Discrimination of Speech From Nonspeech Based on Multiscale Spectro-Temporal Modulations

Nima Mesgarani, *Student Member, IEEE*, Malcolm Slaney, *Senior Member, IEEE*, and Shihab A. Shamma, *Senior Member, IEEE*

***Abstract*—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 nonspeech 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 multilinear 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 (Scheirer and Slaney, 2002 and Kingsbury *et al.*, 2002). The results demonstrate the advantages of the auditory model over the other two systems, especially at low signal-to-noise ratios (SNRs) and high reverberation.**

***Index Terms*—Audio classification and segmentation, auditory model, speech discrimination.**

I. 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 nonspeech 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, nonharmonic and harmonic sounds, and natural silences between words and phonemes. Discriminating speech from nonspeech 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 [5], or noise frame ratios and band periodicity along with K-nearest neighbor and linear spectral pair-vector quantization[6], average zero-crossing rate and energy with a simple threshold to discriminate between speech and music [7], and an optimized dimensionality reduction using distortion discriminant analysis (DDA) [8].

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., maximum *a posteriori* (MAP), Gaussian mixture model (GMM), K nearest neighbor (KNN)] are then employed to achieve a robust performance. The second system is a speech/nonspeech 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 [9]–[12]. 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

Manuscript received February 2, 2004; revised May 20, 2005. This work was supported in part by the National Science Foundation under ITR 1150086075 and a U.S. Air Force STTR under proposal F033-0061, topic number AF03T006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Li Deng.

N. Mesgarani and S. Shamma are with the Electrical and Computer Engineering Department, University of Maryland, College Park, MD 20742 USA (e-mail: mnima@eng.umd.edu; sas@eng.umd.edu).

M. Slaney is with IBM Almaden Research Center, San Jose, CA 95120-6099 USA (e-mail: malcolm@almaden.ibm.com).

Digital Object Identifier 10.1109/TSA.2005.858055

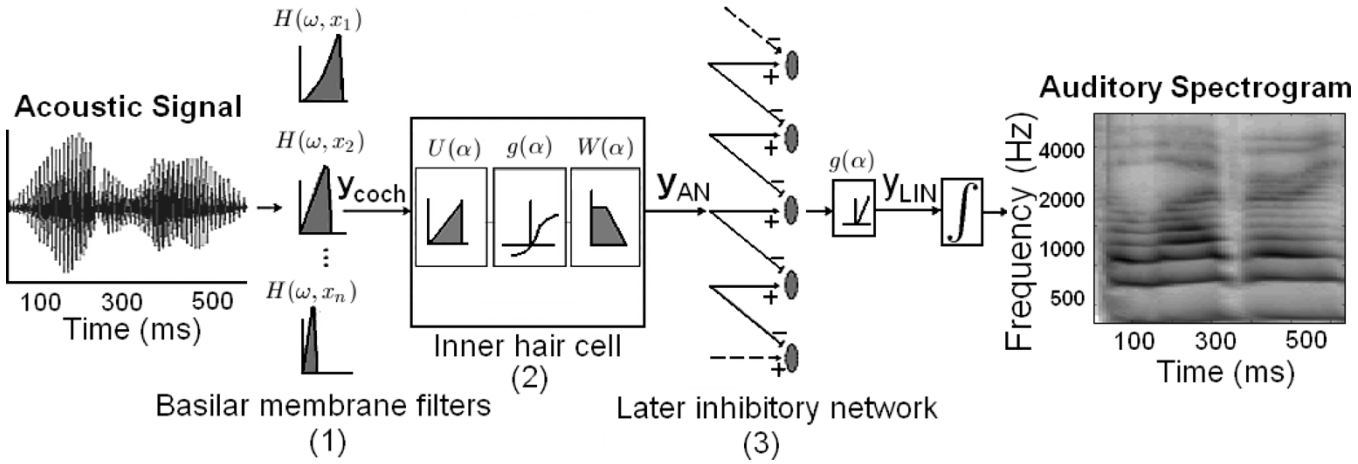

Fig. 1. Schematic of the early stages of auditory processing. (1) Sound is analyzed by a model of the cochlea consisting of a bank of 128 constant-Q bandpass filters with center frequencies equally spaced on a logarithmic frequency axis (tonotopic axis). (2) Each filter output is then transduced into auditory-nerve patterns by a hair cell stage which is modeled as a three-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 $Y_{\rm LIN}$ over a short window, mimicking the further loss of phase-locking observed in the midbrain.

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].

II. Auditory Model

The computational auditory model is based on neurophysiological, biophysical, and psychoacoustical investigations at various stages of the auditory system [9]–[11]. 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 [9]. This stage is responsible for extracting the key features upon which the classification is based.

A. Early Auditory System

The stages of the early auditory model are illustrated in Fig. 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 $h_{\rm cochlea}(t; f)$. The cochlear filter outputs $y_{\rm cochlea}(t, f)$ are then transduced into auditory-nerve patterns $y_{\rm an}(t, f)$ by a hair cell stage which converted cochlear outputs into inner hair cell intracellular potentials. This process is modeled as three-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 [13]. 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 $y_{\rm LIN}(t, f)$. The final output of this stage is obtained by integrating $y_{\rm LIN}(t, f)$ over a short window, $\mu_{\rm midbrain}(t, \tau)$, with time constant $\tau = 8$ ms 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 [9].

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

$$y_{\rm cochlea}(t, f) = s(t) * h_{\rm cochlea}(t; f) \tag{1}$$

$$y_{\rm an}(t, f) = g_{hc}\left(\partial_t y_{\rm cochlea}(t, f)\right) * \mu_{hc}(t) \tag{2}$$

$$y_{\rm LIN}(t, f) = \max\left(\partial_f y_{\rm an}(t, f), 0\right) \tag{3}$$

$$y(t, f) = y_{\rm LIN}(t, f) * \mu_{\rm midbrain}(t; \tau) \tag{4}$$

where $*$ denotes convolution in time.

The above sequence of operations effectively computes a spectrogram of the speech signal (Fig. 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.

B. Central Auditory System

Higher central auditory stages (especially the primary auditory cortex) further analyze the auditory spectrum into more

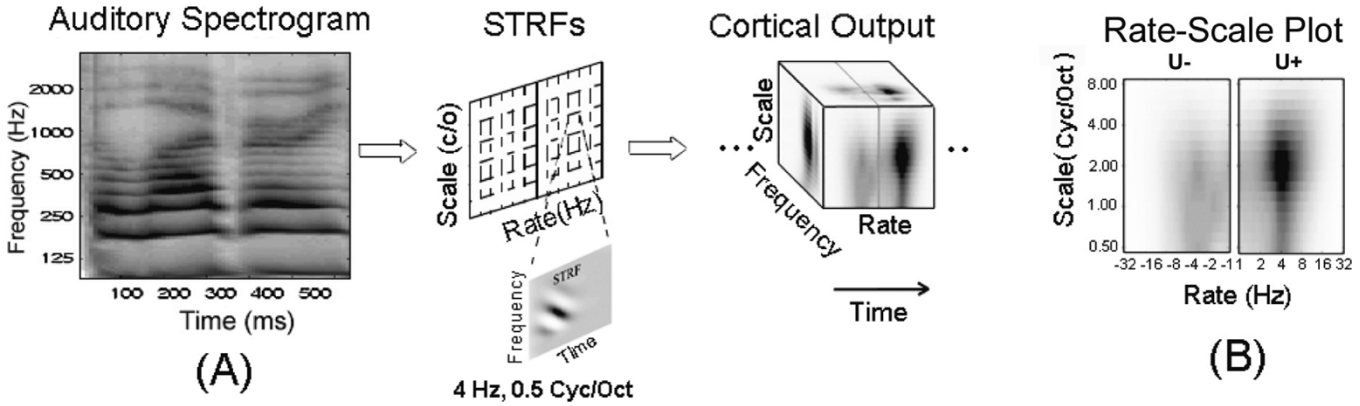

Fig. 2. (a) Cortical multiscale representation of speech. The auditory spectrogram (the output of the early stage) is analyzed by a bank of spectro-temporal modulation selective filters. The spectro-temporal response field (STRF) of one such filter is shown which corresponds to a neuron that responds well to a ripple of 4-Hz 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 (equation (21) and (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_-).

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 two-dimensional 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 hertz) 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 Fig. 2.

We assume a bank of directional selective STRFs (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 STRFs in primary auditory cortex have the quadrant separability property [14]

$$\mathrm{STRF}_+ = \Re\{H_{\mathrm{rate}}(t;\omega,\theta)\cdot H_{\mathrm{scale}}(f;\Omega,\phi)\} \quad (5)$$

$$\mathrm{STRF}_- = \Re\{H^*_{\mathrm{rate}}(t;\omega,\theta)\cdot H_{\mathrm{scale}}(f;\Omega,\phi)\} \quad (6)$$

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}

$$H_{\mathrm{rate}}(t;\omega,\theta) = h_{\mathrm{rate}}(t;\omega,\theta) + j\hat{h}_{\mathrm{rate}}(t;\omega,\theta) \quad (7)$$

$$H_{\mathrm{scale}}(f;\Omega,\phi) = h_{\mathrm{scale}}(f;\Omega,\phi) + j\hat{h}_{\mathrm{scale}}(f;\Omega,\phi) \quad (8)$$

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

$$h_{\mathrm{rate}}(t;\omega,\theta) = h_r(t;\omega)\cos\theta + \hat{h}_r(t;\omega)\sin\theta \quad (9)$$

$$h_{\mathrm{scale}}(f;\Omega,\phi) = h_S(f;\Omega)\cos\phi + \hat{h}_S(f;\Omega)\sin\phi. \quad (10)$$

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

$$h_r(t;\omega) = \omega h_r(\omega t) \quad (11)$$

$$h_s(f;\Omega) = \Omega h_s(\Omega f). \quad (12)$$

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

$$r_+(t,f;\omega,\Omega;\theta,\phi) = y(t,f) *_{t,f} \mathrm{STRF}_+(t,f;\omega,\Omega;\theta,\phi) \quad (13)$$

$$r_-(t,f;\omega,\Omega;\theta,\phi) = y(t,f) *_{t,f} \mathrm{STRF}_-(t,f;\omega,\Omega;\theta,\phi) \quad (14)$$

where $*_{tf}$ 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

$$h_{rw}(t;\omega) = h_r(t;\omega) + j\hat{h}_r(t;\omega) \quad (15)$$

$$h_{sw}(f;\Omega) = h_s(f;\Omega) + j\hat{h}_s(f;\Omega). \quad (16)$$

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

$$z_+(t,f;\Omega,\omega) = y(t,f) *_{tf} [h^*_{rw}(t;\omega)h_{sw}(f;\Omega)] \quad (17)$$

$$z_-(t,f;\Omega,\omega) = y(t,f) *_{tf} [h_{rw}(t;\omega)h_{sw}(f;\Omega)] \quad (18)$$

where $*$ denotes the complex conjugate. The cortical response [(13) and (14)] for all characteristic phases θ and ϕ can be easily obtained from $z_+(\cdot)$ and $z_-(\cdot)$ as follows:

$$r_+(t,f;\omega,\Omega;\theta,\phi) = |z_+|\cos(\angle z_+ - \theta - \phi) \quad (19)$$

$$r_-(t,f;\omega,\Omega;\theta,\phi) = |z_-|\cos(\angle z_- + \theta - \phi) \quad (20)$$

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 STRFs tuned to the same (f, ω, Ω), the ones with $\theta = (1/2)(\angle z_+ + \angle z_-)$ and $\phi = (1/2)(\angle z_+ - z_-)$ 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 the purpose of classification. 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 STRFs to generate the rate-scale plots: [as shown in Fig. 2(b) for speech]

$$u_+(\omega, \Omega) = \sum_t \sum_f |z_+(t, f; \omega, \Omega)| \tag{21}$$

$$u_-(\omega, \Omega) = \sum_t \sum_f |z_-(t, f; \omega, \Omega)|. \tag{22}$$

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 [12], [14], and [15].

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 32 Hz, and spectral filters (scale) from 0.5 to 8.00 cycle/octave, were used to represent the spectro-temporal modulations of the sound.

C. Models of Modulation Filter

The importance of slow temporal modulations of sound in speech intelligibility has been emphasized for a long time [16]. Kingsbury *et al.* [17] showed the advantage of using modulation spectrogram in improving the robustness of automatic speech recognition systems to noise and reverberation. Temporal modulation filter banks inspired by psycoacoustical experiments [18] have been successfully used in a variety of audio processing tasks such as automatic speech recognition [19]. Spectro-temporal features have recently also been used in speech enhancement [20], speech coding [21], and speech recognition to provide more robustness [22].

III. Multilinear Tensor Analysis

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 $(5\,(\text{scale filters}) \times 12\,(\text{rate filters}) \times 128\,(\text{frequency channels}) = 7680)$. Working in this feature space directly is impractical because a sizable number of training samples is required to characterize the space adequately [23]. Traditional dimensionality reduction methods like principal component analysis (PCA) are inefficient for multidimensional 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) [24].

A. 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 [25], while Kroonenberg formulated alternating least-square (ALS) method to implement three mode factor analysis [26]. Lathauwer *et al.* established a generalization of singular value decomposition (SVD) to higher order tensors [24], and also introduced an iterative method for optimizing the best rank $(R_1, R_2, \ldots, R_N)$ approximation of tensors [27]. Tensor algebra and HOSVD have been applied successfully in wide variety of fields including higher-order-only independent component analysis (ICA) [28], face recognition [29], and selective image compression along a desired dimension [30].

A Tensor $A \in R^{I_1 \times I_2 \times \cdots \times I_N}$ is a multi-index array of numerical values whose elements are denoted by $a_{i_1 i_2 \ldots i_N}$. Matrix column vectors are referred to as $\text{mode} - 1$ vectors and row vectors as $\text{mode} - 2$ vectors. The $\text{mode} - n$ vectors of an Nth-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 $\text{mode} - n$ matrix unfolding of $A \in R^{I_1 \times I_2 \times \cdots \times I_N}$ denoted by $A_{(n)}$ is the $(I_n \times I_1 I_2 \ldots I_{n-1} I_{n+1} \ldots I_N)$ matrix whose columns are $n - \text{mode}$ vectors of tensor A.

An Nth-order tensor A has $\text{rank} - 1$ when it is expressible as the outer product of N vectors

$$A = U_1 \circ U_2 \circ \ldots \circ U_N. \tag{23}$$

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

$$R_n = \text{rank}_n(A) = \text{rank}\left(A_{(n)}\right) \tag{24}$$

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

$$(A \times_n U)_{i_1 i_2 \ldots j_n \ldots i_N} = \sum_{i_n} a_{i_1 i_2 \ldots i_n \ldots i_N} u_{j_n i_n} \tag{25}$$

for all index values.

B. 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

$$D = U \cdot S \cdot V^T = S \times_1 U \times_2 V \tag{26}$$

in which U and V are unitary matrices contains the left- and right-singular vectors of D. S is a pseudodiagonal 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 [24]. Every $(I_1 \times I_2 \times \ldots \times I_N)$-tensor A can be written as the product

$$A = S \times_1 U^{(1)} \times_2 U^{(2)} \ldots \times_N U^{(N)} \tag{27}$$

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

$$A_{(n)} = U^{(n)} \cdot S_{(n)} \cdot \Big(U^{(n+1)} \otimes \ldots \otimes U^{(N)} \otimes U^{(1)} \otimes U^{(2)} \otimes \ldots \otimes U^{(n-1)}\Big)^T \tag{28}$$

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

$$A_{(n)} = U^{(n)} \cdot \Sigma^{(n)} \cdot V^{(n)T} \tag{29}$$

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

$$V^{(n)} = \Big(U^{(n+1)} \otimes \ldots \otimes U^{(N)} \otimes U^{(1)} \otimes U^{(2)} \otimes \ldots \otimes U^{(n-1)}\Big) \tag{30}$$

This shows that, at matrix level, the HOSVD conditions lead to an SVD of the matrix unfolding. Lathauwer *et al.* shows [24] 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 basis 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 $\text{rank} - R$ 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 $\text{rank} - (R_1, R_2, \ldots, R_N)$ approximation of a tensor can be obtained by an iterative algorithm in which HOSVD provides the initial values [27].

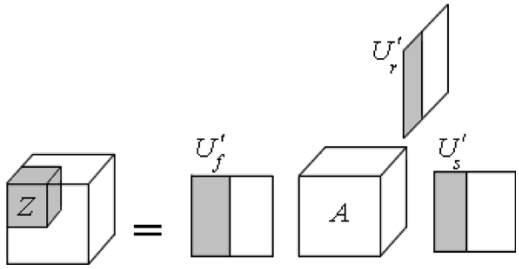

Fig. 3. Illustration of equation (32).

C. Multilinear 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 IV-A which contains 1223 samples from speech and nonspeech 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 $\text{mode} - n$ singular vectors

$$D = S \times_1 U_{\text{frequency}} \times_2 U_{\text{rate}} \times_3 U_{\text{scale}} \times_4 U_{\text{samples}} \tag{31}$$

in which $U_{\text{frequency}}$, 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 to their cortical representation, A, and are then projected onto these truncated orthonormal axes $U'_{\text{freq.}}$, U'_{rate}, U'_{scale} (as shown in Fig. 3)

$$Z = A \times_1 {U'_{\text{freq.}}}^T \times_2 {U'_{\text{rate}}}^T \times_3 {U'_{\text{scale}}}^T \tag{32}$$

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.

D. Classification

Classification was performed using a support vector machine (SVM) [31], [32]. 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 [31]. Radial basis function (RBF) were used as SVM kernel.

IV. Experimental Results

A. 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 [33] which contains short sentences spoken by male and female native English speakers with eight dialects. Two hundred ninety-nine 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 nonspeech 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 [34] (263 for training, 139 for test). Music samples that covered a large variety of musical styles were selected from RWC genre database [35] (349 for training, 185 for test). Environmental sounds were assembled from Noisex [36] and Auroa [37] databases which have stationary and nonstationary 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 nonspeech samples and the test set consisted of 160 speech and 491 nonspeech samples. The length of each utterance in training and test is equal to the selected time window (e.g., one 1-s sample per sound file).[1]

B. Number of Principal Components

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

$$\alpha_{i,j} = \frac{\lambda_{i,j}}{\sum_{k=1}^{N_i} \lambda_{i,k}} \tag{33}$$

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. Fig. 4 shows the number of principal components in each of the three subspaces as a function of threshold on the percentage of contribution. In Fig. 5, 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.

[1]The list of files and offsets is available from the authors.

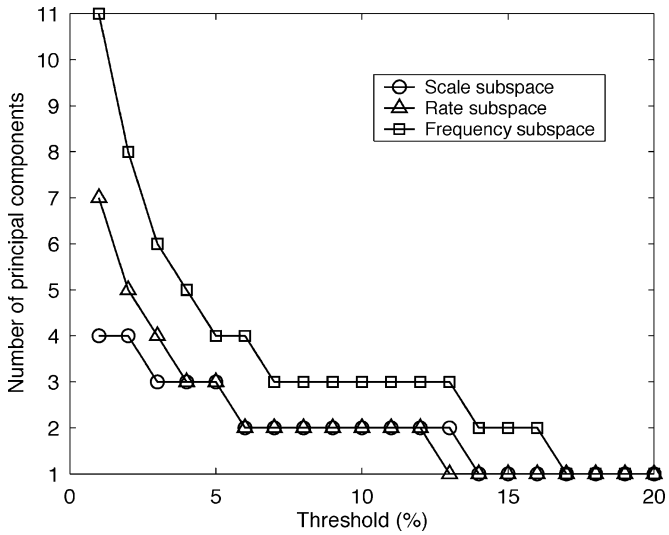

Fig. 4. Total number of retained PCs in each of the subspaces of frequency, rate, and scale as a function of threshold on contribution percentage. The vertical axis indicates the number of PCs in each subspace that have contribution [α from equation (33)] more than the threshold.

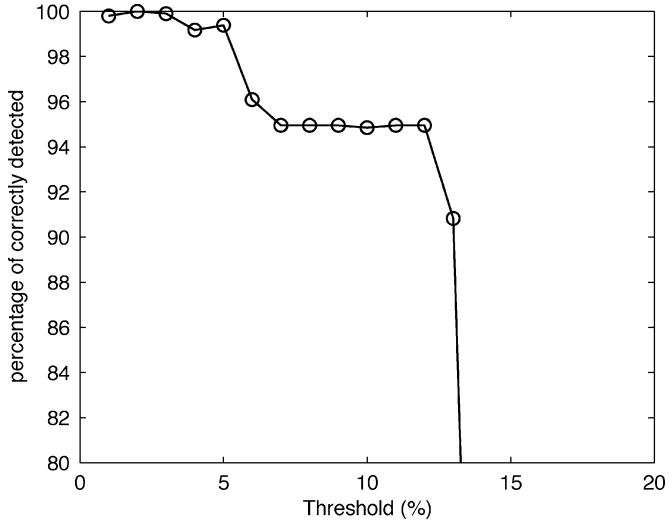

Fig. 5. Percentage of correctly classified samples as a function of threshold on contribution percentage.

C. 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 from music, derived 13 features in time, frequency, and cepstrum domain to represent speech and music. The features were 4-Hz modulation energy, percentage of "low-energy" frames, spectral rolloff point, spectral centroid,

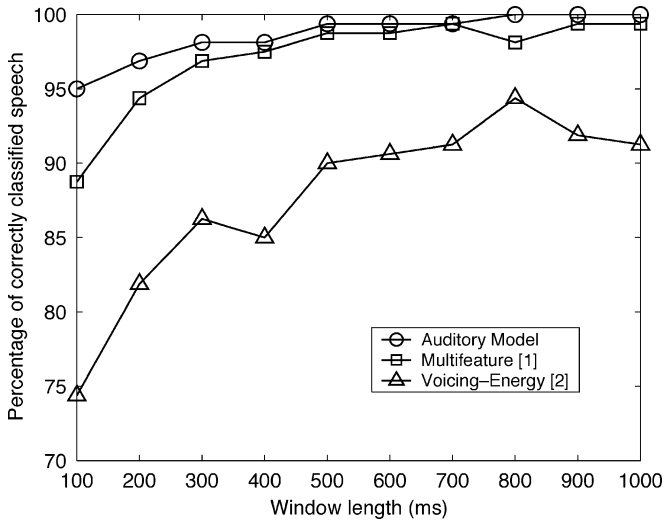

Fig. 6. Effect of window length on the percentage of correctly classified speech.

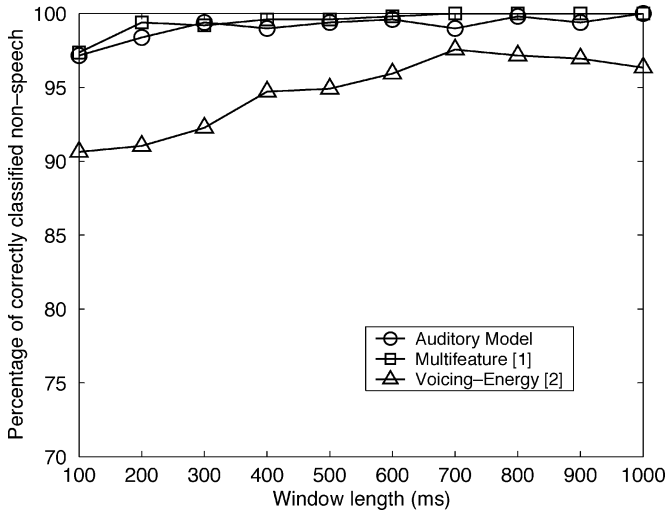

Fig. 7. Effect of window length on the percentage of correctly classified nonspeech.

spectral flux, zero-crossing rate, cepstrum resynthesis residual, and their variances. The 13th feature, pulse metric, was neglected for this comparison since its latency was too long (more than 2 s).

In the 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

TABLE I
PERCENTAGE OF CORRECT CLASSIFICATION FOR WINDOW LENGTH OF ONE SECOND

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

TABLE II
PERCENTAGE OF CORRECT CLASSIFICATION FOR WINDOW LENGTH OF HALF A 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%**

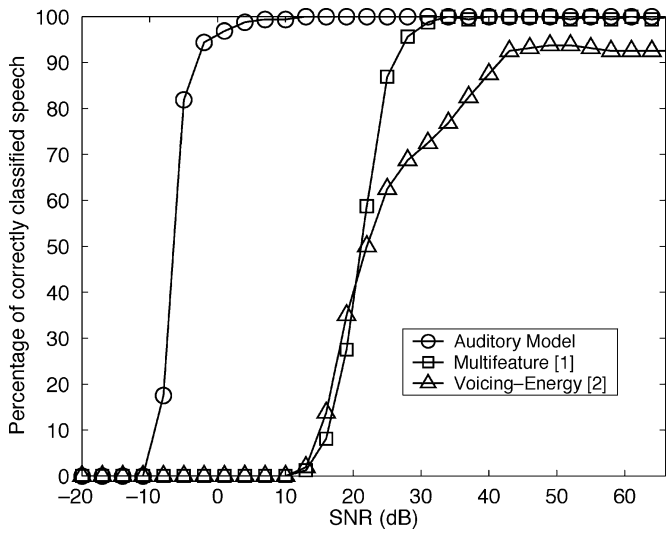

Fig. 8. Effects of white noise on percentage of correctly classified speech for auditory model, multifeature [1], and voicing-energy [2] methods.

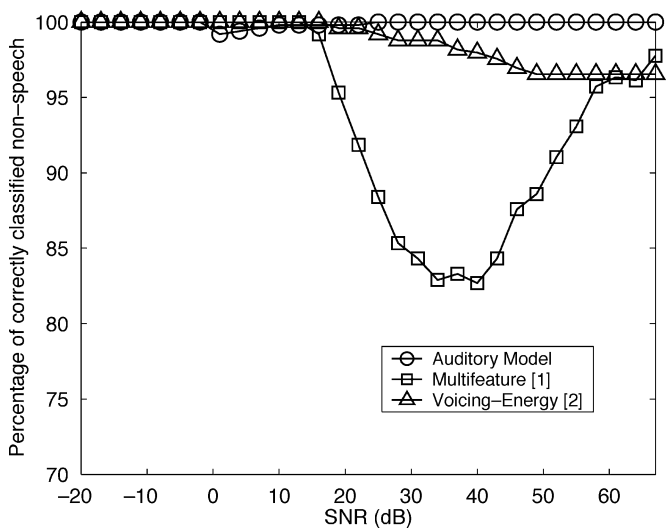

Fig. 9. Effects of white noise on percentage of correctly classified nonspeech for auditory model, multifeature [1], and voicing-energy [2] methods.

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

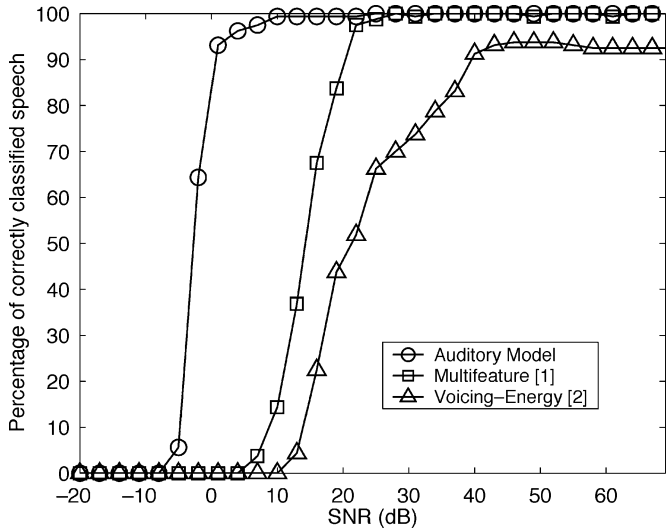

Fig. 10. Effects of pink noise on percentage of correctly classified speech for auditory model, multifeature [1], and voicing-energy [2] methods.

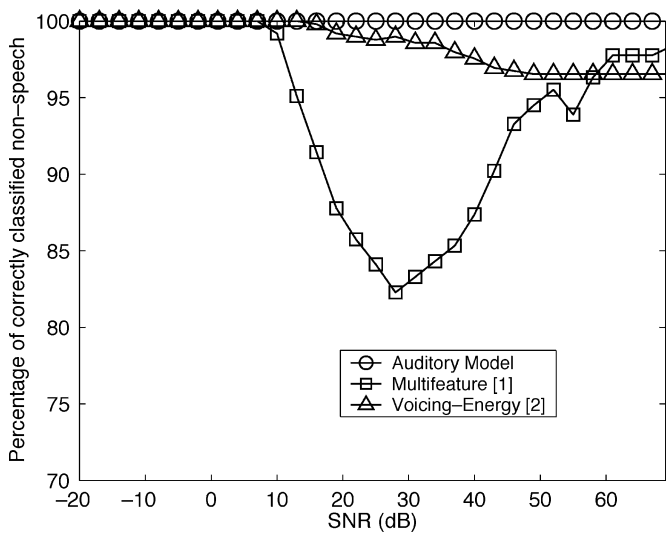

Fig. 11. Effects of pink noise on percentage of correctly classified nonspeech for auditory model, multifeature [1], and voicing-energy [2] methods.

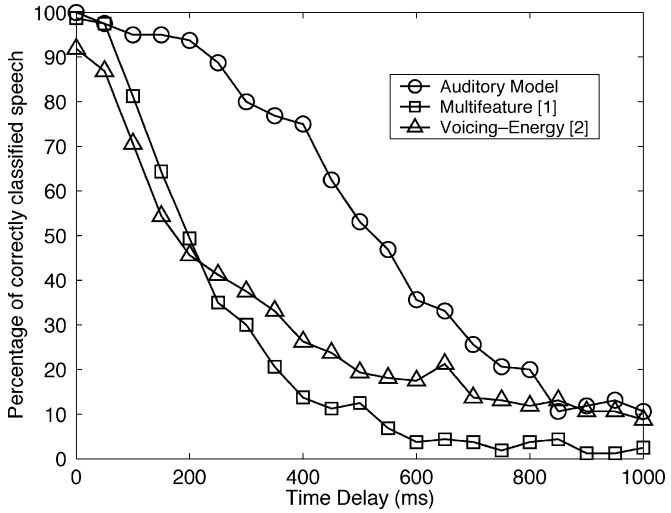

Fig. 12. Effects of reverberation on percentage of correctly classified speech for auditory model, multifeature [1], and voicing-energy [2] methods.

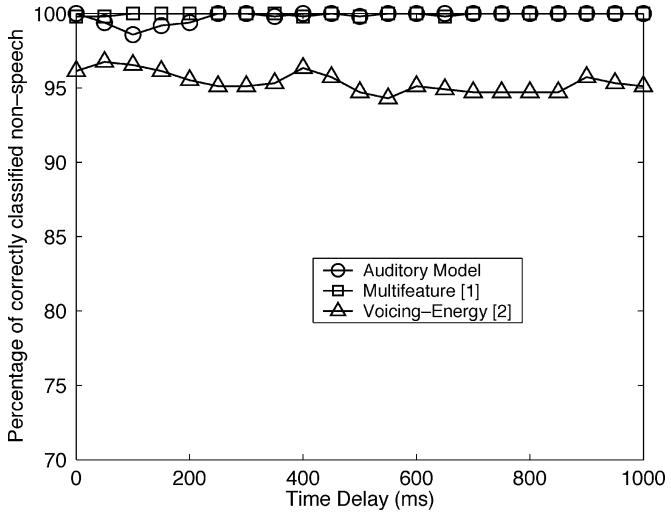

Fig. 13. Effects of reverberation on percentage of correctly classified nonspeech for auditory model, multifeature [1], and voicing-energy [2] methods.

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 detection/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. Figs. 6 and 7 demonstrate the effect of window length on the percentage of correctly classified speech and nonspeech. 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 lengths of 1 and 0.5 s is shown in Tables I and II. 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. Detection of speech at very low SNR is desired in many applications such as speech enhancement in which a robust detection of nonspeech (noise) frames is crucial for accurate measurement of the noise statistics [20]. 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 nonspeech and then tested in three conditions in which speech was distorted with noise or reverberation. In each test, the percentage of correctly detected speech and nonspeech 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

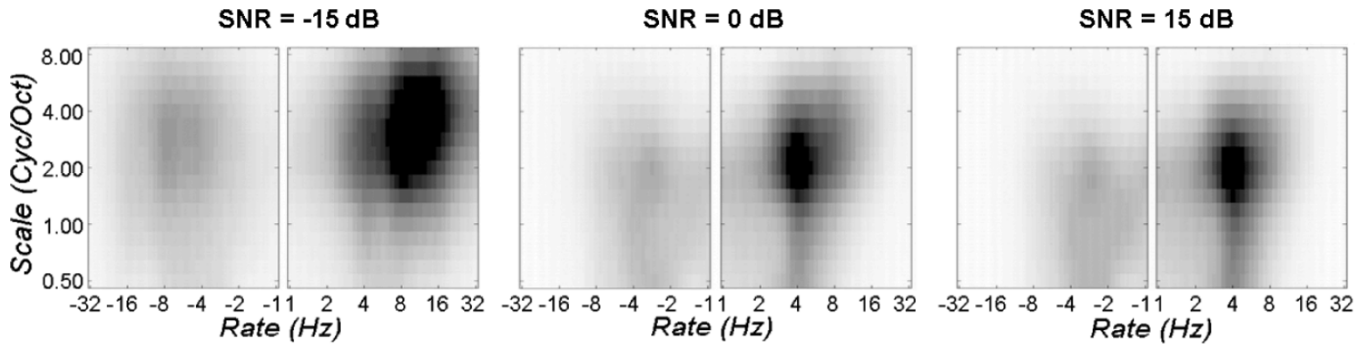

Fig. 14. Effect of *white* noise on average spectro-temporal modulations of speech for SNRs −15, 0, and 15 dB. The spectro-temporal representation of noisy speech preserves the speech specific spectro-temporal features (e.g., near 4 Hz, 2 cycle/octave) even at SNR as low as 0 dB.

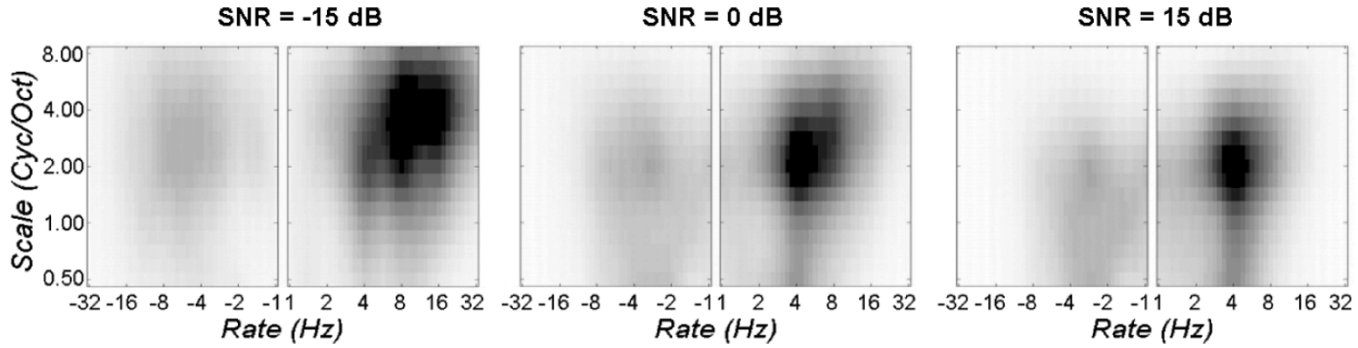

Fig. 15. Effects of *pink* noise on average spectro-temporal modulations of speech for different SNRs −15, 0, and 15 dB. The speech specific spectro-temporal features (e.g. near 4 Hz, 2 cycle/octave) are preserved even at SNR as low as 0 dB.

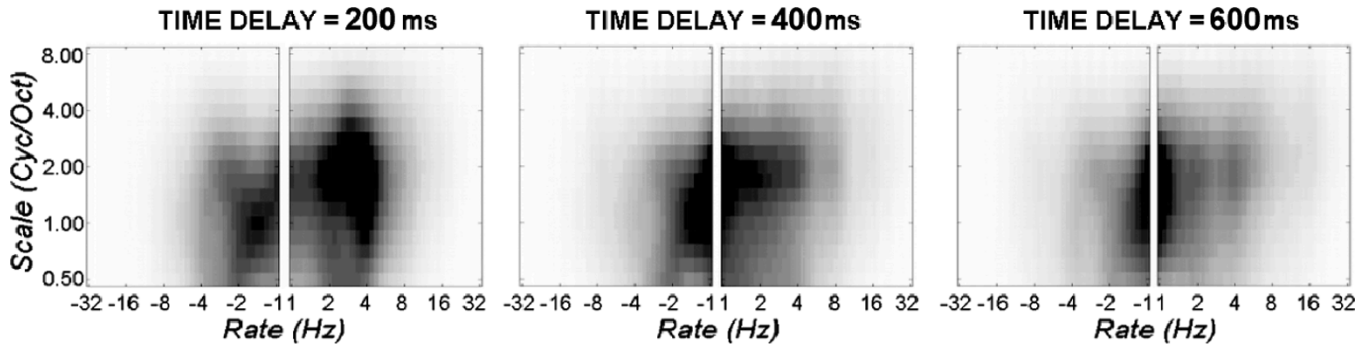

Fig. 16. Effects of *reverberation* on average spectro-temporal modulations of speech for time delays 200, 400, and 600 ms. Increasing the time delay results in gradual loss of high-rate temporal modulations of speech.

not included in the training set as nonspeech samples. SNR was measured from the average power of speech and noise

$$\mathrm{SNR} = 10 \log \frac{Ps}{Pn} \tag{34}$$

Figs. 14 and 15 illustrate the effect of white and pink noise on the average spectro-temporal modulations of speech. The spectro-temporal representation of noisy speech preserves the speech specific features (e.g., near 4 Hz, 2 cycle/octave) even at SNR as low as 0 dB (Figs. 14 and 15, middle). The detection results for speech in white noise (Figs. 8 and 9) demonstrate that while the three systems have comparable performance in clean conditions, the auditory features remain robust down to fairy low SNRs. This pattern is repeated with additive pink noise although performance degradation for all systems occurs at higher SNRs (Figs. 10 and 11) 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 exponential decay. The effect on the average spectro-temporal modulations of speech are shown in Fig. 16. Increasing the time delay results in gradual loss of high-rate temporal modulations of speech. Figs. 12 and 13 demonstrate the effect of reverberation on the classification accuracy.

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

V. Summary and Conclusion

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. The success of these multiscale features for this speech detection task suggests that these features are more worth investigating for speech recognition [38] or noise suppression [39] than conventional approaches based on simple cepstral features.

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., Figs. 14, –16), or indeed under any kind of linear or nonlinear distortion, can be used as an indicator of predicted speech intelligibility [15]. Similarly, the multiscale rate-scale-frequency representation can account for the perception of complex sounds and perceptual thresholds in a variety of settings [40]. Finally, the auditory model can be adapted and expanded for a wide range of applications such as the speech enhancement [20], or the efficient encoding of speech and music [21].

ACKNOWLEDGMENT

The authors would like to thank B. Zook of the Southwest Research Institute for critical contribution and support of this work. They would also like to thank Telluride Neuromorphic Workshop and M. Goto of AIST for his help acquiring the RWC music samples and they would also like to thank the anonymous reviewers for their valuable comments.

REFERENCES

[1] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Int. Conf. Acoustic, Speech and Signal Processing*, vol. 2, Munich, Germany, 1997, p. 1331.

[2] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, "Robust speech recognition in noisy environments: the 2001 IBM SPINE evaluation system," in *Int. Conf. Acoustic, Speech and Signal Processing*, vol. I, Orlando, FL, May 2002, pp. 53–56.

[3] S. Pfeiffer, S. Fischer, and W. Efferlsberg, "Automatic audio content analysis," in *Proc. 4th ACM Int. Multimedia Conf.*, 1996, pp. 21–30.

[4] E. Wold, T. Blum, and D. Keislar *et al.*, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, vol. 3, no. 3, pp. 27–36, Fall 1996.

[5] J. Foote *et al.*, "Content-based retrieval of music and audio," *Proc. SPIE Multimedia Storage and Archiving Systems II*, vol. 3229, pp. 138–147, 1997.

[6] L. Lu, H. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Trans. Speech, Audio, Process.*, vol. 10, no. 7, pp. 504–516, Oct. 2002.

[7] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proc. Int. Conf. Acoustic, Speech and Signal Processing*, vol. 2, Atlanta, GA, May 1996, pp. 993–996.

[8] C. J. C. Burges, J. C. Platt, and S. Jana, "Distortion discriminant analysis for audio fingerprinting," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 165–174, May 2003.

[9] K. Wang and S. A. Shamma, "Spectral shape analysis in the central auditory system," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 382–395, Sep. 1995.

[10] R. Lyon and S. Shamma, "Auditory representation of timbre and pitch," in *Auditory Computation*. New York: Springer-Verlag, 1996, vol. 6, Springer handbook of auditory research, pp. 221–270.

[11] X. Yang, K. Wang, and S. A. Shamma, "Auditory representation of acoustic signals," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 824–839, Mar. 1992. Special issue on wavelet transforms and multi-resolution signal analysis.

[12] 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," *J. Neurophysiology*, vol. 76, no. 5, pp. 3503–3523, 1996.

[13] S. A. Shamma, "Methods of neuronal modeling," in *Spatial and Temporal Processing in the Auditory System*, 2nd ed. Cambridge, MA: MIT Press, 1998, pp. 411–460.

[14] 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," *J. Neurophysiology*, vol. 85, pp. 1220–1234, 2001.

[15] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Commun.*, vol. 41, pp. 331–348, 2003.

[16] H. Dudley, "Remaking speech," *J. Acoustical Soc. Amer.*, vol. 11, no. 2, pp. 169–177, 1939.

[17] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Commun.*, vol. 25, no. 1, pp. 117–132, 1998.

[18] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation: II. Spectral and temporal integration," *J. Acoustical Soc. Amer.*, vol. 102, no. 5, pp. 2906–2919, 1997.

[19] M. Kleinschmidt, J. Tchorz, and B. Kollmeier, "Combining speech enhancement and auditory feature extraction for robust speech recognition," *Speech Commun.*, vol. 34, no. 1–2, pp. 75–91, 2001. Special issue on robust ASR.

[20] N. Mesgarani and S. A. Shamma, "Speech enhancement base on filtering the spectrotemporal modulations," in *Proc. Int. Conf. Acoustic, Speech, Signal Processing*, Philadelphia, PA, Mar. 2005, pp. 1105–1108.

[21] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *Eurasip J. Applied Signal Processing*, no. 7, pp. 668–675, Jun. 2003.

[22] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with gabor feature extraction," in *Proc. Int. Conf. Spoken Language Processing*, Denver, CO, 2002, pp. 25–28.

[23] R. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton Univ. Press, 1961.

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

[25] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, pp. 279–311, 1966.

[26] P. M. Kroonenberg, *Three-Mode Principal Component Analysis*. Leiden, The Netherlands: DSWO, 1982.

[27] L. De Lathauwer, B. De Moore, and J. Vandewalle, "On the best $\mathrm{rank}-1$ and $\mathrm{rank}-(R_1, R_2, \ldots, R_N)$ approximation of higher order tensors," *SIAM J. Matrix Anal. Applicat.*, vol. 21, no. 4, pp. 1324–1342, 2000.

[28] ——, "Dimensionality reduction in higher-order-only ICA," in *IEEE Signal Processing Workshop on Higher Order Statistics*, Banff, AB, Canada, 1997, pp. 316–320.

[29] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear analysis of image ensembles: TensorFaces," in *Proc. Eur. Conf. Computer Vision*, Copenhagen, Denmark, May 2002, pp. 447–460.

[30] ——, "Multilinear subspace analysis of image ensembles," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, Madison, WI, Jun. 2003, pp. 11–93.

[31] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.

[32] T. Joachims, *Making Large-Scale SVM Learning Practical. Advances in Kernel Methods—Support Vector Learning*, B. Scholkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1999.

[33] (1988) *Getting Started With the DARPA TIMIT CDROM: An Acoustic Phonetic Continuous Speech Database*

[34] *BBC Sound Effects Library*, 1984. Original Series, 40 Audio CD Collection. Distributed by Sound Ideas.

[35] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: music genre database and musical instrument sound database," in *Proc. Int. Conf. Music Information Retrieval*, 2003, pp. 229–230.

[36] *The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition*, 1992. Documentation included in the NOISEX-92 CD-ROMs.

[37] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France, Sep. 18–20, 2000, pp. 181–188.
[38] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[39] L. Deng, J. Droppo, and A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 568–580, Nov. 2003.
[40] R. P. Carlyon and S. A. Shamma, "An account of monaural phase sensitivity," *J. Acoust. Soc. Amer.*, vol. 114, no. 1, pp. 333–348, 2003.

Nima Mesgarani received the B.S. degree in electrical engineering from Sharif University of Technology, Tehran, Iran, in 1999. He is currently pursuing the Ph.D. degree in electrical and computer engineering at the University of Maryland, College Park.

His research interests are in neuromorphic models of auditory cortical functions and investigating the usefulness of auditory neuroscientific knowledge for various acoustical signal processing applications, including speech detection, enhancement, and recognition.

Mr. Mesgarani received the George Harhalakis Outstanding Graduate Student Award in 2004.

Malcolm Slaney (SM'01) received the Ph.D. degree from Purdue University, West Lafayette, IN.

He is a Research Staff Member at the IBM Almaden Research Center, San Jose, CA, and a Visiting Instructor at the Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, Stanford, CA. Before IBM, he was with Bell Laboratories, Schlumberger Palo Alto Research, Apple's Advanced Technology Group, and Interval Research. He is a coauthor of the book "Principles of Computerized Tomographic Imaging," which was recently republished by SIAM Press as a "Classic in Applied Mathematics." He is a coeditor of the book "Computational Models of Auditory Function."

Shihab A. Shamma (SM'94) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1980.

He joined the Department of Electrical Engineering at the University of Maryland, College Park, in 1984, where his research has dealt with issues in computational neuroscience and the development of microsensor systems for experimental research and neural prostheses. His primary focus has been on uncovering the computational principles underlying the processing and recognition of complex signals (speech and music) in the auditory system, and the relationship between auditory and visual processing. Other research includes the development of photolithographic microelectrode array for recording and stimulation of neural signals, VLSI implementation of auditory processing algorithms, and development of algorithm for the detection, classification, and analysis of neural activity from multiple simultaneous sources.