Sparse Auditory Reproducing Kernel (SPARK) Features for Noise-Robust Speech Recognition

Amin Fazel, *Student Member, IEEE*, and Shantanu Chakrabartty, *Senior Member, IEEE*

***Abstract*—In this paper, we present a novel speech feature extraction algorithm based on a hierarchical combination of auditory similarity and pooling functions. The computationally efficient features known as "Sparse Auditory Reproducing Kernel" (SPARK) coefficients are extracted under the hypothesis that the noise-robust information in speech signal is embedded in a reproducing kernel Hilbert space (RKHS) spanned by overcomplete, nonlinear, and time-shifted gammatone basis functions. The feature extraction algorithm first involves computing kernel based similarity between the speech signal and the time-shifted gammatone functions, followed by feature pruning using a simple pooling technique ("MAX" operation). In this paper, we describe the effect of different hyper-parameters and kernel functions on the performance of a SPARK based speech recognizer. Experimental results based on the standard AURORA2 dataset demonstrate that the SPARK based speech recognizer delivers consistent improvements in word-accuracy when compared with a baseline speech recognizer trained using the standard ETSI STQ WI008 DSR features.**

***Index Terms*—Auditory HMAX, gammatone functions, reproducing kernel Hilbert space (RKHS), robust speech recognition, sparse features.**

I. Introduction

UNLIKE human audition, the performance of speech-based recognition systems degrades significantly in the presence of noise and background interference [1], [2]. This can be attributed to inherent mismatch between the training and deployment conditions, especially when the characteristics of all possible noise sources are not known in advance. Therefore, in literature several strategies have been proposed that can reduce the effect of this mismatch. They can be broadly categorized into three main groups: 1) speech enhancement techniques that can filter out the noise in the spectral or temporal domain [3]; 2) robust feature extraction techniques that can generate speech features that are invariant to channel conditions; and 3) back-end adaptation techniques that can reduce the effect of training-deployment mismatch by adjusting the parameters of a statistical recognition model [4]–[13]. Even though significant improvements in recognition performance can be expected by the application of the third approach, the overall system performance is still limited by the quality of the speech features. Therefore, this paper focuses on extraction of speech features that are robust to mismatch between training and testing conditions.

Traditionally, speech features used in most of the state-of-the-art speech recognition systems have relied on spectral-based techniques which include Mel-frequency cepstral coefficients (MFCCs) [14], linear predictive coefficients (LPCs) [14]–[16], and perceptual linear prediction (PLP) [17]. Noise-robustness is achieved by modifying these well established techniques to compensate for channel variability. For example, cepstral mean normalization (CMN) [18] and cepstral variance normalization [19] adjust the mean and variance of the speech features in the cepstral domain to reduce the effect of convolutive channel distortion. Another example is the Relative spectra (RASTA) [20] technique which suppresses the acoustic noise by high-pass (or band-pass) filtering of the log-spectral representation of speech. More recently, advanced signal processing techniques like the feature-space non-linear transformation techniques [21], the ETSI advanced front end (AFE) [22], [23], stereo-based piecewise linear compensation (SPLICE) [24] and power-normalized cepstral coefficients (PNCC) [25], have been used to improve the noise-robustness. The AFE approach, for example, integrates several methods to remove the effects of both additive and convolutive noises. A two-stage Mel-warped Wiener filtering, combined with an SNR-dependent waveform processing is used to reduce the effect of additive noise and a blind equalization technique is used to mitigate the channel effects.

An alternate and a promising approach towards extracting noise-robust speech features is to use data-driven statistical learning techniques that do not make strict assumptions on the spectral properties of the speech signal. Examples include kernel based techniques [34], [35] which operate under the premise that robustness in speech signal is encoded in high-dimensional temporal and spectral manifolds which remain intact even in the presence of ambient noise and the objective of the feature extraction procedure is to identify the parameters of the noise-invariant manifold. The procedure used in [34] required solving a quadratic optimization problem for each frame of speech which made the data-driven approach highly computationally intensive. Also, due to its semi-parametric nature, the methods proposed in [34], [35] did not incorporate any *a priori* information available from neurobiological and psycho-acoustical studies, which have been shown to be important for speech recognition [26]–[33]. More recently, it has been demonstrated that cortical neurons use highly efficient and sparse encoding of visual and auditory signals [36]–[38]. The study [38] showed that auditory signals can be sparsely represented by a group of

Manuscript received May 02, 2011; revised September 21, 2011; accepted November 30, 2011. Date of publication December 09, 2011; date of current version February 24, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Haizhou Li.

The authors are with the Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48824 USA (e-mail: fazel@egr.msu.edu; shantanu@egr.msu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2179294

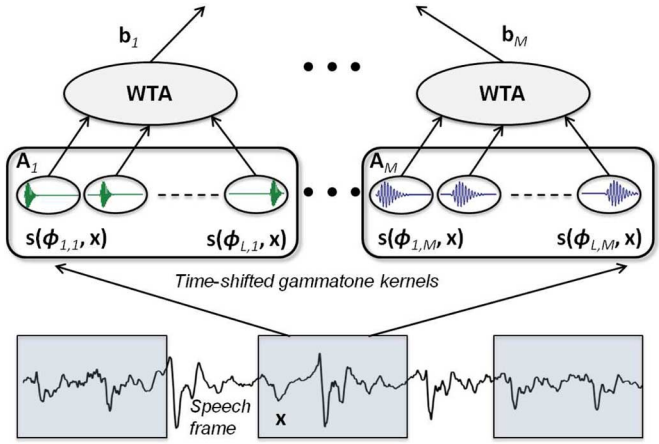

Fig. 1. Hierarchical model of SPARK feature extraction.

basis functions which are functionally similar to gammatone functions which are equivalent to time-domain representations of human cochlea filters, also used in psycho-acoustical studies [39], [40]. Other neurobiological studies [41]–[43] have proposed a hierarchical auditory processing model consisting of spectro-temporal receptive fields (STRFs) [44], [45] that capture information embedded in different frequency, spectral and temporal scales. The results from many of these recent neurobiological and psycho-acoustical studies are being incorporated in small-scale speech recognition systems [46]–[50].

In this paper, we propose a computationally efficient hierarchical auditory feature extraction model using an RKHS based statistical learning approach. The model is summarized in Fig. 1 and consists of two signal-processing layers. The first layer computes the similarity between the sample speech signal and different sets A_1 to A_M of precomputed gammatone basis functions. Each set comprises of time-delayed versions of gammatone functions which emulates an auditory phase-sensitive receptive field. The second layer of the proposed model implements a winner-take-all (WTA) function which selects the largest of the similarity metric from each set A_1 to A_M (see Fig. 1). The WTA-based selection is similar to the popular HMAX algorithm [51] used in vision systems, where edge-sensitive receptor functions are used to compute similarity measures. Based on the hierarchical model for computing SPARK features, there are two main contributions of this paper: 1) Using RKHS functions to determine optimal auditory similarity functions that can capture the high-dimensional speech features; and 2) evaluating the effect of different RKHS parameters on the performance of a SPARK based speech recognition system.

The paper is organized as follows. Section II describes the mathematical basis underlying the SPARK feature extraction algorithm. Section III presents experimental results summarizing the effect of different hyper-parameters and kernel functions when SPARK feature are evaluated for a speech recognition task using the AURORA2 corpus. Section IV concludes the paper with some discussions on the future work. Before we present the SPARK algorithm we summarize some of the mathematical notations that will be used in this paper:

$\mathbf{A}$ (bold capital letters)	denotes a matrix with its elements denoted by a_{ij}, $i = 1, 2, \ldots$; $j = 1, 2, \ldots$ and its row-wise vectors denoted by $\mathbf{a}_i$, $i = 1, 2, \ldots$;
x (normal lowecase letters)	denotes a scalar variable;
$\mathbf{x}$ (bold lowecase letters)	denotes a vector with its elements denoted by x_i, $i = 1, 2, \ldots$;
$x[n]$	denotes a sequence of scalars where $n = 1, 2, \ldots$ denotes a discrete-time index;
$\mathbf{\Psi}(\mathbf{x})$	denotes a vector function whose elements are scalar functions denoted by $\psi_i(\mathbf{x})$, $i = 1, 2, \ldots$;
$\|\mathbf{x}\|_p$	denotes the L_p norm of a vector and is given by $\|\mathbf{x}\|_p = (\sum_i \|x_i\|^p)^{1/p}$;
$\mathbf{A}^T$	denotes the transpose of $\mathbf{A}$;
$\mathbf{x} \cdot \mathbf{y}$	denotes the inner-product between vectors $\mathbf{x}$ and $\mathbf{y}$.

II. SPARK Feature Extraction Algorithm

In this section, we describe the mathematics underlying the SPARK feature extraction procedure. The first part of this analysis will involve deriving the mathematical form of the SPARK similarity functions based on RKHS regression techniques. For the analysis presented in this section, we will assume that a frame of speech signal is extracted using an appropriate windowing function (Hamming or Hanning).

A. SPARK Similarity Functions

As shown in Fig. 1, the similarity function $s : \mathbb{R}^P \times \mathbb{R}^P \rightarrow \mathbb{R}$ is computed between a frame of speech signal $(x[1], x[2], \ldots, x[P])$, compactly denoted by $\mathbf{x} \in \mathbb{R}^P$, and a set of precomputed basis vectors. For SPARK features, the basis vectors are constructed using a set of physiologically inspired gammatone functions $\phi_m(\cdot)$, $m = 1, \ldots, M$ whose discrete-time representation is given by

$$\phi_m[n] = a_m n^{\theta-1} \cos(2\pi f_m n) e^{-2\pi\beta ERB(f_m)n} \quad (1)$$

where f_m is the center frequency parameter, a_m is the amplitude, θ is the order of the gammatone basis, β is the parameter which controls the decay of the envelope along with a monotonic frequency dependent function $ERB(\cdot)$ called equivalent rectangular bandwidth (ERB) scale [55]. One possible form of $\mathrm{ERB}(f_m)$ which has been used in this paper, has been suggested by Glasberg and Moore [56] and takes the form

$$\mathrm{ERB}(f_m) = 0.108 f_m + 24.7. \quad (2)$$

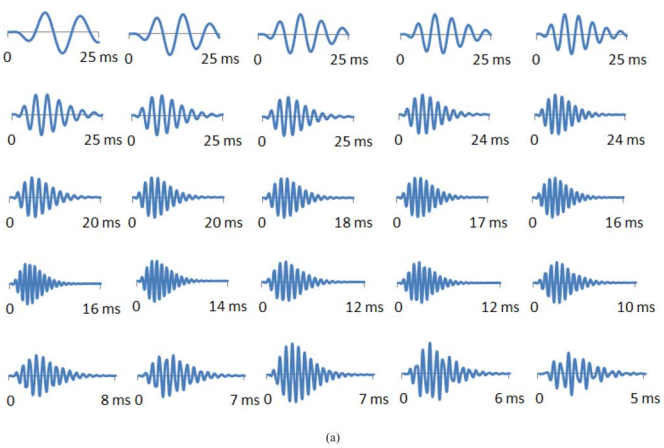

Fig. 2. Set of 25 gammatone kernel basis functions with center frequencies spanning 100 Hz to 4 KHz in the ERB space.

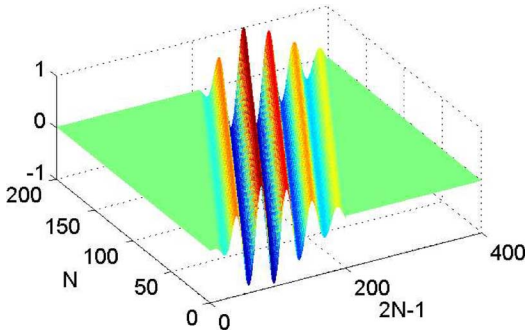

Fig. 3. Three-dimensional plot showing a gammatone function shifted in time by 100 μs.

Also, in this paper we have chosen $\theta = 4$ and $\beta = 1.019$ based on the results reported in [57]. Fig. 2 shows the set of 25 gammatone basis vectors, each with different center frequencies f_m. In the frequency domain, gammatone functions bear close resemblance to cochlear filter-banks due to the following characteristics: 1) nonuniform filter bandwidths where each of the frequency resolution is higher at the lower frequency than at the higher frequency; 2) peak gain of the filter centered at f_m decreases as the level of the input increase, and 3) the cochlea filters are spaced more closely at lower frequencies than at higher frequencies. In [37] and [38], it was shown that natural sounds can be sparsely and hence more compactly represented by a mixture of shift-invariant gammatone-type basis functions. Therefore, in our hierarchical SPARK model, we have chosen a basis set comprising of gammatone functions $\phi_m[n-\tau_{l,m}]$ with different center frequency f_m and with different temporal-shifts $\tau_{l,m}$ (see Fig. 3 which plots a gammatone function time-shifted by a unit time-interval). Incorporating different time-shifts in gammatone functions will be important for extracting phase information in speech signal which has been shown to be effective in extracting the attributes of non-stationary part of speech signals (for, e.g., plosives) [53], [54].

We will compactly represent the discrete-time gammatone function $\phi_m[n-\tau_{l,m}]$ as $\boldsymbol{\phi}_{l,m} \in \mathbb{R}^P$ and correspondingly the similarity function will be given by $s(\boldsymbol{\phi}_{l,m}, \mathbf{x})$. We now define a discrete-time waveform $f[n]$, $n = 1, \ldots, P$ which constructed using the time-shifted basis functions according to

$$f[n] = \sum_{m=1}^{M} \sum_{l=1}^{L} s(\boldsymbol{\phi}_{l,m}, \mathbf{x}) \phi_m[n - \tau_{l,m}]. \tag{3}$$

Our objective will be to determine the form of the similarity functions $s(\boldsymbol{\phi}_{l,m}, \mathbf{x})$ by ensuring that the waveform $f[n]$ is close to the speech waveform $x[n]$ according to some optimization criterion.

Before we present the optimization function, we rewrite the time-domain expressions in a matrix-vector notation as

$$\mathbf{f} = \boldsymbol{\Phi}\mathbf{s} \tag{4}$$

where $\mathbf{f} \in \mathbb{R}^P$ and $\mathbf{s} \in \mathbb{R}^{L \times M}$ is a vector given by $\mathbf{s} = [s_{1,1}, s_{1,2}, \ldots, s_{L,M}]^T$ with its element given by $s_{l,m} = s(\boldsymbol{\phi}_{l,m}, \mathbf{x})$. $\boldsymbol{\Phi} \in \mathbb{R}^P \times \mathbb{R}^{L \times M}$ is a matrix given by $\boldsymbol{\Phi} = [\boldsymbol{\phi}_{1,1}, \ldots, \boldsymbol{\phi}_{L,M}]^T$.

The optimization procedure for SPARK features involves minimizing a cost function $\mathcal{C}$ with respect to $\mathbf{s}$, where $\mathcal{C}$ is given by

$$\mathcal{C} = \lambda ||\mathbf{s}||_2^2 + ||\mathbf{x} - \mathbf{f}||_2^2. \tag{5}$$

The first part of the cost function acts as a regularizer which penalizes large values of $s_{l,m}$, thus favoring similarity measures that are smooth (or penalizes high-frequency components of the similarity function). The second part of the cost function $\mathcal{C}$ is the least-square error function computed between the speech vector and the reconstructed waveform $f[n]$. The hyper-parameter λ in $\mathcal{C}$ controls the tradeoff between the achieving a lower reconstruction error and obtaining smoother similarity function. This tradeoff is similar to the conventional bias-variance trade-off encountered in any functional regression procedure [59]. Equating the derivative

$$\frac{\partial \mathcal{C}}{\partial \mathbf{s}} = 2\lambda \mathbf{s} - 2\boldsymbol{\Phi}^T(\mathbf{x} - \boldsymbol{\Phi}\mathbf{s}) = 0$$

leads to

$$\boldsymbol{\Phi}^T\mathbf{x} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda \mathbf{I})\mathbf{s} \tag{6}$$

where $\mathbf{I}$ denotes an identity matrix. The optimal $\mathbf{s}^*$ can be found to be

$$\mathbf{s}^* = \left[\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda \mathbf{I}\right]^{-1} \boldsymbol{\Phi}^T\mathbf{x}. \tag{7}$$

Equation (7) shows that the optimal similarity function $\mathbf{s}^*$ is expressed in terms of inner-products between different time-shifted gammatone basis $\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \{\boldsymbol{\phi}_{l,m} \cdot \boldsymbol{\phi}_{u,v}\}; l, u = 1, \ldots, L;$ $m, v = 1, \ldots, M$ and between the time-shifted gammatone basis and the input speech vector $\boldsymbol{\Phi}^T\mathbf{x} = \{\boldsymbol{\phi}_{l,m} \cdot \mathbf{x}\}$. Equation (7) shows that the similarity function admits a linear form and involves computing inner-products. We extend this framework to a more general, nonlinear form of similarity functions

by converting the inner-products in (7) into kernel expansions over the gammatone and the speech vectors.

We introduce a nonlinear transformation function $\Psi : \mathbb{R}^P \rightarrow \mathbb{R}^D, D \gg P$ which will map the vectors $\mathbf{x}$ and $\boldsymbol{\phi}_{l,m}$ to a higher dimensional space according to $\mathbf{x} \rightarrow \Psi(\mathbf{x})$ and $\boldsymbol{\phi}_{l,m} \rightarrow \Psi(\boldsymbol{\phi}_{l,m})$. The high-dimensional mapping could consist of cross-correlation terms, for example, $(x[1], x[2], \ldots, x[P]) \rightarrow \big(x[1], x[2], \{x[1]\}^2, \{x[2]\}^2, \{x[1]x[2]\}, \ldots\big)$ which capture nonlinear attributes of the speech signal. Thus, extending (4) to the high-dimensional space, the reconstruction function $\mathbf{f} \in \mathbb{R}^D$ can be written as

$$\mathbf{f} = \Psi(\boldsymbol{\Phi})\mathbf{s} \tag{8}$$

where $\Psi(\boldsymbol{\Phi}) \in \mathbb{R}^D \times \mathbb{R}^{L \times M}$ is a matrix given by $\Psi(\boldsymbol{\Phi}) = [\Psi(\boldsymbol{\phi}_{1,1}), \ldots, \Psi(\boldsymbol{\phi}_{L,M})]^T$. Then, following the regression procedure as described above, the similarity function can be expressed as inner-products in the higher dimensional space according to

$$\mathbf{s}^* = \Big[\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\mathbf{I}\Big]^{-1}\boldsymbol{\Phi}^T\mathbf{x} \xrightarrow{\Psi(\cdot)} \big[\Psi(\boldsymbol{\Phi})^T\Psi(\boldsymbol{\Phi}) + \lambda\mathbf{I}\big]^{-1}\Psi(\boldsymbol{\Phi})^T\Psi(\mathbf{x}). \tag{9}$$

Unfortunately, computing inner-products directly in the high-dimensional space is computationally intensive. The use of reproducing kernels avoids this "curse of dimensionality" by avoiding direct inner-product computation. For example, consider a nonlinear mapping of a two-dimensional vector $\mathbf{y} \in \mathbb{R}^2$ such that $(y_1, y_2) \xrightarrow{\Psi(\cdot)} (1, y_1, y_2, y_1^2, y_2^2, \sqrt{2}y_1y_2)$. The inner-product between two vectors $\mathbf{y}, \mathbf{z} \in \mathbb{R}^2$ in the high-dimensional space can be expressed as $\Psi(\mathbf{y}) \cdot \Psi(\mathbf{z}) = (1 + \mathbf{y} \cdot \mathbf{z})^2$ which requires computing inner-products only in the low-dimensional space, hence, is more computationally tractable. In general, any symmetric positive-definite function $K(\cdot, \cdot)$ (also referred to as the reproducing kernel function) can be expressed as $K(\mathbf{z}, \mathbf{y}) = \Psi(\mathbf{x}).\Psi(\mathbf{y})$ and hence can be used in (9). In literature, many forms of reproducing kernels have been reported, which includes the Gaussian radial basis function or the polynomial spline function [58]. In neurophysiology, kernel functions have also been used for computing similarity measures in neural responses [52]. Equation (9) can be expressed in terms of kernels as

$$\mathbf{s}^* = (\mathbf{K} + \lambda\mathbf{I})^{-1}K(\boldsymbol{\Phi}, \mathbf{x}) \tag{10}$$

where $\mathbf{K} \in \mathbb{R}^{L \times M} \times \mathbb{R}^{L \times M}$ is a RKHS kernel matrix with elements $K(\boldsymbol{\phi}_{l,m}, \boldsymbol{\phi}_{u,v})$. Thus, a generic form of RKHS based similarity function can be expressed as

$$s(\boldsymbol{\phi}_{l,m}, \mathbf{x}) = (\mathbf{K} + \lambda I)^{-1}K(\boldsymbol{\phi}_{l,m}, \mathbf{x}). \tag{11}$$

Note that the matrix inverse in (11) involves only the gammatone basis and hence can be precomputed and stored. Thus, the computation of the SPARK similarity metric involves computing kernels and a matrix-vector multiplication which can be made computationally efficient.

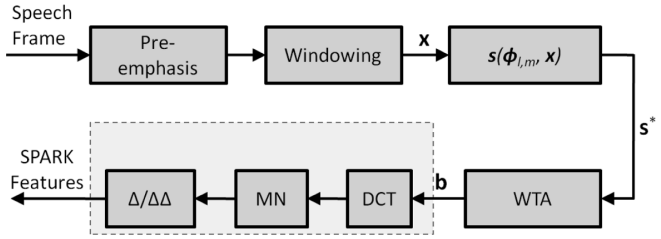

Fig. 4. Signal flow of the SPARK feature extraction algorithm.

B. Feature Pooling

An important consequence of projecting the speech signal onto a gammatone function space (emulating the auditory STRFs) is that the highest scores (in $\|\cdot\|_2$ sense) in the similarity metric vector $\mathbf{s}$ will capture the salient, higher order, and the spectro-temporal aspects of the speech signal. On the other hand, the low-energy components of $\mathbf{s}$ will also capture similarities to noise and channel artifacts. Feature pooling serves two purposes. First, it introduces competitive masking, where only the largest similarity score is chosen. This function emulates the local competitive behavior which has been observed in auditory receptive fields. Also, the functionality bears similarity to the HMAX hierarchical models which has been extensively used in vision based recognition system [51]. The second purpose of feature pooling is to introduce a compressive weighting function (similar to psycho-acoustical responses) which enhances the resolution at low similarity scores and reduces the resolution at high similarity scores. Mathematically, the output $b_m, m = 1, \ldots, M$ resulting from feature pooling is given by

$$b_m = \zeta(\max_{l=1}^{L}(|s_{l,m}|)). \tag{12}$$

where $\zeta(\cdot)$ is the compressive weighing function which could be a logarithmic $\log(\cdot)$ or a power function $(\cdot)^{1/p}, p > 1$. Note that the pooling is performed over a set consisting of time-shifted basis obtained from the same gammatone function.

C. SPARK Feature Extraction Signal-Flow

The flow-chart describing the complete SPARK feature extraction procedure is presented in Fig. 4. The input speech signal is processed by a pre-emphasis filter of the form $x_{pre}(t) = x(t) - 0.97x(t-1)$ after which a 25-ms speech segment is extracted using a Hamming window. The similarity metric vector $\mathbf{s}^* \in \mathbb{R}^{LM}$ is obtained using the procedure described in Section II-A and the sparsified vector $\mathbf{b} \in \mathbb{R}^M$ is obtained based on the pooling procedure described in Section II-B. Fig. 5(a) and (d) shows the spectrograms of utterance "one" for clean and noisy (subway recording) conditions. Fig. 5(b) and (e) shows the similarity metric vector for each 25-ms speech segment shifted by 10 ms over clean and noisy speech utterances. Similarly, Fig. 5(c) and (f) shows the vector $\mathbf{b}$ for the same utterances. Similar to MFCC processing, a discrete cosine transform (DCT) is applied to de-correlate each of the vectors b. Mean normalization is then applied to each of these vectors and the SPARK features are obtained by appending the velocity Δ and acceleration $\Delta\Delta$ coefficients (similar to MFCC processing). To ensure parity in comparison between the MFCC and SPARK-based features, we extracted

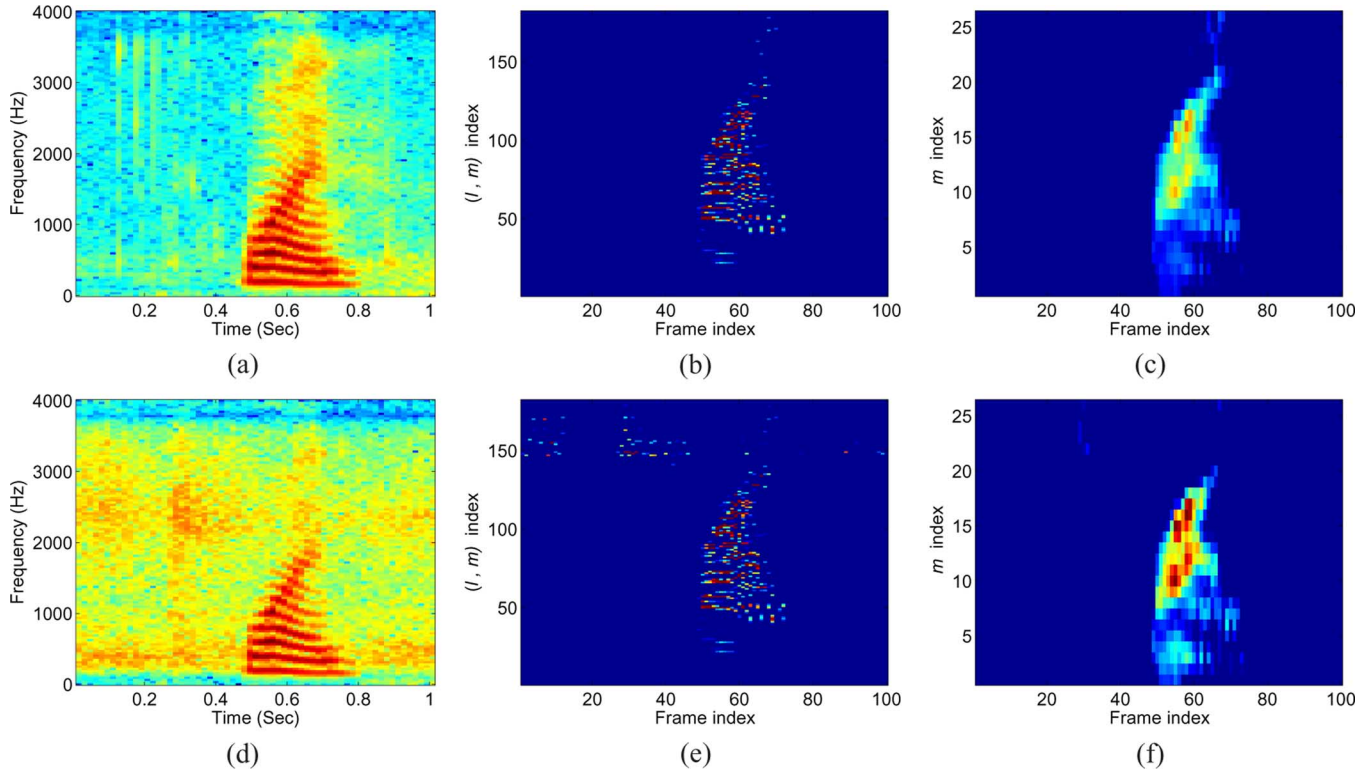

Fig. 5. Colormaps depicting spectrogram, vector $\mathbf{s}^*$, and vector $\mathbf{b}$ for clean utterance (top row) and 20-dB noisy utterance (bottom row) of digit "one."

13 SPARK coefficients and concatenated additional 13 Δ and $\Delta\Delta$ coefficients to form a 39-dimensional feature vector.

III. Experiments and Performance Evaluation

A. Experimental Setup

We have evaluated the SPARK features for the task of noise-robust speech recognition using the AURORA2 dataset [61]. The AURORA2 task involves recognizing English digits in the presence of additive noise and convolutional noise. The task consists of three types of test sets. The first test set (set A) contains 4 subsets of 1001 utterances corrupted by subway, babble, car, and exhibition hall noises, respectively, at different SNR levels. The second set (set B) contains 4 subsets of 1001 utterances corrupted by restaurant, street, airport, and train station noises at different SNR levels. The test set C contains 2 subsets of 1001 sentences, corrupted by subway and street noises and was generated after filtering the speech with an MIRS filter [62] before adding different types of noise.

For all the experiments reported in this paper, a hidden Markov model (HMM)-based speech recognizer has been used [61]. The HMM recognizer was implemented using the hidden Markov toolkit (HTK) package [63]. For each digit a whole word HMM was trained with 16 states per HMM and with three diagonal Gaussian mixture components per state. Additional HMMs were trained for the "sil" and "sp" models.

Next, we summarize the effect of different algorithmic hyper-parameters on the performance of a SPARK-based recognition system.

TABLE I
Effect of Different Time-Shifts on Recognition Performance

	Set A	Set B	Set C
SPARK; Shift=100μs	72.83	73.62	71.97
SPARK; Shift=3ms	72.33	73.02	71.57
SPARK; Shift=4.5ms	71.79	72.48	70.97
SPARK; Shift=7.5ms	70.60	70.63	69.74

B. Effect of the Time-Shift Resolution

As we had described in Section II and shown in Fig. 3, the basis set comprises of time-shifts of gammatone functions. A set of M gammatone functions, each time-shifted L times would produce a total of $L \times M$ basis functions. Thus, reducing L would reduce the number of basis functions and also reduce the computational complexity of the (11). In this experiment, we evaluate the effect of different time-shift resolution on the recognition performance of the system. The results which have been obtained for $K(\mathbf{x}, \mathbf{y}) = \tanh(.01\mathbf{x}\mathbf{y}^T - .01)$ and $\zeta = (\cdot)^{1/13}$ are summarized in Table I.

The result shows that smaller time-shifts (larger value of L) leads to better recognition results, however, at the expense of higher computational complexity. Thus, there exists a tradeoff between L, recognition performance and real-time requirements of the system.

C. Effect of Different Kernel Functions

The generic form of the similarity function $s(\cdot, \cdot)$ is given by (11) and is dependent on the choice of the kernel function $K(\cdot, \cdot)$. In this experiment, we evaluated the effect of different

TABLE II
EFFECT OF DIFFERENT KERNEL FUNCTIONS ON RECOGNITION PERFORMANCE

	Set A	Set B	Set C
SPARK; Exponential kernel, $c = 0.01$	69.83	71.45	69.52
SPARK; Exponential kernel, $c = 1.0$	69.22	71.16	68.24
SPARK; Sigmoid kernel, $a = 0.01, c = 0$	68.35	70.60	68.89
SPARK; Sigmoid kernel, $a = 0.01, c = -0.01$	69.84	71.48	69.54
SPARK; Linear kernel	67.80	69.65	68.30
SPARK; Polynomial kernel, $d = 2$	70.77	71.14	71.07
SPARK; Polynomial kernel, $d = 4$	67.89	68.24	68.05

TABLE III
EFFECT OF COMPRESSIVE WEIGHTING FUNCTION ON RECOGNITION PERFORMANCE

	Set A	Set B	Set C
SPARK; $\zeta(.) = (.)^{1/3}$	64.91	65.60	62.60
SPARK; $\zeta(.) = (.)^{1/11}$	70.91	72.32	70.19
SPARK; $\zeta(.) = (.)^{1/13}$	70.27	71.96	69.68
SPARK; $\zeta(.) = (.)^{1/15}$	69.83	71.24	68.88
SPARK; $\zeta(.) = (.)^{1/17}$	68.83	70.75	68.44
SPARK; $\zeta(.) = (.)^{1/19}$	68.35	70.36	68.10

types of RKHS functions on the recognition performance of the SPARK based system. The results are summarized in Table II for the following kernel functions: (a) linear $K(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$; (b) exponential $K(\mathbf{x}, \mathbf{y}) = \exp(c\mathbf{x} \cdot \mathbf{y})$; (c) sigmoid $K(\mathbf{x}, \mathbf{y}) = \tanh(a\mathbf{x} \cdot \mathbf{y} + c)$; and (d) polynomial $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y})^d$.

The results show that the choice of the kernel function affects the recognition performance, specifically, compared to the case when the linear kernel is used. The improvements in performance demonstrates the utility of exploiting nonlinear features in speech to achieve noise-robustness. Note that the best performance is obtained for a fourth-order polynomial kernel when we fixed $\zeta(\cdot) = (\cdot)^{1/15}$.

D. Effect of Compressive Weighting Function

The compressive weighting function, as described in Section II-B, amplifies the lower values and de-amplifies larger values of the similarity metric. Table III summarizes the effect of different polynomial weighting functions on the performance of the SPARK-based speech recognition system (for $K(\mathbf{x}, \mathbf{y}) = \tanh(.01\mathbf{x}\mathbf{y}^T - .01)$). The results indicate an optimal order of the weighing function that yields the best recognition performance.

E. Effect of Parameter λ

Parameter λ is the regularization parameter which penalizes large values of the similarity metric and in the process makes the solution in (11) more stable. Table IV summarizes the effect of λ on the recognition performance and results show that solutions which penalizes the large values of S yields better recognition performance under noisy conditions.

TABLE IV
EFFECT OF PARAMETER λ ON RECOGNITION PERFORMANCE

	Set A	Set B	Set C
SPARK $\lambda = 0.1$	71.63	72.01	70.59
SPARK $\lambda = 0.01$	72.33	73.02	71.57
SPARK $\lambda = 0.0001$	71.41	72.35	70.25
SPARK $\lambda = 0.00001$	69.18	69.73	67.99
SPARK $\lambda = 0.000001$	64.12	64.79	62.68

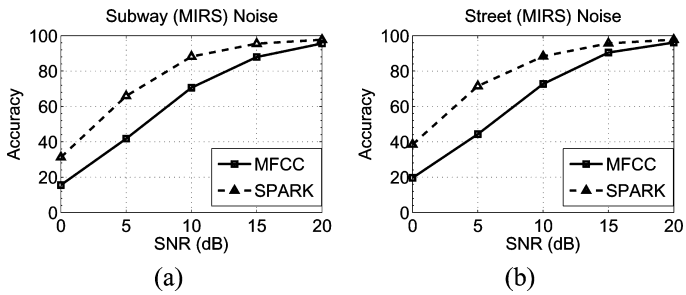

Fig. 6. AURORA2 recognition results obtained under different convolutive noise conditions.

F. Comparison With the Basic ETSI Front-End (MFCC)

The accuracy of the SPARK-based recognition system has been compared against the baseline speech features extracted using the ETSI STQ WI007 DSR front-end [61], [64]. The basic ETSI front-end generates the 39-dimensional MFCC features without any cepstral mean normalization (CMN). Figs. 6 and 7 compare the word recognition-rate obtained by the SPARK (with $\lambda = 0.01$, weighting function of $(\cdot)^{1/15}$, sigmoid kernel, and time-shift of 3.5 ms) and basic ETSI-based recognizers. The results show that the SPARK based recognition system consistently outperforms the benchmark at all SNR levels. The average relative word-accuracy improvement was found to be 33%, 36%, and 27% for set A, set B, and set C of the AURORA2 dataset.

G. Comparison With Gammatone Filter-Bank Based Features

The objective of the next set of experiments was to compare the SPARK features with gammatone filter-bank based features. The signal flow for the gammatone filter-bank features is shown in Fig. 8 which is similar to the MFCC feature extraction procedure except for the use of fourth-order gammatone filters instead of Mel-scale bandpass filters. The center frequencies were placed according to the ERB scale as described in Section II-A and similar to MFCC-based processing, a logarithmic compression, DCT, and CMS procedure is applied to the envelope of the output of each filter-bank. Δ and $\Delta\Delta$ features are then concatenated to obtain the final set of features (labeled GT). Table V summarizes the AURORA2 recognition results obtained using the gammatone filter-bank features. Note that even though these features delivers improved recognition performance over the baseline MFCC-based system [60], the SPARK features yields superior word-accuracy (relative) improvements of 22%, 17%, and 18% for set A, set B, and set C when compared to the gammatone filter-bank features.

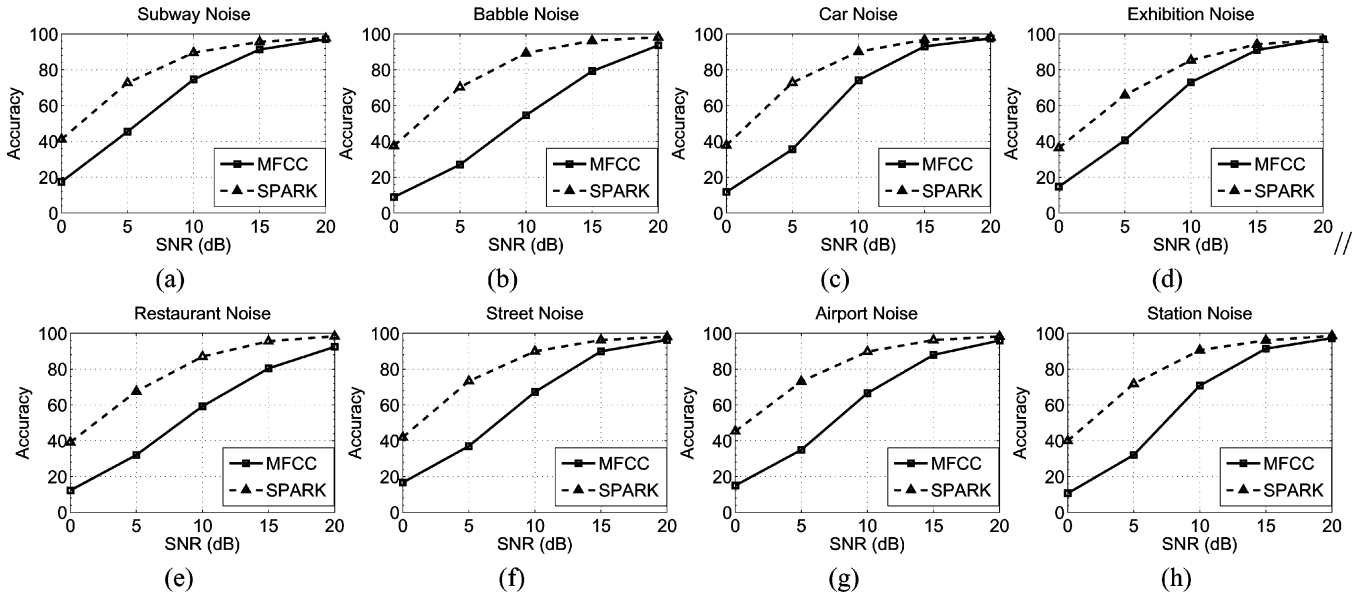

Fig. 7. AURORA2 recognition results obtained under different additive noise conditions.

TABLE V
AURORA2 WORD RECOGNITION RESULTS WHEN GAMMATONE FILTER-BANK (GT) FEATURES ARE USED

	Set A					Set B					Set C		
	Babble	Subway	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway(MIRS)	Street(MIRS)	Average
Clean	99.33	99.23	98.96	99.26	99.20	99.23	99.33	98.96	99.26	99.20	99.14	99.37	99.26
20dB	97.94	96.62	96.99	96.67	97.06	97.97	97.58	97.67	97.81	97.76	96.90	97.49	97.20
15dB	94.71	92.60	93.47	91.89	93.17	95.33	93.65	95.23	94.97	94.80	92.88	93.38	93.13
10dB	83.40	79.61	78.20	77.75	79.74	85.69	82.44	86.85	83.52	84.63	80.04	81.08	80.56
5dB	53.08	50.26	41.57	46.37	47.82	60.12	53.02	59.23	52.92	56.32	50.60	52.09	51.35
0dB	22.43	23.55	19.83	20.67	21.62	27.30	23.61	28.93	24.28	26.03	23.55	22.70	23.13
-5dB	12.76	14.86	12.47	12.22	13.08	12.96	12.85	15.00	13.92	13.68	14.55	12.73	13.64
Average	66.24	65.25	63.07	63.55	64.53	68.37	66.07	68.84	66.67	67.49	65.38	65.55	65.46

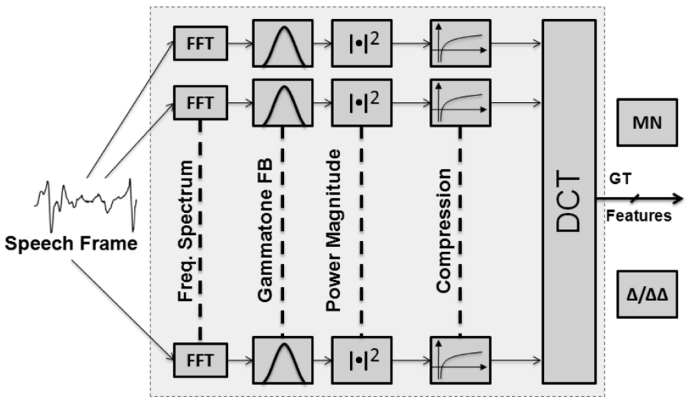

Fig. 8. Signal-flow showing the feature extraction procedure using gammatone filter-bank.

H. Comparison With ETSI AFE

The last set of experiments compared the SPARK features to the state-of-the-art ETSI AFE front-end. The ETSI AFE uses noise estimation, two-pass Wiener filter-based noise suppression, and blind feature equalization techniques. To incorporate an equivalent noise-compensation to the SPARK features, we used a power bias subtraction (PBS) [25] method. PBS method resembles in some ways to the conventional spectral subtraction (SS), but instead of estimating noise from non-speech parts which usually needs a very accurate voice activity detector (VAD), PBS simply subtracts a bias where the bias is adaptively computed based on the level of the background noise. Tables VI and VII compares the performance of ETSI AFE and SPARK+PBS ($\lambda = 0.01$) recognition system under different types of noise. Even though for Set A, the performance improvement of the SPARK+PBS system over the ETSI AFE system is not statistically significant, for Set B and Set C SPARK+PBS system consistently outperforms the ETSI AFE for all types of noise except subway and exhibition noise at low SNR. In fact, SPARK shows an overall relative improvements of 4.69% with respect to the ETSI AFE which is statistically significant.

Table VIII shows a comparative performance of SPARK+PBS features against basic ETSI FE, conventional gammatone filterbank, and ETSI AFE. Even under clean recording conditions, the SPARK+PBS demonstrates improvement over the baseline ETSI AFE system but the

TABLE VI
AURORA2 WORD RECOGNITION RESULTS WHEN ETSI AFE IS USED

	Set A					Set B					Set C		
	Babble	Subway	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway(MIRS)	Street(MIRS)	Average
Clean	99.00	99.08	99.05	99.23	99.09	99.08	99.00	99.05	99.23	99.09	99.08	99.03	99.06
20dB	98.31	97.91	98.48	97.90	98.15	97.97	97.64	98.39	98.36	98.09	97.36	97.70	97.53
15dB	96.89	96.41	97.58	96.82	96.93	95.33	96.74	97.11	96.73	96.48	95.33	95.77	95.55
10dB	92.35	92.23	95.29	92.78	93.16	90.08	92.78	93.47	93.77	92.53	90.24	90.69	90.47
5dB	81.08	83.82	88.49	84.05	84.36	76.27	83.28	84.07	84.57	82.05	79.03	78.17	78.60
0dB	51.90	61.93	66.42	63.28	60.88	51.09	60.07	60.99	62.57	58.68	51.73	52.09	51.91
-5dB	19.71	30.86	30.84	32.86	28.57	18.67	29.87	28.54	29.96	26.76	24.62	25.57	25.10
Average	77.03	80.32	82.31	80.99	80.16	75.50	79.91	80.23	80.74	79.10	76.77	77.00	76.89

TABLE VII
AURORA2 WORD RECOGNITION RESULTS WHEN SPARK AND PBS ARE USED TOGETHER

	Set A					Set B					Set C		
	Babble	Subway	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway(MIRS)	Street(MIRS)	Average
Clean	99.12	99.36	99.19	99.38	99.26	99.36	99.12	99.19	99.38	99.26	99.32	99.09	99.21
20dB	98.70	98.10	98.69	98.15	98.41	98.83	98.37	98.90	98.58	98.67	97.82	98.04	97.93
15dB	97.64	96.41	98.03	96.64	97.18	97.51	97.58	98.30	97.59	97.75	96.41	96.80	96.61
10dB	95.37	92.94	95.47	92.69	94.12	94.32	94.04	96.60	95.06	95.01	92.05	93.59	92.82
5dB	86.61	82.87	88.76	81.67	84.98	82.99	84.22	89.41	86.76	85.85	80.60	82.98	81.79
0dB	58.19	59.26	71.28	56.77	61.38	56.77	60.85	69.52	66.52	63.42	54.81	57.13	55.97
-5dB	21.58	27.97	34.54	25.24	27.33	21.95	27.48	32.03	33.35	28.70	25.02	25.57	25.30
Average	79.60	79.56	83.71	78.65	80.38	78.82	80.24	83.42	82.46	81.24	78.00	79.03	78.52

TABLE VIII
SUMMARY OF RECOGNITION PERFORMANCES OBTAINED FOR THE AURORA2 DATABASE

	Set A	Set B	Set C
ETSI FE WI007	58.67	57.59	60.83
ETSI AFE WI008	80.16	79.10	76.89
Conventional GT	64.53	67.49	65.46
SPARK + PBS	80.38	81.24	78.52

advantage of SPARK+PBS features becomes more apparent under noisy conditions.

IV. CONCLUSION

In this paper, we have presented a framework for extracting noise-robust speech features called sparse auditory reproducing kernel (SPARK) coefficients. The approach follows a computationally efficient hierarchical model where parallel similarity functions (emulating neurobiologically inspired auditory receptive fields) are computed followed by a pooling method (emulating neurobiologically inspired local competitive behavior). In this paper, we have derived an optimal form of the similarity functions which uses reproducing kernels to capture the nonlinear information embedded in the speech signal. Experimental results obtained for the AURORA2 speech recognition tasks demonstrate that the following:

- Under clean recording conditions, the performance of both baseline MFCC and SPARK based systems are comparable with a recognition accuracy of 99.25%. The result is consistent with other state-of-the-art results reported for the AURORA2 dataset.
- The SPARK features demonstrate a more robust performance in the presence of both additive and convolutive noise. We have demonstrated that SPARK can achieve average word recognition rates of 80.38%, 81.24%, and 78.52% for sets A, B, and C of the AURORA2 corpus. We have also shown that for the AURORA2 task, SPARK features combined with the PBS technique consistently out-performs the state-of-the-art ETSI AFE based features [23].

A possible extension to this work will be to investigate whether additional noise-robustness can be achieved by incorporating L_1 metric instead of an L_2 metric in the regression framework (5). We anticipate that this procedure, even though is more computationally intensive, could lead to more noise-robust speech features.

REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 16, pp. 261–291, Apr. 1995.

[2] B. H. Juang and T. H. Chen, "The past, present, and future of speech processing," *IEEE Signal Process. Mag.*, vol. 15, pp. 24–48, May 1998.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 4, pp. 113–120, Apr. 1979.

[4] A. Sankar and C.-H. Lee, "Robust speech recognition based on stochastic matching," in *Proc. ICASSP*, 1995, pp. 121–124.

[5] M. G. Rahim and B.-H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. Speech Audio Process*, vol. 4, no. 1, pp. 19–30, Jan. 1996.

[6] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Comput. Speech Lang.*, vol. 9, pp. 171–185, Apr. 1995.

[7] M. J. F. Gales and S. Young, "An improved approach to the hidden Markov model decomposition of speech and noise," in *Proc. ICASSP*, 1992, pp. 233–236.

[8] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proc. ICSLP*, 2000, pp. 869–872.

[9] Y. Gong, "A method of joint compensation of additive and convolutive distortions for speaker-independent speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 975–983, Sep. 2005.

[10] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions," *Comput. Speech Lang.*, vol. 23, pp. 389–405, Jul. 2009.

[11] H. Liao and M. J. F. Gales, "Adaptive training with joint uncertainty decoding for robust recognition of noisy data," in *Proc. ICASSP*, 2007, pp. 389–392.

[12] J. Li, L. Deng, Y. Gong, and A. Acero, "High-performance HMM adaptation with joint compensation of additive and convolutive distortions via vector Taylor series," in *Proc. ASRU*, 2007, pp. 65–70.

[13] O. Kalinli, M. L. Seltzer, and A. Acero, "Noise adaptive training using a vector Talyor series approach for noise robust automatic speech recognition," in *Proc. ICASSP*, 2009, pp. 3825–3828.

[14] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[15] K. T. Assaleh and R. J. Mammone, "New LP-derived features for speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 630–638, Oct. 1994.

[16] S. Furui, "Cepstral analysis techniques for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[17] H. Hermansky, "Perceptual linear predictive (PLP) analysis for speech," *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, Apr. 1990.

[18] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304–1322, 1974.

[19] S. Molau, F. Hilger, and H. Ney, "Feature space normalization in adverse acoustic conditions," in *Proc. ICASSP*, 2003, pp. 656–659.

[20] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech Audio Process*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[21] M. Padmanabhan and S. Dharanipragada, "Maximum likelihood nonlinear transformation for environment adaptation in speech recognition systems," in *Proc. Eurospeech*, 2001, pp. 2359–2362.

[22] D. Macho, L. Mauuary, B. Noe, Y. M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, and F. Saadoun, "Evaluation of a noise robust DSR front-end on Aurora databases," in *Proc. ICSLP*, 2002, pp. 17–20.

[23] "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms," 2007, ETSI ES 202 050 Vers. 1.1.5.

[24] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large vocabulary speech recognition under adverse acoustic environments," in *Proc. ICSLP*, 2000, pp. 806–809.

[25] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *Proc. Interspeech*, 2009, pp. 28–31.

[26] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Comput. Speech Lang.*, vol. 1, pp. 109–131, 1986.

[27] Y. Shao, S. Srinivasan, Z. Jin, and D. L. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Comput. Speech Lang.*, vol. 24, 2010.

[28] Q. Li and Y. Huang, "Robust speaker identification using an auditory-based feature," in *Proc. ICASSP*, 2010, pp. 4514–4517.

[29] S. Chatterjee and W. B. Kleijn, "Auditory model based design and optimization of feature vectors for automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1813–1825, Aug. 2011.

[30] M. Holmberg, D. Gelbart, and W. Hemmert, "Automatic speech recognition with an adaptation model motivated by auditory processing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 43–49, Jan. 2006.

[31] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 106, pp. 2040–2050, Oct. 1999.

[32] B. Strope and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 451–464, Sep. 1997.

[33] D. S. Kim, S. Y. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 1, pp. 55–69, Jan. 1999.

[34] S. Chakrabartty, Y. Deng, and G. Cauwenberghs, "Robust speech feature extraction by growth transformation in reproducing kernel Hilbert space," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1842–1849, Aug. 2007.

[35] A. Fazel and S. Chakrabartty, "Non-linear filtering in reproducing kernel Hilbert spaces for noise-robust speaker verification," in *Proc. ISCAS*, 2009, pp. 113–116.

[36] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, Jun. 1996.

[37] E. C. Smith and M. S. Lewicki, "Efficient coding of time-relative structure using spikes," *Neural Comput.*, vol. 17, pp. 19–45, Jan. 2005.

[38] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, pp. 978–982, Feb. 2006.

[39] R. Patterson and B. Moore, "Auditory filters and excitation patterns as representations of frequency resolution," *Freq. Select. in Hear.*, pp. 123–177, 1986.

[40] T. Chi, P. Ru, and S. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Amer.*, vol. 118, pp. 887–906, 2005.

[41] C. M. Wessinger, J. VanMeter, B. Tian, J. V. Lare, J. Pekar, and J. P. Rauschecker, "Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging," *J. Cognitive Neurosci.*, vol. 13, pp. 1–7, 2001.

[42] K. Okada, F. Rong, J. Venezia, W. Matchin, I.-H. Hsieh, K. Saberi, J. T. Serences, and G. Hickok, "Hierarchical organization of human auditory cortex: Evidence from acoustic invariance in the response to intelligible speech," *Cereb. Cortex*, vol. 20, pp. 2486–2495, 2010.

[43] A. Boemio, S. Fromm, A. Braun, and D. Poeppel, "Hierarchical and asymmetric temporal sensitivity in human auditory cortices," *Nature Neurosci.*, vol. 8, pp. 389–395, 2005.

[44] F. Theunissen, K. Sen, and A. J. Doupe, "Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds," *J. Neurosci.*, vol. 20, no. 6, pp. 2315–2331, 2000.

[45] J. F. Linden, R. C. Liu, M. Sahani, C. E. Schreiner, and M. M. Merzenich, "Spectrotemporal structure of receptive fields in areas ai and aaf of mouse auditory cortex," *J. Neurophysiol.*, vol. 90, no. 4, pp. 2660–2675, 2003.

[46] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with gabor feature extraction," in *Proc. ICSLP*, 2002.

[47] M. Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition," in *Proc. Eurospeech*, 2003.

[48] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 920–930, May 2006.

[49] T. Ezzat and T. Poggio, "Discriminative word-spotting using ordered spectro-temporal patch features," in *Proc. SAPA Workshop*, Sep. 2008, pp. 35–40.

[50] J. Bouvrie, T. Ezzat, and T. Poggio, "Localized spectro-temporal cepstral analysis of speech," in *Proc. ICASSP*, May 2008, pp. 4733–4736.

[51] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neurosci.*, vol. 2, pp. 1019–1025, 1999.

[52] J. Bouvrie, T. Poggio, L. Rosasco, S. Smale, and A. Wibisono, Generalization and Properties of the Neural Response Mass. Inst. of Technol., 2010, Tech. Rep., MIT-CSAIL-TR-2010-051, CBCL-292.

[53] L. Liu, J. He, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Speech Commun.*, vol. 22, no. 4, pp. 403–417, 1997.

[54] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *Proc. ICASSP*, 2001, pp. 133–136.
[55] M. Slaney, An efficient implementation of the Patterson-Holdsworth auditory filter bank, Apple Computer Tech. Rep., 1993, no. 35.
[56] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, pp. 103–108, 1990.
[57] R. D. Patterson, Holdsworth, I. Nimmo-Smith, and P. Rice, "SVOS final report: The auditory filterbank," APU Rep., 1988, no. 2341.
[58] G. Wahba, "Splines models for observational data," in *Series in Applied Mathematics*. Philadelphia, PA: SIAM, 1990, vol. 59.
[59] F. Girosi, M. Jones, and T. Poggio, "Regularization theory and neural networks architectures," *Neural Comput.*, vol. 7, pp. 219–269, 1995.
[60] R. Schluter, L. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *Proc. ICASSP*, 2007, pp. 649–652.
[61] H. G Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ASR*, 2000, pp. 181–188.
[62] "Transmission Performance Characteristics of Pulse Code Modulation Channels," 1996, ITU-T Recommendation G.712.
[63] "HTK Speech Recognition Toolkit," 2011 [Online]. Available: http://htk.eng.cam.ac.uk/
[64] "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front-End Feature Extraction Algorithm; Compression Algorithms," 2003, ETSI ES 201 108 Version 1.1.3.

Amin Fazel (S'07) received the B.Sc. degree in computer science and engineering from Shiraz University, Shiraz, Iran, in 2002 and the M.Sc. degree in computer engineering from Sharif University of Technology, Tehran, Iran, in 2005. He is currently pursuing the Ph.D. degree in electrical and computer engineering at Michigan State University, East Lansing.

His research interests include speech processing, robust speech/speaker recognition, acoustic source separation, and analog-to-information converters.

Shantanu Chakrabartty (SM'99–M'04–S'09) received the B.Tech. degree from the Indian Institute of Technology, Delhi, in 1996 and the M.S. and Ph.D. degrees in electrical engineering from Johns Hopkins University, Baltimore, MD, in 2002 and 2005, respectively.

He is currently an Associate Professor in the Department of Electrical and Computer Engineering, Michigan State University (MSU), East Lansing. From 1996 to 1999, he was with Qualcomm, Inc., San Diego, CA, and during 2002 he was a Visiting Researcher at The University of Tokyo. His work covers different aspects of analog computing, in particular nonvolatile circuits, and his current research interests include energy harvesting sensors and neuromorphic and hybrid circuits and systems.

Dr. Chakrabartty was a Catalyst foundation fellow from 1999 to 2004 and is a recipient of a National Science Foundation's CAREER award and University Teacher-Scholar Award from MSU. He is currently serving as an Associate Editor for the IEEE TRANSACTIONS OF BIOMEDICAL CIRCUITS AND SYSTEMS, Associate Editor for the *Advances in Artificial Neural Systems* journal and a Review Editor for *Frontiers of Neuromorphic Engineering* journal.