EVALUATION OF MODULATION FREQUENCY FEATURES FOR SPEAKER VERIFICATION AND IDENTIFICATION

M. Markaki, and Y. Stylianou

Computer Science Department, University of Crete
and Institute of Computer Science, FORTH, Greece
email: mmarkaki,yannis@csd.uoc.gr.edu

ABSTRACT

In this paper, we suggest the use of mutual information to explore the information provided by the modulation spectrum for speaker verification and identification purposes. The initial representation is first transformed to a lower-dimensional domain using Higher Order SVD and then, the mutual information between speaker identity and features in the transformed domain is computed. Projection of the relevant features back to the original dimensions reveals the modulation spectral components which discriminate speakers. Simulations carried out on YOHO database show that the relevance of modulation spectral features is speaker-dependent.

1. INTRODUCTION

Speaker verification and recognition systems are commonly based on short term spectrum representations such as Mel frequency cepstral coefficients (MFCC) and linear predictive coding-derived cepstral coefficients (LPCC) since these are computationally efficient and simple to implement features. These features may only be computed at a segmental level (short frames) since they are heavily based on stationary representations of the speech signal like the Fourier transform and the linear prediction theory. In the presence of noise, the performance of speaker verification and recognition systems using frame based features quickly deteriorates. Humans on the other hand are quite robust in recognizing a person in noise because they use high-level information such as speaking style. This is usually referred to as supra-segmental information. This also means that people use much longer time windows than just the usual 10ms window used in frame based approaches, for processing the speech signal. The main obstacle for an automatic speaker verification/recognition system is then how to model and efficiently represent long windows. Speaking style is usually defined by prosodic features such as pitch, duration of words and pauses, articulation rate. Unfortunately, estimation of these features is also vulnerable to noise. As an alternative way, it would be possible to extract longer term information directly from the signal using modulation frequency analysis [1, 2]. Modulation spectral features have been employed for content identification [3], speech detection [4] as well as for speaker recognition and verification in [5, 6]. In [3] subband normalization of modulation spectral features was used to compensate for the change of signal characteristics between training and testing data when time and frequency distortions were imposed. However, using modulation spectra directly for classification poses the disadvantage of extremely high dimensionality. In [6] the discrete cosine transform (DCT) was used in modulation frequency subspace due to its energy compaction property. The authors reported that the fusion gain, when these features were combined with MFCCs, was rather minor for speaker verification and recognition. However, the way that dimensions are reduced in modulation spectral features is essential for the performance of the features. This fact was also recognized by the authors in [6]. Given that dimensionality reduction is required for an efficient training of statistical models the question is then how to select task relevant components of the modulation spectrum.

A theoretical study followed by experimental verification for feature selection in speaker recognition has shown that there is a close connection between classification error probability and mutual information (MI) between speaker identity and features [7]. In this paper, we address the relevance of modulation spectrum to speaker verification and identification using MI. We first define a lower-dimensional feature space using higher order singular value decomposition in the acoustic and modulation frequency subspaces. Reducing the original dimensions facilitates subsequent computation of MI. Moreover, the variance of low-dimensional estimators is often smaller than high-dimensional estimators leading to more accurate results [8]. The relevance of the reduced features for speaker verification or identification and their inter-dependency (redundancy) are quantified through mutual information estimation by conducting simulations on YOHO database [9].

2. MODULATION FREQUENCY ANALYSIS

The modulation frequency analysis framework [10] 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 $|X_k(m)|$ 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 "acoustic" and "modulation" frequency, respectively.

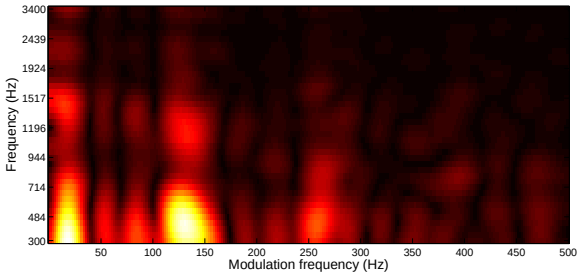


Figure 1: Modulation spectrogram of a 262 ms long segment from a male speaker taken from YOHO [9]. Log magnitude values have been used.

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. Length of the acoustic frequency analysis window $h(n)$ controls the trade-off between resolutions in the frequency and time axes. When $h(n)$ is shorter than a normal pitch period (wideband analysis), the frequency subbands will be wide and the maximum observable modulation frequency will be high enough to provide the pitch information of the speaker. On the other hand, the length of the modulation frequency analysis window $g(m)$ specifies the resolution in the modulation frequency axis.

Figure 1 shows the modulation spectrogram of a 262 ms long segment from a male speaker taken from YOHO [9]. One prominent feature is the pitch energy of the speaker in the modulation frequency dimension ($\sim$ 135 Hz) which is localized in acoustic frequency, peaking at formants [10]. The high energy terms occurring at low modulation frequencies ($\sim 4-30$ Hz) reflect the syllabic and phonetic temporal structure of speech [1].

3. DIMENSIONALITY REDUCTION

Every signal segment in the training database is represented in the acoustic-modulation frequency space as a two-dimensional matrix. "Acoustic" frequency dimensions can be reduced using a bank of triangular shaped mel-frequency filters as usual. Further reduction of both acoustic and modulation frequencies dimensions can be achieved using multilinear algebra [8].

3.1 Multilinear Analysis of Modulation Frequency Features

Joint acoustic and modulation frequencies $B_{mod}[f,t]$ extracted from sound samples in the training database are first mean subtracted (mean values estimated from the whole training set) and stacked producing a data tensor $\mathscr{D}$. Using higher Order SVD (HOSVD) [8], $\mathscr{D}$ can be decomposed to its mode$-n$ singular vectors:

$$\mathscr{D} = \mathscr{S} \times_1 U_{freq} \times_2 U_{mod} \times_3 U_{samples} \tag{3}$$

where U_{freq}, and U_{mod} the orthonormal ordered matrices of the corresponding subspaces of acoustic and modulation frequencies; these contain subspace singular vectors, obtained by unfolding $\mathscr{D}$ along its corresponding modes. Tensor $\mathscr{S}$ is the core tensor with the same dimensions as $\mathscr{D}$. $\mathscr{S} \times_n U$ where $n = 1,2,3$ denotes the $n-$ mode product of tensor

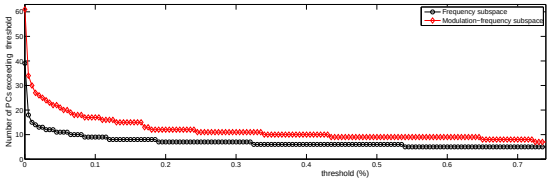


Figure 2: Total number of retained PCs in acoustic and modulation frequency subspace with contribution $\alpha_{n,i}$ greater than a given threshold.

$\mathscr{S} \in R^{I_1 \times I_2 \times I_3}$ by the matrix $U \in R^{J_n \times I_n}$. For $n = 2$ for example, it is an $(I_1 \times J_2 \times I_3)$ tensor given by

$$(\mathscr{S} \times_2 U)_{i_1 j_2 i_3} = \sum_{i_2} s_{i_1 i_2 i_3} u_{j_2 i_2}. \tag{4}$$

Ordering of $n-$mode singular values $\sigma_{i_n}^{(n)}$ implies that the "energy" of tensor $\mathscr{D}$ is concentrated in the singular vectors $U_i^{(n)}$ with the lowest values of i. We can truncate each singular matrix by setting a threshold and keeping only the principal axes in each mode which contribute above this threshold. The contribution of the j^{th} principal component (PC) of subspace S_i with eigenvalue $\lambda_{i,j}$, is defined as:

$$\alpha_{i,j} = \frac{\lambda_{i,j}}{\sum_{j=1}^{N_i} \lambda_{i,j}} \tag{5}$$

where N_i is the dimension of S_i. Figure 2 presents the number of PCs in both subspaces as a function of $\alpha_{i,j}$ which contribute more than a given threshold for the training set described in Section 5. For higher thresholds, there are more PCs in the modulation frequency subspace whose contribution exceeds threshold.

Joint acoustic and modulation frequencies $B_{mod}[f,t]$ extracted from new sound samples are first mean subtracted (mean values estimated from the whole training set) before they are projected on the truncated orthonormal axes of interest, U'_{freq} and U'_{mod_freq}

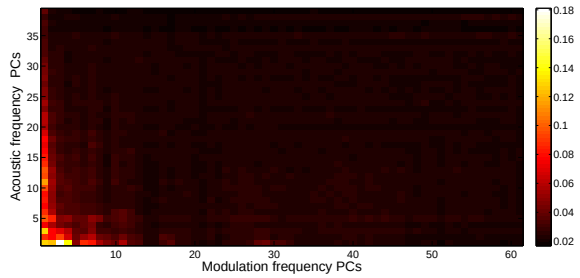
$$Z = B \times_1 {U'_{freq}}^T \times_2 {U'_{mod\_freq}}^T \tag{6}$$

The resulting matrix Z whose dimension is equal to the product of retained singular vectors in each mode contains thus the multilinear PCs of a sound sample.

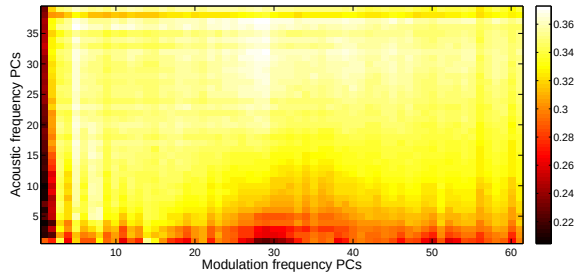
4. MUTUAL INFORMATION BASED FEATURE SELECTION

The *maximal relevance* (MaxRel) feature selection criterion 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 $I(x_j;c)$ are selected [11].

The mutual information between two random variables x_i and x_j is defined as the KL-divergence between their joint probability density function (pdf) $P_{ij}(x_i,x_j)$ and the marginal pdf's $P_i(x_i)$, $P_j(x_j)$.



(a)



(b)

Figure 3: (a) MI of reduced modulation spectral features for the discrimination of 69 speakers. (b) Features redundancy estimated as the median of MI values between pairs of reduced features.

Estimating $I[P_{ij}]$ from a finite sample requires regularization of $P_{ij}(x_i, x_j)$. We have simply 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 [12]. There is an interaction between the precision of features quantization and the sample size dependence of the MI estimates. We study first how the MI between two variables varies as a function of this resolution in order to select the quantizer step size, b^*. We define b^* according to a procedure described in [12]: when data are shuffled, mutual information $I_\infty^{shuffle}(b)$ should be near zero for $b < b^*$ while it increases for $b > b^*$. On the other hand, $I_\infty(b)$ increases with b and converges to the true mutual information near b^*.

5. SIMULATIONS

We have evaluated features of the modulation spectrogram of speech signals for text independent speaker verification and identification tasks. The YOHO database [9] was used in these experiments. All 96 training utterances of the first 53 male speakers and the first 16 female speakers in the enrollment sessions have been used. Each phrase is a sequence of three two-digit numbers read. The data has a telephone bandwidth of 3.8 kHz but no telephone transmission degradations [9]. Silence frames within each utterance were segmented out using an adaptive, energy-based thresholding algorithm [13].

For this application, we considered wideband modulation frequency analysis according to [10]; in that work, derived features such as a speaker's pitch in modulation frequency could be used to localize the speaker in acoustic frequency for single channel speaker separation (see Figure 1). Hence, we analyzed the modulation spectral content of 262 ms long frames of $x(n)$ at 64 ms intervals. The algorithm parameters were set to $M = 8$, $K = 512$, $L = 38$, $I = 512$ and $h(n)$ and $g(m)$ were a 24-point and 78-point sinewindow. Acoustic frequencies were reduced from 257 down to 40, by combining 40 mel-scale bands. The mean was subtracted from each subband envelope before modulation frequency estimation, in order to reduce the interference of large DC components. One uniform modulation frequency vector was produced in each one of the 40 subbands consisting of 257 elements up to 500 Hz.

We assess the relevance to the speaker verification task of projections (principal components) of features with contribution $\alpha_{i,j} > 0.01\%$ based on (5). These are the first 39 PCs in the acoustic frequency subspace and the first 61 PCs in the modulation frequency subspace. We have set the quantizer step size, b^*, to 8.

For HOSVD and MI estimation of the "reduced" features, we divided our database into 69 equal size partitions, one for each speaker. For every speaker, we used 192 speech frames for HOSVD (due to computer memory limitations) and 1536 frames for MI estimation. We computed MI in two different ways:

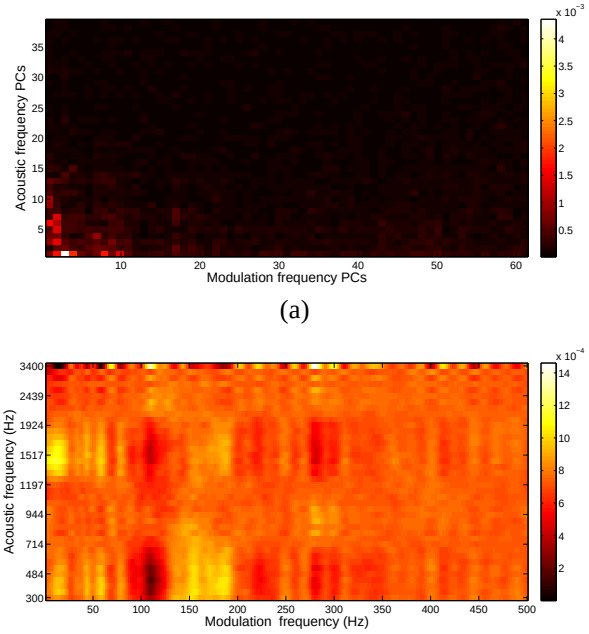
- by considering 69 classes, that is, as many as the speakers in our training database (Figure 3); we refer to features selected according to this definition as "global" features
- by considering a binary class variable, corresponding to the speaker to be verified/recognized vs all the others (Figures (4,5).

We also estimated the MI between pairs of features in order to assess their "redundancy" [11]. For every feature, the median value of its MI to every other feature is displayed in Fig. 3(b). The most relevant features depicted in Fig. 3(a) are also among the least redundant.

As expected, the most relevant features differ in each case. Comparing Fig. 3 to Fig. 4 and Fig. 5, we observe that the speaker-relevant features obviously differ between speakers as well as with the "global" features.

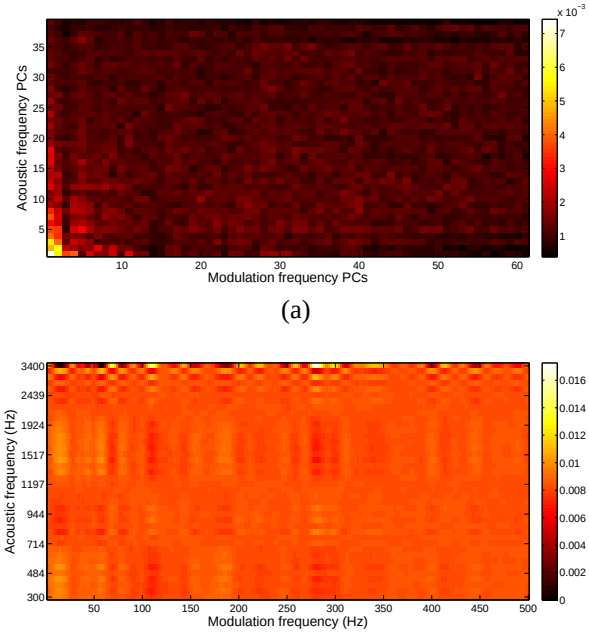
There is a large variability regarding relevance of modulation features for speaker verification. Some speakers will be "easier" to classify using modulation features than others. When we compare Fig. 4 with Fig. 5, which correspond to 2 male speakers from YOHO, we can observe that there are clearly more features with higher MI on average in the case of the 2nd speaker.

Moreover, in Fig. 4 and Fig.5 we have projected the speaker-relevant information back to the original space to visualize the characteristic modulation spectral features of both speakers. By inspecting both Fig. 4 and Fig. 5, we observe different patterns of energy allocation in different frequency bands. These rather reflect acoustic phenomena such as the nature of glottalization - irregural or not - of the particular speakers [14]. In Fig. 4 the most prominent patterns of energy allocation for the 1st speaker correspond to rather intuitive speaker-specific characteristics such as the pitch energy at $\sim 150-200$ Hz modulation frequency, localized at two wide acoustic frequency bands. Pitch-related energy patterns are not prominent in the case of the 2nd speaker (Fig. 5).



(a) (b)

Figure 4: (a) MI of reduced features for verification of 1st speaker. (b) Projection back to the original space.



(a) (b)

Figure 5: (a) MI of features for 2nd speaker verification. (b) Projection back in the original space.

6. DISCUSSION

Our results show that a speaker verification system based on modulation spectral features could be built on *speaker-specific* features. These might reflect intuitively distinctive features of a speaker such as his/her pitch, a particular manner of speaking, or the nature of glottalization [14]. Amplitude-modulation features can capture glottal source differences in normal speech; variation in realization of glottalization of a normal speaker, appears to an extreme degree in dysphonic speech [14]. As the speaker-dependent variability of the mutual information of these features implies, the degree of their significance to speaker recognition and verification and the fusion gain with MFCCs will vary accordingly: it might be minor for some speakers and greater for others with more "atypical" speech. Future work will focus on the experimental verification of these results using databases with channel mismatch and noise and atypical voices. In the latter case, combination of modulation spectrum with MFCC features might be proven beneficial.

REFERENCES

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

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

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

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

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

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

[7] T. Eriksson, S. Kim, H.-G. Kang, and C. Lee, "An information-theoretic perspective on feature selection in speaker recognition," *IEEE Signal Processing Letters*, vol. 12, pp. 500–503, July 2005.

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

[9] J.P. Jr. Campbell, "Testing with the yoho cd-rom voice verification corpus," *Proc. ICASSP*, vol. 1, pp. 341–344, 1995.

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

[11] 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 Analysis and Machine Intelligence*, vol. 27, pp. 1226–1238, 2005.

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

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

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