Discrimination of Speech from Nonspeeech in Broadcast News Based on Modulation Frequency Features

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 content based speech discrimination algorithm in broadcast news based on the time-varying information provided by the modulation spectrum. 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. We further select the optimal principal axes in each subspace based on mutual information. Projection of modulation spectral features in these axes results in a compact feature set at a very low cost for subsequent classification with SVMs. We present experimental comparison between our algorithm and MFCCs using the same classifier and dataset.

Index Terms: audio classification, modulation spectrum, speech discrimination, feature selection, mutual information.

1. Introduction

Speech/non-speech segmentation can be formulated as a pattern recognition problem where the optimal features and the classifier built on them are application-dependent. In broadcast news nonspeech consists of music, various sound sources, and silence although its duration is usually reduced to a minimum. Methods that work well on speech/music discrimination usually do not handle efficiently other non-speech classes. It has been shown that for successful audio segmentation and classification, the classification unit has to be a segment, i.e., a sequence of frames rather than a single frame [1, 2, 3].

Reviewing relevant past work, many approaches in the literature have examined various features and classifiers. Mel-frequency cepstral coefficients - the most commonly used features in speech and speaker recognition systems - have also been successfully applied in audio indexing tasks [1, 2].

In this work we compare modulation spectral features [4] to MFCC features. 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. However the use of modulation spectral features in pattern classification is prevented by their large dimensionality. An efficient way to address this issue is a generalization of SVD to tensors (Higher Order SVD [5]) - a technique which has been applied in auditory-based features with multiple scales of time and spectral resolution [3]. Joint acoustic and modulation frequencies can be projected on the retained singular vectors in each subspace to obtain the multilinear principal components (PCs) of the sound samples. Next we spot near-optimal PCs for classification among those contributing more than 1% through an incremental search method based on mutual information [7].

The organization of the paper is as follows: Section 2 briefly reviews the modulation frequency analysis framework. The multilinear dimensionality reduction method and the mutual information-based feature selection are presented in Section 3. In Section 4 we describe the experimental setup, the database and the results. Finally in Section 5 we present our conclusions.

2. Modulation Frequency Analysis

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. 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; 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 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.

3. Description of the method

3.1. 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 data tensor. A generalization of SVD to tensors referred to as Higher Order SVD (HOSVD) [5] enables the decomposition of a tensor D to

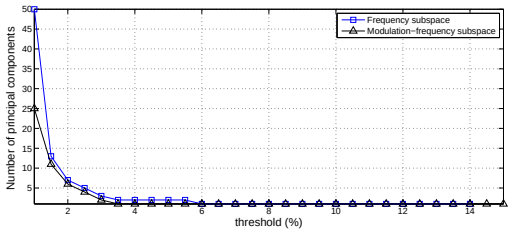

Figure 1: *Total number of retained PCs in each subspace as a function of threshold on contribution percentage. The vertical axis indicates the number of PCs in each subspace that have contribution (eq.5) greater than the threshold*

its mode$-n$ singular vectors:

$$D = S \times_1 U_{frequency} \times_2 U_{mod_freq} \times_3 U_{samples} \quad (3)$$

where $U_{frequency}$, and U_{mod_freq} are the orthonormal ordered matrices of the corresponding subspaces of acoustic and modulation frequencies; these contain subspace singular vectors, obtained by unfolding D along its corresponding modes. Samples subspace matrix, $U_{samples}$, is ignored. Tensor S is the core tensor with the same dimensions as D. $S \times_n U$ where $n = 1, 2, 3$ denotes the $n-$ mode product of tensor $S \in R^{I_1 \times I_2 \times I_3}$ 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

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

Each singular matrix 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 subspace S_i whose corresponding eigenvalue is $\lambda_{i,j}$, is defined as:

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

where N_i is the dimension of S_i - 65 for acoustic frequency and 125 for modulation frequency. Figure 1 presents the number of PCs in these two subspaces as a function of $\alpha_{i,j}$.

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

$$Z = B \times_1 {U'_{freq}}^T \times_2 {U'_{mod_freq}}^T \quad (6)$$

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.

Next, we detect the near-optimal projections (principal components) of features among those contributing more than 1% based on mutual information [7]. That is, we examine the relevance to the target class of the first 50 PCs in the acoustic frequency subspace and the first 25 PCs in the modulation frequency subspace.

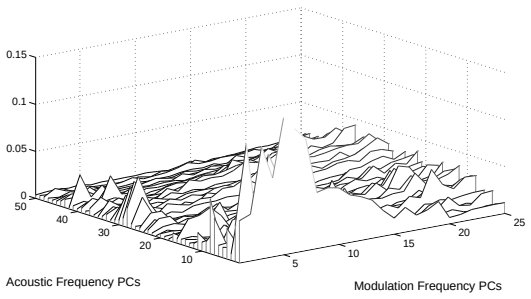

Figure 2: *Mutual information between each of the first* 50×25 *PCs and the speech/non-speech class variable.*

3.2. Mutual Information Estimation

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 interdependence between those variables:

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

MI is invariable to any invertible transformation of the individual variables [8].

It is well-known that MI estimation from observed data is non-trivial when (all or some of) the variables involved are continuous-valued. Estimating $I[P_{ij}]$ from a finite sample requires regularization of $P_{ij}(x_i, x_j)$. The simplest regularization is to define 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]. The precision of features quantization also affects the sample size dependence of MI estimates [8]. Entropies are systematically underestimated and mutual information is overestimated according to:

$$I_{est}(b, N) = I_\infty(b) + A(b)/N + C(b, N) \quad (7)$$

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 (7) become important. We have defined 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 $b > b^*$. $I_\infty(b)$ on the other hand increases with b and converges to the true mutual information near b^*. Although the given sample size ($N = 7200$) permits even greater values for b^*, we have set $b^* = 12$ for reasons of computational efficiency. Figure 2 presents MI estimates between each of the first 50×25 PCs and the speech/non-speech class variable for the training set. The subspace spanned by the first 3 acoustic frequency PCs and the first 13 modulation frequency PCs appear to be the most relevant with much lower peaks elsewhere.

3.3. Max-Relevance and Min-Redundancy

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 [7]. "Minimal-redundancy-maximal-relevance" (mRMR) criterion, on the other hand, spots near-optimal features for classification optimizing the following condition:

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j;c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j;x_i) \right] \quad (8)$$

where $I(x_j;x_i)$ is the mutual information between features x_j and x_i and S_{m-1} is the initially given set of $m-1$ features. The m^{th} feature selected from the set $X - S_{m-1}$ maximizes relevance *and* reduces redundancy. The computational complexity of both incremental search methods is $O(|S|M)$ [7].

In our case, the HOSVD technique has already addressed redundancy reduction. Therefore, mutual information $I(x_j;x_i)$ between pairs of reduced features is rather small. Nevertheless, we used both methods to select n sequential feature sets $S_1 \subset \ldots \subset S_k \subset \ldots \subset S_n$.

3.4. System evaluation

Classification of segments was performed using support vector machines. SVMs find the optimal boundary that separates two classes maximizing the margin between separating boundary and closest samples to it (support vectors). We have used SVM-light [6] with a Radial-Basis-Functions kernel. We have defined an hierarchy of classes similar to [2] for resolving conflicts that arise due to the overlap of segments: frames are classified as non-speech if they are part of any segment that was classified as non-speech; otherwise, they are classified as speech.

We evaluate system performance on the test set using the detection error trade-off curve (DET) between false rejection rate (or speech miss probability) and false acceptance rate (or false alarm probability). Since both classes have equal prior probabilities in our data sets and the costs of miss and false alarm probabilities are considered equally important, the minimum value of the detection cost function, DCF_{opt}, is;

$$DCF_{opt} = \min\left(\frac{P_{miss}+P_{false}}{2}\right). \quad (9)$$

4. Experiments

4.1. Data Collection

We tested the algorithms 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 different conditions such as studios or outdoors; also, some of the speech data have been transmitted over telephone channels. Non-speech data consists of music (25 %), outdoors noise (moving cars, crowd noise, etc), claps, and very noisy unintelligible speech due to many speakers talking simultaneously (speech babble). Music content consists of the audio signals at the beginning and the end of TV shows as well as songs from music CDs. Audio data are all mono channel and 16 bit per sample, with 16 kHz sampling frequency. The database has been manually segmented and labeled at Computer Science Department, UoC. Speech signals have been partitioned into 30 minutes 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.

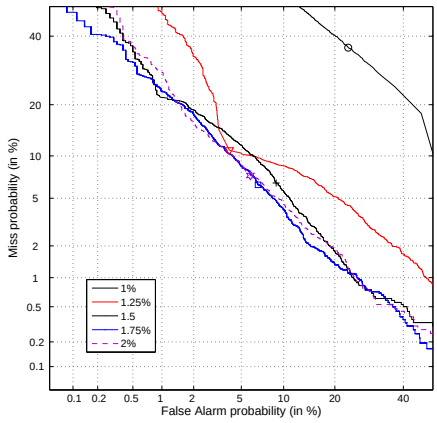

Figure 3: *Detection Error Trade-off curves and minimum detection cost function (markers) for the validation set (see Section 4.1) when retaining PCs with contribution greater than:* 1% *(circle),* 1.25% *(triangle),* 1.5% *(+),* 1.75% *(square) and* 2% *(hexagram)*

4.2. Feature Extraction and Classification

The modulation spectrogram has been calculated using Modulation Toolbox [11]. For every 500 ms block modulation spectrum features were generated using a 128 point spectrogram with a Gaussian window. The envelope in each subband was detected by a magnitude square operator. To reduce the interference of large dc components of the subband envelope, the mean was subtracted before modulation frequency estimation. 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 $U'_{mod_{freq}}$ according to eq. (6).

Figure 3 presents the Detection Error Trade-off (DET) curves and minimum detection cost function (DCF_{opt}) for the validation set when retaining PCs with contributions greater than 1%, 1.25%, 1.5%, 1.75% and 2% (see Figure 1). The dimensionality of the reduced features is respectively $50 \times 25 = 1250$, $21 \times 16 = 336$, $13 \times 11 = 143$, $9 \times 7 = 63$ and $7 \times 6 = 42$. DCF_{opt} is better for the configurations of 63 and 42 dimensions: 6.49% and 6.61%, respectively. increase in dimensionality beyond 143 features induces poor generalization as seen in Figure 3 whereas for less than 42 features, the performance is significantly worse (not shown).

It is known that "the m best features are not the best m features" [7]. As seen in Figure 4, MaxRel feature sets produce smaller errors than mRMR features, for every $k \in [1,n]$, so in the rest of the paper the MaxRel approach will only be used. We combine Max-Rel with a wrapper using the backward feature selection scheme into a two-stage algorithm according to [7]. The first 31 most relevant features led to a $DCF_{opt} = 5.69\%$; we recall that the best DCF_{opt} was 6.49% using the $[9\times7]$ PCs. By setting our initial feature set to S_{31} then, we exclude one feature at a time from the current feature set S_k and estimate the respective error rate $e_{k-1} = DCF_{opt_{k-1}}$; the feature that leads to the greatest error reduction e_{k-1} which is not worse than e_k,

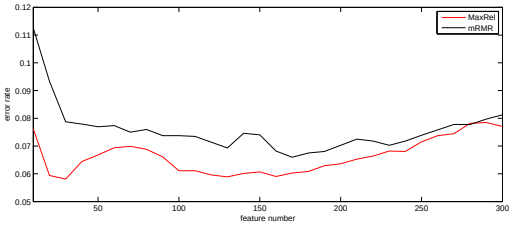

Figure 4: *SVM classifier equal error rate using mRMR and MaxRel features.*

is removed. The procedure terminates when we have considered every feature in the row without no gain in classification error. For Max-Rel and the validation dataset, the wrapper obtained the lowest error $DCF_{opt} = 5.1\%$ by selecting 22 out of the 31 features.

We also present a comparison to results for the same dataset, using mean and standard deviation of MFCCs for segment parameterization [12]. 13th order MFCCs were extracted from 25 ms audio frames with a 10 ms frame rate. Critical-band analysis of the power spectrum with a set of triangular band-pass filters was performed as usual; also an auditory-like spectrogram was derived by applying equal-loudness pre-emphasis and cube-root intensity-loudness compression according to Hermansky [10]. The mean and standard deviation of MFCCs over 50 frames resulted in a 26-element feature vector per 500 ms segment.

Figure 5 and Table 1 present the DET curves and the respective optimal values of DCF, P_{miss} and P_{false} for the systems tested. MaxRel^{+} denotes the combination of backward selection with MaxRel features, which retained 22 out of the first 31 MaxRel features. MFCCs* are MFCC features extracted after loudness equalization and cube root compression according to [10]; the results for both MFCCs* and common MFCC features are based on the same dataset using SVM classifier and have been previously reported in [12]. "Fusion" refers to the combined feature set of MaxRel^{+} and MFCCs*; the concatenated 48-features vector improves the best DCF_{opt} by $\sim 20\%$.

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

	MFCC	MFCC*	MaxRel^{+}	63 PCs	Fusion
DCF	9.54	4.96	5.10	6.49	3.99
P_{miss}	6.24	4.07	4.47	6.31	3.06
P_{false}	12.84	5.84	5.72	6.67	4.92

5. Conclusions

We presented a novel and alternative to the commonly used MFCC feature set for the discrimination of speech from non-speech sounds of broadcast news. We have found that the proposed feature set performs much better than the simple MFCC features while they perform equally well with the perceptually enhanced MFCCs. Moreover, the fusion with perceptual MFCCs further improves classification accuracy which indicates that the two feature sets provide complimentary information for the speech signal.

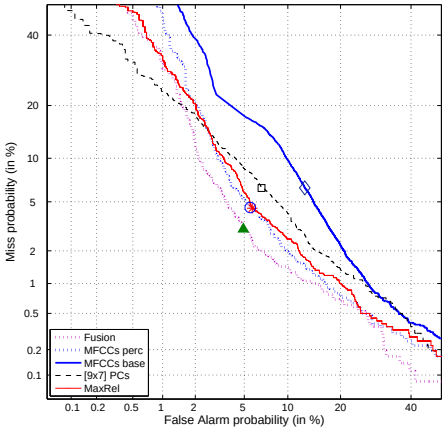

Figure 5: *DET curves and DCF_{opt} for mean and variance of 13 "perceptual" $(\circ)$ or base MFCCs $(\diamond)$, $[9 \times 7]$ PCs $(\square)$, 22 MaxRel features $(*)$ and their fusion with MFCCs* $(\triangle)$.*

6. Acknowledgements

This work was partially funded by the General Secretariat of Research and Technology (GSRT) grant 05AKMON106.

7. References

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

[2] Aronowitz, H., "Segmental modeling for audio segmentation", Proc. ICASSP 2007, Hawaii, USA, 2007.

[3] 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.

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

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

[6] 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.

[7] 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.

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

[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] Hermansky, H., Hanson, B. & Wakita, H., "Perceptually based linear predictive analysis of speech", Proc. ICASSP 1985, pp. 509-512, 1985.

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

[12] Markaki, M., Karpov, A., Apostolopoulos, E., Astrinaki, M., Stylianou, Y. & Ronzhin, A., "A hybrid system for Audio segmentation and speech-endpoint detection of broadcast news", SPECOM 2007 Proceedings, $12th$ International Conference on Speech and Computer, vol. 2, pp:691–696, 2007.