31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Using Modulation Spectra for Voice Pathology Detection and Classification

Maria Markaki and Yannis Stylianou

Abstract— **In this paper, we consider the use of Modulation Spectra for voice pathology detection and classification. To reduce the high-dimensionality space generated by Modulation spectra we suggest the use of Higher Order Singular Value Decomposition (SVD) and we propose a feature selection algorithm based on the Mutual Information between subjective voice quality and computed features. Using SVM with a radial basis function (RBF) kernel as classifier, we conducted experiments on a database of sustained vowel recordings from healthy and pathological voices. For voice pathology detection, the suggested approach achieved a detection rate of 94.1% and an Area Under the Curve (AUC) score of 97.8%. For voice pathology classification, an average detection rate and AUC of 88.6% and 94.8%, respectively, was achieved in classifying polyp against keratosis leukoplakia, adductor spasmodic dysphonia and vocal nodules.**

I. INTRODUCTION

Many studies in voice function assessment try to identify acoustic measures or cues that highly correlate with pathological voice qualities (also referred to as voice alterations). Organic pathologies that affect vocal folds usually modify their morphology resulting in abnormal vibration patterns and increased turbulent airflow at the level of the glottis [6]. Examples of acoustic parameters trying to quantify the glottal noise include pitch, jitter, shimmer, amplitude perturbation quotient (APQ), pitch perturbation quotient (PPQ) and glottal to noise excitation (GNE)([7], [22], [18] and references within).

Some of the suggested features require accurate estimation of fundamental frequency, which is not a trivial task in the case of certain 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) [8] or by inverse filtering of speech [21] [23] where an estimate of the excitation waveform to the vocal tract model 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) [7]. SFF and SFR can be considered as a measure of the noise masking formants and harmonics, respectively [22].

The estimation of the glottal signal or the residual signal (excitation to the vocal tract) or measurements of glottal activity (for example, by using EGG) are quite questionable. Assuming that speech signal is produced based on linear systems theory, then it is expected that perturbations at the glottal level will affect the spectral properties of the recorded speech signal. In this case, an estimation of the glottal signal can be avoided. Then, however, another difficult problem is created; that of features identification in the speech signal, which reflect the activity of the glottal source. There have been suggested both parametric and non parametric approaches for this, and in general, these approaches can be referred to 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 [3]. The non parametric approaches are based on magnitude spectrum of speech where short-term mel frequency cepstral coefficients (MFFC) are widely used in representing the magnitude spectrum in a compact way [1] [2] [9]. 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 [22], or likelihood scores using Gaussian Mixture Models (GMM) [1] [9] and Hidden Markov Models (HMM) [2]. Also neural networks and Support Vector Machines based classifiers have been suggested [10] [12].

There have been a few approaches 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 [22]. In [23] after applying an iterative residual signal estimator features like jitter have been computed. Jitter provided the best classification score between pathologies (54.8%)(21 pathologies). In [2], an HMM approach, using MFCC, provided an average score of correct classification of 70% (5 pathologies). A recent study for discrimination of voice pathologies was carried out via adaptive growth of Wavelet Packet tree, based on the criterion of Local Discriminant Bases (LDB) [12]. A genetic algorithm was employed to select the best feature set and then a Support Vector Machines (SVM) based classifier was used.

In this work we suggest the use of modulation spectra for detection of voice pathologies [11], [5]. Modulation spectral features have been employed for single-channel speaker separation [4], as well as for speech and speaker recognition [13]. Modulation spectra may be seen as a non-parametric way to represent the modulations in speech. Moreover, it offers an implicit way to fuse the various

M. Markaki and Y. Stylianou are with the Department of Computer Science, University of Crete, 71409 Crete, Greece mmarkaki,yannis@csd.uoc.gr

Y. Stylianou is with the Institute of Computer Science, FORTH, Crete, Greece styliano@ics.forth.gr

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

phenomena observed during speech production, in a compact way. Still, modulation spectra contain a large amount of features, posing serious problems for the classification algorithms. The initial representation is first transformed to a lower-dimensional domain using Higher Order SVD. For features relevant to pathology selection, the mutual information between voice quality and features mapped in the transformed domain, is estimated. Specifically we use the *maximal relevance* (MaxRel) feature selection criterion which simply selects the features most relevant to a target class. Projection of the relevant features back to the original space reveals the modulation spectral components which can discriminate normal from abnormal voices, or different voice pathologies. Simulations carried out on MEEI database [17] show that the modulation spectral features are useful in assessing vocal impairment as well as making finer classifications than the normal versus abnormal classification. In the following, we will refer to our method as Modulation Spectra and Maximal Relevance (MSMR) method.

II. Modulation Spectra and Maximal Relevance - MSMR

A. Modulation Spectra

The most common modulation frequency analysis framework [4] for a discrete signal $x(n)$, initially employs a short-time Fourier transform (STFT) $X_k(m)$

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

where $W_K = e^{-j(2\pi/K)}$ and $h(n)$ is the acoustic frequency analysis window with a hop size of M samples. 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:

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

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 Fig. 1 modulation spectrogram of a 262 ms long frame from a normal male speaker from MEEI database [17] is shown. We can observe the modulation frequency localization of strongest formant and the acoustic frequency localization of pitch energy.

In Fig. 2 two examples of voice pathologies are shown for (a) a male speaker with vocal polyps and (b) for a woman with adductor spasmodic dysphonia.

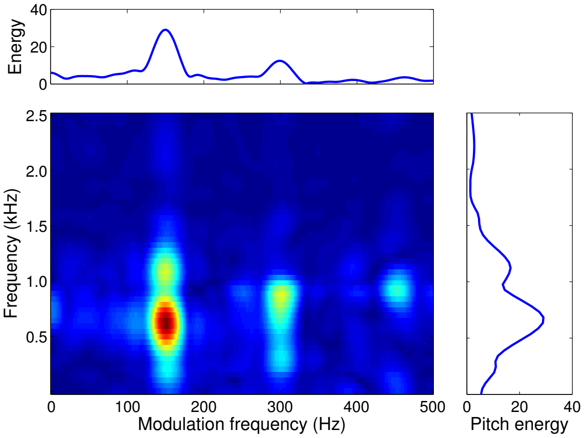


Fig. 1. Modulation spectrogram of sustained vowel /AH/ by a 34 years old normal male speaker.

B. Dimensionality reduction

Modulation spectra were computed in a frame-by-frame basis using relatively long windows in time (262 ms) which were overlapping. Each modulation spectrum consisted of $I_1 = 257$ acoustic frequencies and $I_2 = 257$ modulation frequencies, resulting therefore in a 257×257 image per frame. The modulation spectra computed in each frame were mean subtracted and then, they were stacked to produce a a third order tensor $\mathscr{D} \in R^{I_1 \times I_2 \times I_3}$, where I_3 is the number of frames in the training dataset.

We used a generalization of SVD to tensors referred to as Higher Order SVD (HOSVD) [15] which enables the decomposition of tensor $\mathscr{D}$ to its $n-$mode singular vectors (or, principal components). Ordering of these $n-$mode singular values implies that the “energy” of tensor $\mathscr{D}$ 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. The contribution of the j^{th} principal component (PC) of the acoustic or modulation frequency-subspace S_i whose corresponding eigenvalue is $\lambda_{i,j}$, is defined as:

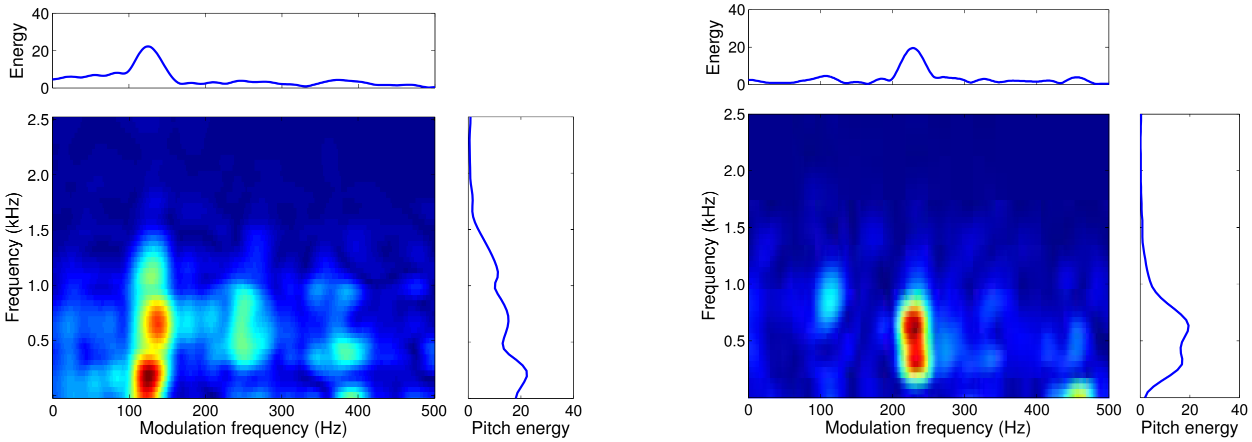
$$\alpha_{i,j} = \frac{\lambda_{i,j}}{\sum_{j=1}^{N_i} \lambda_{i,j}} \qquad (3)$$

where N_i is the dimension of S_i - 257 for acoustic frequency and 257 for modulation frequency.

Next, we detected the near-optimal projections (PCs) of features among those contributing more than 0.2% to the “energy” of $\mathscr{D}$. That is, we examined the relevance to the target class of the first 34 PCs in the acoustic frequency and the first 34 PCs in the modulation frequency subspace.

C. Feature Selection based on Mutual Information

We selected among the $34 \times 34 = 1156$ features the ones which were more relevant to a given classification task using mutual information (MI). Specifically we used the *maximal relevance* (MaxRel) [20] feature selection criterion which



(a) Vocal Polyp

(b) Adductor spasmodic dysphonia

Fig. 2. Modulation spectrogram on sustained vowel /AH/ of : (a) a 33 years old male speaker with vocal polyps (∼ 129 Hz fundamental frequency), (b) a 49 years old woman with adductor spasmodic dysphonia - the maximum is not located at the pitch value in the second case (∼ 113.6 Hz fundamental frequency).

simply selects the features most relevant to the target class c. Relevance is usually 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 $I(x_j;c)$ were selected in every case.

III. Results

We have evaluated features of the modulation spectrogram of sustained vowel /AH/ from MEEI, for voice pathology detection and classification tasks. For the pathology detection experiments, a subset of the database (53 normophonic, 173 dysphonic speakers) was used in order to cover as many as possible disorders while at the same time the normophonic and dysphonic classes to have similar age and sex distributions [19]. For voice pathology classification, we selected from the whole MEEI database the same subset of pathologies as the one used in [12]: vocal fold polyp, adductor spasmodic dysphonia, keratosis leukoplakia, and vocal nodules. There were 88 such cases in the database. Five persons exhibited two of the above pathologies at the same time and they were excluded. All the tests were conducted on signals sampled at 25 kHz. For classifier, we used SVM with a radial basis function (RBF) kernel. We used 4-fold stratified cross-validation, repeated 400 times. The classifier was trained on the 75% of normal and pathological speakers then tested using the rest 25% and provided a decision per segment. Then, for utterance classification the median of the decisions over its segments was computed.

For evaluation, we used detection error trade-off (DET) curves since DET curves present more accurately than Receiver Operating Characteristic (ROC) curves the performance of the different assessment systems at the low error operating points [16]. The optimal detection accuracy (DCF_{opt}) occurs when the sum of Type I and Type II errors is minimum. For the voice pathology detection task, we achieved a detection rate $DCF_{opt} = 94.08\%$ (± 0.86) using the $m = 25$ most relevant features (which corresponds to an Area Under the Curve (AUC), when using ROC curves, of $AUC = 97.75\%$; see Table I).

TABLE I

Classification scores per task using m features selected by the MaxRel criterion for the suggested approach (MSMR). The corresponding best discrimination rates on the same datasets referred to in the literature are listed in the last column of the table.

	MSMR			DR (%)
	DCF_{opt} (%)	AUC (%)	m	
Normal/Pathol	94.08 ± 0.86	97.75	25	94.07 ± 3.28 [9]
Polyp/Adductor	88.33 ± 2.64	95.74	60	82.5 [12]
Polyp/Keratosis	86.11 ± 5.52	93.61	80	81.8 [12]
Polyp/Nodules	91.25 ± 3.13	95.03	20	87.5 [12]

In addition, Table I presents the classification per pathology scores in terms of DCF_{opt} and AUC in percent. Specifically we show the results for classification of polyp against keratosis leukoplakia, adductor spasmodic dysphonia and vocal nodules. We also provide the standard deviation for DCF_{opt} and the optimum number of features as this is selected using the MaxRel criterion. For comparison purposes, Table I shows the best discrimination rates (DR) obtained on each task using the *same data* by Godino-Llorente et al. [9] and Hosseini et al. [12], respectively. In [9] the authors use Gaussian mixture models and short-term mel cepstral parameters for pathological voice quality assessment. In [12] the voice pathology classification system is based on local discriminant wavelet packet basis; a Genetic Algorithm is employed for feature selection and a SVM with a RBF kernel as classifier.

IV. Conclusions

We suggested the use of Modulation Spectra for voice pathology detection and voice pathology classification. Furthermore, we suggested a Maximal Relevance feature selection algorithm based on the mutual information between subjective voice quality and measured features. Using recordings of sustained vowels from MEEI database showed that the modulation spectral features are useful in assessing vocal impairment as well as making finer classifications than the normal versus abnormal classification. An average score of over 90% for voice pathology classification was achieved in classifying polyp against keratosis leukoplakia, adductor spasmodic dysphonia and vocal nodules.

References

[1] 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, 2002.

[2] A.A.Dibazar, T.W.Berger and S.S.Narayanan, ”Pathological Voice Asessment”, *in IEEE, 28th Eng. in Med. and Biol. Soc.*, NY, USA, 2006, pp 1669-1673.

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

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

[5] L. Atlas and S.A. Shamma, Joint Acoustic and Modulation Frequency, *EURASIP Journal on Applied Signal Processing*, vol. 7, 2003, pp 668-675.

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

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

[8] A. Fourcin and E. Abberton, Hearing and phonetic criteria in voice measurement: Clinical applications, *Logopedics Phoniatrics Vocology*, April 2007, pp 1-14.

[9] J.I. 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, 10, 2006, pp 1943-1953.

[10] J.I. 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, 2, 2004, pp 380-384.

[11] H. Hermansky, Should recognizers have ears?, *Speech Communication*, vol. 25, 1998, pp 3-27.

[12] P.T. 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.

[13] T. Kinunnen, ”Joint acoustic-modulation frequency for speaker recognition”, *in Proc. ICASSP*, vol. 1, 2006, pp. 665-668.

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

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

[16] A. Martin, G.R. Doddington, T. Kamm, M. Ordowski and M. Przybocki, ”The DET curve in assessment of detection task performance”, *Proc. Eurospeech ’97*, vol. IV, 1997, pp. 1895-1898.

[17] Kay Elemetrics, *Elemetrics Disordered Voice Database (Version 1.03)*, 1994.

[18] V. Parsa and D.G. Jamieson, Identification of Pathological Voices using Glottal Noise measures, *J. Speech, Language, Hearing Res.*, vol. 43, 2, 2000, pp 469-485.

[19] K. Umapathy, S. Krishnan, V. Parsa and D.G. Jamieson, Discrimination of pathological voices using time-frequency approach, *IEEE Trans. Biomed. Eng.*, vol. 52, 3, 2005, pp 421-430.

[20] H. Peng and 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, 2005, pp 1226-1238.

[21] M.D. Plumbe, T.F. Quatieri and D.A. Reynolds, Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification, *IEEE Trans. Speech Audio Process.*, vol. 7, 1999, pp 569-587.

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

[23] 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, 2000, pp 96-104.