Dimensionality Reduction of Modulation Frequency Features for Speech Discrimination

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

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

Abstract

We describe a dimensionality reduction method for modulation spectral features, which keeps the time-varying information of interest to the classification task. Due to the varying degrees of redundancy and discriminative power of the acoustic and modulation frequency subspaces, we first employ a generalization of SVD to tensors (Higher Order SVD) to reduce dimensions. Projection of modulation spectral features on the principal axes with the higher energy in each subspace results in a compact feature set. We further estimate the relevance of these projections to speech discrimination based on mutual information to the target class. Reconstruction of modulation spectrograms from the "best" 22 features back to the initial dimensions, shows that modulation spectral features close to syllable and phoneme rates as well as pitch values of speakers are preserved.

Index Terms: modulation spectrum, multilinear algebra, feature selection, mutual information, speech discrimination

1. Introduction

Dynamic information provided by the modulation spectrum capture fast and slower time-varying quantities such as pitch, phonetic and syllabic rates of speech, tempo of music, etc [1]. The use of modulation spectral features for pattern classification is prevented by their dimensionality. Methods addressing this problem have proposed reducing acoustic frequencies using critical band filtering, and modulation frequencies using a continuous wavelet transform instead of a Fourier transform [2].

A different approach to dimensionality reduction of modulation spectral features was presented in [3]. We employed a 3^{rd} order generalization of singular value decomposition (HOSVD)[4] and projected features on the singular vectors of acoustic and modulation frequency subspaces with the higher energy. HOSVD has been also previously applied in auditory-based features with multiple scales of time and spectral resolution [5].

If HOSVD addresses the varying degrees of redundancy of the acoustic and modulation frequency subspaces, mutual information (MI) estimation can be used to assess their discriminative power. By first projecting the high-dimensional data to a lower order manifold, we can approximate the statistical dependence of these projections to the target class (speech versus non-speech, i.e., noise, music, speech babble) with reduced computational effort .

In [3] we showed that these reduced features exhibited comparable classification performance to that of "perceptual" MFCCs [6]. Fusion of both features further decreased the classification error by $\sim 20\%$ which supports the hypothesis that they provide non-redundant information to that encoded by MFCCs. Standard MFCCs represent the spectral envelope variation during a small window - hence, their mean value and standard deviation over a much longer window is commonly used in audio classification [7]. "Perceptual" MFCCs approximate more basic concepts from psychophysics of human hearing besides the critical-band resolution, such as the unequal sensitivity at different frequencies, and the power law relation between the intensity of sound and its perceived loudness. Both operations reduce the spectral-amplitude variation of the critical band spectrum [6].

In this work we investigate the information content of these tranformed features which justifies their improved performance. We first refer to the modulation frequency analysis framework most commonly used [1]. The multilinear dimensionality reduction method and the mutual information-based feature selection are presented in Section 3. In Section 4 we discuss the practical implementation of mutual information estimation. In Section 5 we compare the reduced rank approximation with the reconstruction of modulation spectrogram from the "best" 22 features to show the joint acoustic and modulation frequencies of interest to speech discrimination. Finally in Section 6 we present our conclusions.

2. Modulation Frequency Analysis

For a discrete signal $x(n)$, a short-time Fourier transform (STFT) $X_k(m)$ is initially employed

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

where $W_K = e^{-j(2\pi/K)}$ and $h(n)$ is the acoustic frequency analysis window. The mean of each subband amplitude envelope - defined as $|X_k(m)|$ - is subtracted to remove static information. Next, a Fourier transform detects the frequency content of $|X_k(m)|$:

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

where $g(m)$ is the modulation frequency analysis window; k and i are referred to as the "acoustic" and "modulation" frequency, respectively. Tapered windows $h(n)$ and $g(m)$ are used to reduce the sidelobes of both frequency estimates.

A modulation spectrogram representation then, displays modulation spectral energy $|X_l(k,i)|$ 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. The degree of overlap between successive windows sets the upper limit of the subband sampling rate during the modulation transform. We have chosen a short $h(n)$ so that frequency subbands are wide and maximum observable modulation frequency permits to resolve the pitch of an adult speaker ($\sim$ 250 Hz) [10].

3. Multilinear Analysis of Modulation Frequency Features

Every signal segment in the training database is represented in the acoustic-modulation frequency space as a two-dimensional matrix. By stacking all training matrices we obtain a third order tensor. Matrix representation of a third order tensor $\mathcal{A} \in R^{I_1 \times I_2 \times I_3}$ is particularly useful for computations: we can simply stack all column (row, ...) vectors of the tensor one after another. “Unfolding” of the $(I_1 \times I_2 \times I_3)$-tensor $\mathcal{A}$ then gives a $(I_1 \times I_2 I_3)$-matrix $A_{(1)}$, a $(I_2 \times I_3 I_1)$-matrix $A_{(2)}$, and a $(I_3 \times I_1 I_2)$-matrix $A_{(3)}$. In a $I_c \times I_a I_b$ unfolding, index i_a is assumed to vary more slowly than i_b [4].

3.1. The higher order singular value decomposition

A multilinear generalization of SVD to tensors referred to as Higher Order SVD (HOSVD) [4] enables the decomposition of the data tensor A to its mode$-n$ singular vectors:

$$A = S \times_1 U_{frequency} \times_2 U_{mod_{freq}} \times_3 U_{samples} \qquad (3)$$

where $U_{frequency}$ and $U_{mod_{freq}}$ are unitary matrices with the singular vectors of the corresponding subspaces: $U_{frequency}$ is the matrix of left singular vectors of the matrix unfolding $A_{(1)}$ and $U_{mod_{freq}}$ is the matrix of left singular vectors of $A_{(2)}$. Non-vanishing singular values $\sigma_{i_1}^{(1)}$, $\sigma_{i_2}^{(2)}$ of $A_{(1)}$ and $A_{(2)}$ depict the column ($1-$mode) and row ($2-$mode) rank of $\mathcal{A}$. (Here, we simply ignore samples subspace matrix, $U_{samples}$).

Tensor S is the core tensor with the same dimensions as A and $\mathcal{S} \times_n U$ denotes the $n-$mode product of $\mathcal{S} \in R^{I_1 \times I_2 \times I_3}$ by the matrix $U \in R^{J_n \times I_n}$. ; e.g., for $n = 2$ multiplication of $\mathcal{S}$ by U produces an $(I_1 \times J_2 \times I_3)$-tensor with entries:

$$(S \times_2 U)_{i_1 j_2 i_3} \equiv \sum_{i_2} s_{i_1 i_2 i_3} u_{j_2 i_2}. \qquad (4)$$

Ordering of $n-$mode singular values $\sigma_{i_n}^{(n)}$ implies that the “energy” of tensor $\mathcal{A}$ is concentrated in the singular vectors $U_i^{(n)}$ with the lowest values of i. Let $\hat{\mathcal{A}}$ a rank-(R_1, R_2) approximation of $\mathcal{A}$ obtained by discarding the smallest n-mode singular values $\sigma_{R_n+1}^{(n)}, \ldots, \sigma_{I_n}^{(n)}$. The least-squares error is bounded as:

$$||\mathcal{A} - \hat{\mathcal{A}}||^2 \leq \sum_{i_1=R_1+1}^{I_1} \sigma_{i_1}^{(1)^2} + \sum_{i_2=R_2+1}^{I_2} \sigma_{i_2}^{(2)^2} \qquad (5)$$

Joint acoustic & modulation frequencies $B \in R^{I_1 \times I_2}$ extracted from audio signals are normalized by the standard deviation over the training set & projected on the truncated orthonormal axes, $\hat{U}_{freq}$, $\hat{U}_{mod}$:

$$Z = B \times_1 \hat{U}_{freq}^T \times_2 \hat{U}_{mod}^T = \hat{U}_{freq}^T.B.\hat{U}_{mod} \qquad (6)$$

Z is an $(R_1 \times R_2)-$matrix, where R_1, R_2 is the number of retained principal components (PCs) in each mode. We can project Z back into the full $I_1 \times I_2$-dimensional space to get the rank-(R_1, R_2) approximation of B:

$$\hat{B} = Z \times_1 \hat{U}_{freq} \times_2 \hat{U}_{mod} = \hat{U}_{freq}.Z.\hat{U}_{mod}^T \qquad (7)$$

Next, we detect the “relevant” projections of features among those contributing more than a threshold to the “energy” of $\mathcal{A}$. The contribution $\alpha_{n,i}$ of the i^{th} basis vector $U_i^{(n)}$ in the n-mode space of $\mathcal{A}$ is related to its eigenvalue $\sigma_i^{(n)}$:

$$\alpha_{n,i} = \frac{\sigma_i^{(n)}}{\sum_{i=1}^{I_n} \sigma_i^{(n)}} \qquad (8)$$

4. Feature Selection based on MI

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 [8].

4.1. Mutual Information Estimation

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 $P_i(x_i)$, $P_j(x_j)$.

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 [9]. Quantization qualitatively has a similar effect to that of adding noise. 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. Entropies are systematically underestimated and mutual information is overestimated according to:

$$I_{est}(b, N) = I_\infty(b) + A(b)/N + C(b, N) \qquad (9)$$

where I_∞ is the extrapolation to infinite sample size and the term $A(b)$ increases with b [9]. There is a critical value, b^*, beyond which the term $C(b, N)$ in (9) become important. We define b^* according to a procedure described in [9]: 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. Experiments on Speech Discrimination

We tested the method described in section 3 on audio data collected from Greek TV programs (TV++) and music CDs. Speech data consists of broadcast news and TV shows recorded in studios, outdoors, or transmitted over telephone lines. Non-speech data consists of music (25 %), outdoors noise (moving cars, crowd noise, etc), claps, and speech babble. All audio data are mono channel, 16 bit per sample, with 16 kHz sampling frequency. Signals have been partitioned into 30 minutes

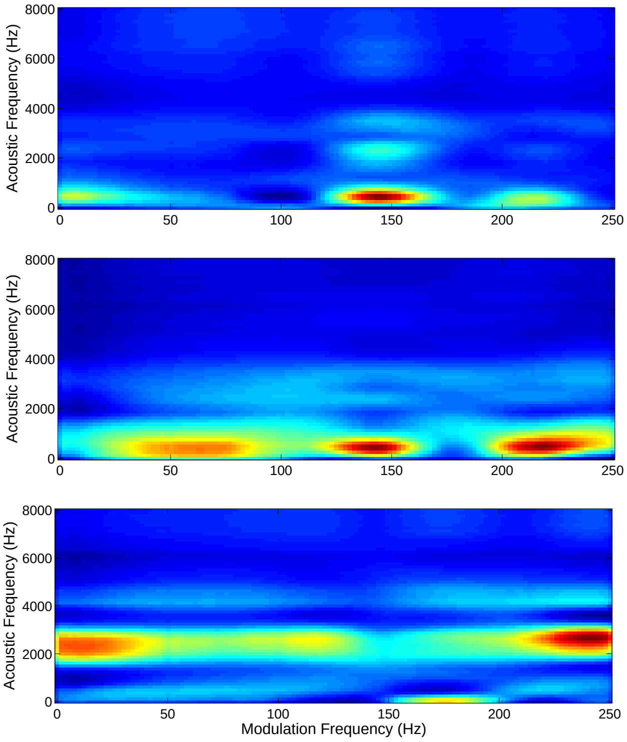

Figure 1: $|X_l(k,i)|$ *rank*$-(9,7)$ *approximation for* 500 *ms of a speech and two non-speech signals (music).*

for training, 30 minutes for validation, and 60 minutes for testing. Each file has been partitioned into 500 ms segments for long-term feature analysis. We extract evenly spaced overlapping segments every 250 ms producing 7200 samples for training and validation, and 14400 samples for testing.

The modulation spectrogram has been calculated using Modulation Toolbox [10]. For every 500 ms block, modulation spectrum features were generated using a 128 point spectrogram with a Gaussian window. One uniform modulation frequency vector was produced in each one of the 65 subbands. Due to a window shift of 32 samples, each modulation frequency vector consists of 125 elements up to 250 Hz. All features were normalized by their corresponding standard deviation estimated from the entire training set to reduce their dynamic range. They were projected on the truncated orthonormal axes U'_{freq}, and $U'_{modfreq}$ according to eq. (6).

Each singular matrix was truncated by setting a predetermined threshold so as to retain only the desired number of principal axes in each mode (eq. 8). Figure 1 presents examples of rank-$(9,7)$ approximations of modulation spectra of 500 ms of speech and non-speech (music) signals where we kept singular vectors contributing more than 1.75% to respective subspace (eq. 8). These were the projections producing the lower classification error [3]. Reconstruction to initial dimensions highlights the modulation spectral features with greatest energy: modulations corresponding to pitch ($\sim$ 140 Hz) and syllabic and phonetic rates ($<$ 40 Hz) in speech; pitch-like energies in 1^{st} music signal and energy oscillations in higher frequency bands in the 2^{nd} music signal.

In order to estimate MI we first project each sample $B \in R^{I_1 \times I_2}$ on the manifold of rank-(R_1, R_2) tensors using equation 6. We have set $R_1 = 50$ and $R_2 = 25$ corresponding to singular vectors with contribution greater than 1% (eq. 8). As Figure 2 shows, mutual information between features in the truncated subspaces of $\mathcal{A}$ is almost zero for most of them - that is, redundancy between them is minimal as we should expect because of the HOSVD process. MI after shuffling data ($\triangle$) is of course zero. Also experiments with the training set showed that only 33 out of the 1250 projected features (2.64%) have mutual information to the target class more than 0.04 bits. Figure 3 presents these MI estimates: the subspace spanned by the first 3 acoustic frequency PCs and the first 13 modulation frequency PCs appear to be the most relevant. We point out that singular values criterion would keep more acoustic frequency PCs than modulation PCs.

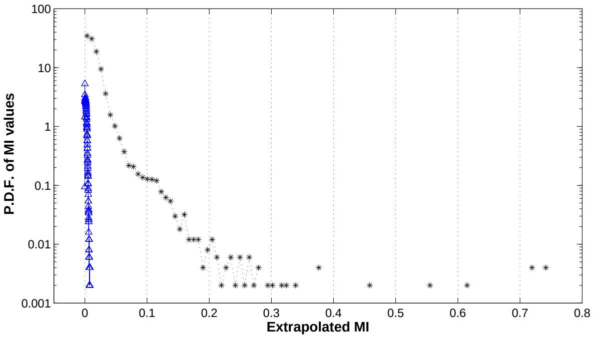

Figure 2: *PDF in a logarithmic scale of MI values obtained when the training dataset is projected onto the first* 50×25 *PCs, before* ($\star$) *and after reshuffling* ($\triangle$).

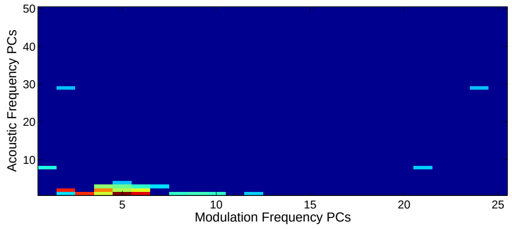

Figure 3: *Mutual information between projections of features on the first* 50×25 *PCs and the speech/non-speech class variable- only the 33 "best" features with MI* > 0.04 *bits are shown.*

Further, we determined potentially redundant features among the 33 most relevant ones with a wrapper using the backward feature selection scheme according to [8]: by setting our initial feature set to S_{33}, we exclude one feature at a time from the current feature set S_k and estimate the respective error rate e_{k-1}; the feature that leads to the greatest error reduction e_{k-1} which is not worse than e_k, is removed. The procedure terminates when we have considered every feature in the row without no gain in classification error.

For SVM classifier [11] and the validation dataset, the wrapper obtained the lowest error (minimum detection cost function for equal costs of miss and false alarm):

$$DCF_{opt} = \min(P_{miss} + P_{false})/2 \qquad (10)$$

by selecting 22 out of the 31 features (referred to as MaxRel^{+} in Table 1). We also used mean and standard deviation of standard

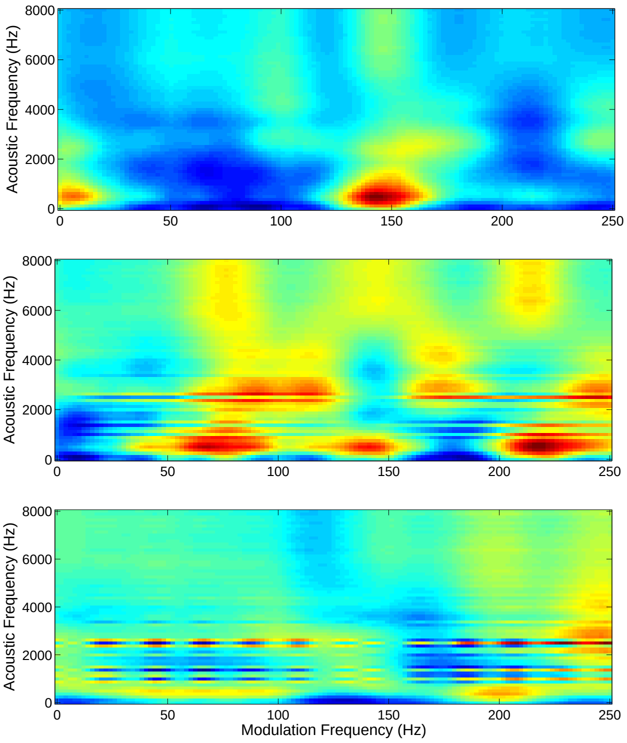

Figure 4: *22 features approximation for the same speech and music signals as in Fig. 1. Energy at modulations corresponding to pitch (*$\sim$ *140 Hz) and syllabic and phonetic rates (*$<$ *40 Hz) remain prominent in speech.*

MFCCs and perceptual MFCCs* for segment parameterization, each resulting in a 26-element feature vector. Table 1 presents the DCF_{opt} values for the systems tested using SVM and the same data set. For comparison, we also report the best DCF_{opt} when using the first (R_1, R_2) projections, which was 6.49% for the $[9 \times 7]$ PCs. The last column refers to the fusion of MFCCs* with MaxRel$^+$ features which further reduced DCF_{opt} down to 3.99%, i.e., a $\sim 20\%$ relative improvement.

Table 1: DCF_{opt}, $P_{miss_{opt}}$ *and* $P_{false_{opt}}$ *on test set*

	$(9, 7)$	MFCCs	MFCCs*	MaxRel$^+$	fusion
DCF	6.49	9.54	5.03	5.10	3.99
P_{miss}	6.31	6.24	4.53	4.47	3.06
P_{false}	6.67	12.84	5.53	5.72	4.92

Figure 4 depicts the 22 features approximation for the same speech and music signals as in Fig. 1. Energy at modulations corresponding to pitch ($\sim$ 140 Hz) and syllabic and phonetic rates ($<$ 40 Hz) remain prominent in speech. Pitch-like energy in 1^{st} music signal is also preserved. The 2^{nd} music signal with most of its energy concentrated in higher frequency bands, is severely blurred under this approximate representation.

6. Discussion

Previous studies have shown the importance of joint acoustic and modulation frequency concept in signal analysis and synthesis, as well as single-channel talker separation and coding applications ([1, 2]). We presented a dimensionality reduction method for modulation spectral features which could be tailored to various classification tasks. HOSVD efficiently addresses the differing degrees of redundancy in acoustic and modulation frequency subspaces. By projecting features on a lower dimensional subspace, we significantly reduce computational load of MI estimation. On the other hand, the HOSVD step has already significantly reduced features redundancy (see Fig. 2). Detection of remaining redundant features among the most relevant ones can be easily accomplished then using a wrapper [8].

The set of 22 features that result, performs much better than the standard MFCCs features while they perform equally well with the perceptually enhanced MFCCs* features. Moreover their fusion further lowers speech discrimination error by $\sim$ 20% (Table 1). It is worthwhile noting here that the combination of modulation spectrum with cepstral representation is analogous to a two-dimensional spectro-temporal transform [5]. Comparing Figures (1) and (4), we notice that reconstruction of audio signals from the "best" 22 features produces a biased (as opposed to a least-squares error) representation: modulations that characterize speech at the lower acoustic frequency bands, corresponding to syllable and phonemic rates and the pitch of different speakers, are enhanced. Modulations which are only localized at the higher frequency bands, are diminished. Subsequently, the classification task has been greatly simplified.

7. Acknowledgements

This work was partially funded by GSRT grant 05AKMON106

8. References

[1] Atlas, L. & Shamma, S.A., "Joint Acoustic and Modulation Frequency", EURASIP Journal on Applied Signal Processing, 7:668–675, 2003.

[2] Sukittanon, S., Atlas, L. & Pitton, J.W., "Modulation-Scale Analysis for Content Identification", IEEE Trans. Audio, Speech and Language Proc., vol.52, 10:3023–3035, 2004.

[3] Markaki, M., & Stylianou, Y., "Discrimination of speech from nonspeech in broadcast news based on modulation frequency features", Proc. ISCA, Speech Analysis and Processing for Knowledge Discovery, June 2008 (accepted).

[4] De Lathauwer, L., & Vandewalle, J., "Dimensionality reduction in higher-order signal processing and rank$-(R_1, R_2, \ldots, R_N)$ reduction in multilinear algebra", Linear Algebra and its Applications, vol. 391, pp. 31–55, 2004.

[5] Mesgarani, N., Slaney, M. & Shamma, S.A., "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations", IEEE Trans. Audio, Speech and Language Proc., 14:920–930, 2006.

[6] Hermansky, "Perceptual linear predictive (PLP) analysis of speech", J.A.S.A., vol.87 4:1738–1752, 1990.

[7] Lu, L., Zhang, H.J. & Li, S., "Content-based audio classification and segmentation by using support vector machines", Multimedia Systems 8: 482-492, 2003.

[8] Peng, H., Long, F. & Ding, C., "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy", IEEE Trans. Pattern Analysis and Machine Intelligence, vol 27, 8:1226–1238, 2005.

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

[10] Modulation Toolbox : "http://www.ee.washington.edu/ research/isdl/projects/modulationtoolbox"

[11] Joachims, T., "Making large-scale SVM Learning Practical" *in Advances in Kernel Methods - Support Vector Learning*, B. Scholkopf, C. Burges, and A. Smola, eds., MIT-Press, 1999.