Voice Pathology Detection and Discrimination based on Modulation Spectral Features

Maria Markaki, *Student Member, IEEE,* and Yannis Stylianou, *Member, IEEE*

***Abstract*—In this paper, we explore the information provided by a joint acoustic and modulation frequency representation, referred to as Modulation Spectrum, for detection and discrimination of voice disorders. The initial representation is first transformed to a lower-dimensional domain using higher order singular value decomposition (HOSVD). From this dimension-reduced representation a feature selection process is suggested using an information theoretic criterion based on the Mutual Information between voice classes (i.e., normophonic/dysphonic) and features. To evaluate the suggested approach and representation, we conducted cross-validation experiments on a database of sustained vowel recordings from healthy and pathological voices, using support vector machines (SVM) for classification. For voice pathology detection, the suggested approach achieved a classification accuracy of $94.1 \pm 0.28\%$ (95% confidence interval), which is comparable to the accuracy achieved using cepstral based features. However, for voice pathology classification the suggested approach significantly outperformed the performance of cepstral based features.**

***Index Terms*—pathological voice detection, modulation spectrum, Higher Order SVD, mutual information, pathological voice, pathology classification.**

I. Introduction

Many studies have focused on identifying acoustic measures that highly correlate with pathological voice qualities (also referred to as voice alterations). Using acoustic analysis, we seek to objectively evaluate the degree of voice alterations in a noninvasive manner. Organic pathologies that affect vocal folds usually modify their morphology in a diffuse or a nodular manner. Consequently, abnormal vibration patterns and increased turbulent airflow at the level of the glottis might be observed [1]. Acoustic parameters that quantify the glottal noise include fundamental frequency, jitter, shimmer, amplitude perturbation quotient (APQ), pitch perturbation quotient (PPQ), harmonics to noise ratio (HNR), normalized noise energy (NNE), voice turbulence index (VTI), soft phonation index (SPI), frequency amplitude tremor (FATR), glottal to noise excitation (GNE)([2], [3], [4] and references within).

Some of the suggested features require accurate estimation of the fundamental frequency which is not a trivial task in the case of certain vocal pathologies. Moreover, since these features refer to the glottal activity, an estimation of the glottal airflow signal is required. This can be obtained either by electroglottography (EGG) [5] or by inverse filtering of speech [6] [7] where an estimate of the glottal airflow signal is obtained. Based on the second approach, spectral related features have been defined such as the spectral flatness of the inverse filter (SFF) and the spectral flatness of the residue signal (SFR) [2]. Flatness is defined as the ratio of the geometric mean of the spectrum to its arithmetic mean (usually in dB) [2]. The more noise-like a speech signal is, the larger is the "flatness" of its magnitude spectrum [8]. SFF and SFR can be considered as a measure of the noise masking formants and harmonics, respectively [3].

Apart from the above measurements, there is a great interest in applying methods from the non-linear time series analysis to speech signals, trying to quantify in a compact way the high degree of abnormalities observed during sustained phonation when dysphonia is present. Correlation dimension and second-order dynamical entropy measures [9], Lyapunov exponents [10], higher-order statistics [11], and measures based on time-delay state-space recurrence and detrended fluctuation analysis [12] have also been used in classifying normophonic from dysphonic speakers. For an extended summary on non-linear approaches for voice pathology detection, the interested reader is referred to [12].

Assuming that the speech signal production is based on the well-known source-filter theory, then it is expected that perturbations at the glottal level (source signal) will affect the spectral properties of the recorded speech signal. In this case, the estimation of the glottal signal is not necessary. Nevertheless, another difficult problem is raised; the estimation of appropriate features from the speech signal which are connected with properties of the glottal signal. Alternatively both parametric and non-parametric approaches have been suggested in this respect, these being generally referred as *Waveform Perturbation* methods (even if they only work with a partial information of the waveform, i.e., magnitude spectrum, frequency perturbations, etc.). The parametric approaches are based on the source-filter theory for the speech production and on the assumptions made for the glottal signal (i.e., impulse train, noise-like) [13] [14]. The non-parametric approaches are based on the magnitude spectrum of speech where short-term mel frequency cepstral coefficients (MFFC) are widely used in representing the magnitude spectrum in a compact way [15] [16] [17] [18]. The non-parametric approaches also include time-frequency representations as the one suggested in [19].

Correlation of the various suggested features and representations with voice pathology is evaluated using techniques like linear multiple regression analysis [3], or likelihood scores using Gaussian Mixture Models (GMM) [15] [17] and Hidden Markov Models (HMM) [16]. Also neural networks and Sup-

M. Markaki and Y. Stylianou are with the Multimedia Informatics Lab, Computer Science Dept. University of Crete, Greece; email:mmarkaki,yannis@csd.uoc.gr.

Y. Stylianou is with the Institute of Computer Science, FORTH, Crete, Greece.

port Vector Machine classifiers have been suggested [18] [20].

While there are many suggested features and systems for voice pathology detection in the literature, there have been a few attempts towards separating different kinds of voice pathologies. Linear Prediction-derived measures were found inadequate for making a finer distinction than the normal/pathological voice discrimination in [3]. In [7], after applying an iterative residual signal estimator, features like jitter have been computed. Jitter provided the best classification score between pathologies (54.8% for 21 pathologies). In [16], an HMM approach using MFCC provided an average score of correct classification of 70% (5 pathologies, multi classification experiment).

In [21] a vocal-fold paralysis recognition system using amplitude-modulation and MFCC features combined with GMM, provided an Equal Error Rate (EER) of $\sim 30\%$ in the best case. A recent study for the discrimination of voice pathology signals was carried out using adaptive growth of Wavelet Packet tree, based on the criterion of Local Discriminant Bases (LDB) [20]. A genetic algorithm was employed to select the best feature set and then a Support Vector Machine (SVM) classifier was used. An average detection score of 83.9% was reported in classifying vocal polyps against adductor spasmodic dysphonia, keratosis leukoplakia, and vocal nodules.

In this work, we suggest the use of modulation spectra for detection and classification of voice pathologies [22], [23]. Modulation spectral features have been employed for single-channel speaker separation [24], for speech and speaker recognition [25], [26] as well as for content-based audio identification [27] and speech detection [28]. There are a few works which make use of modulation spectra for voice pathology detection [21] [29], [30]. Modulation spectra may be seen as a non-parametric way to represent the modulations in speech. Modulation spectra offer an implicit way to fuse the various phenomena observed during speech production, such as the harmonic structure during voiced phonation etc. [24]. This is achieved by describing the joint distribution of energy across different acoustic and modulation frequencies. The long-term ($\sim 200-300$ms) information that modulation spectrum represents poses a serious challenge to classification algorithms because of its high dimensionality. Past research has addressed the problem of reducing modulation spectral feature dimensions by simple averaging [21], or using modulation scale analysis, a joint representation of the acoustic and modulation frequency with nonuniform bandwidth [27]. In [31], a bank of mel-scale filters has been applied along the acoustic frequency dimension, and discrete cosine transform (DCT) along the modulation frequency axis.

In this paper, we compute modulation spectra using simple Fourier transform in both frequency axes (acoustic and modulation). Moreover, in this paper we approach the dimensionality reduction of the acoustic and modulation frequency subspaces in the framework of multilinear algebra. Since the acoustic and modulation spectra are characterized by varying degrees of redundancy, we address dimensionality reduction separately in each subspace using higher order singular value decomposition (HOSVD) [32]. The Mutual Information (MI) measurement based on Information Theory [33] can subsequently analyze the relation between the compact lower dimensional features and classes (i.e., voice disorders).

In Section II, the modulation frequency analysis framework is briefly described. Section III motivates the use of modulation frequency analysis for voice pathology detection and classification, by providing examples of this joint frequency representation computed for speech signals generated by normophonic and dysphonic speakers. For this purpose, speech examples from the Massachusetts Eye and Ear Infirmary Voice and Speech Laboratory (MEEI) database [34] are considered. In Section IV, the lower-dimensional feature space where feature extraction/selection will be eventually performed, is defined. In Section V, the Mutual Information (MI) estimation procedure is presented and in Section VI, the pattern classification algorithm and the performance analysis measures used in the paper, are explained. In Section VII, a general description of MEEI [34] database is provided along with its subsets used in the classification experiments. In the first experiment the ability of modulation frequency features to distinguish between normal and pathological voices is investigated. Next, we investigate the ability of modulation spectra and the suggested feature selection algorithm to make distinctions that are finer than the normal/pathological dichotomy. Specifically we address the binary discrimination between vocal fold polyp, adductor spasmodic dysphonia, keratosis leukoplakia, vocal nodules, as well as between paralysis and all the above voice disorders. Finally, conclusions are drawn and future directions are indicated in Section VIII.

II. Modulation Frequency Analysis

The most common modulation frequency analysis framework for a discrete signal $x(n)$, initially employs a short-time Fourier transform (STFT) [23] [24], while other joint time-frequency representation may also be used [35]. In this paper, the STFT is used, which is computed as:

$$\begin{aligned} X_m(k) &= \sum_{n=-\infty}^{\infty} h(mM-n)x(n)W_{I_1}^{kn}, \qquad (1)\\ k &= 0,\ldots,I_1-1, \end{aligned}$$

where I_1 denotes the number of frequency bins in the acoustic frequency axis, $W_{I_1}=\exp\left(-j\pi/I_1\right)$, M is the shift parameter in the computation of the STFT, and $h(n)$ is the acoustic frequency analysis window. The mean is subtracted from each subband envelope $|X_m(k)|$ before modulation frequency estimation, in order to reduce the interference of large DC components (of subband envelopes). Next, a second STFT is applied along the time dimension of the spectrogram to perform frequency analysis (modulation frequency estimation) of subband envelopes:

$$\begin{aligned} X_l(k,i) &= \sum_{m=-\infty}^{\infty} g(lL-m)|X_m(k)|W_{I_2}^{im}, \qquad (2)\\ i &= 0,\ldots,I_2-1, \end{aligned}$$

where I_2 is the number of frequency bins along the modulation frequency axis, $W_{I_2}=\exp\left(-j(f_M/F_s)\ \pi/I_2\right)$, with f_M and

F_s denoting the maximum modulation frequency we search for, and the sampling frequency, respectively, L is the shift parameter of the second STFT, and $g(m)$ is the modulation frequency analysis window. Tapered windows $h(n)$ and $g(m)$ are used to reduce the sidelobes of both frequency estimates. The magnitude of the acoustic-modulation frequency representation computed in eq. (2) is referred to as modulation spectrogram. It displays the modulation spectral energy $|X_l(k,i)| \in R^{I_1 \times I_2}$ (magnitude of the subband envelope spectra) in the joint acoustic/modulation frequency plane. Length of the analysis window $h(n)$ controls the trade-off between resolutions in the acoustic and modulation frequency axes [24]. When $h(n)$ is short (wideband analysis) the frequency subbands will be wide and the maximum observable modulation frequency will be high. When $h(n)$ is long (narrowband analysis) the frequency subbands will be narrow and the maximum observable modulation frequency will be low. Also, the degree of overlap between successive windows sets the upper limit of the subband sampling rate during the modulation transform.

III. Modulation Spectral Patterns in Normal and Dysphonic Voices

We have evaluated features of the modulation spectrogram of sustained vowel /AH/ for voice pathology detection and classification tasks. As explained in the work of Vieira et al [36], sustained vowel phonations at comfortable levels of fundamental frequency and loudness are useful from a clinical point of view. In addition, the time domain acoustic signal of /AH/ exhibits larger and sharper peaks than the other vowels; these signal features are well correlated to the electroglottal graph (EGG) parameters.

Fig. 1a shows the modulation spectrogram $|X_l(k,i)|$ of a 262 ms long frame from sustained phonation speech samples of the vowel /AH/ uttered by a normal male speaker from the MEEI database [34]. Apparently these phonations do not possess the syllabic and phonetic temporal structure of speech. Hence, the higher energy values are not concentrated at the lower modulation frequencies which are typical in running speech, $\sim 1-20$ Hz [25]. Instead, since we used an analysis window $h(n)$ that was shorter than the expected lowest pitch period, the highest energy terms usually occur at the fundamental frequency of the speaker (~ 150 Hz in the example shown in Fig. 1a) and its harmonics in the modulation frequency axis (up to 500 Hz). Fundamental frequency energy appears localized at the first two formants of vowel /AH/ along the acoustic frequency axis (their range is $\sim 677 \pm 95$ Hz and $\sim 1083 \pm 118$ Hz). Fig. 1b displays the mean modulation spectrum, and fundamental frequency distribution of 40 normal speakers from MEEI, with equal number of male and female subjects. All modulation spectra have been normalized to 1 prior to averaging. The two main clusters reflect the fundamental frequency distribution of male (range: 146 ± 24.4 Hz) and female talkers (244 ± 30 Hz). The second cluster contains more energy than the first cluster, since it also comprises energy from the first harmonic of the fundamental frequency of male speakers. Regarding the vertical coordinates of clusters, most energy is concentrated around the first two formants of /AH/. Overall, modulation spectral representations of normal vowel phonations are quite similar to each other, exhibiting a clear harmonic structure.

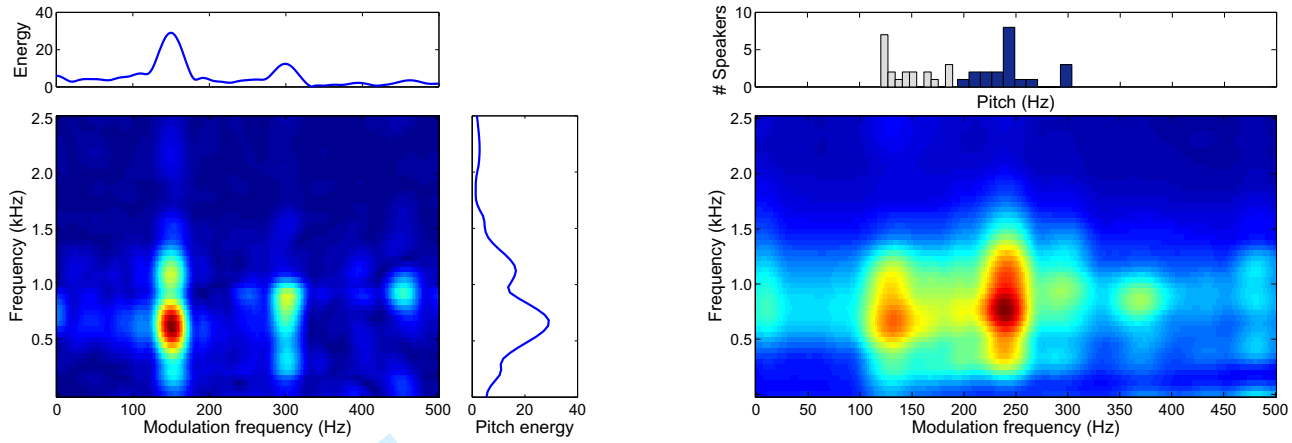
These patterns of amplitude modulations are expected to be distorted when voice pathology is present - providing therefore cues for its detection and classification. Fig. 2 and 3 depict modulation spectra $|X_l(k,i)|$ of sustained vowels produced by patients with various voice pathologies: vocal polyps, adductor spasmodic dysphonia, keratosis and vocal nodules. A comprehensive description of these pathologies is provided in [1]. Polyps are solid or fluid filled growths arising from the vocal fold mucosa. They affect vibration of vocal folds depending on their size and location. In adductor spasmodic dysphonia vocal folds suddenly squeeze together very tightly and in effect the voice breaks, stops, or strangles. Keratosis refers to a lesion on the mucosa of the vocal folds, appearing as a white patch. Nodules are swellings below the epithelium of vocal folds; they might prevent the vibration of the vocal folds either by causing a gap between the two vocal folds - which lets air to escape - or by stiffening the mucosal tissue.

Compared to the normal ones (see Fig. 1), pathological modulation spectra lack a uniform harmonic structure and appear more "spread" and "flattened" across the acoustic frequency axis. Main differences can be spotted near the low acoustic frequency bands where the first formant of /AH/ is located ($\sim$ 500 Hz). In the polyp case (Fig. 2a), the maximum energy is located below the first formant in the acoustic frequency axis, close to its fundamental frequency in the modulation frequency axis ($\sim$ 220 Hz). In the case of the speaker with adductor spasmodic dysphonia, we also observe the strong modulations of the first formant by the fundamental frequency ($\sim 230Hz$) of the speaker. However, in this case, there is important energy in a frequency lower than the 1st formant (280 Hz) which is also modulated by the fundamental frequency. For this speaker, there are strong subharmonics. Fig. 2b shows then that there are noticeable modulations (although not as strong as for the fundamental frequency) of the 2nd formant ($\sim 900Hz$) by these subharmonics ($\sim 115Hz$) (see Fig. 2b). Some differences are also observed at larger modulation frequencies, which correspond to the harmonics of these fundamental frequency values (Fig. 2a, 2b and 3b). High energy might appear at modulations lower than $\sim$ 30 Hz, near the first formant as in the case of keratosis (Fig. 3a); there is also high energy beyond the second formant ($\sim$ 1100 Hz) located near the fundamental frequency value in the modulation axis ($\sim$ 134 Hz).

In short, the high resolution of modulation spectral representation yields quite distinctive patterns depending on the type and the severity of voice pathology allowing thus a finer than normal/abnormal distinction. The following section describes the multilinear analysis of modulation frequency features in order to map them to a lower-dimensional domain.

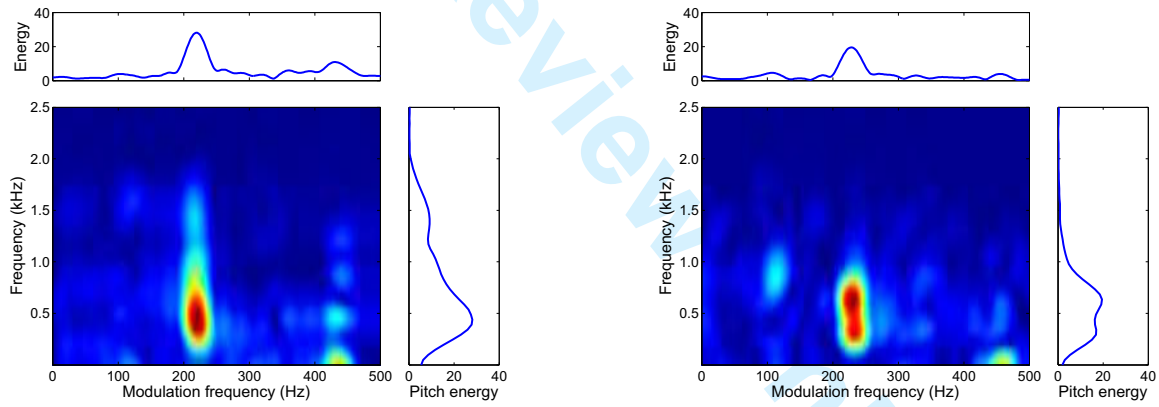
IV. Multilinear Analysis of Modulation Frequency Features

Every signal segment is represented in the acoustic-modulation frequency space as a two-dimensional matrix. Let



(a) (b)

Fig. 1. (a) Modulation spectrogram of sustained vowel /AH/ by a 34 years old normal male speaker ($\sim$ 150 Hz fundamental frequency). The two side plots present the slices intersecting at the point of maximum energy; its coordinates coincide with the fundamental frequency and the first formant of /AH/ ($\sim$ 590 Hz). Vertical plot displays the localization of fundamental frequency energy at vowel formants along the acoustic frequency axis; the upper horizontal plot presents the energy localization of first formant at the fundamental frequency and its harmonics along the modulation frequency axis. (b) Mean values for the modulation spectra of 40 normal speakers from MEEI database [34]. The number of male equals the number of female subjects. All modulation spectra have been normalized to 1 prior to averaging. Upper horizontal plot displays the histogram of fundamental frequency values of male (grey) and female normal speakers (black).



(a) Vocal Polyps (b) Adductor spasmodic dysphonia

Fig. 2. Modulation spectrogram of (a) a 39 years old woman with vocal polyps ($\sim$ 220 Hz fundamental frequency), (b) a 49 years old woman with adductor spasmodic dysphonia ($\sim$ 230 Hz fundamental frequency).

I_3 denote the number of signal segments contained in the training set. Thus, I_3 can be seen as a dimension of time (we recall that I_1 and I_2 correspond to the acoustic and modulation frequency dimensions, respectively). The mean value is then computed over I_3, and it is subtracted from all the modulation spectra in the training set. The zero-mean modulation spectra are then stacked, creating the data tensor $\mathcal{D} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$. A generalization of Singular Value Decomposition (SVD) algorithm to tensors referred to as Higher Order SVD (HOSVD) [32] enables the decomposition of tensor $\mathcal{D}$ to its mode$-n$ singular vectors:

$$\mathcal{D} = \mathcal{S} \times_1 \mathbf{U}_{af} \times_2 \mathbf{U}_{mf} \times_3 \mathbf{U}_s \tag{3}$$

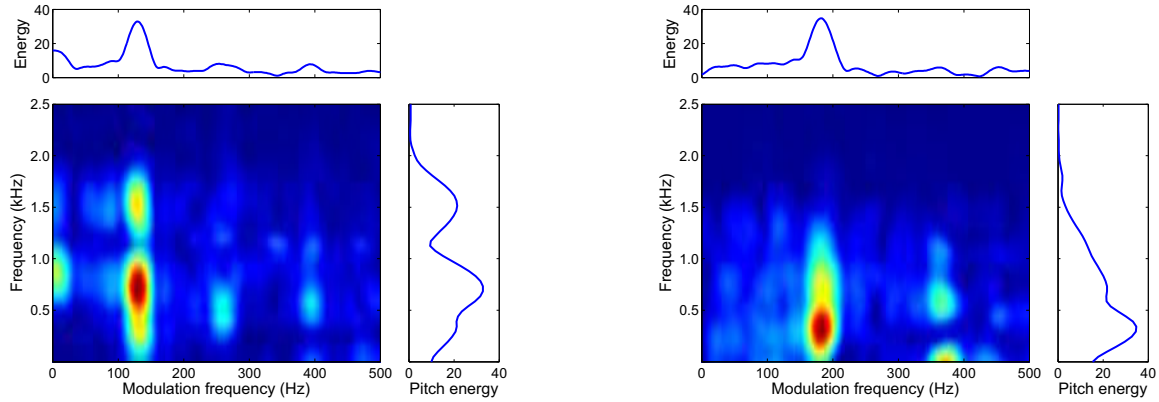
where $\mathcal{S}$ is the core tensor with the same dimensions as $\mathcal{D}$; $\mathcal{S} \times_n \mathbf{U}^{(n)}$, $n = 1, 2, 3$, denotes the $n-$mode product of $\mathcal{S} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ by matrix $\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times I_n}$. For $n = 2$ for example, $\mathcal{S} \times_2 \mathbf{U}^{(2)}$ is an $(I_1 \times I_2 \times I_3)$ tensor given by

$$\left(\mathcal{S} \times_2 \mathbf{U}^{(2)}\right)_{i_1 i_2 i_3} \stackrel{\text{def}}{=} \sum_{i_2} s_{i_1 i_2 i_3} u_{i_2 i_2}. \tag{4}$$

$\mathbf{U}_{af} \in \mathbb{R}^{I_1 \times I_1}$, $\mathbf{U}_{mf} \in \mathbb{R}^{I_2 \times I_2}$ are the unitary matrices of the corresponding subspaces of acoustic and modulation frequencies; $\mathbf{U}_s \in \mathbb{R}^{I_3 \times I_3}$ is the samples subspace matrix. These $(I_n \times I_n)$ matrices $\mathbf{U}^{(n)}$, $n = 1, 2, 3$, contain the n-mode singular vectors (SVs):

$$\mathbf{U}^{(n)} = \left[U_1^{(n)}\ U_2^{(n)}\ \ldots\ U_{I_n}^{(n)}\right]. \tag{5}$$

Each matrix $\mathbf{U}^{(n)}$ can directly be obtained as the matrix of left singular vectors of the "matrix unfolding" $\mathbf{D}_{(n)}$ of $\mathcal{D}$ along the corresponding mode [32]. Tensor $\mathcal{D}$ can be unfolded



(a) Keratosis (b) Vocal Nodules

Fig. 3. Modulation spectrogram of (a) a 50 years old female speaker with keratosis leukoplakia ($\sim$ 135 Hz fundamental frequency). (b) a 38 years old female speaker with vocal nodules ($\sim$ 185 Hz fundamental frequency).

to the $I_1 \times I_2 I_3$ matrix $\mathbf{D}_{(1)}$, the $I_2 \times I_3 I_1$ matrix $\mathbf{D}_{(2)}$, or the $I_3 \times I_1 I_2$ matrix $\mathbf{D}_{(3)}$. The n-mode singular values correspond to the singular values found by the SVD of $\mathbf{D}_{(n)}$.

The contribution $\alpha_{n,j}$ of the j^{th} n-mode singular vector $U_j^{(n)}$ is defined as a function of its singular value $\lambda_{n,j}$:

$$\alpha_{n,j} = \lambda_{n,j} \,/\, \sum_{j=1}^{I_n} \lambda_{n,j} \tag{6}$$

By setting a threshold in the contribution of each singular vector, the R_n with $n = 1, 2$ singular vectors (SVs) can be retained for which the contribution exceeds that threshold. Thus, the truncated matrices $\hat{\mathbf{U}}^{(1)} \equiv \hat{\mathbf{U}}_{af} \in \mathbb{R}^{I_1 \times R_1}$ and $\hat{\mathbf{U}}^{(2)} \equiv \hat{\mathbf{U}}_{mf} \in \mathbb{R}^{I_2 \times R_2}$ are obtained. Joint acoustic and modulation frequencies $\mathbf{B} \equiv |X_l(k,i)| \in \mathbb{R}^{I_1 \times I_2}$ extracted from audio signals are projected on $\hat{\mathbf{U}}_{af}$ and $\hat{\mathbf{U}}_{mf}$ [32]:

$$\mathbf{Z} = \mathbf{B} \times_1 \hat{\mathbf{U}}_{af}^T \times_2 \hat{\mathbf{U}}_{mf}^T = \hat{\mathbf{U}}_{af}^T . \mathbf{B} . \hat{\mathbf{U}}_{mf} \tag{7}$$

where $\mathbf{Z}$ is an $(R_1 \times R_2)-$matrix, and R_1, R_2 denote the number of retained SVs in the acoustic and modulation frequency subspace, respectively.

The modulation spectra can be approximated then in a lower-dimensional space producing a compact feature set suitable for classification. According to the "maximum contribution" criterion, the number of retained components (or SVs) in each subspace can be determined by analyzing the "discriminative" contribution of each component. By including only the components whose contribution is larger than a threshold, we proceed to compute the cross-validation classification error (EER) as a function of this threshold in order to determine the optimal components.

HOSVD addresses features redundancy by selecting mutually independent features. However, these are not necessarily the most discriminative features. Thus we suggest to detect the near-optimal projections of features among the retained singular vectors. Based on mutual information [33], the relevance to the target class of the first R_1 SVs in the acoustic frequency subspace and the first R_2 SVs in the modulation frequency subspace, is examined.

V. Feature Selection Based on Maximum Relevance

The mutual information between two random variables x_i and x_j is defined in terms of their joint probability density function (pdf) $P_{ij}(x_i, x_j)$ and the marginal pdf's $P_i(x_i)$, $P_j(x_j)$. Mutual information (MI) $I[P_{ij}]$ is a natural measure of the inter-dependency between those variables:

$$I[P_{ij}] = \int dx_i \int dx_j P_{ij}(x_i, x_j) \log_2 \left[\frac{P_{ij}(x_i, x_j)}{P_i(x_i) P_j(x_j)} \right] \tag{8}$$

MI is invariant to any invertible transformation of the individual variables [33].

Estimating $I(x_i; x_j)$ from a finite sample requires regularization of $P_{ij}(x_i, x_j)$ [37]. We have simply quantized the continuous space of acoustic features by defining b discrete bins along each axis. An adaptive quantization (variable bin length) is adopted so that the bins are equally populated and the coordinate invariance of the MI is preserved [37]. There is an interaction between the precision of features quantization and the sample size dependence of the MI estimates. The optimal b^* is defined according to a procedure described in [37]: when data are shuffled, mutual information should be near zero for a smaller number of bins ($b < b^*$) while it increases for more bins ($b > b^*$).

The *maximal relevance* (maxRel) feature selection criterion simply selects the most relevant features to the target class c [38]. Relevance is defined as the mutual information $I(x_j; c)$ between feature x_j and class c. 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 selected [38]. Next the cross-validation classification error for an increasing number of these sequential features needs to be computed, in order to determine the optimal size of feature set, m.

VI. Pattern Classification and Performance Analysis

Eight binary classification tasks were defined that exploit the patterns of energy distribution in modulation spectra:

normal vs abnormal phonation, a full pairwise comparison between four voice disorders (vocal polyps, adductor spasmodic dysphonia, keratosis, vocal nodules), and paralysis vs the combined previous four disorders.

Classification performance was computed when vector components were selected based on maximum contribution (maxContrib) (eq.6), or maximum relevance (maxRel) criteria. Pattern classification was carried out using Support Vector Machine (SVM) classifiers. SVM find the optimal boundary that separates two classes maximizing the margin between separating boundary and closest samples to it (support vectors) [39]. In this work, SVMlight [39] with a Radial-Basis-Functions kernel was used. Tests with linear SVM with or without spherical normalization were also conducted. This is a modified stereographic projection recommended before classification of high dimensional vectors using linear SVM [40].

A 4-fold stratified cross-validation was used, which was repeated 40 times. The classifier was trained on the 75% of speakers of both classes, then tested using the remaining 25%. MI estimation using (randomly chosen) 75% of each dataset during 4-fold stratified cross-validation gives almost identical results with MI estimation based on the full dataset. Training and testing was based on 262ms segments; utterance classification was then computed using the median of the decisions over its segments.

The system performance was evaluated using the detection error trade-off curve (DET) between false rejection rate (or miss probability) and false acceptance rate (or false alarm probability) [41]. The rates of each type of errors depend upon the value of a threshold, T. The optimal detection accuracy (DCF_{opt}) occurs when T is set such that the total number of errors is minimized. DCF_{opt} reflects performance at a single operating point on the detection error trade-off (DET) curve. The Equal Error Rate (EER) refers to the point at the DET curve where the false-alarm probability equals the miss probability. DET curves present more accurately than Receiver Operating Characteristic (ROC) curves the performance of the different assessment systems at the low error operating points [41]. We depict representative DET curves, and report on DCF_{opt}, EER, and area under the ROC curve (AUC) for the classification tasks, along with their corresponding 95% confidence intervals. Please note that the curves and measures refer to the average of the 40 runs.

VII. Experiments

A. Database

The database we used was designed to support the evaluation of voice pathology assessment systems; it was developed by Massachusetts Eye and Ear Infirmary Voice and Speech Laboratory and it is referred to as MEEI database [34]. The database contains sustained vowel samples of ~ 3 sec duration from 53 normal talkers and of ~ 1 sec duration from 657 pathological talkers with a wide variety of organic, neurological, traumatic, psychogenic and other voice disorders. The database also includes voice samples of ~ 12 sec duration of the same subjects reading text from "Rainbow passage".

For the first test case, we used the sustained vowel phonations from a subset of MEEI, referred to as MEEI$_{sub}$, first defined in [4]. MEEI$_{sub}$ includes 53 normal and 173 pathological speakers with similar age and sex distributions avoiding therefore any bias by these two factors. Pathological class includes many different voice disorders. Since the ratio of the normal to pathological talkers in MEEI$_{sub}$ (~ 0.3) is quite close to the inverse ratio of the respective vowel durations, the number of segments in each class is close enough: 2240 samples of normal voices, vs 1864 samples of pathological ones. Statistics of this subset of MEEI database are provided in Table I.

TABLE I
Normal and pathological talkers [4]

	Number		Mean age (years)		Standard deviation (years)	
Talkers	Male	Female	Male	Female	Male	Female
Normal	21	32	38.8	34.2	8.5	7.9
Pathological	70	103	41.8	37.4	9.3	8.1

TABLE II
Number and Sex of Patients Included in Medical Diagnosis Categories

Medical diagnosis	No. of males	No. of females	No. of segments
Vocal Nodules	1	19	212
Vocal Polyp	12	8	220
Keratosis	15	11	253
Adductor	3	19	232
Paralysis	35	36	781

For voice disorder discrimination, two different kinds of experiments were performed. The first series of experiments consisted of discrimination between a pair of different pathologies. For comparison purposes, the same subset of pathologies as the one considered in [20] was selected: vocal fold polyp, adductor spasmodic dysphonia, keratosis leukoplakia, and vocal nodules. A full pairwise classification was performed as opposed to [20] where only the binary discrimination of vocal fold polyp against the three other pathologies has been reported. There were 88 such cases in the whole MEEI database; only 49 out of these speakers were included in MEEI$_{sub}$ dataset. There was a co-occurence of two pathologies at the same person in 5 cases, making a total of 83 subjects The last experiment consisted of the discrimination of vocal fold paralysis from all the above mentioned pathologies. There were 71 paralysis cases in MEEI with no co-occurence of the other four disorders (refer to Table II for statistics). These were compared to 71 cases characterized by at least one of the four disorders.

Most of the selected recordings had a sampling rate of 25 kHz; files with a 50 kHz sampling rate were antialias-filtered and downsampled to 25 kHz. Each file was partitioned into 262 ms segments for long-term feature analysis; evenly spaced overlapping segments were extracted every 64 ms similar to [24]. This frame rate can capture the time variation of amplitude modulation patterns evident in each frequency band.

B. Feature Extraction and Classification

Modulation spectra were computed using the Modulation Toolbox [42] throughout all experiments. Wideband modulation frequency analysis was considered so that an adult

speaker's fundamental frequency could be resolved in the modulation frequency axis [24]. Hence, the variables in eq. (2) and (3) were set as following: $M = 25$ samples (1 ms time-shift at 25 kHz sampling frequency), $L = 38$ samples, $I_1 = 257$, and $I_2 = 257$; $h(n)$ and $g(m)$ were a 75-point (or, 3 ms) and 78-point Hamming window, respectively. One uniform modulation frequency vector was produced in each one of the 257 subbands. Due to the 1 ms time-shift (window shift $M = 25$ samples) each modulation frequency vector consisted of 257 (up to π) elements up to 500 Hz.

For the computation of the singular matrices for HOSVD, a random subset of 25 normophonic and 25 dysphonic speakers was selected once. Using 1s from each speaker, and considering segments of 262ms for the computation of modulation spectra, with a shift of 64ms, 12 modulation spectra matrices of dimension 257×257 each, were generated per speaker. Stacking the $12 \times 50 = 600$ modulation spectra matrices for all the speakers in the above subset, produced the data tensor $\mathcal{D} \in \mathbb{R}^{257\times257\times600}$. Before applying HOSVD, the mean value of the tensor was computed and then subtracted from the tensor.

The singular matrices $\mathbf{U}^{(1)} \equiv \mathbf{U}_{af} \in \mathbb{R}^{257\times257}$ and $\mathbf{U}^{(2)} \equiv \mathbf{U}_{mf} \in \mathbb{R}^{257\times257}$ were directly obtained by SVD of the "matrix unfoldings" $\mathbf{D}_{(1)}$ and $\mathbf{D}_{(2)}$ of $\mathcal{D}$ respectively. The singular vectors which exceeded a contribution threshold of 0.2% were retained in each mode (eq. 6), resulting in the truncated singular matrices $\hat{U}_{af} \in \mathbb{R}^{257\times34}$ and $\hat{U}_{mf} \in \mathbb{R}^{257\times34}$. It is worth noting that the above process to compute the truncated singular matrices using HOSVD was performed only once. HOSVD is the most costly process in our system since it consists of the SVD of the two data matrices $\mathbf{D}_{(1)}$ and $\mathbf{D}_{(2)}$, with dimension $N \times k$ each. Note that the computational complexity of SVD transform is $O(Nk^2)$. N is either the acoustic frequency dimension or the modulation frequency dimension; respectively, k is the product of the modulation or the acoustic frequency dimension multiplied by the size of the training dataset (i.e., $k = 257 \times 600$ in this case). The truncated matrices were saved and used for all the detection and classification experiments. Features were projected on $\hat{U}_{af}$ and $\hat{U}_{mf}$ according to eq. (7) resulting in matrices $\mathbf{Z} \in \mathbb{R}^{34\times34}$; these were subsequently reshaped into vectors before MI estimation, feature selection, and SVM classification.

For the data discretization involved in MI estimation, the number of discrete bins along each axis was set to $b^* = 8$ according to the procedure described in [37]. Through a sequential search, the top m features in the descent ordering of $I(x_j; c)$ - i.e., the most relevant features - were selected in every case [38]. We computed the cross-validation classification error (EER) for an increasing number of these sequential features in order to determine the optimal size of feature set, m.

Fig. 4 and 5, present the MI estimates between reduced features and the class variable in the four (out of 8) different classification tasks. In the normal vs pathological case and the polyp vs nodules case, the MI of the most relevant features is ~ 0.35 and ~ 0.3 bits, respectively, and the number of relevant features is small. For polyp/adductor discrimination MI is ~ 0.2 bits whereas for polyp/keratosis discrimination MI is ~ 0.14 bits. For adductor/nodules, adductor/keratwsis and keratwsis/nodules discrimination, the corresponding values of MI are ~ 0.18, ~ 0.25 and ~ 0.28 bits, respectively. However, the MI is significantly lower for the discrimination of paralysis against the other four disorders: its maximum value is only ~ 0.06 bits. This is due to the fact that the non-paralysis signals include several other disorders (four at least) so there is not an homogeneity in the non-paralysis class. Hence, it is very difficult in this case to find optimum features in terms of relevance as in the other binary classification cases. The absolute scale of MI is actually a predictor of the performance of the classification system based on the maximum relevance feature selection scheme as it will be shown next [33].

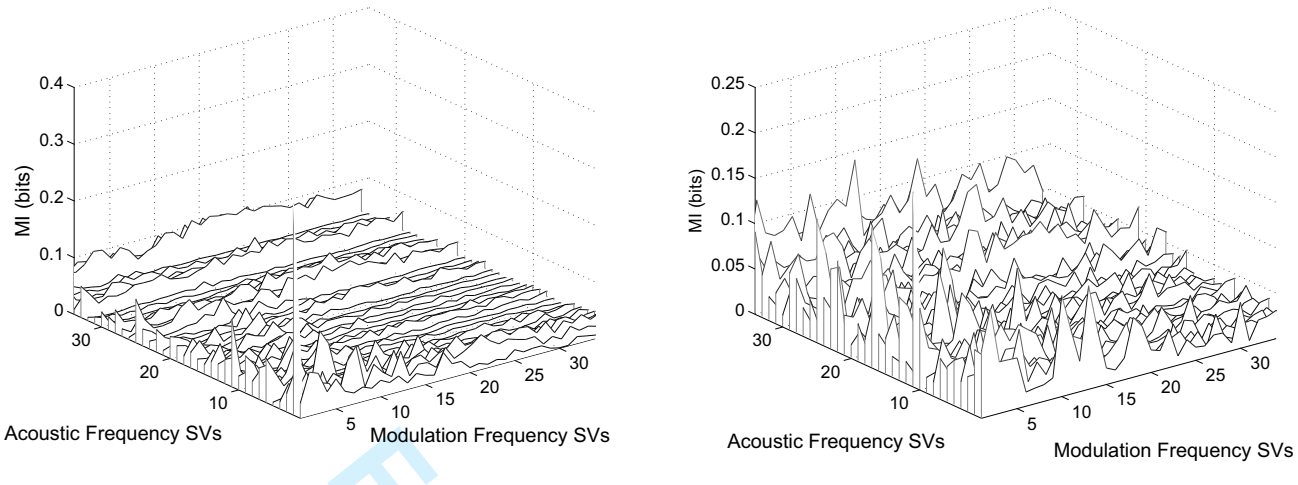
In Table III, we present AUC, DCF_{opt}, and EER for the dysphonia detection task, both for segments and utterances along with their corresponding 95% confidence intervals. For the cases of maximum relevance (maxRel) and maximum contribution criterion (maxContrib), the optimum number of features is also provided in parenthesis. For comparison purposes, we present the performance of another system obtained for utterances on the *same data* based on short term mel-cepstral parameters (defined as in [17]) and the same SVM classifier (denoted as MFCC-SVM in Table III). We also present the AUC and the DCF_{opt} of the system described in Godino et al. [17] based on Gaussian Mixture Models (GMM) and MFCC parameters using approximately the same subset of MEEI (this is denoted as MFCC-GMM in Table III). Although the results reported in [17] are better in terms of AUC, the authors have used a somewhat different cross-validation procedure and have kept 147 pathological signals out of the 173 ones which are included in the MEEI subset used in this work [4].

The best system that was based on maxRel used 20 features whereas the best system based on maxContrib used $[7 \times 13] = 91$ features. In Fig. 6, we compare the performance of the systems using the same SVM classifier in terms of DET curves. The system that has been built on most relevant features is a little superior compared to the other systems, especially in the lower false alarm or miss probability regions.

Similar to normal vs pathological discrimination, for the pathology discrimination task the features were first reduced by projecting them on the singular vectors extracted from the same normal and pathological subjects referred to in Table I. The idea was to improve the generalization ability of our pathology classification system. There were less training vectors during the 4-fold cross-validation in all classification tasks. We also tested both strategies for choosing the suitable levels of detail of this representation: maximum contribution and maximum relevance.

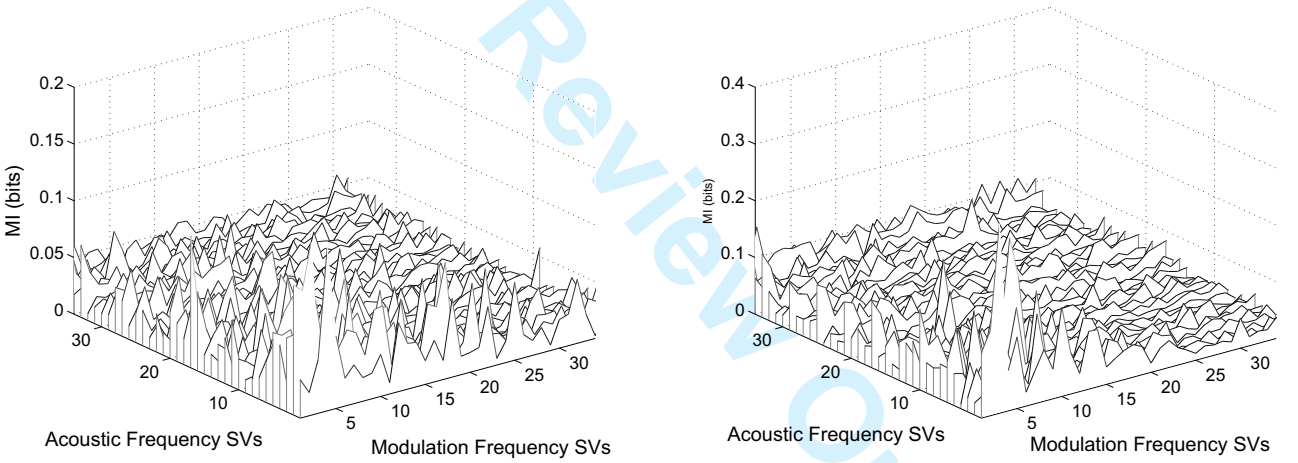
Different kernels and spherical normalization [40] yielded marginal differences in classification performance: in general, results were better using RBF kernel than linear kernel. Spherical normalization enhanced results for linear SVMs and large number of features, but this trend was not observed for RBF kernel.

Tables IV, V, VI provide the classification per pathology scores in terms of AUC, DCF_{opt} and EER and the correspond-



(a) (b)

Fig. 4. Mutual information (MI) values (a) for the normal vs pathological voice classification task; (b) for the polyp vs adductor classification task.



(a) (b)

Fig. 5. Mutual information (MI) values (a) for the polyp vs keratosis classification task; (b) for the polyp vs nodules classification task.

TABLE III

AREA UNDER THE ROC CURVE (AUC), EFFICIENCY (DCF_{opt}) AND EQUAL ERROR RATE (EER) FOR DISCRIMINATION OF NORMAL AND PATHOLOGICAL TALKERS USING MODULATION SPECTRA AND MFCC FEATURES WITH THE SAME SVM CLASSIFIER (95% CONFIDENCE INTERVALS). THE LAST ROW IN THE TABLE REFERS TO THE CORRESPONDING AUC AND DCF_{opt} FOR THE SAME TASK USING MFCC FEATURES AND GMM AS REPORTED IN [17].

	Segment (262ms)			Utterance		
	AUC	DCF_{opt} (%)	EER (%)	AUC	DCF_{opt} (%)	EER (%)
max Relevance (20)	0.9656±0.0032	90.43 ± 0.15	8.63±0.57	0.9775±0.0028	94.08 ± 0.28	6.29±0.67
max Contribution [7×13]	0.9544±0.0036	89.70 ± 0.23	9.18±0.41	0.9633±0.0035	92.67 ± 0.08	7.50±0.28
MFCC-SVM (40)	0.9626±0.0032	89.60±0.41	10.01±0.54	0.9666±0.0029	91.48 ± 0.37	8.47±0.57
MFCC-GMM [17]	-	-	-	0.9997	94.07 ± 1.05	-

ing 95% confidence intervals. For simplicity, only the scores per utterance (or per speaker) are provided. The optimum number of features as this is selected using the maximum relevance or maximum contribution criterion is also presented. For comparison purposes, we report the best discrimination rates (DR) obtained on the *same data* for three classification tasks by Hosseini et al. [20] using SVM on Fisher distance and Genetic Algorithms for feature selection in Table IV (it is denoted as FD-GA). Tables V, VI also present the classification performance of systems based on the standard MFCC features

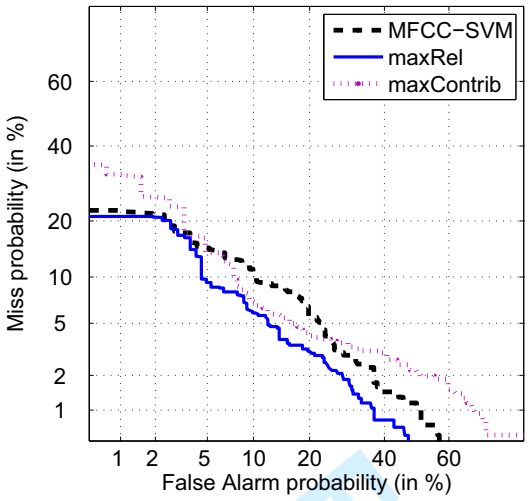


Fig. 6. DET curves for the dysphonia detection system using $[7 \times 13]$ dimensions according to maximum contribution criterion (red dashed), the system based on the 20 most relevant features (blue solid) and MFCC features (black dotted) with the same SVM classifier.

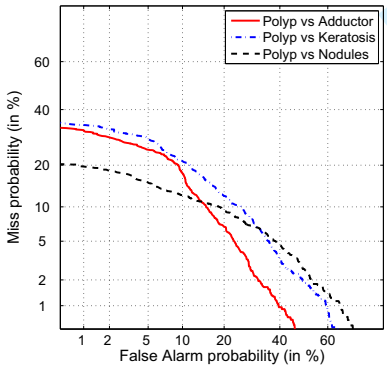


Fig. 7. DET curves with 4-fold cross-validation using modulation spectral features and SVMs for discrimination between polyp/adductor, polyp/keratosis and polyp/nodules cases in MEEI.

and the same SVM classifier for the other four voice pathology discrimination tasks. Fig. 7 presents the DET curves of the system based on most relevant modulation spectral features and SVM for three binary pathology classification tasks.

In every pathology discrimination task, the modulation spectral features were superior to MFCC (see Tables V, VI; the results using MFCC for the tasks in Table IV were not included because of lack of space). Except for the paralysis/non-paralysis case (see Table VI), classification performance was better when we used most relevant (maxRel) features than features with greatest eigenvalue contribution (maxContribution).

As it was noticed before, the absolute scale of MI could almost "predict" the classification performance of the system based on the maximum relevance feature selection scheme [33]. The MI was significantly lower for the discrimination of paralysis against the other four disorders: its maximum value was only ~ 0.06 bits. There is a trade-off between features relevance and features redundancy in each feature selection technique [38]. When the relevance of individual features towards a classification task is very low then, the minimal redundancy (or, "maximal contribution") criterion obviously prevails. The best EER in the paralysis / non-paralysis discrimination task was $15.45 \pm 0.56\%$ using the $[8 \times 15$ components with maximum contribution vs $27.99 \pm 0.81\%$ (95% confidence intervals) using the 200 most relevant modulation spectral features (Table VI). For comparison, the authors in [21] reported an EER of $\sim 30\%$ for the discrimination of paralysis from other voice disorders in MEEI (binary task) based on amplitude modulation features.

VIII. Discussion and Conclusions

We have evaluated features of the modulation spectrogram of sustained vowel /AH/ for voice pathology detection and classification. Our results show that modulation spectral features are well suited to voice pathology assessment and discrimination tasks.

In order to extract a compact set of features out of this multidimensional representation, we first removed "redundancy" at the first step of our processing, using HOSVD. HOSVD was performed on the same dataset of normal and pathological talkers for all classification tasks. The efficiency scores for pathologies discrimination would be better if we had performed HOSVD on pathological samples only. Still we wanted to build a system that could proceed from normal vs pathological discrimination to voice disorder classification, based on features projected on the same principal axes. Features relevance to each task was assessed based on MI estimation.

Classification experiments with MEEI database [34] confirmed that the absolute scale of MI can indeed "predict" the performance of the system based on the maximum relevance feature selection scheme [33]. There is a trade-off between features relevance and features redundancy in each feature selection technique [38]. When the relevance of individual features towards a classification task is very low then, the minimal redundancy (or, "maximal contribution") criterion obviously prevails. Hence in the last classification task (paralysis/non-paralysis), the maximum contribution features outperformed the maximum relevance features.

It was shown in [30] that Modulation Spectra can appropriately normalized in order to successfully address the detection of dysphonic voices in new, unseen, databases. However, Normalized Modulation Spectra have not been applied yet to the task of disorders classification for new databases. Currently we are looking for a new database with enough examples from each disorder in order to conduct experiments with Normalized Modulation Spectra. A very important problem in voice disorders is the quantification of the degree of voice pathology (i.e., degree of breathiness, roughness and hoarseness). The results presented in [43] using modulation spectra for quantifying hoarseness were very encouraging. As a future plan, we would like to quantify the degree of voice pathology for the other cases too, but using more databases that the one used in [43]. Moreover, regarding future plans, analysis of continuous speech samples could be used instead of sustained vowels. Acoustic features derived from continuous speech provide

TABLE IV

AREA UNDER THE ROC CURVE (AUC), EFFICIENCY (DCF_{opt}) AND EQUAL ERROR RATE (EER) PER DISORDER USING MODULATION SPECTRAL FEATURES AND SVM (95% CONFIDENCE INTERVALS). THE CORRESPONDING BEST DISCRIMINATION RATES FOR THE SAME TASKS USING FD-GA [20] ARE LISTED IN THE LAST COLUMN OF THE TABLE.

	max Relevance			max Contribution			FD-GA [20]
	AUC	DCF_{opt} (%)	EER (%)	AUC	DCF_{opt} (%)	EER (%)	DCF_{opt} (%)
Polyp / Adductor	0.9585±0.0087	91.23±1.10 (60)	8.25±1.3792	0.9309±0.0084	81.23±2.11 [17 × 24]	13.25±1.32	82.5
Polyp / Keratosis	0.9359±0.0058	84.77±1.42 (80)	15.71±1.0848	0.6279±0.0204	57.26±0.25 [17 × 24]	40.25±1.85	81.8
Polyp / Nodules	0.9428±0.0073	91.66 ± 1.14 (20)	11.25±1.2064	0.8802±0.0127	86.03 ± 1.50 [6 × 10]	16.44±1.29	87.5

TABLE V

AREA UNDER THE ROC CURVE (AUC), EFFICIENCY (DCF_{opt}) AND EQUAL ERROR RATE (EER) FOR DISCRIMINATION OF DIFFERENT KIND OF DYSPHONIAS USING MODULATION SPECTRAL FEATURES AND MFCC FEATURES WITH THE SAME SVM CLASSIFIER (95% CONFIDENCE INTERVALS).

	Adductor / Nodules			Adductor / Keratosis		
	AUC	DCF_{opt} (%)	EER (%)	AUC	DCF_{opt} (%)	EER (%)
max Relevance	0.9578±0.0064	92.09±0.92 (95)	8.44±1.09	0.9949±0.0017	95.77±0.92 (90)	2.17±0.62
max Contribution	0.7981±0.0147	75.21±1.53 12x19	25.46±1.30	0.8844±0.0113	72.15±1.19 12x20	17.21±1.51
MFCC	0.6728±0.0147	63.91±1.03 (20)	37.12±1.32	0.7188±0.0123	66.65±1.81 (20)	36.70±1.47

TABLE VI

AREA UNDER THE ROC CURVE (AUC), EFFICIENCY (DCF_{opt}) AND EQUAL ERROR RATE (EER) FOR DISCRIMINATION OF DIFFERENT KIND OF DYSPHONIAS USING MODULATION SPECTRAL FEATURES AND MFCC FEATURES WITH THE SAME SVM CLASSIFIER (95% CONFIDENCE INTERVALS).

	Keratosis / Nodules			Paralysis / Other		
	AUC	DCF_{opt} (%)	EER (%)	AUC	DCF_{opt} (%)	EER (%)
max Relevance	0.9527±0.0053	89.11± 1.33 (97)	13.21±1.18	0.7648±0.0078	70.09±1.05 (200)	27.99±0.81
max Contribution	0.9265±0.0106	86.59± 0.45 [12 × 20]	15.23±1.57	0.9063±0.0052	82.14±0.85 [8 × 15]	15.45±0.56
MFCC	0.7286±0.0183	67.88±0.91 (20)	31.39±1.77	0.6504±0.0081	64.02±0.62 (60)	38.68±0.65

information about the voice source, vocal tract and articulators, shedding light on more aspects of a pathological voice quality. In that case, we expect that higher (acoustic) frequency bands in the modulation spectra would also contain highly discriminating patterns for vocal pathologies assessment. Different Time-Frequency (TF) distributions could also be used in the first stage of modulation frequency analysis instead of the STFT spectrogram, offering better resolution [35]. Also, alternative time-frequency transformations, such as decomposition based approaches, proposed in a previous study [19], could also used.

REFERENCES

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

[2] S. Davis, "Computer evaluation of laryngeal pathology based on inverse filtering of speech," SCRL Monograph Number 13, Speech Communications Research Laboratory, Santa Barbara, CA, 1976.

[3] R. Prosek, A. Montgomery, B. Walden, and D. 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. Jamieson, "Identification of pathological voices using glottal noise measures," *J. Speech, Language, Hearing Res.*, vol. 43, no. 2, pp. 469–485, Apr. 2000.

[5] A. Fourcin and E. Abberton, "Hearing and phonetic criteria in voice measurement: Clinical applications," *Logopedics Phoniatrics Vocology*, pp. 1–14, Apr. 2007.

[6] M. P. T. Quatieri and D. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 569–587, 1999.

[7] M. Rosa, J.C.Pereira, and M.Grellet, "Adaptive estimation of residue signal for voice pathology diagnosis," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 1, pp. 96–104, Jan 2000.

[8] S. Marple, *Digital spectral analysis with applications*. NJ: Prentice-Hall, 1987.

[9] Y. Zhang, C. McGilligan, L. Zhou, M. Vig, and J. Jiang, "Nonlinear dynamic analysis of voices before and after surgical excision of vocal polyps," *Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2270–2277, 2004.

[10] A. Giovanni, M. Ouaknine, and J. Triglia, "Determination of largest Lyapunov exponents of vocal signal: Application to unilateral laryngeal paralysis," *Journal of Voice*, vol. 13(3), pp. 341–354, 1999.

[11] J. Alonso, J. de Leon, I. Alonso, and M. Ferrer, "Automatic detection of pathologies in the voice by hos based parameters," *Journal on Applied Signal Processing*, vol. 4, pp. 275–284, 2001.

[12] M. Little, P. McSharry, S. Roberts, D. Costello, and I.M.Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *BioMedical Engineering, Published online, doi:10.1186/1475-925X-6-23*, Jun. 2007.

[13] J. Deller, J. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*. NY: McMillan, 1993.

[14] A. Askenfelt and B. Hammarberg, "Speech waveform perturbation analysis revisited," *Speech Transmission Laboratory - Quartely Progress and Status Report*, vol. 22, no. 4, pp. 49–68, 1981.

[15] A.A.Dibazar and S.S.Narayanan, "A system for automatic detection of pathological speech," in *36th Asilomar Conf. Signal, Systems, and Computers*, Asilomar, CA, USA, Oct. 2002.

[16] A.A.Dibazar, T.W.Berger, and S.S.Narayanan, "Pathological voice asess-

ment," in *IEEE, 28th Eng. in Med. and Biol. Soc.*, NY, NY, USA, Aug. 2006, pp. 1669–1673.

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

[18] J. Godino-Llorente and P. Gómez-Vilda, "Automatic detection of voice impairments by means of short-time cepstral parameters and neural network-based detectors," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 2, pp. 380–384, Feb. 2004.

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

[20] P. Hosseini, F. Almasganj, T. Emami, R. Behroozmand, S. Gharibrade, and F. Torabinezhad, "Local discriminant wavelet packet basis for voice pathology classification," in *2nd Intern. Conf. on Bioinformatics and Biomedial Eng. (ICBBE)*, May 2008, pp. 2052–2055.

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

[22] H. Hermansky, "Should recognizers have ears?" *Speech Communication*, vol. 25, pp. 3–27, Aug. 1998.

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

[24] S. Schimmel, L. Atlas, and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," in *Proc. ICASSP*, vol. 4, 2007, pp. 605–608.

[25] S. Greenberg and B. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," in *Proc. ICASSP*, vol. 3, 1997, pp. 1647–1650.

[26] T. Kinnunen, "Joint acoustic-modulation frequency for speaker recognition," in *Proc. ICASSP*, vol. 1, 2006, pp. 665–668.

[27] S. Sukittanon, L. Atlas, and J. Pitton, "Modulation-scale analysis for content identification," *IEEE Trans. Speech Audio Process.*, vol. 52, no. 10, pp. 3023–3035, 2004.

[28] M. Markaki and Y. Stylianou, "Dimensionality reduction of modulation frequency features for speech discrimination," in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 646–649.

[29] ——, "Using modulation spectra for voice pathology detection and classification," in *Proceedings of IEEE EMBC'09*, Minneapolis, Minnesota, U.S.A., 2009.

[30] ——, "Normalized modulation spectral features for cross-database voice pathology detection," in *Proc. Interspeech*, Brighton, U.K., 2009.

[31] T. Kinnunen, K. Lee, and H. Li, "Dimension reduction of the modulation spectrogram for speaker verification," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2008.

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

[33] T. Cover and J. Thomas, *Elements of Information Theory*. New York: John Wiley and Sons, 1991.

[34] M. Eye and E. Infirmary, "Elemetrics Disordered Voice Database (Version 1.03)," Voice and Speech Lab, Boston, MA, Oct. 1994, kay Elemetrics Corp.

[35] L. Cohen, *Time-Frequency Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[36] M. Vieira, F. MClnnes, and M. Jack, "On the influence of laryngeal pathologies on acoustic and electroglottographic jitter measures," *J.A.S.A.*, vol. 111, no. 2, pp. 1045–1055, 2002.

[37] N. Slonim, G. Atwal, G. Tkacik, and W. Bialek, "Estimating mutual information and multi-information in large networks," 2005. [Online]. Available: http://arxiv.org/abs/cs.IT/0502017

[38] 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 Anal. Mach. Intell.*, vol. 27, pp. 1226–1238, 2005.

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

[40] V. Wan and S. Renals, "Speaker verification using sequence discriminant support vector machines," *IEEE Trans. Audio, Speech and Language Proc.*, vol. 13, no. 2, pp. 203–210, 2005.

[41] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech '97*, vol. IV, 1997, pp. 1895–1898.

[42] "Modulation toolbox." [Online]. Available: http://www.ee.washington.edu/research/isdl/projects/modulationtoolbox

[43] M. Markaki and Y. Stylianou, "Modulation spectral features for objective voice quality assessment," in *Proc. IEEE ISCCSP*, Limassol, Cyprus, 2010.