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

Maria Markaki[a], Yannis Stylianou[a,b]

[a] *Computer Science Department, University of Crete, Greece*
[b] *Institute of Computer Science, FORTH, Greece*

Abstract

In audio content analysis, the discrimination of speech and non-speech is the first processing step before speaker segmentation and recognition, or speech transcription. Speech/non-speech segmentation algorithms usually consist of a frame based scoring phase using MFCC features, combined with a smoothing phase. In this paper, a content based speech discrimination algorithm is designed to exploit long-term information inherent in modulation spectrum. In order to address 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 set of features with minimum redundancy. We further estimate the relevance of these projections to speech discrimination based on mutual information to the target class. This system is built upon a segment based SVM classifier in order to recognize the presence of voice activity in audio signal. Detection experiments using Greek and U.S. English broadcast news data composed of many speakers in various acoustic conditions suggest that the system provides complementary information to state-of-the-art mel-cepstral features.

Key words: speech discrimination, modulation spectrum, mutual information, higher order singular value decomposition

1. Introduction

The increasingly larger volumes of audio that are amassing nowadays, require a pre-processing in order to remove information-less content before storing. Usually the first stage of processing partitions the signal into primary components such as speech, and non-speech before speaker segmentation and recognition, or speech transcription.

Reviewing relevant past work, many approaches in the literature have examined various features and classifiers. In telephone speech adaptive methods such as short-term energy-based methods, first measure the energy of each frame in the file and then set the speech detection threshold relative to the maximum energy level. A simple energy level detector that is very efficient in high signal-to-noise ratio (SNR) conditions would fail in lower SNR or when music and noise are present (which also contain substantial energy). In [28] a real-time speech/music classification system was presented based on zero-crossing rate and short-term energy over a 2.4 sec segment of broadcast FM radio. Scheirer and Slaney [29] proposed another real-time speech/music discriminator using thirteen features in time, frequency and cepstrum domain for

modeling speech and music and different classification schemes over 2.4 sec segments. Methods based on such low level perceptual features are considered less efficient when a window smaller than 2.4 sec is used, or when more audio classes such as environmental sounds are taken into account [16].

Mel-frequency cepstral coefficients (MFCC) - the most commonly used features in speech and speaker recognition systems - have been successfully applied in audio indexing task [1, 4, 16]. For applications in which the audio is also transcribed, these features are available at no additional computational cost for direct audio search. Each audio frame can be represented with either just the "static" cepstra or also augmenting the representation with the first and second order time derivatives to capture dynamic features in the audio stream. It has been extensively documented that it is difficult to accurately discriminate speech from nonspeech given a single frame [1, 16, 22]. Speech/non-speech segmentation algorithms usually consist of a frame based scoring phase using MFCC features, combined with a smoothing phase. The general approach used for audio segmentation is based on Maximum Likelihood (ML) classification of a frame with Gaussian mixture models (GMMs) using MFCC features [4]. The smoothing of likelihoods, when using the GMM framework, assumes that the feature vectors of neighboring frames are independent given a certain class; this smoothing is commonly applied by the GMM-based algorithms either for speech-nonspeech and audio classification or for speaker recognition [4, 26]. In [12], SVM classifier was used based on cepstral features; median smoothing of SVM output scores over 1 sec segments improved frame-based classification accuracy by $\sim$ 30%. The performance of SVM-based system on different domains was more consistent or even better than GMMs based on the same cepstral features [12].

In [16, 32, 1], the classification entity is a sequence of frames (a segment) rather than a single frame. In [16, 32], segments were parameterized by the mean value and standard deviation of frame-based features over a much longer window.Audio classification was performed using SVMs in [16], and GMMs in [32]. In [1], a segment based classifier was built unifying both frame based scoring phase and the smoothing phase. Audio segments were modeled as supervectors through a segment based generative model and each class (speech, silence, music) was modeled by a distribution over the supervector space. Classification of speech/non-speech classes proceeded then using either GMMs or SVMs [1].

In this work we first compare and then combine the speech discrimination ability of cepstral features to that of modulation spectral features [8, 2]. Dynamic information provided by the modulation spectrum captures fast and slower time-varying quantities such as pitch, phonetic and syllabic rates of speech, tempo of music, etc [8, 2]. In [24], it was suggested that these high level modulation features could be combined with standard mel-cepstral features to enhance speaker recognition performance. Hence these features could be available at no additional computational cost for direct audio search (as MFCC).

Still, the use of modulation spectral features for pattern classification is prevented by their dimensionality. Methods addressing this problem have proposed critical band filtering to reduce acoustic frequencies, and a continuous wavelet transform instead of a Fourier transform [33], or a discrete cosine transform [13] for modulation frequencies. In [24], dimensionality reduction was performed either by averaging across modulation filters or across acoustic frequency bands.

We adopt a different approach towards dimensionality reduction of this two-dimensional representation. We employ a higher order generalization of singular value decomposition (HOSVD) to tensors [7], and retain the singular vectors of acoustic and modulation frequency subspaces with the higher energy. Joint acoustic and modulation frequencies are projected on the retained singular vectors in each subspace to obtain the multilinear principal components (PCs) of the

sound samples. In this way the varying degrees of redundancy of the acoustic and modulation frequency subspaces are efficiently addressed. This technique has been successfully applied in auditory-based features with multiple scales of time and spectral resolution in [22].

Truncation of singular vectors based on their energy addresses features redundancy; to assess their discriminative power, we need an estimate of their mutual information (MI) to the target class (speech versus non-speech, i.e., noise, music, speech babble) [6]. By first projecting the high-dimensional data to a lower order manifold, we can approximate the statistical dependence of these projections to the class variable with reduced computational effort. We spot near-optimal PCs for classification among those contributing more than an energy threshold through an incremental search method based on mutual information [23].

In Section 2, we overview a modulation frequency analysis framework which is commonly used [2]. The multilinear dimensionality reduction method and the mutual information-based feature selection are presented in Section 3. In the same Section we also discuss the practical implementation of mutual information estimation based on the joint probability density function for two variables and its marginals. In Section 4, we describe the experimental setup, the database and the results using the proposed features, mel cepstral features and the concatenation of both feature sets. Finally, in Section 5 we present our conclusions.

2. Modulation Frequency Analysis

The most common modulation frequency analysis framework [8, 2] for a discrete signal $x(n)$, initially computes via the discrete Fourier transform (DFT) the discrete short-time Fourier transform (DSTFT) $X_k(m)$, m denoting the frame number and k the DFT frequency sample:

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

where $W_K = e^{-j(2\pi/K)}$, $h(n)$ the (acoustic) frequency analysis window and M the hopsize (in number of samples). Subband envelope detection - defined as the magnitude $|X_k(m)|$ or square magnitude $|X_k(m)|^2$ of the subband - and their frequency analysis (with DFT) are performed next, to yield the modulation spectrum with a uniform modulation frequency decomposition:

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

where $W_I = e^{-j(2\pi/I)}$, $g(m)$ is the modulation frequency analysis window and L the corresponding hopsize (in number of 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 sidelobes of both frequency estimates.

The modulation spectrogram representation then, displays modulation spectral energy $|X_l(k,i)| \in R^{I_1 \times I_2}$ 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 subtracting their mean value (computed over the training set of I_3 samples) and stacking all training matrices we obtain the data tensor $\mathcal{D} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$. A generalization of SVD to tensors referred to as Higher Order SVD (HOSVD) [7] enables the decomposition of tensor $\mathcal{D}$ to its n-mode singular vectors:

$$\mathcal{D} = \mathcal{S} \times_1 \mathbf{U}_{freq} \times_2 \mathbf{U}_{mod} \times_3 \mathbf{U}_{samples} \tag{3}$$

where $\mathcal{S}$ is the core tensor with the same dimensions as $\mathcal{D}$; $\mathcal{S} \times_n \mathbf{U}^{(n)}$, $n = 1, 2, 3$, denotes the $n-$mode product of $\mathcal{S} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ by the matrix $\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times I_n}$. For $n = 2$ for example, $\mathcal{S} \times_2 \mathbf{U}^{(2)}$ is an $(I_1 \times I_2 \times I_3)$ tensor given by

$$\left(\mathcal{S} \times_2 \mathbf{U}^{(2)}\right)_{i_1 i_2 i_3} \stackrel{\text{def}}{=} \sum_{i_2} s_{i_1 i_2 i_3} u_{i_2 i_2}. \tag{4}$$

$\mathbf{U}_{freq} \in R^{I_1 \times I_1}$, $\mathbf{U}_{mod} \in \mathbb{R}^{I_2 \times I_2}$ are the unitary matrices of the corresponding subspaces of acoustic and modulation frequencies; $\mathbf{U}_{samples} \in R^{I_3 \times I_3}$ is the samples subspace matrix. These $(I_n \times I_n)$ matrices $\mathbf{U}^{(n)}$, $n = 1, 2, 3$, contain the n-mode singular vectors (SVs):

$$\mathbf{U}^{(n)} = \left[U_1^{(n)}\ U_2^{(n)}\ \ldots\ U_{I_n}^{(n)}\right]. \tag{5}$$

Each matrix $\mathbf{U}^{(n)}$ can directly be obtained as the matrix of left singular vectors of the "matrix unfolding" $\mathbf{D}_{(n)}$ of $\mathcal{D}$ along the corresponding mode [7]. Tensor $\mathcal{D}$ can be unfolded to the $I_1 \times I_2 I_3$ matrix $\mathbf{D}_{(1)}$, the $I_2 \times I_3 I_1$ matrix $\mathbf{D}_{(2)}$, or the $I_3 \times I_1 I_2$ matrix $\mathbf{D}_{(3)}$. The n-mode singular values correspond to the singular values found by the SVD of $\mathbf{D}_{(n)}$.

We define the contribution $\alpha_{n,j}$ of the j^{th} n-mode singular vector $U_j^{(n)}$ as a function of its singular value $\lambda_{n,j}$:

$$\alpha_{n,j} = \lambda_{n,j} \,/\, \sum_{j=1}^{I_n} \lambda_{n,j} \quad \text{or} \quad \alpha_{n,j} = \lambda_{n,j} \,/\, \sqrt{\sum_{j=1}^{I_n} \lambda_{n,j}^2} \tag{6}$$

We set a threshold and retain only the R_n singular vectors with contribution exceeding that threshold in modes $n = 1, 2$. We thus obtain the truncated matrices $\hat{\mathbf{U}}^{(1)} \equiv \hat{\mathbf{U}}_{freq} \in \mathbb{R}^{I_1 \times R_1}$ and $\hat{\mathbf{U}}^{(2)} \equiv \hat{\mathbf{U}}_{mod} \in \mathbb{R}^{I_2 \times R_2}$. Joint acoustic & modulation frequencies $\mathbf{B} \equiv |X_l(k, i)| \in \mathbb{R}^{I_1 \times I_2}$ extracted from audio signals are normalized by their standard deviation over the training set and projected on $\hat{\mathbf{U}}_{freq}$ and $\hat{\mathbf{U}}_{mod}$ [7]:

$$\mathbf{Z} = \mathbf{B} \times_1 \hat{\mathbf{U}}_{freq}^T \times_2 \hat{\mathbf{U}}_{mod}^T = \hat{\mathbf{U}}_{freq}^T \mathbf{B} \hat{\mathbf{U}}_{mod} \tag{7}$$

$\mathbf{Z}$ is an $(R_1 \times R_2)-$matrix, where R_1, R_2 is the number of retained SVs in the acoustic and modulation frequency subspace. We can project $\mathbf{Z}$ back into the full $I_1 \times I_2$-dimensional space to get the rank-(R_1, R_2) approximation of $\mathbf{B}$ [7]:

$$\hat{\mathbf{B}} = \mathbf{Z} \times_1 \hat{\mathbf{U}}_{freq} \times_2 \hat{\mathbf{U}}_{mod} = \hat{\mathbf{U}}_{freq} . \mathbf{Z} . \hat{\mathbf{U}}_{mod}^T \tag{8}$$

HOSVD addresses features redundancy by selecting mutually independent features; these are not necessarily the most discriminative features. We proceed then to detect the near-optimal projections of features among those contributing more than a threshold. Based on mutual information [6], we examine the relevance to the target class of the first R_1 SVs in the acoustic frequency subspace and the first R_2 SVs in the modulation frequency subspace.

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 inter-dependency between those variables:

$$I[P_{ij}] = \int dx_i \int dx_j P_{ij}(x_i, x_j) \log_2 \left[\frac{P_{ij}(x_i, x_j)}{P_i(x_i) P_j(x_j)} \right] \tag{9}$$

MI is invariable to any invertible transformation of the individual variables [6].

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 [31]. The precision of features quantization also affects the sample size dependence of MI estimates [6]. Entropies are systematically underestimated and mutual information is overestimated according to:

$$I_{est}(b, N) = I_\infty(b) + A(b)/N + C(b, N) \tag{10}$$

where I_∞ is the extrapolation to infinite sample size and the term $A(b)$ increases with b [31]. There is a critical value, b^*, beyond which the term $C(b, N)$ in (10) become important. We have defined b^* according to a procedure described in [31]: 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^*.

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

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j; c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i) \right] \tag{11}$$

where $I(x_j; x_i)$ is the mutual information between features x_j and x_i, i.e., redundancy, 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 $O(|S|M)$ [23].

In our case the HOSVD technique has already addressed redundancy reduction; mutual information $I(x_j; x_i)$ between pairs of "packed" features is significantly smaller than MI between original features. Hence we used MaxRel method to select n sequential feature sets $S_1 \subset \ldots \subset S_k \subset \ldots \subset S_n$ and computed the respective equal error rate (EER) using SVM classifier and the validation data set.

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) [11]. We have used SVMlight [11] with a Radial-Basis-Functions kernel.

We evaluate system performance on the validation and the test set using the Detection Error Trade-off curve (DET) [21]. The DET curves depict the false rejection rate (or miss probability) of the speech detector versus its false acceptance rate (or false alarm probability). DET curves are quite similar to the Receiver Operating Characteristic (ROC) curves, except that the detection error probabilities are plotted on a nonlinear scale. This scale transforms the error probabilities by mapping them to the corresponding Gaussian deviates. Thus DET curves are straight lines when the underlying distributions are Gaussian. This makes DET plots more intelligible than ROC plots [21]. We have used the matlab files that NIST has made available for producing DET curves with the matlab software package [21].

Since the costs of miss and false alarm probabilities are considered equally important, the minimum value of the detection cost function, DCF_{opt}, is:

$$DCF_{opt} = \min\left(P_{miss} * P_{speech} + P_{false} * P_{non-speech}\right). \tag{12}$$

where P_{speech} and $P_{non-speech}$ are the prior probabilities of speech and non-speech class respectively. We also report the equal-error rate (EER) - the point of DET curve where the false alarm probability equals the miss probability.

4. Experiments

4.1. Data Collection

We first tested the methods described in section 3 on audio data recorded from broadcasts of Greek TV programs (ERT3). The database was manually segmented and labeled at CSD. The labeled dataset used in these experiments consists of 6 hours; it is available upon request from the first author.

Audio data are all mono channel and 16 bit per sample, with 16 kHz sampling frequency. Speech data consists of broadcast news and TV shows recorded in different conditions such as studios or outdoors, under quiet conditions or with background noise; also, some of the speech data have been transmitted over telephone channels. Non-speech data consists of music (mainly audio signals at the beginning and the end of TV shows, or music accompanying talks of political candidates), outdoors noise from moving cars, beeps, crowd, claps, or very noisy unintelligible speech due to many speakers talking simultaneously (speech babble). We used 7 broadcast shows for training, with minimum duration of $\sim$ 6 min, and maximum duration of $\sim$ 1 hour (1 and a half hour in total). Fifteen shows were used for testing with minimum duration of $\sim$ 6 min and maximum duration of $\sim$ 1 hour ($\sim$ 4 and a half hours in total). Each file was partitioned into 500 ms segments for long-term feature analysis. We extracted evenly spaced overlapping segments every 250 ms for speech and every 50 ms for non-speech (in order to maximize non-speech data).

We also conducted experiments on the NIST RT-03 evaluation data distributed by LDC (LDC2007S10). The dataset we used consisted of six audio files with 30 minutes duration each, recorded in February 2001 from U.S. radio or TV broadcast news shows, from ABC, CNN, NBC, PRI, and VOA. For parameter tuning, we performed 5-fold cross-validation experiments on a subset of $\sim$ 1 hour of this data; system performance was evaluated on the rest of data.

4.2. Feature Extraction and Classification

The modulation spectrogram was calculated using Modulation Toolbox [3]. 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 consisted of 125 elements up to 250 Hz. Feature calculation runtime is $O(N \log_2 N)$, since the estimation of modulation spectral features consists of two FFTs.

The mean value was computed over the training set and subtracted from all matrices; stacking of the training matrices produced the data tensor $\mathcal{D} \in \mathbb{R}^{65\times125\times7200}$. The singular matrices $\mathbf{U}^{(1)} \equiv \mathbf{U}_{freq} \in \mathbb{R}^{65\times65}$ and $\mathbf{U}^{(2)} \equiv \mathbf{U}_{mod} \in \mathbb{R}^{125\times125}$ were directly obtained by SVD of the "matrix unfoldings" $\mathbf{D}_{(1)}$ and $\mathbf{D}_{(2)}$ of $\mathcal{D}$ respectively. By retaining the singular vectors that exceeded a contribution threshold of 1% in each mode (eq. 6), resulted in the truncated singular matrices $\hat{U}_{freq} \in \mathbb{R}^{65\times24}$ and $\hat{U}_{mod} \in \mathbb{R}^{125\times29}$. Features were projected on $\hat{U}_{freq}$ and $\hat{U}_{mod}$ according to eq. (8) resulting in matrices $\mathbf{Z} \in \mathbb{R}^{24\times29}$; these were subsequently reshaped into vectors before MI estimation, feature selection and SVM classification. All features were normalized by their corresponding standard deviation estimated from the entire training set to reduce their dynamic range before classification (their mean value has already been set to zero before projecting them to the truncated singular matrices).

HOSVD is the most costly process in our system but it is performed only once. HOSVD consists of the SVD of two data matrices $N \times k$ each composed of N k-dimensional vectors; computational complexity of SVD transform is $O(Nk^2)$. N is either the acoustic frequency dimension or the modulation frequency dimension; respectively, k is the product of the modulation or the acoustic frequency dimension multiplied by the size of the training dataset.

Figure (1) presents the contribution of the first 25 singular vectors $U_j^{(1)}$ and $U_j^{(2)}$, $j = 1, \ldots, 25$, in the acoustic and modulation frequency subspaces, respectively. Ordering of the $n-$mode singular values $\lambda_{n,j}$ implies that the "energy" of modulation spectral representation is concentrated in the lower j-indices. In addition, Figure (1) shows that variance in the acoustic frequency subspace slightly exceeds that in the modulation frequency subspace; rather more acoustic frequency SVs should be retained for "best rank approximation" of a modulation spectral representation.

For the data discretization involved in MI estimation, the number of discrete bins along each axis was set to $b^* = 8$ according to the procedure described in [31]. Figure 2 compares the relevance of features in the original and reduced representation. The number of relevant features in the original representation is large, posing a serious drawback to any classifier: 1147 out of the 8125 features (14.12%) have mutual information to the target class more than 0.04 bits. As Figure 2a depicts, the most relevant among the original features are mainly distributed along the modulation frequency axis: they span the ranges of the lower syllabic and phonetic rates of speech ($\sim 4 - 30$ Hz) as well as the range of pitch of the majority of speakers, i.e., up to ~ 200 Hz). They also appear confined to the lower acoustic frequency bands up to ~ 2500 Hz.

The HOSVD redundancy reduction method has reduced dimensions in each subspace separately. Therefore, the differential relevance of the two subspaces is preserved in the compressed representation as MI estimation reveals. Figure (2b) presents MI estimates between each of the first 25 singular vectors and the speech/non-speech class variable for the training set. The subspace spanned by the first two acoustic frequency singular vectors (SVs) and the first 15 modulation frequency SVs appear to be the most relevant to speech-non-speech discrimination with

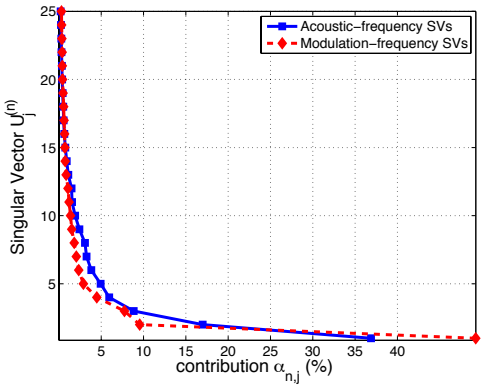
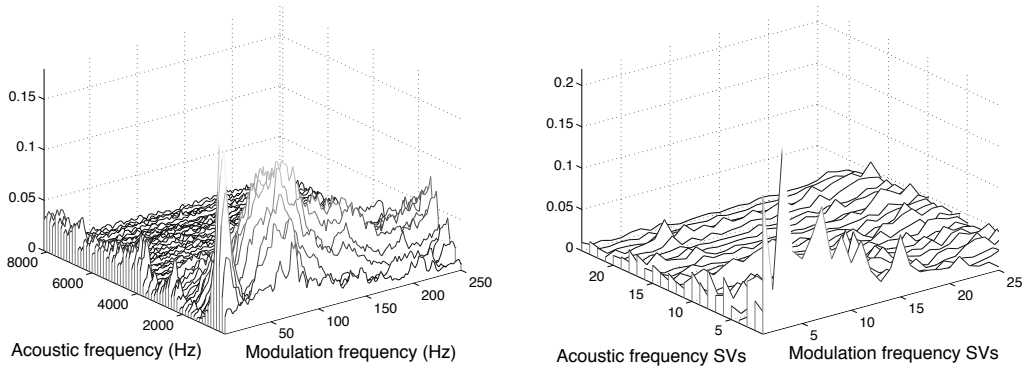


Figure 1: Contribution $\alpha_{n,j}$ of the first 25 singular vectors (SVs) $U_j^{(1)}$, $U_j^{(2)}$, $j = 1, \ldots, 25$, to the acoustic and modulation frequency subspaces, respectively.



(a) (b)

Figure 2: Relevance of the original and compressed modulation spectral features: (a) Mutual information (MI) between the acoustic and modulation frequencies (65×125 dimensions) and the speech/non-speech class variable. (b) MI between the first 25 singular vectors in each subspace and the speech/non-speech class variable.

much lower peaks elsewhere. According to MI criterion, then, variance in modulation frequency subspace is more relevant to the classification task. In addition, the number of relevant features is significantly reduced in the compressed representation: only 27 out of the 696 "packed" features (3.94%) have mutual information to the target class more than 0.04 bits. Still the maximum value of relevance to the classification task is increased.

In Figure 3 we compare the SVM classifier EER on the validation data set when using features selected either in terms of contribution or relevance. According to the maximum contribution criterion, we retained singular vectors with contributions varying between 0.5% up to 6% (eq. 6). The dimensionality of the reduced features varied between $18 \times 18 = 324$ dimensions up to $3 \times 3 = 9$ dimensions, respectively. EER was lowest for the configuration of $13 \times 12 = 156$ dimensions; increase in dimensionality beyond 156 features induced poor generalization whereas for less than

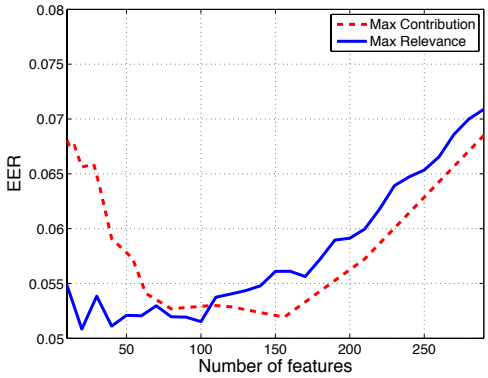
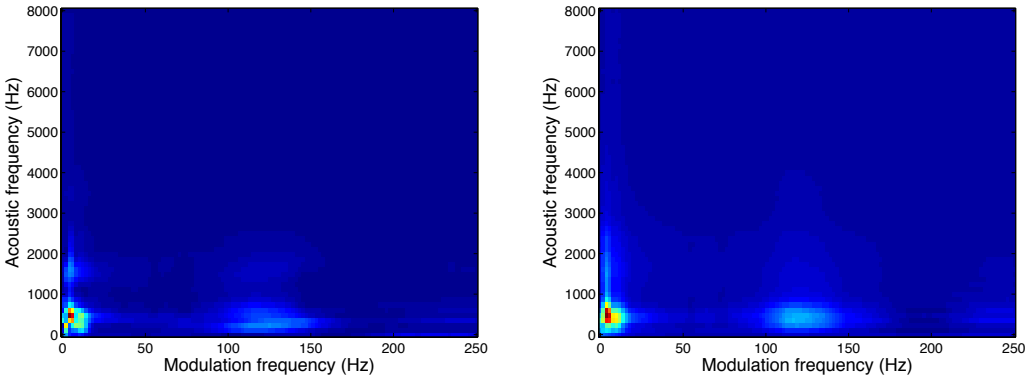


Figure 3: SVM classifier equal error rate (EER) as a function of features selected in terms of relevance or contribution.

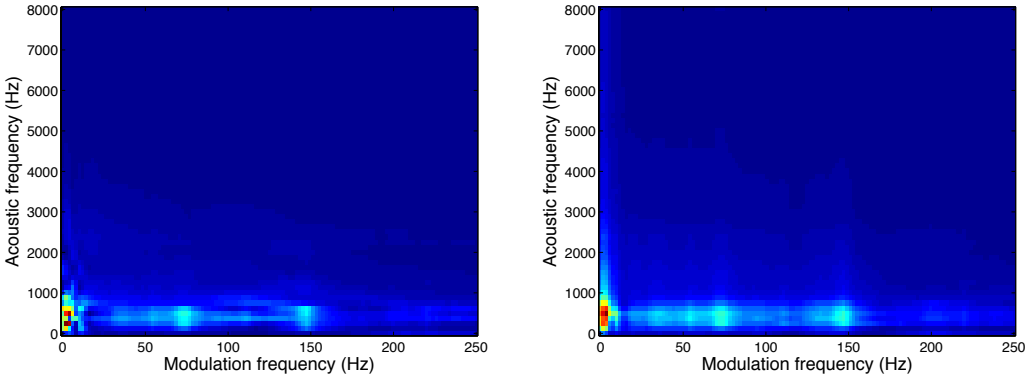


(a) (b)

Figure 4: (a) Rank$-(13, 12)$ approximation (eq. 8) of $|X_l(k, i)|$ for 500 ms of a speech signal. (b) 21 features approximation for the same speech signal. Energy at modulations corresponding to pitch ($\sim$ 120 Hz) and syllabic and phonetic rates ($<$ 40 Hz) remain prominent.

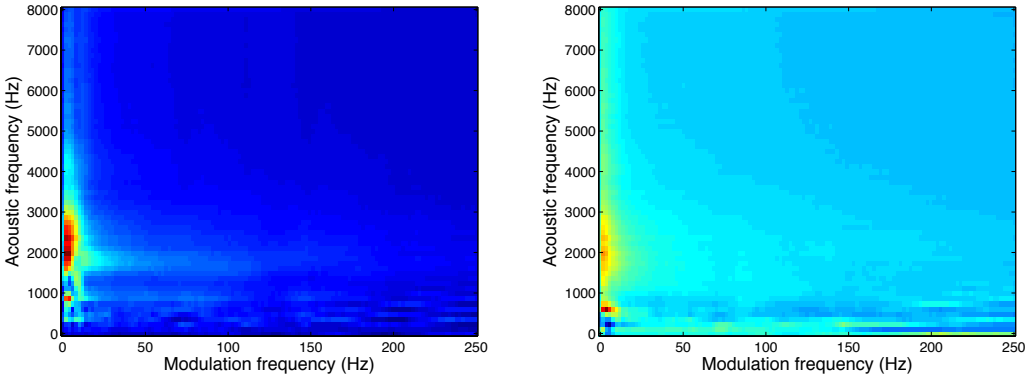
$9\times6 = 54$ features, the performance became progressively worse. Under the maximum relevance selection criterion, just 21 features yielded the best classification performance in terms of EER.

Figures 4, 5, 6 depict the rank$-(13, 12)$ approximation of modulation spectra (eq. 8) as well as their reconstruction from the 21 most relevant features for speech, music and noise signals, respectively. Energy at modulations that characterize speech at the lower acoustic frequency bands, corresponding to syllable and phonemic rates ($<$ 40 Hz) and the pitch of speaker, remain prominent in both representations of speech (Fig. 4). In Fig. 5, the energy at modulations corresponding to harmonics characterize the music signal (at the beginning of a TV show). The approximate representations of the noise signal (claps and crowd noise outdoors) in Fig. 6, depict most of its energy localized in higher frequency bands, and concentrated in lower modulation frequencies.



(a) (b)

Figure 5: (a) Rank$-(13, 12)$ approximation of $|X_l(k, i)|$ for 500 ms of a music signal. (b) 21 features approximation for the same music signal; the characteristic patterns are not lost.



(a) (b)

Figure 6: (a) Rank$-(13, 12)$ approximation of $|X_l(k, i)|$ for 500 ms of a noise signal (claps and crowd noise outdoors). (b) 21 features approximation for the same signal.

4.3. Combining Modulation and Cepstral Features

Speech/Non-Speech discrimination systems for broadcast news are typically based on the mel-frequency cepstral coefficients that are also routinely used in speech and speaker recognition systems. The features used in the baseline system consist of 12th-order Mel frequency cepstral coefficients (MFCCs), log-energy, along with their first and second differences to capture dynamic features in the audio stream [4]. This makes a frame-based feature vector of 39 elements (13×3) The features were extracted from 30 ms audio frames with a 10 ms frame rate, i.e. every 10 ms the signal was multiplied using a Hamming window of 30 ms duration. Critical-band analysis of the power spectrum with a set of triangular band-pass filters was performed as usual. For each frame, equal-loudness pre-emphasis and cube-root intensity-loudness compression were applied

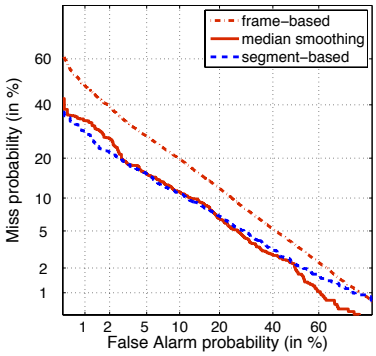


Figure 7: Detection Error Trade-off (DET) curves for frame- and segment-based SVM classification using cepstral features, and median smoothing of the frame-level scores; a small subset of training/validation set from the greek broadcast news shows has been used.

according to Hermansky [9]. The general approach used is maximum-likelihood classification with Gaussian mixture models (GMMs) trained on labeled training data. Still in [12] it was reported that the performance of SVM on different domains was more consistent than GMMs based on the same MFCC features. Therefore, in the subsequent experiments we will use the MFCC-based features with SVM classifiers. This will make easier the comparison between the suggested features and the MFCC-based features. Moreover, we will discuss the fusion of the two sets of features.

In [12], it was found that smoothing the SVM output scores when frame-based features are used, improves the final score in terms of EER (an improvement of about 30% was reported in [12] as compared to the frame-based results prior to smoothing). In [16, 32], segment-based MFCC features were considered. For segments of 500ms, the mean and the standard deviation of 50 frame-based MFCC feature vectors were the segment-based features [16, 32] (i.e., a 78-element feature vector).

We decided to compare the frame-based and segment-based SVM classifiers. We performed 2-fold cross-validation on a subset of the Greek training data set (two broadcast shows of total duration 17 minutes, with 26 speakers). Figure 7 presents the DET curves for frame-based and segment-based SVM classification results. Applying smoothing, using a median filter, on the frame-based SVM classification results, the frame-based approach is highly improved (solid line in Fig.7). Actually it provides on average equivalent result to the segment-based MFCC features. The major disadvantage, however, of any of the frame-based MFCC features approach, is that the computation time for the training and testing of SVM classifier, is much bigger as compared to the segment-based MFCC features. Therefore, we will only consider the segment-based MFCC features for comparison purposes with the suggested modulation spectral features.

Different approaches to information fusion exist [27]: information can be combined prior to the application of any classifier (pre-classification fusion), or after the decisions of the classifier have been obtained (post-classification fusion). Pre-classification fusion refers to feature level fusion in the case of data from a single sensor (such as single channel audio data). When the feature vectors are homogeneous, such as the MFCC features of successive frames of a speech or non-speech audio segment, a single feature vector can be calculated from the mean and standard deviation of the individual feature vectors as in [16, 32]. When different feature extraction

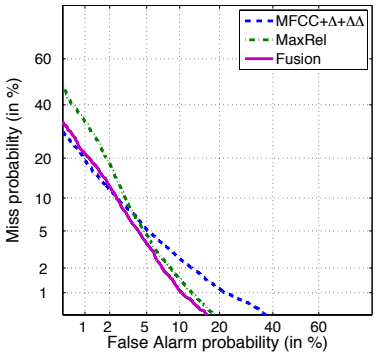


Figure 8: DET curves for segment-based SVM classification using cepstral features (MFCC+Δ + ΔΔ), the 21 most relevant features (MaxRel), and the concatenated feature vector (Fusion) for the same training and testing sets from greek broadcast news shows.

algorithms are applied on the input data, the non-homogeneous feature vectors that incur can be concatenated to produce a single feature vector [27]. On the other hand, post-classification fusion can be accomplished either at the matching score level or at the decision level as explained in [10]. According to [10], integration at the feature level is preferable since the features contain richer information about the input data than the matching scores or output decisions of a classifier/matcher. We simply concatenated the different feature vectors into a single representation of the input pattern.

Table 1: $\hat{DCF}$, $\hat{P}_{miss}$ *and* $\hat{P}_{false}$ *on test set from Greek shows*

	[13, 12]	MFCCs+Δ + ΔΔ	MaxRel	fusion
EER	5.19	4.79	5.06	4.45
$\hat{DCF}$	5.12	4.63	5.05	4.35
$\hat{P}_{miss}$	4.73	3.20	4.84	2.50
$\hat{P}_{false}$	5.50	6.06	5.27	6.19

Figure 8 presents the DET curves and Table 1 the respective EER, and the optimal values of $\hat{DCF}$, $\hat{P}_{miss}$ and $\hat{P}_{false}$ for the systems tested using SVM and the same training data set from greek broadcast news shows. MaxRel denotes the system based on the first 21 most relevant features. The last column refers to the fusion of cepstral with MaxRel features; the concatenated (78+21=99)-features vector further reduced $\hat{DCF}$ down to 4.35%. For comparison, we also report the best EER and $\hat{DCF}$ when using the first (R_1, R_2) projections, which were 5.19% and 5.12% respectively for the $[13 \times 12]$ PCs. MaxRel system is better at the low miss probability regions of the DET curve; cepstral features on the other hand yield better classification performance at the low false alarm regions. Fusion of the two feature sets then follows the best of performances across the whole DET curve.

4.3.1. Results on the NIST RT-03 Data

To train our system on US English, we used about 1 hour from U.S. broadcast news from the NIST RT-03 evaluation data (LDC2007S10). Parameter tuning was performed using 5-fold

cross-validation along with SVM classifier. Figure 9 presents the SVM classifier equal error rate (EER) as a function of the most relevant modulation spectral features alone, or using them in combination with MFCC features. The EER was minimum when using the 52 most relevant modulation spectral features. On the other hand, using a concatenated feature vector, best performance was achieved through the combination of the **16** most relevant modulation spectral features with MFCC features.Probably, there is some redundancy between modulation spectral features and the augmented MFCC parameters (when Δ and $\Delta\Delta$ are included).

Figure 10 presents the respective DET curves and Table 2 the EER, and the optimal values of $D\hat{C}F$, $\hat{P}_{miss}$ and $\hat{P}_{false}$ for the test set. When using cepstral features alone, EER was 3.78% and $D\hat{C}F$ was 3.65%. MaxRel denotes the system based on the first 52 maximal relevance modulation spectra (MRMS) features, which yielded an EER of 4.98% and a $D\hat{C}F$ of 4.88%. Fusion in the last column refers to the concatenation of the augmented MFCC and the 16 MRMS feature vectors (78 + 16 = 94 features). Fusion reduced the EER to 3.14% and $D\hat{C}F$ to 2.97% which is an improvement of $\sim$ 17% and $\sim$ 19%, respectively, over the augmented MFCC.

Performance of speech detection systems on broadcast news audio in other NIST datasets, typically corresponds to a P_{miss} of $\sim$ 1.5% and a P_{false} of 1% $-$ 2% [4, 34, 35]. Here, we report a P_{miss} value of $\sim$ 2.91% and a P_{false} value of $\sim$ 3.12%, which are both higher than the corresponding published values. We believe that this difference is due to the fact that we used just two classes (speech/nonspeech) while in general more classes are considered (speech plus music, speech and noise etc., see references in [34]). The use of more classes will minimize the false rejection of speech (i.e., P_{miss}) when noise or music is also present with speech, because these extra classes could be subsequently reclassified as speech [34]. In addition, several hours of data are commonly used for training of a speech/nonspeech detector [1, 35] whereas we only used about one hour of data.

Table 2: *$D\hat{C}F$, $\hat{P}_{miss}$ and $\hat{P}_{false}$ for testing on NIST RT-03*

	MFCCs+Δ + $\Delta\Delta$	MaxRel	fusion
EER	3.78	4.98	3.14
$D\hat{C}F$	3.65	4.88	2.97
$\hat{P}_{miss}$	3.38	4.62	2.91
$\hat{P}_{false}$	4.40	5.60	3.12

Comparing Tables 1, 2, we conclude that system performance is better in terms of EER and accuracy in the NIST database than in the Greek broadcast audio data. By inspection of the DET curves in Figures 8, 10, we notice that the lower false alarm regions of the DET curve correspond to higher P_{miss} (false speech rejection) in the Greek dataset than in NIST; on the other hand, P_{false} is lower in the Greek dataset for the lower miss probability regions. This difference in performance could be explained by the different content of the U.S. English and Greek TV shows, i.e., the variability in speech and non-speech classes in every database. Besides, the concatenation of features yields greater improvement over cepstral features in the NIST database (accuracy $\sim$ 19%, EER $\sim$ 17%) than in Greek broadcast audio data (accuracy $\sim$ 6%, EER $\sim$ 7%).

5. Conclusions

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

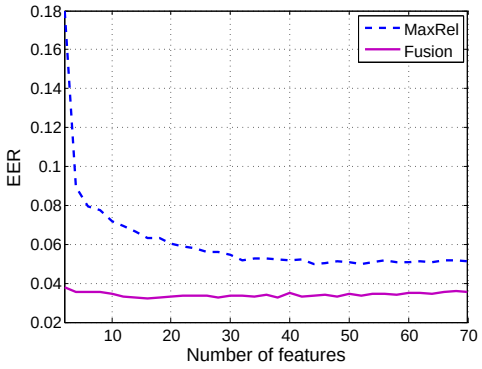


Figure 9: SVM classifier equal error rate (EER) as a function of most relevant modulation spectral features alone, or using them in combination with MFCC features for the U.S. English validation dataset.

applications ([2, 30, 33]). 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. Using HOSVD alone would lead to feature selection based minimal redundancy irrespective of their discriminative power [23].

The set of most relevant features exhibited rather comparable classification performance to that of state-of-the-art mel cepstral features (see Figures 8& 10). Feeding the fused feature set into the same SVM classifier that we used before, further decreased the classification error across the DET curve which supports the hypothesis that modulation spectral features provide non-redundant information to that encoded by MFCCs (Tables 1& 2).

The suggested features span a segment of 500 ms which is roughly equivalent to two syllables

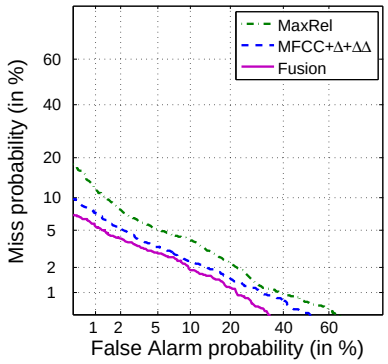


Figure 10: DET curves for segment-based SVM classification using the 52 most relevant features (MaxRel), the augmented MFCC features, and Fusion (concatenation of 16 MaxRel with augmented MFCC feature vectors) for the U.S. English test dataset.

duration; hence, they can capture sound patterns present in a language and that is how they complement MFCC features. On the other hand, this is a non-desirable aspect when we want to use the same system for different languages since further training may be necessary.

Modulation spectra have found important applications to classification tasks such as content identification [33], speaker recognition [13, 24], etc. We expect that modulation based features will be very important in detecting dysphonic voices [17, 20].

References

[1] Aronowitz, H., 2007. Segmental modeling for audio segmentation. Proc. ICASSP 2007, Hawaii, USA, pp. 393–396.

[2] Atlas, L., Shamma S.A., 2003. Joint Acoustic and Modulation Frequency. EURASIP Journal on Applied Signal Processing 7, 668–675.
[3] Atlas, L., Schimmel, S., 2005. Modulation Toolbox for Matlab: "http://isdl.ee.washington.edu/projects/modulationtoolbox/"
[4] Barras, C., Zhu, X., Meignier, S., Gauvain, J.-L., 2006. Multistage speaker diarization of broadcast news. IEEE Trans. Audio, Speech and Language Proc. 14 (5), 1505–1512.
[5] Boakye, K., Stolcke, A., 2006. Improved speech activity detection using cross-channel features for recognition of multiparty meetings. Proc. ICSLP 2006, 1962–1965.
[6] Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory, John Wiley and Sons, New York.
[7] De Lathauwer, L., De Moor, B., Vandewalle, J., 2000. A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. 21, 1253–1278.
[8] Greenberg, S., Kingsbury, B., 1997. The modulation spectrogram: in pursuit of an invariant representation of speech. Proc. ICASSP 1997, 3, 1647–1650.
[9] Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. JASA 87(4), 1738–1752.
[10] Jain, A., Nandakumar, K., Ross, A., 2005. Score normalization in multimodal biometric systems. Pattern Recognition 38, 2270–2285.
[11] Joachims, T., 1999. Making large-scale SVM Learning Practical, in: Scholkopf, B., Burges, C., Smola, A. (Eds.), Advances in Kernel Methods - Support Vector Learning, MIT Press, Cambridge, USA, pp. 41–56.
[12] Kinnunen, T., Chernenko, E., Tuononen, M., Franti, P., Li, H., 2007. Voice Activity Detection using MFCC features and Support Vector Machine. Proc. SPECOM 2007.
[13] Kinnunen, T., Lee, K.A., Li, H., 2008. Dimension Reduction of the Modulation Spectrogram for Speaker Verification. Proc. Odyssey: The Speaker and Language Recognition Workshop, Stellenbosch, South Africa.
[14] Kittler, J., Hatef, M., Duin, R., Matas, J., 1998. On combining classifiers. IEEE Trans. Pattern Anal. and Machine Intel. 20 (3), 226–239.
[15] Lu, L., Zhang, H.J., Jiang, H., 2002. Content analysis for audio classification and segmentation. IEEE Trans. Speech and Audio Proc. 10(7), 504–516.
[16] Lu, L., Zhang, H.J., Li, S., 2003. Content-based audio classification and segmentation by using support vector machines. Multimedia Systems 8, 482–492.
[17] Malyska, N., Quatieri, T.F., Sturim, D., 2005. Automatic Dysphonia Recognition Using Biologically Inspired Amplitude-Modulation Features. Proc. ICASSP 2005, 873–876..
[18] Markaki, M., Stylianou, Y., 2008. Discrimination of Speech from Nonspeeech in Broadcast News Based on Modulation Frequency Features. Proc. ISCA Tutorial and Research Workshop (ITRW 2008).
[19] Markaki, M., Stylianou, Y., 2008. Dimensionality Reduction of Modulation Frequency Features for Speech Discrimination. Proc. Interspeech 2008.
[20] Markaki, M., Stylianou, Y., 2009. Using Modulation Spectra for Voice Pathology Detection and Classification. Proc. IEEE EMBC'09.
[21] Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., 1997. The Det Curve In Assessment Of Detection Task Performance, 1895–1898.
[22] Mesgarani, N., Slaney, M., Shamma S.A., 2006. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. IEEE Trans. Audio, Speech and Language Proc. 14, 920–930.
[23] Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Analysis and Machine Intelligence 27 (8), 1226–1238.
[24] Quatieri, T.F., Malyska, N., Sturim, D.E., 2003. Auditory Signal Processing as a basis for Speaker Recognition. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk Mountain, NY.
[25] Redi, L., Shattuck-Hufnagel, S., 2001. Variation in the realization of glottalization in normal speakers. J. Phonetics 29, 407–429.
[26] Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian Mixture Models. Digit. Signal Processing 10 (1), 19–41.
[27] Sanderson, C., Paliwal, K.K., 2002. Information fusion and person verification using speech and face information. Research Paper IDIAP-RR 02-33, IDIAP.
[28] Saunders, J., 1996. Real-time discrimination of broadcast speech/music. Proc. ICASSP 1996, 993–996.
[29] Scheirer, E., Slaney, M., 1997. Construction and evaluation of a robust multifeature music/speech discriminator. Proc. ICASSP 1997, 1331–1334.
[30] Schimmel, S.M., Atlas, L.E., Nie., K., 2007. Feasibility of single channel speaker separation based on modulation frequency analysis. Proc. ICASSP 2007, 605–608.
[31] Slonim, N., Atwal, G.S. , Tkacik, G., Bialek, W., 2005. Estimating mutual information and multi-information in large networks. arXiv:cs.IT/0502017.
[32] Spina, M.S., Zue, V.W., 1996. Automatic transcription of general audio data: Preliminary analysis. Proc. ICSLP 1996, 594–597.
[33] Sukittanon, S., Atlas, L., Pitton, J.W., 2004. Modulation-Scale Analysis for Content Identification. IEEE Trans.

Audio, Speech and Language Proc. 52 (10), 3023-3035.
[34] Tranter, S.E., Reynolds, D.A., 2006. An overview of Automatic Speaker Diarization Systems. IEEE Trans. Audio, Speech and Language Proc. 14 (5), 1557-1565.
[35] Wooters, C., Fung, J., Peskin, B., Anguera, X., 2004. Towards robust speaker segmentation: The ICSI-SRI Fall 2004 Diarization System. Proc. Fall 2004 Rich Transcription Workshop 2004.