</th>
<th>P-B1</th>
<th>P-B2</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-B0</td>
<td>33</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>S-B1</td>
<td>16</td>
<td>99</td>
<td>10</td>
</tr>
<tr>
<td>S-B2</td>
<td>1</td>
<td>16</td>
<td>45</td>
</tr>
</tbody>
</table>

Table 2: Performance per breathiness score in terms of correct classification rate (CCR %) of phonations in PdA [14].

<table>
<thead>
<tr>
<th>Score 0</th>
<th>Score 1</th>
<th>Score 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>66.00</td>
<td>83.19</td>
<td>81.82</td>
<td>79.02</td>
</tr>
</tbody>
</table>

IV. DISCUSSION

In this paper we have proposed a method for objective assessment of breathy voice quality, based on modulation spectra. We used a method for dimensionality reduction and feature selection on a database of sustained vowels. Using mutual information, we could locate the most relevant frequency bands at the “formant zone”, i.e. lower than 3000 Hz. Based on different feature sets, two NB classifiers were tested and found to be optimal in the discrimination of different classes. By combining them, a global classification rate of 79.02% was achieved.

Future work will address additional GRASB parameters using a database of reading text. We will explore the discriminative ability of consonant classes as well in the objective assessment of different voice qualities. In addition, benchmarking against more standard approaches like those used for the automatic speaker recognition [6] will be performed.

ACKNOWLEDGEMENTS

The authors would like to thank J.I. Godino-Llorente of the Department of Circuits & Systems Engineering, Universidad Politécnica de Madrid, for the availability of PdA.

REFERENCES


