Normalized Modulation Spectral Features for Cross-Database Voice Pathology Detection

Maria Markaki[1], *Yannis Stylianou*[1,2]

[1]Computer Science Dept, University of Crete, Greece
[2]Institute of Computer Science, FORTH, Crete, Greece
mmarkaki@csd.uoc.gr, yannis@csd.uoc.gr

Abstract

In this paper, we employ normalized modulation spectral analysis for voice pathology detection. Such normalization is important when there is a mismatch between training and testing conditions, or in other words, employing the detection system in real (testing) conditions. 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 subjective voice quality and computed features, which leads to an adaptive to the classification task modulation spectra representation. For voice pathology detection, the adaptive modulation spectra is combined with an SVM classifier. To simulate the real testing conditions; one for training and the other for testing. We address the difference of signal characteristics between training and testing data through subband normalization of modulation spectral features. Simulations show that feature normalization enables the cross-database detection of pathological voices even when training and test data are different.

Index Terms: pathologic voice detection, modulation spectrum, feature normalization, mutual information, SVD.

1. Introduction

Many studies in voice function assessment try to identify acoustic measures or cues that highly correlate with pathological voice qualities. Organic pathologies modify the morphology of vocal folds resulting in abnormal vibration patterns and increased turbulent airflow at the level of the glottis [1]. 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)[2] [3] [4]. 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].

Perturbations at the glottal level will also affect the spectral properties of the recorded speech signal. There are both parametric and non parametric approaches for identifying the abnormal glottal activity based on analysis of speech signals. The parametric approaches are based on the source filter theory for the speech production and on the assumptions made for the glottal signal [7]. 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 [8, 9]. Non parametric approaches also include time-frequency representations [10].

Modulation spectra may be seen as a non-parametric way to represent the frequency-band-dependent amplitude modulations in speech [11, 12]. In our recent works we suggested the use of modulation spectra for speech detection [13] and for detection and classification of voice pathologies [14]. Modulation spectral analysis produces a high-dimensional feature space, which is inconvenient for detection or classification purposes. In [13] the initial high dimensional representation was first transformed to a lower-dimensional space using Higher Order SVD [15]. To further enhance the lower dimensional space taking into account the classification task, the mutual information between the features and the class variable was measured [13]. This is usually referred to as feature selection [16] and it leads to an adaptive to the classification task modulation spectrum representation. In [14], this representation was tested on voice pathology detection and classification using sustained vowels recordings from the Massachusetts Eye and Ear Infirmary (MEEI) Disordered Voice Database [17]. Using a support vector machine (SVM) classifier, it was shown that a high classification performance can be obtained; specifically, a detection rate of 94.1% and an Area Under the Curve (AUC) of 97.8% was achieved for voice pathology detection [14].

The above detection results were obtained using a 4-fold stratified cross-validation scheme repeated 40 times. It is then interesting to check the performance of the trained detector on unknown (completely unseen) data, in the sense that these data are not just part (of the testing set) of the initial database. Unseen data may have been recorded under different conditions and independently from those of the initial database which was used for training. For this purpose, we used a second 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 [18]. Similar to MEEI, PdA contains recordings of sustained vowels (/a/) and was developed for voice function assessment purposes. Testing the optimal detector which was developed on MEEI on recordings from PdA, we found that the performance of the detector was significantly decreased.

Apparently, this degradation was caused by the difference of the environmental characteristics - channel transmission effects, noises, etc. - of the training and testing data. Past research has addressed the sensitivity of features to data mismatch with feature normalization [19]. Feature normalization scales or warps the components of the fixed feature vector in order to make both training and testing features independent of environmental characteristics. In this work, and as a first step towards a robust voice pathology detector, we implement subband normalization of modulation spectral features that makes them insensitive to time and frequency distortions according to [20]. After a brief overview of modulation spectral analysis in Sec-

tion 2, we describe the normalization we employ and its effects in Section 3. We validate our approach with cross-database detection experiments in Section 4 and we provide conclusions and future directions in Section 5.

2. Modulation Spectra

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

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

with $k = 0, \ldots, K-1$, where $W_K = e^{-j(2\pi/K)}$ and $h(n)$ is the acoustic frequency analysis window with a hopsize 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 $|X_k(m)|$ of the subband - is performed next by computing a second STFT:

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

with $i = 0, \ldots, I-1$, and where $g(m)$ is the modulation frequency analysis window; 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.

Modulation spectra are computed in a frame-by-frame basis using relatively long windows in time (262 ms). This provides one matrix per frame of $I_1 \times I_2$, where I_1 and I_2 denote the acoustic and modulation frequencies, respectively. The modulation spectra computed in each frame are stacked to produce a tensor $\mathcal{D}$. Matrix representation of a third order tensor $\mathcal{D} \in R^{I_1 \times I_2 \times I_3}$, where I_3 is the time dimension, is particularly useful for computations, however it contains a large amount of features, posing serious problems for the classification algorithms. We use Higher Order SVD (HOSVD) in order to decompose tensor $\mathcal{D}$ to its $n-$mode singular vectors [13]. Ordering of the $n-$mode singular values implies that the "energy" of tensor $\mathcal{D}$ is concentrated in the singular vectors with the lowest indices. Each singular matrix containing the $n-$mode singular vectors, can be truncated then by setting a predetermined threshold so as to retain only the desired number of principal axes in each mode. Cross-validation permits to determine an optimal energy threshold for classification.

After reducing dimensions, we select features which are more relevant to a given classification task using mutual information (MI). Specifically we used the *maximal relevance* (MaxRel) feature selection criterion which simply selects the features most relevant to the target class c [13]. 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.

3. 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 (2):

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

where $\{\bar{\cdot}\}$ 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:

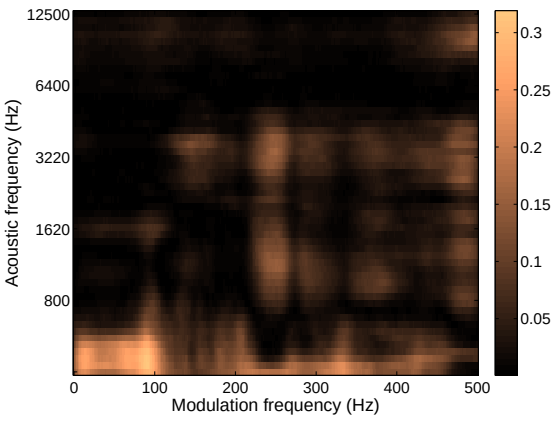
$$X_{l,sub}(k,i) \quad = \quad \frac{X_l(k,i)}{\sum_i X_l(k,i)} \qquad (4)$$

Previous work [20] has shown that this subband normalization makes modulation spectral features insensitive to convolutional noise and time distortions such as time scaling and shifting.

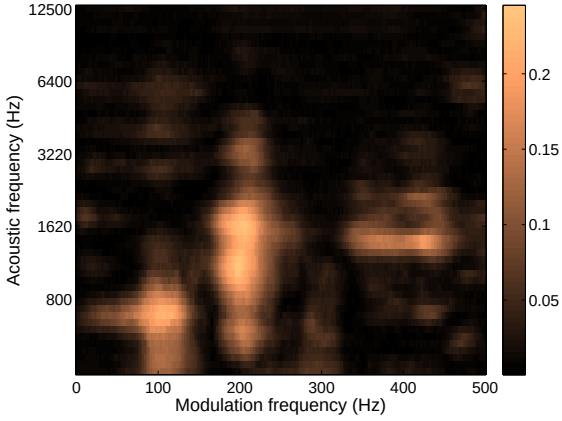
As a first test of the normalization effects on features sensitivity, we assess the relevance of features to voice pathology detection in MEEI and PdA databases, before and after normalization. As previously, relevance is defined as the mutual information (MI) $I(x_j; c)$ between feature x_j and class c. In general, MI between two random variables x_i and x_j is defined as the KL-divergence between their joint probability density functions (pdf) $P_{ij}(x_i, x_j)$ and the marginal pdf $P_i(x_i)$ and $P_j(x_j)$ [21]. Estimating $I(x_i; x_j)$ from a finite sample requires regularization of $P_{ij}(x_i, x_j)$. We quantized the continuous alphabet of acoustic features by defining b discrete bins along each axis. We make an adaptive quantization (variable bin length) so that the bins are equally populated and the coordinate invariance of the MI is preserved [13]. In the case of modulation spectrum representation, the distribution of the MI for a set of features and a given class can be visualized as a picture. In Fig. 1a and Fig. 1b, the distribution of MI between the selected features and pathologic voices class is depicted, for the MEEI and the PdA databases, respectively, *before* normalization of features. It is then obvious from these two sub-figures that the two distributions of MI are quite different. This means that training a detector on one database and test it on the other database, will result in a very poor detection performance. Fig. 1c and Fig. 1d the corresponding distribution of MI for both databases *after* feature normalization is depicted (Fig. 1c for MEEI and Fig. 1d for PdA). We observed that after applying the suggested feature normalization the maximum value of MI per database lowers almost by half. This will lead to lower performance detector for each database. However, and compared to the upper panels of the same figure, we observe that the distribution of MI in MEEI is quite comparable to the one obtained in PdA. This means that a robust to unseen data detector is now possible to develop.

4. Experiments

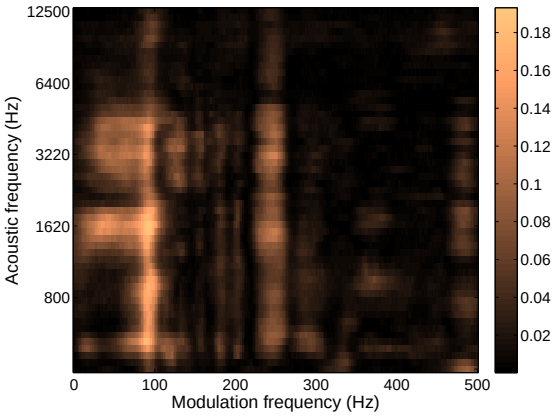
We can proceed then to train voice pathology detection system in MEEI and test it on PdA or vice versa. For training a detector on MEEI, we use a subset of the MEEI in order to cover as many as possible disorders while at the same time the normophonic and dysphonic classes to have similar age and sex distributions. Specifically we used the subset defined in [10] where 53 normophonic and 173 dysphonic speakers from the MEEI database were used. Following the same procedure, we identified a training subset for the PdA database having the same distribution characteristics as those in the training subset in MEEI, which contains 100 normophonic and 100 dysphonic speakers. All the tests were conducted on signals sampled at 25 kHz.



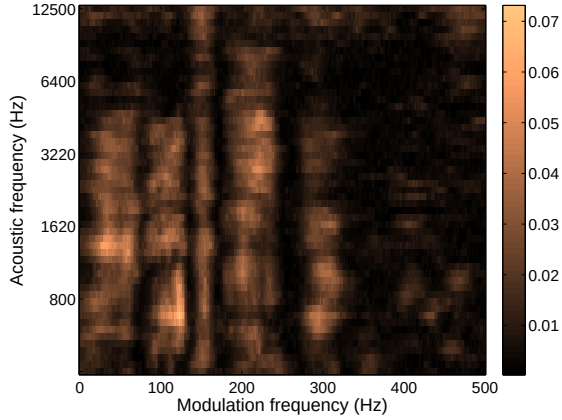
(a) MEEI



(b) PdA



(c) MEEI



(d) PdA

Figure 1: Relevance (MI) between modulation spectral features and pathologic voice class *without normalization* (a) in MEEI, and (b) in PdA and *after normalization* in (c) in MEEI, and (d) in PdA.

For each of the training subsets (for MEEI and for PdA), an optimum detector was obtained. In each case, 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 π). Therefore, each modulation spectrum consisted of $I_1 = 53$ acoustic frequencies and $I_2 = 257$ modulation frequencies, resulting therefore in an 53×257 image per frame. The modulation spectra computed in each frame were mean subtracted and then they were stacked to produce a third order tensor $\mathcal{D} \in R^{I_1 \times I_2 \times I_3}$, where I_3 is the number of frames in the training dataset. Applying the High Order SVD algorithm described previously, the near-optimal projections or principal axes (PCs) of features were detected among those contributing more than 0.1% to the "energy" of $\mathcal{D}$. For MEEI, we detected 44 PCs in the acoustic frequency and 29 PCs in the modulation frequency subspace. This resulted in a reduced space of $44 \times 29 = 1276$ features. For PdA, the corresponding reduced space had dimensions of $53 \times 36 = 1908$. Next, the features which were more correlated to the voice pathology detection task were selected for each database, using the Maximal Relevance criterion (MaxRel). For details about the application of the MaxRel criterion on this task please refer to [13]. The top m features were selected for each database. The optimum detector for MEEI was obtained by considering the $m = 25$ most relevant features. For PdA, the optimum detector was obtained for $m = 68$. The detection results in terms of Detection Rate (DR) and Area Under the Curve (AUC) (when we consider ROC curves for evaluating the performance of the detector) per database are provided in Table 1. We observe that detection results are better for MEEI than for PdA.

Table 1: Non-normalized modulation spectrum features and 4-fold stratified cross-validation repeated 40 times. Detection Rate (DR) in % and Area Under the Curve (AUC) in %, using $m = 25$ for MEEI, and $m = 68$ for PdA.

	DR (%)	AUC (%)
MEEI	94.1	97.8
PdA	81.2	90.2

Let us now consider the optimum detector defined in one database and perform detections in the other database, simulating then the unseen data case. Results are shown in Table

2 where only the Detection Rate (DR) is provided. We see that indeed the performance of both detectors decreased significantly. The detector trained on MEEI, in particular, exhibited random classification performance on PdA. As mentioned in [12], a problem with MEEI database is that some of the normal speakers were recorded at different sites and over potentially different channels than the pathological voices This could explain the better performance of voice pathology detection system on MEEI - as well as the larger degradation when the same system is tested on a different database.

Table 2: Detection Rate (DR) in % using non-normalized modulation spectrum features using optimum detectors: D_{MEEI} as defined only in MEEI ($m = 25$) and D_{PdA} ($m = 68$).

	D_{MEEI}	D_{PdA}
MEEI	94.1	62.3
PdA	51.1	81.2

To increase the robustness of the optimum detectors per database, we performed feature normalization as this is described in the previous section. We observed that after feature normalization, the optimum number m (used in the MaxRel criterion) was significantly increased for both databases. Specifically, for MEEI we found $m = 450$, and $m = 125$ for PdA. The corresponding results after features normalization are listed in Table 3. Comparing the results in Tables (2) and (2), we see that the performance of the optimal detectors was significantly improved in the cross-database evaluation case. We observe that the performance of the optimum detector given one database (i.e., D_{MEEI} for MEEI) is slightly worse in the case of using normalized features as compared to non-normalized features. This is expected, since, as we have seen in Fig.1, the maximum value of MI per database was lowered almost by half when normalized features was used (as compared to the MI when non-normalized features are used). Nevertheless, the overall performance of the optimal detectors was improved and therefore, a more robust detector can now be defined (i.e., a system that is optimum based on PdA performance is preferable over the one that is optimum for classification in MEEI.)

Table 3: Detection Rate (DR) in % using normalized modulation spectrum features and optimum detectors: D_{MEEI} as defined only in MEEI ($m = 450$) and D_{PdA} ($m = 125$).

	D_{MEEI}	D_{PdA}
MEEI	92.7	80.8
PdA	76.1	82.7

5. Conclusion

In this paper we showed that subband normalization of modulation spectral features can compensate for the mismatch of environmental conditions during training and testing. We evaluated the normalized modulation spectral features for voice pathology detection using two different databases (MEEI and PdA) and performing cross-database performance evaluation. Results show that the current normalization procedure lowers the MI between features and detection task, but overall there is a significant increase in the robustness of the optimal detectors. Future work includes the investigation of other normalization procedures for increasing further the robustness of the detectors.

6. 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 the PdA database.

7. References

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

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

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

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

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

[6] M. Rosa, J.C.Pereira, and M.Grellet. Adaptive estimation of residue signal for voice pathology diagnosis. *IEEE Trans. Biomed. Eng.*, 47(1):96–104, Jan 2000.

[7] A. Askenfelt and B. Hammarberg. Speech waveform perturbation analysis revisited. *Speech Transmission Laboratory - Quartely Progress and Status Report*, 22(4):49–68, 1981.

[8] A.A.Dibazar, T.W.Berger, and S.S.Narayanan. Pathological voice asessment. In *IEEE, 28th Eng. in Med. and Biol. Soc.*, pages 1669–1673, NY, NY, USA, August 2006.

[9] J.I. Godino-Llorente, P. Gómez-Vilda, and M. Blanco-Velasco. Dimensionality reduction of a pathological voice quality assessment system based on GMMs and short-term cepstral parameters. *IEEE Trans. Biomed. Eng.*, 53(10):1943–1953, October 2006.

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

[11] S.M. Schimmel, L.E. Atlas, and K. Nie. Feasibility of single channel speaker separation based on modulation frequency analysis. In *Proc. ICASSP*, volume 4, pages 605–608, 2007.

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

[13] M. Markaki and Y. Stylianou. Dimensionality reduction of modulation frequency features for speech discrimination. In *Proc. Interspeech*, pages 646–649, 2008.

[14] M. Markaki and Y. Stylianou. Using modulation spectra for voice pathology detection and classification. In *Proc. EMBS*, to appear, 2009.

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

[16] 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.*, 27:1226–1238, 2005.

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

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

[19] T.F. Quatieri. *Discrete Time Speech Signal Processing*. Prentice-Hall, Upper Saddle River, NJ, 2002.

[20] S. Sukittanon, L. Atlas, and J.W. Pitton. Modulation-scale analysis for content identification. *IEEE Trans. Speech Audio Process.*, 52(10):3023–3035, 2004.

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