

ABSTRACT

Title of Thesis: DISCRIMINATION OF SPEECH FROM
 NON-SPEECH BASED ON MULTISCALE
 SPECTRO-TEMPORAL MODULATIONS

 Nima Mesgarani, Master of Science, 2005

Thesis directed by: Professor Shihab Shamma
 Department of Electrical and Computer Engineering

We describe a content-based audio classification algorithm based on novel multiscale spectro-temporal modulation features inspired by a model of auditory cortical processing. The task explored is to discriminate speech from non-speech consisting of animal vocalizations, music and environmental sounds. Although this is a relatively easy task for humans, it is still difficult to automate well, especially in noisy and reverberant environments. The auditory model captures basic processes occurring from the early cochlear stages to the central cortical areas. The model generates a multidimensional spectro-temporal representation of the sound, which is then analyzed by a multi-linear dimensionality reduction technique and classified by a Support Vector Machine (SVM). Generalization of the system to signals in high level of additive noise and reverberation is evaluated and compared to two existing approaches [1] [2]. The results demonstrate the advantages of the auditory model over the other two systems, especially at low SNRs and high reverberation.

DISCRIMINATION OF SPEECH FROM NON-SPEECH BASED ON
MULTISCALE SPECTRO-TEMPORAL MODULATIONS

by

Nima Mesgarani

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Science
2005

Advisory Committee:

Professor Shihab Shamma, Advisor
Professor P. S. Krishnaprasad
Professor Carol Y. Espy-Wilson

© Copyright by
Nima Mesgarani
2005

TABLE OF CONTENTS

List of Figures	iii
1 Introduction	1
2 Auditory Model	3
2.1 Early auditory system	3
2.2 Central auditory system	5
2.3 Models of modulation filter	7
3 Multilinear Tensor Analysis	8
3.1 Basic tensor definitions	8
3.2 Multilinear SVD and PCA	9
3.3 Multi-linear analysis of cortical representation	10
3.4 Classification	11
4 Experimental results	12
4.1 Audio Database	12
4.2 Number of principal components	12
4.3 Comparison and results	13
5 Summary and Conclusions	20
Bibliography	21

LIST OF FIGURES

2.1	Schematic of the early stages of auditory processing. (1) Sound is analyzed by a model of the cochlea consisting of a bank of 128 constant-Q bandpass filters with center frequencies equally spaced on a logarithmic frequency axis (tonotopic axis). (2) Each filter output is then transduced into auditory-nerve patterns by a hair cell stage which is modeled as a 3-step operation: a highpass filter (the fluid-cilia coupling), followed by an instantaneous nonlinear compression (gated ionic channels) and then a lowpass filter (hair cell membrane leakage). (3) Finally, a lateral inhibitory network detects discontinuities in the responses across the tonotopic axis of the auditory nerve array by a first-order derivative with respect to the tonotopic axis and followed by a half-wave rectification. The final output of this stage (auditory spectrogram) is obtained by integrating Y_{LIN} over a short window, mimicking the further loss of phase-locking observed in the midbrain.	4
2.2	(A) The cortical multi-scale representation of speech. The auditory spectrogram (the output of the early stage) is analyzed by a bank of spectro-temporal modulation selective filters. The spectro-temporal response field (STRF) of one such filter is shown which corresponds to a neuron that responds well to a ripple of 4Hz rate and 0.5 cycle/octave scale. The output from such a filter is computed by convolving the STRF with the input spectrogram. The total output as a function of time from the model is therefore indexed by three parameters: scale, rate, and frequency. (B) Average rate-scale modulation of speech obtained by summing over all frequencies and averaging over each time window (equations 2.21 and 2.22). The right panel with positive rates is the response of downward filters (u_+) and the right panel with negative rates is the upward ones (u_-).	5
3.1	Illustration of equation 3.10	11
4.1	Total number of retained Principal Components (PCs) in each of the subspaces of frequency, rate and scale as a function of threshold on contribution percentage. The Y axis indicates the number of PCs in each subspace that have contribution (α from equation 4.1) more than the threshold.	13
4.2	Percentage of correctly classified samples as a function of threshold on contribution percentage.	13
4.3	Effect of window length on the percentage of correctly classified speech	15
4.4	Effect of window length on the percentage of correctly classified non-speech	15

4.5	Effects of white noise on percentage of correctly classified speech for auditory model, multifeature[1] and voicing-energy[2] methods.	16
4.6	Effects of white noise on percentage of correctly classified non-speech for auditory model, multifeature[1] and voicing-energy[2] methods.	16
4.7	Effects of pink noise on percentage of correctly classified speech for auditory model, multifeature[1] and voicing-energy[2] methods.	17
4.8	Effects of pink noise on percentage of correctly classified non-speech for auditory model, multifeature[1] and voicing-energy[2] methods.	17
4.9	Effects of reverberation on percentage of correctly classified speech for auditory model, multifeature[1] and voicing-energy[2] methods.	18
4.10	Effects of reverberation on percentage of correctly classified non-speech for auditory model, multifeature[1] and voicing-energy[2] methods.	18
4.11	Effect of <i>white</i> noise on average spectro-temporal modulations of speech for SNRs -15dB , 0dB and 15dB . The spectro-temporal representation of noisy speech preserves the speech specific spectro-temporal features (e.g. near 4Hz , 2Cyc/Oct) even at SNR as low as 0 dB	19
4.12	Effects of <i>pink</i> noise on average spectro-temporal modulations of speech for different SNRs -15dB , 0dB and 15dB . The speech specific spectro-temporal features (e.g. near 4Hz , 2Cyc/Oct) are preserved even at SNR as low as 0 dB	19
4.13	Effects of <i>reverberation</i> on average spectro-temporal modulations of speech for time delays 200ms , 400ms and 600ms . Increasing the time delay results in gradual loss of high-rate temporal modulations of speech.	19

Chapter 1

Introduction

Audio segmentation and classification have important applications in audio data retrieval, archive management, modern human-computer interfaces, and in entertainment and security tasks. In speech recognition systems designed for real world conditions, a robust discrimination of speech from other sounds is a crucial step. Speech discrimination can also be used for coding or telecommunication applications where non-speech sounds are not of interest and hence bandwidth is saved by not transmitting them or by assigning them a low resolution code. Finally, as the amount of available audio data increases, manual segmentation of audio sounds has become more difficult and impractical and alternative automated procedures are much needed.

Speech is a sequence of consonants and vowels, non-harmonic and harmonic sounds, and natural silences between words and phonemes. Discriminating speech from non-speech is often complicated by the similarity of many sounds to speech, such as animal vocalizations. As with other pattern recognition tasks, the first step in this audio classification is to extract and represent the sound by its relevant features. To achieve good performance and generalize well to novel sounds, this representation should be able both to capture the discriminative properties of the sound, and to resist distortion under various noisy conditions.

Research into content-based audio classification is relatively new. Among the earliest is the work of Pfeiffer *et al.* [3] where a 256 phase-compensated gammaphone filter bank was used to extract audio features that mapped the sound to response probabilities. Wold *et al.* [4] adopted instead a statistical model of time-frequency measurements to represent perceptual values of the sound. A common alternative approach involves the extraction of different higher level features to classify audio, such as Mel-frequency cepstral coefficients (MFCCs) along with a vector quantizer (Foote [5]), or noise frame ratios and band periodicity along with K-nearest neighbor and linear spectral pair-vector quantization (Lu [6]), and average zero-crossing rate and energy with a simple threshold to discriminate between speech and music (Saunders [7]).

Two more elaborate systems have been proposed, against which we shall compare our system. The first is proposed by Scheirer and Slaney [1] in which thirteen features in time, Frequency, and cepstrum domain are used to model speech and music. Several classification techniques (e.g., MAP, GMM, KNN) are then employed to achieve a robust performance. The second system is a speech/non-speech segmentation technique [2] in which frame-by-frame maximum autocorrelation and log-energy features are measured, sorted and then followed by linear discriminant analysis and a diagonalization transform.

The novel aspect of our proposed system is a feature set inspired by investigations of various stages of the auditory system [8][9][10][11]. The features are computed using a model of the auditory cortex that maps a given sound to a high-dimensional representation of its spectro-temporal

modulations. A key component that makes this approach practical is a multilinear dimensionality reduction method that by making use of multimodal characteristic of cortical representation, effectively removes redundancies in the measurements in each subspace separately, producing a compact feature vector suitable for classification (section III).

We shall briefly review the auditory model in section II and then outline in section III the mathematical foundation of the analysis of the auditory model's outputs. In section IV, experimental results and performance evaluation of our proposed system are presented, followed by a comparison against two different approaches that represent the best of breed in the literature [1][2].

The computational auditory model is based on neurophysiological, biophysical, and psychoacoustical investigations at various stages of the auditory system [8][9][10]. It consists of two basic stages. An early stage models the transformation of the acoustic signal into an internal neural representation referred to as an auditory spectrogram. A central stage analyzes the spectrogram to estimate the content of its spectral and temporal modulations using a bank of modulation selective filters mimicking those described in a model of the mammalian primary auditory cortex [8]. This stage is responsible for extracting the key features upon what the classification is based.

2.1 Early auditory system

The stages of the early auditory model are illustrated in Figure 2.1. The acoustic signal entering the ear produces a complex spatio-temporal pattern of vibrations along the basilar membrane of the cochlea. The maximal displacement at each cochlear point corresponds to a distinct tone frequency in the stimulus, creating a tonotopically-ordered response axis along the length of the cochlea. Thus, the basilar membrane can be thought of as a bank of constant-Q highly asymmetric bandpass filters ($Q = 4$) equally spaced on a logarithmic frequency axis. In brief, this operation is an affine wavelet transform of the acoustic signal $s(t)$. This analysis stage is implemented by a bank of 128 overlapping constant-Q (QERB = 5.88) bandpass filters with center frequencies (CF) that are uniformly distributed along a logarithmic frequency axis (f), over 5.3 octaves (24 filters/octave). The impulse response of each filter is denoted by $h_{cochlea}(t; f)$. The cochlear filter outputs $y_{cochlea}(t, f)$ are then transduced into auditory-nerve patterns $y_{an}(t, f)$ by a hair cell stage which converted cochlear outputs into inner hair cell intra-cellular potentials. This process is modeled as a 3-step operation: a highpass filter (the fluid-cilia coupling), followed by an instantaneous nonlinear compression (gated ionic channels) $g_{hc}(\cdot)$, and then a lowpass filter (hair cell membrane leakage) $\mu_{hc}(t)$. Finally, a Lateral Inhibitory Network (LIN) detects discontinuities in the responses across the tonotopic axis of the auditory nerve array [12]. The LIN is simply approximated by a first-order derivative with respect to the tonotopic axis and followed by a half-wave rectifier to produce $y_{LIN}(t, f)$. The final output of this stage is obtained by integrating $y_{LIN}(t, f)$ over a short window, $\mu_{midbrain}(t, \tau)$, with time constant $\tau = 8$ msec mimicking the further loss of phase-locking observed in the midbrain. This stage effectively sharpens the bandwidth of the cochlear filters from about $Q = 4$ to 12 [8].

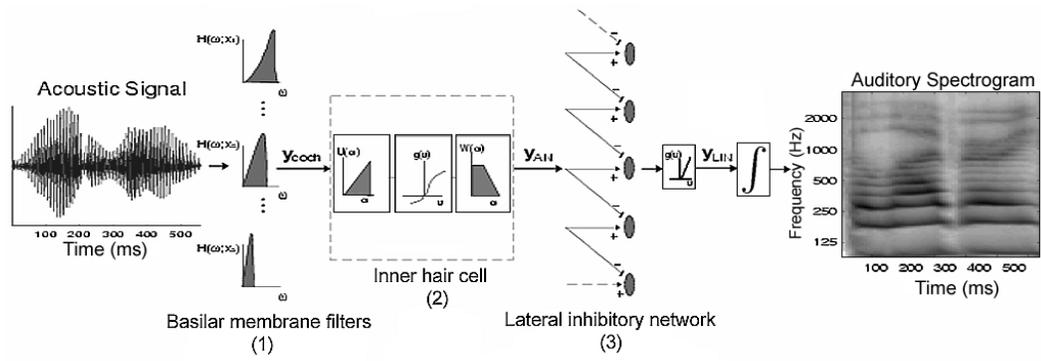


Figure 2.1: Schematic of the early stages of auditory processing. (1) Sound is analyzed by a model of the cochlea consisting of a bank of 128 constant-Q bandpass filters with center frequencies equally spaced on a logarithmic frequency axis (tonotopic axis). (2) Each filter output is then transduced into auditory-nerve patterns by a hair cell stage which is modeled as a 3-step operation: a highpass filter (the fluid-cilia coupling), followed by an instantaneous nonlinear compression (gated ionic channels) and then a lowpass filter (hair cell membrane leakage). (3) Finally, a lateral inhibitory network detects discontinuities in the responses across the tonotopic axis of the auditory nerve array by a first-order derivative with respect to the tonotopic axis and followed by a half-wave rectification. The final output of this stage (auditory spectrogram) is obtained by integrating Y_{LIN} over a short window, mimicking the further loss of phase-locking observed in the midbrain.

The mathematical formulation for this stage can be summarized as followed:

$$y_{cochlea}(t, f) = s(t) * h_{cochlea}(t; f) \quad (2.1)$$

$$y_{an}(t, f) = g_{hc}(\partial_t y_{cochlea}(t, f)) * \mu_{hc}(t) \quad (2.2)$$

$$y_{LIN}(t, f) = \max(\partial_f y_{an}(t, f), 0) \quad (2.3)$$

$$y(t, f) = y_{LIN}(t, f) * \mu_{midbrain}(t; \tau) \quad (2.4)$$

where $*$ denotes convolution in time.

The above sequence of operations effectively computes a spectrogram of the speech signal (Figure 2.1, right) using a bank of constant-Q filters, with a bandwidth tuning Q of about 12 (or just under 10% of the center frequency of each filter). Dynamically, the spectrogram also encodes explicitly all temporal *envelope modulations* due to interactions between the spectral components that fall within the bandwidth of each filter. The frequencies of these modulations are naturally limited by the maximum bandwidth of the cochlear filters.

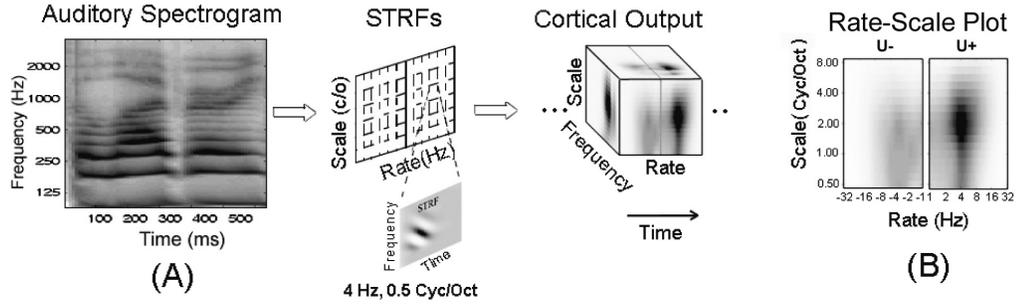


Figure 2.2: (A) The cortical multi-scale representation of speech. The auditory spectrogram (the output of the early stage) is analyzed by a bank of spectro-temporal modulation selective filters. The spectro-temporal response field (STRF) of one such filter is shown which corresponds to a neuron that responds well to a ripple of 4Hz rate and 0.5 cycle/octave scale. The output from such a filter is computed by convolving the STRF with the input spectrogram. The total output as a function of time from the model is therefore indexed by three parameters: scale, rate, and frequency. (B) Average rate-scale modulation of speech obtained by summing over all frequencies and averaging over each time window (equations 2.21 and 2.22). The right panel with positive rates is the response of downward filters (u_+) and the right panel with negative rates is the upward ones (u_-).

2.2 Central auditory system

Higher central auditory stages (especially the primary auditory cortex) further analyze the auditory spectrum into more elaborate representations, interpret them, and separate the different cues and features associated with different sound percepts. Specifically, the auditory cortical model employed here is mathematically equivalent to a two-dimensional affine wavelet transform of the auditory spectrogram, with a spectro-temporal mother wavelet resembling a 2-D spectro-temporal Gabor function. Computationally, this stage estimates the spectral and temporal modulation content of the auditory spectrogram via a bank of modulation-selective filters (the wavelets) centered at each frequency along the tonotopic axis. Each filter is tuned ($Q = 1$) to a range of temporal modulations, also referred to as rates or velocities (ω in Hz) and spectral modulations, also referred to as densities or scales (Ω in *cycles/octave*). A typical Gabor-like spectro-temporal impulse response or wavelet (usually called Spectro-temporal Response Field - STRF) is shown in Figure 2.2.

We assume a bank of directional selective STRF's (downward $[-]$ and upward $[+]$) that are real functions formed by combining two complex functions of time and frequency. This is consistent with physiological finding that most STRF's in primary auditory cortex have the quadrant

separability property [13].

$$STRF_+ = \Re\{H_{rate}(t; \omega, \theta) \cdot H_{scale}(f; \Omega, \phi)\} \quad (2.5)$$

$$STRF_- = \Re\{H_{rate}^*(t; \omega, \theta) \cdot H_{scale}(f; \Omega, \phi)\} \quad (2.6)$$

where \Re denotes the real part, $*$ the complex conjugate, ω and Ω the velocity (*Rate*) and spectral density (*Scale*) parameters of the filters, and θ and ϕ are characteristic phases that determine the degree of asymmetry along time and frequency respectively. Functions H_{rate} and H_{scale} are analytic signals (a signal which has no negative frequency components) obtained from h_{rate} and h_{scale} :

$$H_{rate}(t; \omega, \theta) = h_{rate}(t; \omega, \theta) + j\hat{h}_{rate}(t; \omega, \theta) \quad (2.7)$$

$$H_{scale}(f; \Omega, \phi) = h_{scale}(f; \Omega, \phi) + j\hat{h}_{scale}(f; \Omega, \phi) \quad (2.8)$$

where $\hat{\cdot}$ denotes Hilbert transformation. h_{rate} and h_{scale} are temporal and spectral impulse responses defined by sinusoidally interpolating seed functions $h_r(\cdot)$ (second derivative of a Gaussian function) and $h_s(\cdot)$ (Gamma function), and their Hilbert transforms:

$$h_{rate}(t; \omega, \theta) = h_r(t; \omega)\cos\theta + \hat{h}_r(t; \omega)\sin\theta \quad (2.9)$$

$$h_{scale}(f; \Omega, \phi) = h_s(f; \Omega)\cos\phi + \hat{h}_s(f; \Omega)\sin\phi \quad (2.10)$$

The impulse responses for different scales and rates are given by dilation

$$h_r(t; \omega) = \omega h_r(\omega t) \quad (2.11)$$

$$h_s(f; \Omega) = \Omega h_s(\Omega f) \quad (2.12)$$

Therefore, the spectro-temporal response for an input spectrogram $y(t, f)$ is given by

$$r_+(t, f; \omega, \Omega; \theta, \phi) = y(t, f) *_{t,f} STRF_+(t, f; \omega, \Omega; \theta, \phi) \quad (2.13)$$

$$r_-(t, f; \omega, \Omega; \theta, \phi) = y(t, f) *_{t,f} STRF_-(t, f; \omega, \Omega; \theta, \phi) \quad (2.14)$$

where $*_{t,f}$ denotes convolution with respect to both t and f . Its useful to compute the spectro-temporal response $r_{\pm}(\cdot)$ in terms of the output magnitude and phase of the downward (+) and upward (−) selective filters. For this, the temporal and spatial filters, h_{rate} and h_{scale} can be equivalently expressed in the wavelet-based analytical forms $h_{rw}(\cdot)$ and $h_{sw}(\cdot)$ as:

$$h_{rw}(t; \omega) = h_r(t; \omega) + j\hat{h}_r(t; \omega) \quad (2.15)$$

$$h_{sw}(f; \Omega) = h_s(f; \Omega) + j\hat{h}_s(f; \Omega) \quad (2.16)$$

The complex response to downward and upward selective filters, $z_+(\cdot)$ and $z_-(\cdot)$ is then defined as:

$$z_+(t, f; \Omega, \omega) = y(t, f) *_{t,f} [h_{rw}^*(t; \omega)h_{sw}(f; \Omega)] \quad (2.17)$$

$$z_-(t, f; \Omega, \omega) = y(t, f) *_{tf} [h_{rw}(t; \omega)h_{sw}(f; \Omega)] \quad (2.18)$$

where $*$ denotes the complex conjugate. The cortical response (equations 2.13 and 2.14) for all characteristic phases θ and ϕ can be easily obtained from $z_+(\cdot)$ and $z_-(\cdot)$ as follows:

$$r_+(t, f; \omega, \Omega; \theta, \phi) = |z_+| \cos(\angle z_+ - \theta - \phi) \quad (2.19)$$

$$r_-(t, f; \omega, \Omega; \theta, \phi) = |z_-| \cos(\angle z_- + \theta - \phi) \quad (2.20)$$

Where $|\cdot|$ denotes the magnitude and $\angle \cdot$ the phase. The magnitude and the phase of z_+ and z_- have a physical interpretation: at any time t and for all the STRF's tuned to the same (f, ω, Ω) , the ones with $\theta = \frac{1}{2}(\angle z_+ + \angle z_-)$ and $\phi = \frac{1}{2}(\angle z_+ - \angle z_-)$ symmetries have the maximal downward and upward responses of $|z_+|$ and $|z_-|$.

These maximal responses, the magnitude of z_+ and z_- , are used throughout the paper for classification purpose. Where the spectro-temporal modulation content of the spectrogram is of particular interest, we obtain the summed output from all filters with identical modulation selectivity or STRF's to generate the rate-scale plots: (as shown in Figure 2.2B for speech)

$$u_+(\omega, \Omega) = \sum_t \sum_f |z_+(t, f; \omega, \Omega)| \quad (2.21)$$

$$u_-(\omega, \Omega) = \sum_t \sum_f |z_-(t, f; \omega, \Omega)| \quad (2.22)$$

The final view that emerges is that of a continuously updated estimate of the spectral and temporal modulation content of the auditory spectrogram. All parameters of this model are derived from physiological data in animals and psychoacoustical data in human subjects as explained in detail in [11][13][14].

Unlike conventional features, our auditory based features have multiple scales of time and spectral resolution. Some respond to fast changes while others are tuned to slower modulation patterns; A subset are selective to broadband spectra, and others are more narrowly tuned. For this study, temporal filters (Rate) ranging from 1 to 32Hz, and spectral filters (Scale) from 0.5 to 8.00 Cycle/Octave, were used to represent the spectro-temporal modulations of the sound.

2.3 Models of modulation filter

The importance of slow temporal modulations of sound in speech intelligibility has been emphasized for a long time [15]. Kingsbury et.al. [16] showed the advantage of using modulation spectrogram in improving the robustness of ASR systems to noise and reverberation. Temporal modulation filter banks inspired by psychoacoustical experiments [17] have been successfully used in a variety of audio processing tasks such as automatic speech recognition [18]. Spectro-Temporal features have recently also been used in speech enhancement [19], speech coding [20] and speech recognition to provide more robustness [21].

The output of auditory model is a multidimensional array in which modulations are presented along the four dimensions of time, frequency, rate and scale. For our purpose here, the time axis is averaged over a given time window which results in a three mode tensor for each time window with each element representing the overall modulations at corresponding frequency, rate and scale. In order to obtain a good resolution, sufficient number of filters in each mode are required. As a consequence, the dimensions of the feature space are very large (5 (scale filters) \times 12 (rate filters) \times 128 (frequency channels)=7680). Working in this feature space directly is impractical because a sizable number of training samples is required to characterize the space adequately [22]. Traditional dimensionality reduction methods like principal component analysis (PCA) are inefficient for multi-dimensional data because they treat all the elements of the feature space similarly without considering the varying degrees of redundancy and discriminative contribution of each mode.

Instead, it is possible using multidimensional PCA to tailor the amount of reduction in each subspace independently of others based on the relative magnitude of corresponding singular values. Furthermore, it is also feasible to reduce the amount of training samples and computational load significantly since each subspace is considered separately. We shall demonstrate here the utility of a generalized method for the PCA of multidimensional data based on higher-order singular-value decomposition (HOSVD) [23].

3.1 Basic tensor definitions

Multilinear algebra is the algebra of tensors. Tensors are generalizations of scalars (no indices), vectors (single index), and matrices (two indices) to an arbitrary number of indices. They provide a natural way of representing information along many dimensions. Substantial results have already been achieved in this field. Tucker first formulated the three-mode data model [24], while Kroonenberg formulated alternating least-square (ALS) method to implement three mode factor analysis [25]. Lathauwer *et al.* established a generalization of singular value decomposition (SVD) to higher order tensors [23], and also introduced an iterative method for optimizing the best rank (R_1, R_2, \dots, R_N) approximation of tensors [26]. Tensor algebra and HOSVD have been applied successfully in wide variety of fields including Higher-order-only ICA [27], face recognition [28] and selective image compression along a desired dimension [29].

A Tensor $A \in R^{I_1 \times I_2 \times \dots \times I_N}$ is a multi-index array of numerical values whose elements are denoted by $a_{i_1 i_2 \dots i_N}$. Matrix column vectors are referred to as mode-1 vectors and row vectors as mode-2 vectors. The mode- n vectors of an N th order tensor A are the vectors with I_n components obtained from A by varying index I_n while keeping the other indices fixed. Matrix representation

of a tensor is obtained by stacking all the columns (rows, . . .) of the tensor one after the other. The mode- n matrix unfolding of $A \in R^{I_1 \times I_2 \times \dots \times I_N}$ denoted by $A_{(n)}$ is the $(I_n \times I_1 I_2 \dots I_{n-1} I_{n+1} \dots I_N)$ matrix whose columns are n -mode vectors of tensor A .

An N th-order tensor A has $rank-1$ when it is expressible as the outer product of N vectors:

$$A = U_1 \circ U_2 \circ \dots \circ U_N \quad (3.1)$$

The rank of an arbitrary N th-order tensor A , denoted by $r = rank(A)$ is the minimal number of $rank-1$ tensors that yield A in a linear combination. The n -rank of $A \in R^{I_1 \times I_2 \times \dots \times I_N}$ denoted by r_N , is defined as the dimension of the vector space generated by the *mode* - n vectors

$$R_n = rank_n(A) = rank(A_{(n)}) \quad (3.2)$$

The n -mode product of a tensor $A \in R^{I_1 \times I_2 \times \dots \times I_N}$ by a matrix $U \in R^{J_n \times I_n}$, denoted by $A \times_n U$, is an $(I_1 \times I_2 \times \dots \times J_n \dots \times I_N)$ -tensor given by

$$(A \times_n U)_{i_1 i_2 \dots j_n \dots i_N} = \sum_{i_n} a_{i_1 i_2 \dots i_n \dots i_N} u_{j_n i_n} \quad (3.3)$$

for all index values.

3.2 Multilinear SVD and PCA

Matrix Singular-Value Decomposition orthogonalizes the space spanned by column and rows of the matrix. In general, every matrix D can be written as the product

$$D = U \cdot S \cdot V^T = S \times_1 U \times_2 V \quad (3.4)$$

in which U and V are unitary matrices contains the left- and right-singular vectors of D . S is a pseudo-diagonal matrix with ordered singular values of D on the diagonal.

If D is a data matrix in which each column represents a data sample, then the left singular vectors of D (matrix U) are the principal axes of the data space. Keeping only the coefficients corresponding to the largest singular values of D (Principal Components or PCs) is an effective means of approximating the data in a low-dimensional subspace. To generalize this concept to multidimensional data, we consider a generalization of SVD to tensors [23]. Every $(I_1 \times I_2 \times \dots \times I_N)$ -tensor A can be written as the product

$$A = S \times_1 U^{(1)} \times_2 U^{(2)} \dots \times_N U^{(N)} \quad (3.5)$$

in which $U^{(n)}$ is a unitary matrix containing left singular vectors of mode- n unfolding of tensor A , and S is a $(I_1 \times I_2 \times \dots \times I_N)$ tensor which has the properties of all-orthogonality and ordering. The matrix representation of the HOSVD can be written as

$$\begin{aligned} A_{(n)} &= U^{(n)} \cdot S_{(n)} \cdot (U^{(n+1)} \otimes \dots \\ &\quad \otimes U^{(N)} \otimes U^{(1)} \otimes U^{(2)} \otimes \dots \otimes U^{(n-1)})^T \end{aligned} \quad (3.6)$$

in which \otimes denotes the Kronecker product. The above equation can also be expressed as

$$A_{(n)} = U^{(n)} \cdot \Sigma^{(n)} \cdot V^{(n)T} \quad (3.7)$$

in which $\Sigma^{(n)}$ is a diagonal matrix made by singular values of $A^{(n)}$ and

$$V^{(n)} = (U^{(n+1)} \otimes \dots \otimes U^{(N)} \otimes U^{(1)} \otimes U^{(2)} \otimes \dots \otimes U^{(n-1)}) \quad (3.8)$$

This shows that, at matrix level, the HOSVD conditions lead to an SVD of the matrix unfolding. Lathauwer *et al.* shows [23] that the left-singular matrices of the different matrix unfolding of A correspond to unitary transformations that induce the HOSVD structure which in turn ensures that the HOSVD inherits all the classical space properties from the matrix SVD.

HOSVD results in a new ordered orthogonal bases for representation of the data in subspaces spanned by each mode of the tensor. Dimensionality reduction in each space is obtained by projecting data samples on principal axes and keeping only the components that correspond to the largest singular values of that subspace. However, unlike the matrix case in which the best rank- R approximation of a given matrix is obtained from the truncated SVD, this procedure does not result in optimal approximation in the case of tensors. Instead, the optimal best rank- (R_1, R_2, \dots, R_N) approximation of a tensor can be obtained by an iterative algorithm in which HOSVD provides the initial values [26].

3.3 Multi-linear analysis of cortical representation

The auditory model transforms a sound signal to its corresponding time-varying cortical representation. Averaging over a given time window results in a cube of data in rate-scale-frequency space. Although the dimension of this space is large, its elements are highly correlated making it possible to reduce the dimension significantly using a comprehensive data set, and finding new multilinear and mutually orthogonal principal axes that approximate the real space spanned by these data. The assembled training set is described in detail in section 4.1 which contains 1223 samples from speech and non-speech classes. The resulting data tensor D , obtained by stacking all training tensors is a $5 \times 12 \times 128 \times 1223$ tensor. Next, tensor D is decomposed to its mode- n singular vectors:

$$D = S \times_1 U_{frequency} \times_2 U_{rate} \times_3 U_{scale} \times_4 U_{samples} \quad (3.9)$$

in which $U_{frequency}$, U_{rate} and U_{scale} are orthonormal ordered matrices containing subspace singular vectors, obtained by unfolding D along its corresponding modes. Tensor S is the core tensor with the same dimensions as D .

Each singular matrix is then truncated by setting a predetermined threshold so as retain only the desired number of principal axes in each mode. New sound samples are first transformed

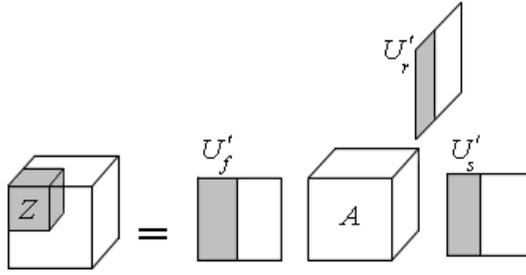


Figure 3.1: Illustration of equation 3.10

to their cortical representation, A , and are then projected onto these truncated orthonormal axes U'_{freq} , U'_{rate} , U'_{scale} (as shown in Figure 3.1):

$$Z = A \times_1 U'_{freq}{}^T \times_2 U'_{rate}{}^T \times_3 U'_{scale}{}^T \quad (3.10)$$

The resulting tensor Z whose dimension is equal to the total number of retained singular vectors in each mode thus contains the multilinear cortical principal components of the sound sample. Z is then vectorized and normalized by subtracting its mean and dividing by its norm to obtain a compact feature vector for classification.

3.4 Classification

Classification was performed using a Support Vector Machine (SVM) [30][31]. SVMs find the optimal boundary that separates two classes in such a way as to maximize the margin between separating boundary and closest samples to it (support vectors). This in general results in improving generalization from training to test data [30]. Radial basis function (RBF) were used as SVM kernel.

4.1 Audio Database

An audio database was assembled from five publicly available corpora. Details of the database are as follows.

Speech samples were taken from TIMIT Acoustic-Phonetic Continuous Speech Corpus [32] which contains short sentences spoken by male and female native English speakers with 8 dialects. 299 different sentences spoken by different speakers (male and female) were selected for training and 160 different sentences spoken by different speakers (male and female) were selected for test purpose. Sentences and speakers in training and test sets were also different.

To make the non-speech class as comprehensive as possible, sounds from animal vocalizations, music and environmental sounds were assembled together. Animal vocalization were taken from BBC Sound Effects audio CD collection [33] (263 for training, 139 for test). Music samples that covered a large variety of musical styles were selected from RWC Genre Database [34] (349 for training, 185 for test). Environmental sounds were assembled from Noisex [35] and Auroa [36] databases which have stationary and non-stationary sounds including white and pink noise, factory, jets, destroyer engine, military vehicles, cars and several speech babble recorded in different environments like restaurant, airport and exhibition (312 for training, 167 for test).

The training set included 299 speech and 924 non-speech samples and the test set consisted of 160 speech and 491 non-speech samples. The length of each utterance in training and test is equal to the selected time window (e.g. one one-second sample per sound file).

4.2 Number of principal components

The number of retained principal components (PCs) in each subspace is determined by analyzing the contribution of each PC to the representation of associated subspace. The contribution of j th principal component of subspace S_i whose corresponding eigenvalue is $\lambda_{i,j}$ is defined as

$$\alpha_{i,j} = \frac{\lambda_{i,j}}{\sum_{k=1}^{N_i} \lambda_{i,k}} \quad (4.1)$$

where N_i denotes the dimension of S_i (128 for frequency, 12 for rate and 5 for scale). The number of PCs in each subspace then can be specified by including only the PCs whose α is larger than some threshold. Figure 4.1 shows the number of principal components in each of the three subspaces as a function of threshold on the percentage of contribution. In Figure 4.2, the classification accuracy is demonstrated as a function of threshold. Based on this analysis, the minimum number of principal components to achieve 100% accuracy was specified to be 7 for frequency, 5 for rate and 4 for scale subspace which includes PCs that have contribution of 3.5% or more.

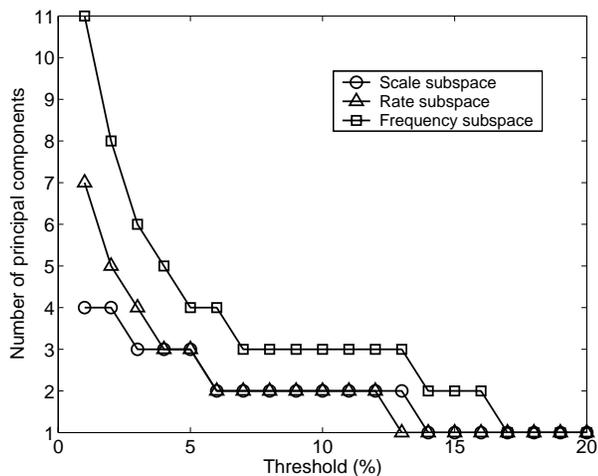


Figure 4.1: Total number of retained Principal Components (PCs) in each of the subspaces of frequency, rate and scale as a function of threshold on contribution percentage. The Y axis indicates the number of PCs in each subspace that have contribution (α from equation 4.1) more than the threshold.

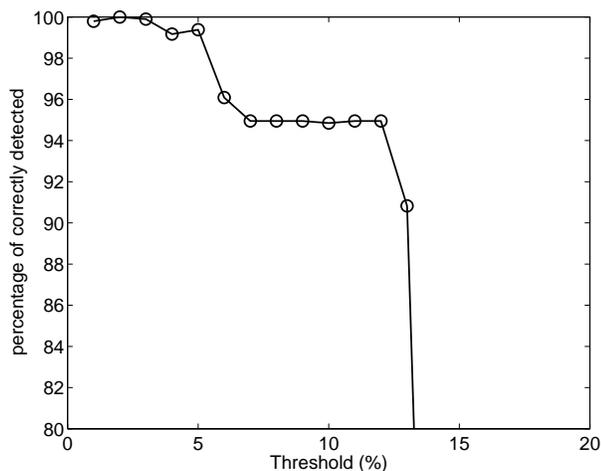


Figure 4.2: Percentage of correctly classified samples as a function of threshold on contribution percentage.

4.3 Comparison and results

To evaluate the robustness and the ability of system to generalize to unseen noisy conditions, we conducted a comparison with two state-of-the-art studies, one from generic-audio analysis community by Scheirer and Slaney [1] and one from automatic-speech-recognition community by Kingsbery *et al.* [2].

Multifeature [1]: The first system, which was originally designed to distinguish speech

	Auditory Model	Multifeature[1]	Voicing-Energy [2]
Correct Speech	100%	99.3%	91.2%
Correct Non-speech	100%	100%	96.3%

Table 4.1: Percentage of correct classification for window length of one second

	Auditory Model	Multifeature[1]	Voicing-Energy[2]
Correct Speech	99.4%	98.7%	90.0%
Correct Non-speech	99.4%	99.5%	94.9%

Table 4.2: Percentage of correct classification for window length of half a second

from music, derived thirteen features in time, frequency and cepstrum domain to represent speech and music. The features were 4Hz modulation energy, percentage of "low-energy" frames, spectral rolloff point, spectral centroid, spectral flux, zero-crossing rate, cepstrum resynthesis residual and their variances. The thirteenth feature, pulse metric, was neglected for this comparison since its latency was too long (more than two seconds).

In original system, two models were formed for speech and music in the feature space. Classification was performed using a likelihood estimate of a given sample for each model. To eliminate performance differences due to the use of different classifiers, an SVM with an RBF kernel was used in all comparisons. Our implementation of the system was first evaluated on the original database and similar or better results were obtained with SVM compared to the original publication [1].

Voicing-Energy [2]: A second system was tested that was based on an audio segmentation algorithm from the ASR work [2]. In the proposed technique, the feature vector used in the segmentation incorporated information about the degree of voicing and frame-level log-energy value. Degree of voicing is computed by finding the maximum of autocorrelation in a specified range, whereas log-energy was computed for every short frame of sound weighted with a hanning window. Several frames of these features were then concatenated and sorted in increasing order, and the resulting feature vector was reduced to two dimensions by a linear discriminant analysis followed by diagonalizing transform. The reason for sorting the elements was to eliminate details of temporal evolutions which were not relevant for this task. Our evaluation of Kingsbury’s system suggested that direct classification of the original sorted vector with an SVM classifier similar to the other two systems outperformed the one in reduced dimension. For this reason, the classification was performed in the original feature space.

Our auditory model and the two benchmark algorithms from the literature were trained and tested on the same database. One of the important parameters in any such speech detec-

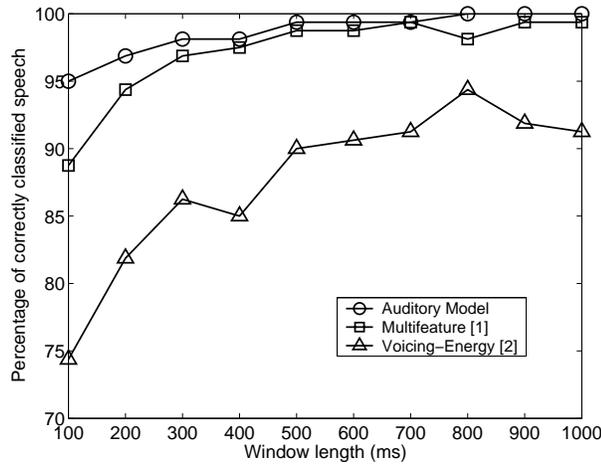


Figure 4.3: Effect of window length on the percentage of correctly classified speech

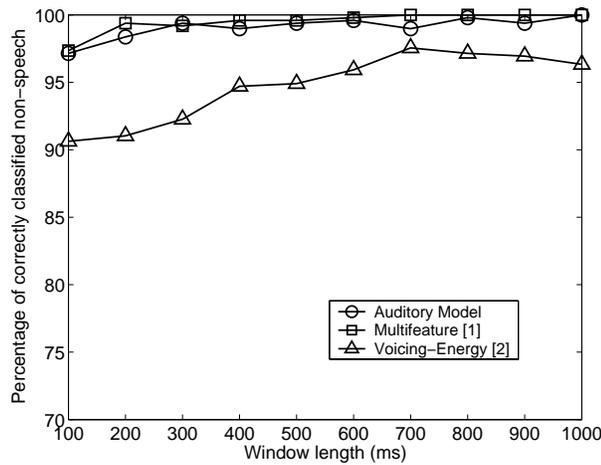


Figure 4.4: Effect of window length on the percentage of correctly classified non-speech

tion/discrimination task is the time window or duration of the signal to be classified, because it directly affects the resolution and accuracy of the system. Figures 4.3 and 4.4 demonstrate the effect of window length on the percentage of correctly classified speech and non-speech. In all three methods, some features may not give a meaningful measurement when the time window is too short. The classification performance of the three systems for two window length of 1 and 0.5 second is shown in Tables 4.1 and 4.2. The accuracy of all three systems improve as the time window increases.

Audio processing systems designed for realistic applications must be robust in a variety of conditions because training the systems for all possible situations is impractical. A series of tests were conducted to evaluate the generalization of the three methods to unseen noisy and reverberant sound. Classifiers were trained solely to discriminate clean speech from non-speech and then

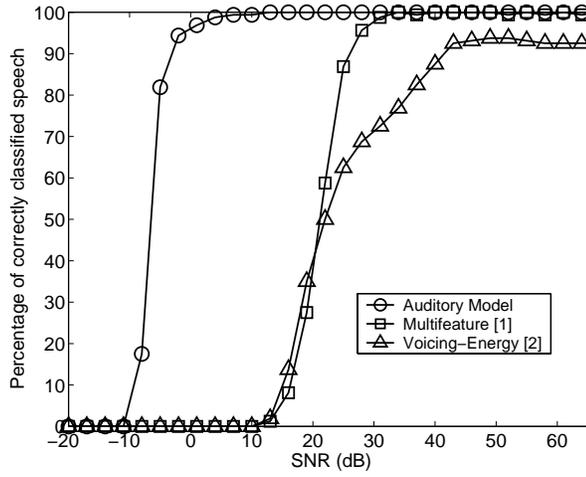


Figure 4.5: Effects of white noise on percentage of correctly classified speech for auditory model, multifeature[1] and voicing-energy[2] methods.

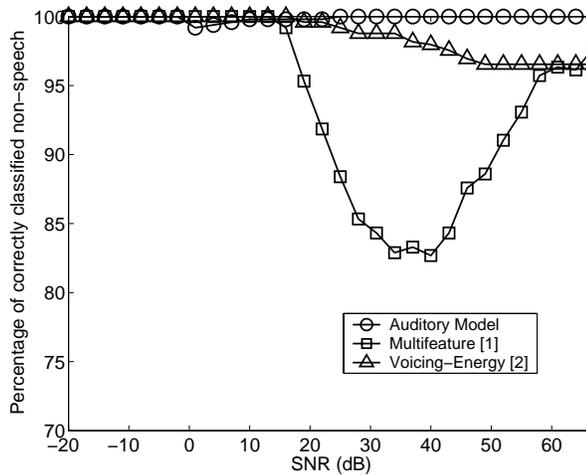


Figure 4.6: Effects of white noise on percentage of correctly classified non-speech for auditory model, multifeature[1] and voicing-energy[2] methods.

tested in three conditions in which speech was distorted with noise or reverberation. In each test, the percentage of correctly detected speech and non-speech was considered as the measure of performance. For the first two tests, white and pink noise were added to speech with specified signal to noise ratio (SNR). White and pink noise were not included in the training set as non-speech samples. SNR was measured from the average power of speech and noise:

$$SNR = 10 \log \frac{P_s}{P_n} \quad (4.2)$$

Figures 4.11 and 4.12 illustrate the effect of white and pink noise on the average spectro-temporal modulations of speech. The spectro-temporal representation of noisy speech preserves

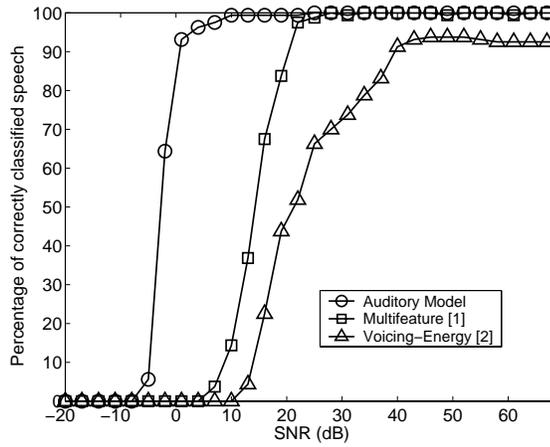


Figure 4.7: Effects of pink noise on percentage of correctly classified speech for auditory model, multifeature[1] and voicing-energy[2] methods.

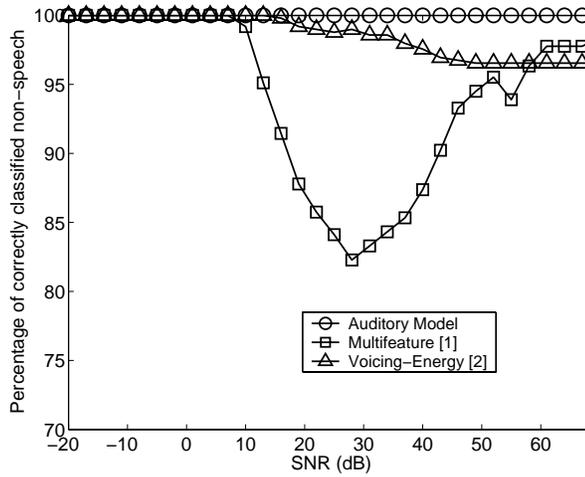


Figure 4.8: Effects of pink noise on percentage of correctly classified non-speech for auditory model, multifeature[1] and voicing-energy[2] methods.

the speech specific features (e.g. near 4Hz, 2Cyc/Oct) even at SNR as low as 0 dB (Figures 4.11 and 4.12, middle). The detection results for speech in white noise (Figures 4.5, 4.6) demonstrate that while the three systems have comparable performance in clean conditions, the auditory features remain robust down to fairly low SNRs. This pattern is repeated with additive pink noise although performance degradation for all systems occurs at higher SNRs (Figures 4.7 and 4.8) because of more overlap between speech and noise energy.

Reverberation is another widely encountered distortion in realistic applications. To examine the effect of different levels of reverberation on the performance of these systems, a realistic reverberation condition was simulated by convolving the signal with a random gaussian noise with

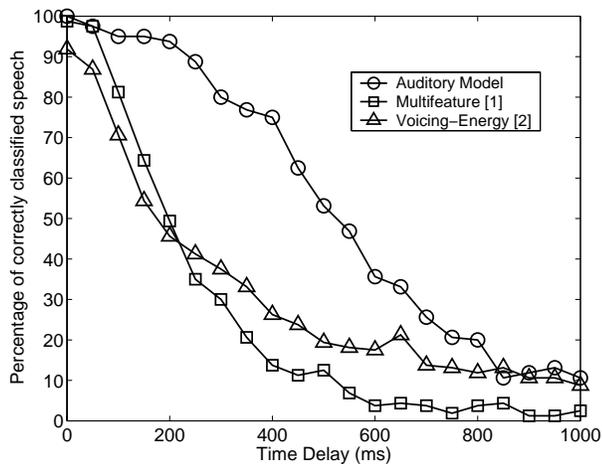


Figure 4.9: Effects of reverberation on percentage of correctly classified speech for auditory model, multifeature[1] and voicing-energy[2] methods.

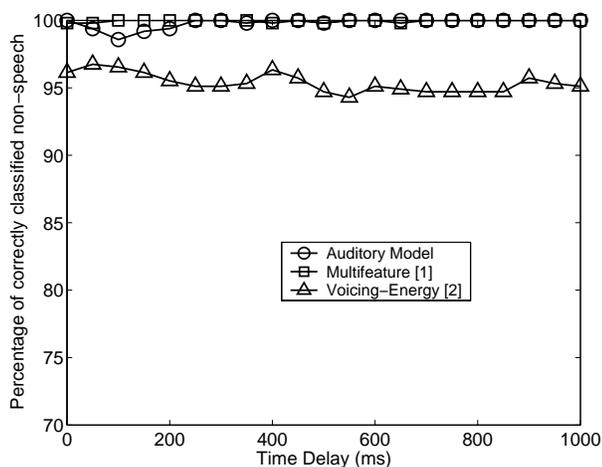


Figure 4.10: Effects of reverberation on percentage of correctly classified non-speech for auditory model, multifeature[1] and voicing-energy[2] methods.

exponential decay. The effect on the average spectro-temporal modulations of speech are shown in Figure 4.13. Increasing the time delay results in gradual loss of high-rate temporal modulations of speech. Figures 4.9 and 4.10 demonstrate the effect of reverberation on the classification accuracy.

On the whole, these tests demonstrate the significant robustness of the auditory model.

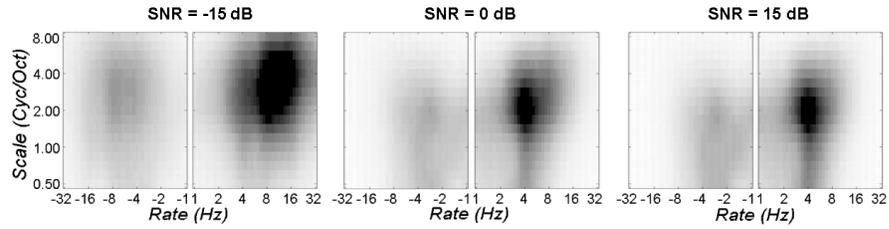


Figure 4.11: Effect of *white* noise on average spectro-temporal modulations of speech for SNRs -15dB , 0dB and 15dB . The spectro-temporal representation of noisy speech preserves the speech specific spectro-temporal features (e.g. near 4Hz , 2Cyc/Oct) even at SNR as low as 0dB .

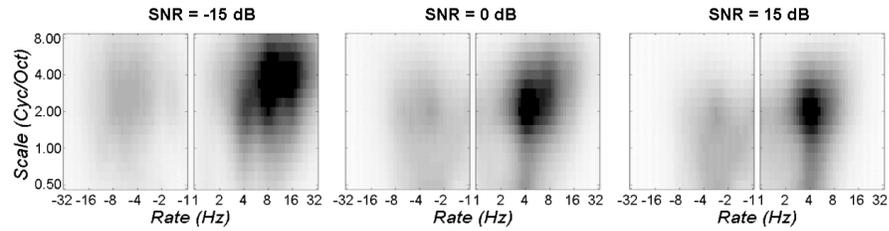


Figure 4.12: Effects of *pink* noise on average spectro-temporal modulations of speech for different SNRs -15dB , 0dB and 15dB . The speech specific spectro-temporal features (e.g. near 4Hz , 2Cyc/Oct) are preserved even at SNR as low as 0dB .

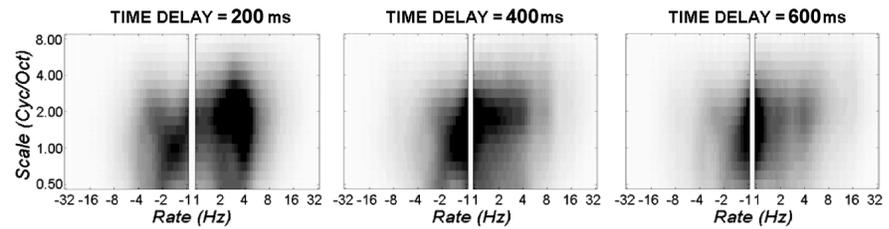


Figure 4.13: Effects of *reverberation* on average spectro-temporal modulations of speech for time delays 200ms , 400ms and 600ms . Increasing the time delay results in gradual loss of high-rate temporal modulations of speech.

A *Spectro-Temporal Auditory method* for audio classification and segmentation has been described, tested and compared to two state-of-the-art alternative approaches. The method employs features extracted by a biologically inspired auditory model of auditory processing in the cortex. Unlike conventional features, auditory based features have multiple-scales of time and spectral resolution. The drawback of such a representation is its high dimensionality, and hence to utilize it, we applied an efficient multilinear dimensionality reduction algorithm based on HOSVD of multimodal data.

The performance of the proposed auditory system was tested in noise and reverberation, and compared favorably with alternative systems, thus demonstrating that the proposed system generalizes well to novel situations, an ability that is generally lacking in many of today's audio and speech recognition and classification systems.

This work is but one in a series of efforts at incorporating multiscale cortical representations (and more broadly, perceptual insights) in a variety of audio and speech processing applications. For example, the deterioration of the spectro-temporal modulations of speech in noise and reverberation (e.g. Figures 4.11,4.12 and 4.13), or indeed under any kind of linear or non-linear distortion, can be used as an indicator of predicted speech intelligibility [14]. Similarly, the multi-scale rate-scale-frequency representation can account for the perception of complex sounds and perceptual thresholds in a variety of settings [37]. Finally, the auditory model can be adapted and expanded for a wide range of applications such as the speech enhancement [19], or the efficient encoding of speech and music [20].

BIBLIOGRAPHY

- [1] E. Scheirer, and M. Slaney, *Construction and evaluation of a robust multifeature speech/music discriminator*, International Conference on Acoustic, Speech and Signal Processing, Munich, Germany, 1997.
- [2] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, *Robust speech recognition in noisy environments: The 2001 IBM SPINE evaluation system*, International Conference on Acoustic, Speech and Signal Processing, vol. I, Orlando, FL, pp. 53-56, May 2002.
- [3] S. Pfeiffer, S. Fischer, and W. Efferlsberg, *Automatic audio content analysis*, in proc. 4th ACM International Multimedia Conference, pp.21-30, 1996.
- [4] E. Wold, T. Blum, and D. Keislar *et al.*, *Content-based classification, search, and retrieval of audio*, IEEE multimedia, pp. 27-36, Fall 1996.
- [5] J. Foote, *Content-based retrieval of music and audio*, In C.-C. J. Kuo et al., editor, Multimedia Storage and Archiving Systems II, Proc. of SPIE, Vol. 3229, pp. 138-147, 1997.
- [6] L. Lu, H. Zhang, and H. Jiang, *Content analysis for audio classification and segmentation*, IEEE Transaction on Speech, Audio and Signal Processing, Vol. 10, No. 7, 2002.
- [7] J. Saunders, *Real-time discrimination of broadcast speech/music*, Proc. International Conference on Acoustic, Speech and Signal Processing, vol. II, Atlanta, GA, pp. 993-996, May 1996.
- [8] K. Wang, and S. A. Shamma, *Spectral shape analysis in the central auditory system*, IEEE Transaction on Speech and Audio Processing, vol. 3 (5), pp. 382-395, 1995.
- [9] R. Lyon, and S. Shamma, *Auditory representation of timbre and pitch*, Auditory computation, vol. 6 of Springer handbook of auditory research, Springer-Verlag, New York, pp. 221-270, 1996.
- [10] X. Yang, K. Wang, and S. A. Shamma, *Auditory representation of acoustic signals*, IEEE Transaction on Information Theory, 38 (2), pp. 824-839, (Special issue on wavelet transforms and multi-resolution signal analysis), 1992.
- [11] N. Kowalski, D. A. Depireux, and S. A. Shamma, *Analysis of dynamic spectra in ferret primary auditory cortex I. Characteristics of single-unit response to moving ripple spectra*, Journal of Neurophysiology, vol. 76 (5), pp. 3503-3523, 1996.
- [12] S. A. Shamma, *Methods of neuronal modeling*, In: Spatial and temporal processing in the auditory system, second edition, MIT press, Cambridge, MA, pp. 411-460, 1998.

- [13] D. A. Depireux, J. Z. Simon, D. J. Klein, and S. A. Shamma, *Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex*, Journal of Neurophysiology vol. 85, pp. 1220-1234, 2001.
- [14] M. Elhilali, T. Chi, and S. A. Shamma, *A spectro-temporal modulation index (STMI) for assessment of speech intelligibility*, Speech communication, vol. 41, pp. 331-348, 2003.
- [15] H. Dudley, *Remaking speech*, Journal of Acoustical Society of America. 11(2), pp. 169-177, 1939.
- [16] B. Kingsbury, N. Morgan, and S. Greenberg, *Robust speech recognition using the modulation spectrogram*, Speech Communication, vol. 25, no. 1, pp. 117-132, 1998.
- [17] T. Dau, B. Kollmeier, and A. Kohlrausch, *Modeling auditory processing of amplitude modulation: II. Spectral and temporal integration*, Journal of Acoustical Society of America, 102(5), pp. 2906-2919, 1997.
- [18] M. Kleinschmidt, J. Tchorz, and B. Kollmeier *Combining Speech Enhancement and Auditory Feature Extraction for Robust Speech Recognition*, Speech Communication 34(1-2): Special Issue on Robust ASR, pp. 75-91, 2001.
- [19] N. Mesgarani and S. A. Shamma, *Speech enhancement base on filtering the spectrotemporal modulations*, International Conference on Acoustic, Speech and Signal processing, Philadelphia, PA, March 2005.
- [20] L. Atlas, S. A. Shamma, *Joint acoustic and modulation frequency*, Eurasip Journal on Applied Signal Processing, No. 7, pp. 668-675, June 2003.
- [21] M. Kleinschmidt, and D. Gelbart. *Improving Word Accuracy with Gabor Feature Extraction*, International Conference on Spoken Language Processing, Denver, CO, 2002.
- [22] R. Bellman, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, 1961.
- [23] L. De Lathauwer, B. De Moor, J. Vandewalle, *A multilinear singular value decomposition*, SIAM Journal of Matrix Analysis and Applications, vol. 21, pp. 1253-1278, 2000.
- [24] L. R. Tucker, *Some mathematical notes on three-mode factor analysis*, Psychometrika, vol. 31, pp. 279-311, 1966.
- [25] P. M. Kroonenberg, *Three-mode principal component analysis*, DSWO Press, Leiden, Netherlands, 1982.
- [26] L. De Lathauwer, B. De Moore, and J. Vandewalle. *On the best rank-1 and rank- (R_1, R_2, \dots, R_N) approximation of higher order tensors*, SIAM Journal of Matrix Analysis and Applications, vol. 21(4):1324-1342, 2000.

- [27] L. De Lathauwer, B. De Moor, J. Vandewalle, *Dimensionality reduction in higher-order-only ICA*, IEEE signal processing workshop on Higher Order Statistics, Banff, Alberta, Canada, pp.316-320, 1997.
- [28] M. A. O. Vasilescu and D. Terzopoulos, *Multilinear analysis of image ensembles: TensorFaces*, European Conference on Computer Vision, Copenhagen, Denmark. pp. 447-460, May 2002.
- [29] M. A. O. Vasilescu and D. Terzopoulos, *Multilinear subspace analysis of image ensembles*, IEEE Conference on Computer Vision and Pattern Recognition, Madison, WI, June 2003.
- [30] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.
- [31] T. Joachims, *Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning*, B. Scholkopf, C. Burges and A. Smola (ed.), MIT-Press, 1999.
- [32] S. Seneft, and V. Zue, *Transcription and alignment of the timit database*, in Getting Started with the DARPA TIMIT CDROM: An Acoustic Phonetic Continuous Speech Database, J. S. Garofolo, Ed. National Institute of Standards and Technology (NIST), Gaithersburg, MD, 1988.
- [33] BBC Sound Effects Library, Original Series, 40 Audio CD Collection, Distributed by Sound Ideas, (c) 1984.
- [34] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, *RWC Music Database: Music Genre Database and Musical Instrument Sound Database*, International Conference on Music Information Retrieval, pp. 229230, 2003.
- [35] A. Varga, H. J. M. Steenneken, M. Tomlinson, and D. Jones, *The NOISEX-92 study on the effect of additive noise on automatic speech recognition*, Documentation included in the NOISEX-92 CD-ROMs, 1992.
- [36] H. G. Hirsch, and D. Pearce, *The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions*, ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium; Paris, France, September 18-20, 2000.
- [37] R. P. Carlyon, S. A. Shamma, *An account of monaural phase sensitivity*, Journal of Acoustic Society of America, vol. 114(1), pp. 333-48, 2003.