Robust Speech Feature Extraction by Growth Transformation in Reproducing Kernel Hilbert Space

Shantanu Chakrabartty, *Member, IEEE*, Yunbin Deng, *Member, IEEE*, and Gert Cauwenberghs, *Senior Member, IEEE*

Abstract—**The performance of speech recognition systems depends on consistent quality of the speech features across variable environmental conditions encountered during training and evaluation. This paper presents a kernel-based nonlinear predictive coding procedure that yields speech features which are robust to nonstationary noise contaminating the speech signal. Features maximally insensitive to additive noise are obtained by growth transformation of regression functions that span a reproducing kernel Hilbert space (RKHS). The features are normalized by construction and extract information pertaining to higher-order statistical correlations in the speech signal. Experiments with the TI-DIGIT database demonstrate consistent robustness to noise of varying statistics, yielding significant improvements in digit recognition accuracy over identical models trained using Mel-scale cepstral features and evaluated at noise levels between 0 and 30-dB signal-to-noise ratio.**

Index Terms—**Feature extraction, growth transforms, noise robustness, nonlinear signal processing, reproducing kernel Hilbert Space, speaker verification.**

I. Introduction

WHILE most current speech recognizers give acceptable recognition accuracy for clean speech, their performance degrades when subjected to noise present in practical environments [1]. For instance, it has been observed that additive white noise severely degrades the performance of Mel-cepstra-based recognition systems [1], [2]. This performance degradation has been attributed to unavoidable mismatch between training and recognition conditions. Therefore, in literature, several approaches have been presented for alleviating the effects of mismatch. These methods can be broadly categorized as follows:

- noise estimation and filtering methods that reconditions the speech signal based on noise characteristics [2];
- online model adaptation methods for reducing the effect of mismatch in training and test environments [3];
- robust feature extraction methods [4], which includes techniques based on human auditory modeling [5], [6].

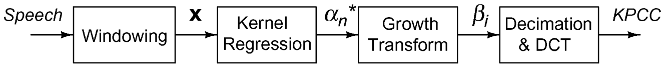

Fig. 1. Signal flow in KPCC feature extraction.

An excellent survey of techniques for improving performance of speech recognition systems under noisy environments can be found in [1]. This paper describes a novel feature extraction algorithm based on nonlinear processing of the speech signal. Termed *kernel predictive coding cepstra (KPCC)* [7], the procedure consists of two key steps, as summarized in Fig. 1: 1) estimation of a nonlinear function that captures robust higher order statistics in a segment of speech signal and 2) mapping of nonlinear function parameters onto a computationally tractable lower-dimensional manifold using *growth transformations*. *Growth transformation* is an iterative procedure for optimizing homogeneous polynomial functions of probability mass functions [13]. The technique has been used in discriminative hidden Markov model (HMM) training using maximum mutual information (MMI) [14], where it has been extended to optimizing nonhomogeneous rational functions. In this paper, estimation of nonlinear function is performed using regression techniques over a reproducing kernel Hilbert space (RKHS) [9]. RKHS regression have been extensively studied in the context of regularization theory [11], support vector machines [12], and for detection/estimation of covariance functionals [10]. Combining RKHS regression with *growth transformation* endows the proposed KPCC feature extraction algorithm with the following robustness properties.

1) The algorithm uses a semiparametric function estimation procedure without making any prior assumption on noise statistics.
2) The algorithm uses kernel methods to extract features that are nonlinear, thus utilizing higher-order statistical correlations in speech which are robust to corruption by noise.
3) Robust parameter estimation is ensured by imposing smoothness constraints based on regularization principles.
4) The features extracted are self-calibrated and normalized, which reduces mismatch between training and testing conditions.

In this paper, a step-by-step derivation of the KPCC algorithm is described along with some of its mathematical properties (Sections II and III). In Section IV, robustness of the KPCC algorithm is demonstrated by training a simple HMM-based recognizer and comparing the results with an equivalent system trained on Mel frequency cepstral coefficient (MFCC)-based features. Section V provides concluding remarks and with possible extensions of the KPCC algorithm.

Manuscript received May 3, 2006; revised March 8, 2007. This work was supported in part by a grant from the Catalyst Foundation, the National Science Foundation (NFS) under Grants IIS-0209289 and IIS-0434161, and the Office of Naval Research/Defense Advanced Research Projects Agency under Grant N00014-00-C-0315. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alex Acero.

S. Chakrabartty is with the Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48824 USA.

Y. Deng is with LumenVox, LLC, San Diego, CA 92123 USA.

G. Cauwenberghs is with the Section of Neurobiology, Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093 USA.

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.2007.899285

II. Theory

The theory of the KPCC feature extraction algorithm uses concepts from inner-product spaces, and in particular RKHS, which for the sake of completeness, is described briefly in this section. For detailed treatment of RKHS and its properties, the readers are referred to [8], [9], and [20].

A. Kernel Regression

The first step in the KPCC feature extraction algorithm is a nonlinear functional estimation procedure that extracts higher order statistics from speech signals. Given a stationary discrete time speech signal represented by $x[n] \in \mathbb{R}$, with $n = 1, \ldots, N$ denoting time indices, the aim of nonlinear regression step is to estimate a function $f : \mathbb{R}^P \to \mathbb{R}$, that can predict $\hat{x}[n]$ based on previous P samples $x[n-1], \ldots, x[n-P-1]$. We will use a vector notation to concisely represent previous P samples at time-instant n as $\mathbf{x}[n] = [x[n-1], \ldots, x[n-P-1]]^T$. The estimator function f will be assumed to be an element of a Hilbert space $f \in \mathcal{H}$, where inner-product between two functional elements $f, g \in \mathcal{H}$ will be represented as $\langle f, g\rangle_{\mathcal{H}}$. In the estimation procedure that follows, we will employ topological properties of Hilbert spaces for which readers are referred to [8] for detailed analysis. Using the property of Hilbert spaces the estimator function $f \in \mathcal{H}$ can be decomposed (*Riez's decomposition theorem* [8]) into a weighted sum of countable orthonormal basis functions $\phi_i(.) : \mathbb{R}^P \to \mathbb{R}, i \in \mathbb{N}$ as

$$f(\mathbf{x}[n]) = \sum_{i=1}^{\infty} b_i \phi_i(\mathbf{x}[n]) \tag{1}$$

with $b_i \in \mathbb{R}$. The orthonormal property of the basis functions can be expressed as $\langle \phi_i(.), \phi_j(.)\rangle_{\mathcal{H}} = (1/\Lambda_i); \forall i = j$ and equals zero otherwise. $\Lambda_i > 0$ represents a parameter of the inner-product operation. An example of a Hilbert space that admits such a decomposition is a discrete cosine transform (DCT) where the orthogonal basis functions are $\phi_i(\mathbf{x}) = \cos(\Omega_i^T \mathbf{x})$ defined at vector frequencies Ω_i. According to the decomposition given by (1), estimation of the function f is equivalent to estimation of coefficients b_i such that a reconstruction error between the input signal $x[n]$ and its estimated value, computed over a finite time window $n = P+1, \ldots, N$, is minimized. In this paper, an Euclidean metric has been chosen for computing reconstruction error as

$$\min_{f \in \mathcal{H}} R_e(f) = \sum_{n=P+1}^{N} (x[n] - f(\mathbf{x}[n]))^2 . \tag{2}$$

Since the number of parameters b_i in (1) is infinite, the minimization problem given by (2) is underspecified. Thus, any function f will over-fit the time-series data and, hence, capture the unwanted high-frequency noise components. Smoothness constraints are therefore imposed on the function f by augmenting the cost function in (2) using a regularization factor as

$$\min_{f \in \mathcal{H}} C(f) = \lambda \|f\|_H^2 + R_e(f). \tag{3}$$

The regularizer $\|f\|_{\mathcal{H}}^2$ will penalize large signal excursions by constraining the functional norm, and thus will avoid overfitting to the time-series data. This is equivalent to filtering with the aim of estimating smooth functions that can eliminate noise and at the same time preserve salient speech recognition features. λ in (3) is a parameter that determines the tradeoff between reconstruction error and the smoothness of the estimated regression function f (also known as bias-variance tradeoff in estimation theory). The regularizer can be reformulated in terms of parameters b_i using orthonormal property of the basis functions $\phi_i(.)$ as

$$\begin{aligned} \|f\|_{\mathcal{H}}^2 &= \langle f, f\rangle_{\mathcal{H}} && (4)\\ &= \left\langle \sum_{i=1}^{\infty} b_i\phi_i(.), \sum_{j=1}^{\infty} b_j\phi_j(.) \right\rangle_{\mathcal{H}} && (5)\\ &= \sum_{i,j=1}^{\infty} b_i b_j \langle \phi_i(.), \phi_j(.)\rangle_{\mathcal{H}} && (6)\\ &= \sum_{i=1}^{\infty} b_i^2/\Lambda_i . && (7) \end{aligned}$$

where linear property of inner-product operation $\langle ., .\rangle_{\mathcal{H}}$ has been used.

The optimization problem given by (3) can now be written as

$$\min_{b_i} \left[\lambda \sum_{i=1}^{\infty} b_i^2/\Lambda_i + \sum_{n=P+1}^{N} \left(x[n] - \sum_{i=1}^{\infty} b_i \phi_i(\mathbf{x}[n]) \right)^2 \right] . \tag{8}$$

The first-order condition for the optimization problem is obtained by equating the derivative of (8) with respect to parameter b_i to zero which leads to

$$b_i = \Lambda_i/\lambda \sum_{n=P+1}^{N} \alpha_n \phi_i(\mathbf{x}[n]) \tag{9}$$

where α_i represents *regression coefficients*, given by

$$\begin{aligned} \alpha_n &= x[n] - \sum_{i=1}^{\infty} b_i\phi_i(\mathbf{x}[n]) && (10)\\ &= x[n] - f(\mathbf{x}[n]) && (11) \end{aligned}$$

which is the reconstruction error for speech sample at time instant $n = P+1, \ldots, N$. The function f evaluated at any vector $\mathbf{y} \in \mathbb{R}^P$ can be written in terms of α_n as

$$f(\mathbf{y}) = \sum_{n=P+1}^{N} \alpha_n \sum_{i=1}^{\infty} \Lambda_i \phi_i(\mathbf{x}[n]) \phi_i(\mathbf{y}). \tag{12}$$

A brute force evaluation of (12) would require computation of individual basis functions $\phi_i(.)$. However, several functions of type $K : \mathbb{R}^P \times \mathbb{R}^P \to \mathbb{R}$ exist that can be expressed as

$$K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} \Lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{y}). \tag{13}$$

Such functions have been extensively studied in literature in the context of covariance kernels [10] but are commonly referred to as *reproducing kernels* [20]. The name comes from intrinsic property of space defined by K (also known as reproducing kernel Hilbert space), that exhibits a unique reproducing property as $K(\mathbf{x},\mathbf{y}) = \langle K(\mathbf{x},.), K(.,\mathbf{y})\rangle_{\mathcal{H}}$. Within the literature of covariance functions, the $\phi_i(.)$ are referred to as eigenfunctions with Λ_i being the corresponding eigenvalues. Other popular reproducing kernels have been extensively studied in machine learning literature [12] which include polynomial kernels $K(\mathbf{x},\mathbf{y}) = (\mathbf{x}\cdot\mathbf{y})^\nu; \nu = 1,2,\ldots,$ and Gaussian kernels $K(\mathbf{x},\mathbf{y}) = \exp(-\sigma(\mathbf{x}-\mathbf{y})\cdot(\mathbf{x}-\mathbf{y}))$. Interested readers are referred to [20] for exhaustive treatment of reproducing kernels and RKHS. The function $f(\mathbf{y})$ in (12) can now be written in terms of kernel functions as

$$f(\mathbf{y}) = \sum_{n=P+1}^{N} \alpha_n K\left(\mathbf{x}[n],\mathbf{y}\right). \tag{14}$$

The condition given by (11) can be written as

$$\alpha_n = x[n] - \sum_{m=P+1}^{N} \alpha_m K\left(\mathbf{x}[m],\mathbf{x}[n]\right) \tag{15}$$

which is a first-order condition for a kernel regression given in a matrix-vector form as

$$W(\boldsymbol{\alpha};\mathbf{K}) = 1/2\lambda\boldsymbol{\alpha}^T(\mathbf{K}+\lambda I)\boldsymbol{\alpha} - \boldsymbol{\alpha}^T\mathbf{X}. \tag{16}$$

with $\boldsymbol{\alpha} = \{\alpha_n\}$, $\mathbf{K}$ representing the kernel matrix with elements $K_{ij} = K(\mathbf{x}[i],\mathbf{x}[j])$, and $\mathbf{X}$ represents the time-series matrix with row vector given by $\mathbf{X}_i = \mathbf{x}[i-1]$. Based on the first-order condition given by (15), the solution is given by

$$\boldsymbol{\alpha}^* = \lambda(\lambda I + \mathbf{K})^{-1}\mathbf{X}. \tag{17}$$

Unfortunately, the regression coefficients α^* estimated over a window of size $N-P$ time samples cannot be used directly as recognition features because of the following.

1) The dimension of the regression vectors is equal to the regression window size and thus is not scalable.
2) The regression function given by (14) is contained in high-dimensional space and captures noise statistics within the regression window. From (15), it follows that $\alpha_n^* = x[n] - f^*(\mathbf{x}[n])$ where $f^*(\mathbf{x}[n]) = \sum_{m=P+1}^{N}\alpha_m^* K(\mathbf{x}[n],\mathbf{x}[m])$. Therefore, the regression coefficients α^* represent reconstruction error including noise. Fig. 2(a) and (b) show a colormap of regression coefficients computed over consecutive speech frames for utterances “zero” and “five.”
3) The regression coefficients α_n^* are not normalized, and strongly depend on ambient conditions.

The next step in KPCC feature extraction procedure is to project the regression function $f(\mathbf{x})$ onto a low-dimensional manifold, by capturing only the salient statistics of the regression function. This mapping is performed using a *growth transformation* procedure which is described in the following section.

B. Kernel Parametrization and Growth Transformation

The second step in KPCC feature extraction algorithm is mapping of nonlinear function $f(\mathbf{x})$ onto a tractable low-dimensional manifold. This is achieved by parameterizing the kernel function $K(.,.)$ according to

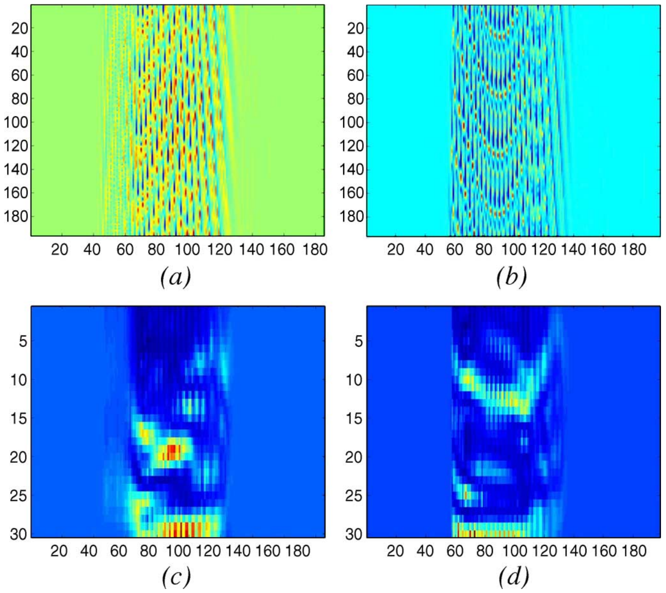

Fig. 2. Colormaps depicting regression vectors $\boldsymbol{\alpha}$ obtained for utterances of (a) digit 0 and (b) digit 5. Predictive coefficient vector $\boldsymbol{\beta}$ for same (c) digit 0 and (d) digit 5. Vertical axes denote the dimension of the vector, and horizontal axes denote time-frames.

$$K\left(\mathbf{x}[n],\mathbf{x}[m]\right) = \psi\left(\sum_{i=1}^{P}\beta_i x[n-i]x[m-i] + \gamma\right) \tag{18}$$

where $\psi(.)$ denotes any function that satisfies the RKHS properties given by (13). Examples of $\psi(.)$ include exponential functions $\psi(.) = \exp(.)$ or polynomial functions given by $\psi(.) = (.)^\nu$, $\nu = 1,2,\ldots$. $\boldsymbol{\beta} = \beta_i$, $i = 1,\ldots,P$, and γ represent parameters of the low-dimensional manifold onto which the regression function f will be mapped. The parametrization endows the kernel with the following properties.

- The kernel function $K(.,.)$ and hence the cost function $W(\boldsymbol{\alpha};\mathbf{K})$ in (16) is polynomial in $\boldsymbol{\beta}$.
- The kernel contains higher-order correlation terms of the discrete signal $x[n]$ and its delayed versions. In this sense, the kernel expansion is similar to linear predictive coding (LPC), where the coefficients β_i weigh the correlation across samples at time lags i, although the relationship is nonlinear through the map $\psi(\cdot)$ in the kernel.

To ensure proper calibration, we will enforce normalization constraints on $\beta_i \geq 0$ as $\sum_{i=1}^{P}\beta_i = 1$. The values of β_i will be determined such that it can capture higher order statistical information embedded in the regression function $f(\mathbf{x})$. For this purpose, we will determine the normalized direction that leads to maximum increase in the cost function $W(\boldsymbol{\alpha};\mathbf{K})$ given in (16). Maximization over the coefficients β_i will identify the dimensions in the input vector $\mathbf{x}$ that are least predictable, thereby differentiating between noisy and systematic components in the input.

To proceed with the maximization step of $W(\boldsymbol{\alpha};\mathbf{K})$ with respect to $\boldsymbol{\beta}$, first note that the cost function $W(\boldsymbol{\alpha};K)$ is polynomial in normalized parameters β_i. Thus, growth transformations can be applied for maximizing the cost function $W(\boldsymbol{\alpha};\mathbf{K})$. Growth transformation was first introduced in [13] for optimizing homogeneous polynomial functions defined on probability measures. In [14], the framework was extended to nonhomogenous polynomials with nonnegative coefficients. This formulation has been extensively used for designing large-vocabulary speech recognition systems based on maximum mutual information. The key result in using growth transformation for optimizing a polynomial cost function is the iterative map described by the following theorem.

Theorem 1 (Gopalakrishnan et al.): Let $H(\{P_{ik}\})$ be a polynomial of degree d in variables P_{ik} in domain $D : P_{ik} \geq 0, \sum_{k=1}^{q_i} P_{ik} = 1,\ i = 1,\ldots,N,\ k = 1,\ldots,q_i$ such that $\sum_{k=1}^{q_i} P_{ik}(\partial H/\partial P_{ik})(P_{ik}) \neq 0, \forall i$. Define an iterative map according to the following recursion:

$$\widehat{P}_{ik} \leftarrow \frac{P_{ik}\left(\frac{\partial H}{\partial P_{ik}}(P_{ik})+\Gamma\right)}{\sum_{k=1}^{q_i} P_{ik}\left(\frac{\partial H}{\partial P_{ik}}(P_{ik})+\Gamma\right)} \tag{19}$$

where $\Gamma \geq Sd(N+1)^{d-1}$ with S being the smallest coefficient of the polynomial $H(\{P_{ik}\})$. Then $\{\widehat{P}_{ik}\} \in D$ and $H(\{\widehat{P}_{ik}\}) > H(\{P_{ik}\})$.

Using the growth transformations procedure described in [14] over the manifold $M : \sum_i \beta_i = 1, \beta_i \geq 0$ the coefficients $\boldsymbol{\beta}$ are remapped according to

$$\beta_i \leftarrow \frac{\beta_i \partial W(\boldsymbol{\alpha}^*;\mathbf{K})/\partial\beta_i + D}{\sum_k \left(\beta_k \partial W(\boldsymbol{\alpha}^*;\mathbf{K})/\partial\beta_k + D\right)}. \tag{20}$$

Fig. 2(c) and (d) show a colormap of coefficients β computed over consecutive speech frames for utterances “zero” and “five.” The parameter D is a smoothing constant that determines the degree of deviation of the new parameters with respect to the old parameters and plays an important role in noise robustness. Fig. 3 shows colormaps depicting kernel predictive coefficients β for an instance of utterance *five*, computed using different values of D. The plot shows that lower values of D makes the coefficients $\boldsymbol{\beta}$ sensitive to low-energy speech frames (fricatives and silent frames), as evident from Fig. 3(a).

Insight into the role of the transformation (20) can be gained through the following mathematical manipulation:

$$\frac{\partial W(\boldsymbol{\alpha}^*;\mathbf{K})}{\partial\beta_i} = \frac{1}{2}\sum_{n,m} \alpha_n^* \alpha_m^* K'_{ij} x[n-i]x[m-i] \tag{21}$$

$$= \frac{1}{2}\left\| \sum_m \alpha_m^* x[m-i] K'\left(.,\mathbf{x}[m]\right) \right\|_{\mathcal{H}'}^2 \tag{22}$$

where $\mathbf{K}'$ is the derivative of kernel $\mathbf{K}$. For the kernels under consideration, $\mathbf{K}'$ also has a reproducing property over a corresponding Hilbert space $\mathcal{H}'$. For the specific choice $K(\mathbf{x},\mathbf{y}) = \exp(\langle \mathbf{x},\mathbf{y}\rangle)$, both RKHS representations coincide $\mathcal{H}' = \mathcal{H}$ and, thus, the coefficients β_i in (20) are proportional to the norm of the function $h_i \in \mathcal{H}$ where

$$h_i(\mathbf{x}) = \sum_{m=1}^{N-P} \alpha_m^* x[m+P-i] K\left(\mathbf{x},\mathbf{x}[m+P]\right). \tag{23}$$

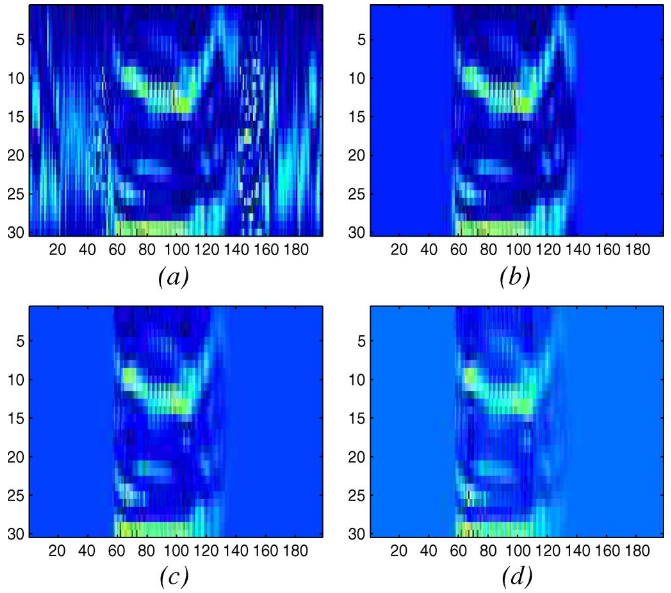

Fig. 3. Colormaps depicting predictive coefficient vectors $\boldsymbol{\beta}$ for utterance five computed using different values of parameter D (a) $D = 0$, (b) $D = 0.01$, (c)$D = 1$, and (d) $D = 100$. Vertical axes denote the dimension of the vector, and horizontal axes denote time frames.

The function $h_i(\mathbf{x})$ represents sum of kernel functions (similarity functions) evaluated between speech vectors $\mathbf{x}[m+P]$ and $\mathbf{x}$, weighted by a factor obtained as a product of reconstruction error α_m^* at time-instant $m+P$ [see (11)] and the data sample at a time lag i. Thus, the higher the correlation between reconstruction error with data samples at time lag i, the higher the norm of h_i and hence the larger the value of β_i. As an interesting limiting case, for $\lambda \to \infty$ and for a linear kernel of type $K(\mathbf{x}[n],\mathbf{x}[m]) = (\sum_{i=1}^{P} \beta_i x[n-i]x[m-i] + \gamma)$, according to (20), the coefficients $\beta_i \propto (\sum_m x[m]x[m-i])^2$ relate directly to the L_2 norm of the autocorrelation function. This affirms that the transformation (20) subsumes linear predictive coding [16] and produces more general, nonlinear features.

Each iteration of the map (20) increases the weight β_i corresponding to the time delay i for which the signal value $x[m-i]$ correlates strongly with the reconstruction error. Thus, any subsequent iterations of (20) favors noisy coefficients and degrades the performance of the features. Therefore, in this paper, the KPCC feature extraction algorithm uses only a single iteration of the map (20).

III. KPCC Feature Extraction Algorithm

A formulation of KPCC feature extraction procedure consists of a kernel regression step to estimate the regression vectors α^* followed by growth transformation steps to obtain KPCC coefficients β_i. The feature extraction steps are illustrated in Fig. 1 and are described as follows.

1) Extract speech samples by shifting a rectangular window of size N by W intervals; the values of N and W are determined by the sampling frequency.
2) For a given order P, choose an initial value of parameters β_i, $i = 1,\ldots,P$. In the experiments described in the paper, the initial values were chosen according to the profile $\beta_i = c + h\sin(i\pi/P)$, akin to the liftering

profile in Mel-scale filterbank cepstral coefficient feature extraction [16]. The initial value of β_i controls the shape of the regression window and, hence, the form of the regression function.

3) Obtain the kernel matrix K by applying the map $\exp(.)$ to (18) over the data window. Train the dual objective (16) by assigning optimal coefficients α_n^* according to (17). The optimal coefficients directly determine the regression function according to (14). Fig. 2(a) and (b) shows an image plot of optimal coefficients for digits 0 and 5 sampled at 20 KHz.
4) Perform growth transformation (20) to obtain new estimates of β_i, at the optimum value of α_n^*.
5) Average and decimate the coefficients β_i along $i = 1, \ldots P$ to reduce the number of features. Fig. 2(c) and (d) shows the image plot of the prediction coefficients β_i corresponding to digits 0 and 5.
6) Perform discrete-cosine transformation (DCT) on the reduced coefficients to obtain the final KPCC features. As in MFCC feature extraction, the first DCT coefficient and higher-order coefficients are discarded since they carry little information relevant to speech.

A. Computational Complexity

Functional estimation step given by (17) and quadratic computation step given by (22) are the computationally intensive part of the KPCC feature extraction algorithm. However, similar to linear-predictive feature extraction, KPCC feature extraction algorithm utilize the structure of the kernel matrix to efficiently solve (17). If stationarity of higher order speech features is assumed, the kernel matrix exhibits a Toeplitz property with $K_{m-n} \doteq K(\mathbf{x}[i+m], \mathbf{x}[i+n]), \forall i$. Under such conditions, *Yule–Walker* recursions [16] can be used for solving (17). When higher order stationarity cannot be assumed, similar to LPC feature extraction auto-covariance like methods, for instance Cholesky decompositions [17] can be used to solve (17). Unfortunately, KPCC features require estimation of $N-P$ regression coefficients α_n^*, which makes the algorithm slower than conventional LPC features that use low-order all-pole filters. Quadratic computation, on the other hand, requires evaluation of (22) for which several efficient implementation have been reported [17]. Using these results, the complexity of the quadratic computation step can be estimated to be approximately $P(N-P)\log(N-P)$, which models evaluation of P predictive coefficients β_i, $i = 1, \ldots, P$. Since the focus of this paper is to introduce the KPCC feature extraction algorithm, we have resorted to a brute force matrix inversion implementation. Subsequent papers will elaborately discuss possible efficient algorithmic implementation of KPCC feature extraction.

IV. Experiments and Results

For all experiments, KPCC features were extracted using a 20-ms window shifted by 10 ms, with kernel regression order $P = 60$, and with $c = 0.3$, $h = 0.5$, $\lambda = 0.5$, $D = 1$, and $\gamma = 0.3$. These parameter values were chosen based on several recognition experiments that yielded superior performance. The 60 growth features β_i were averaged and decimated to 30 coefficients. Without loss of generalization, it has been assumed that the input signal is rescaled such that $x[n] \leq 1, \forall n$. Fig. 4 compares the KPCC predictive coefficients for digits 0-4 spoken by three different speakers. The first two columns correspond to the same speaker. Similarly, Fig. 5 compares the KPCC spectrum for digits 5–9. A visual inspection of Figs. 4 and 5 show that the KPCC features corresponding to the same digit bears similarity across different speakers. This provides motivation for integrating KPCC features with inference models to build speech recognition systems.

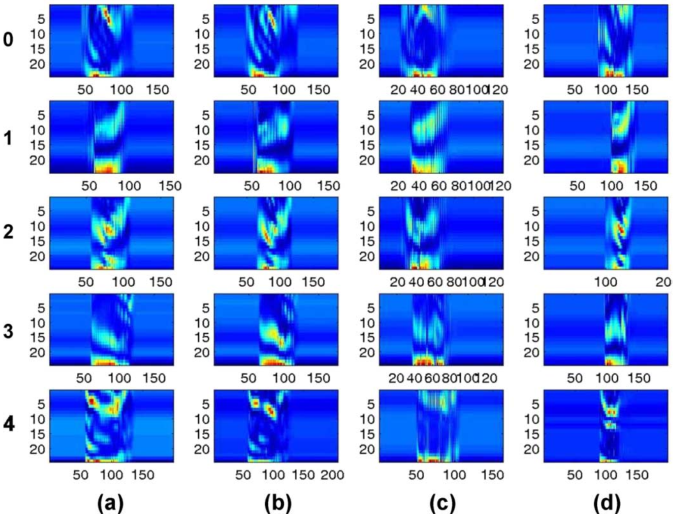

Fig. 4. KPCC predictive coefficients for digits 0–4 corresponding to three different speakers. The feature maps in (a) and (b) correspond to the same speaker, whereas (c) and (d) correspond to different speakers.

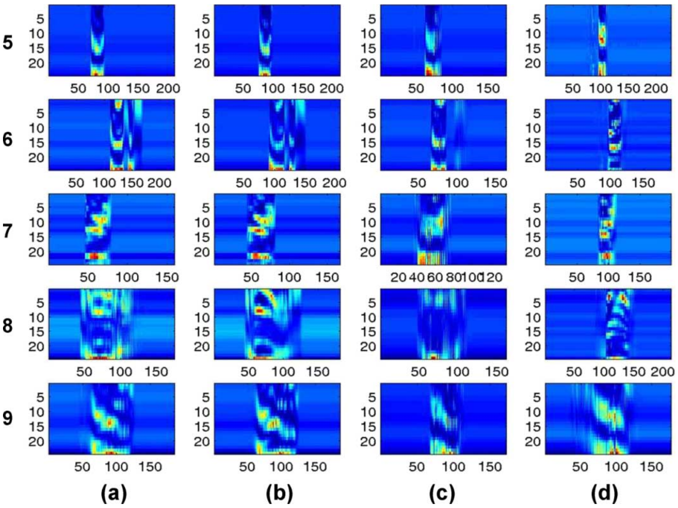

Fig. 5. KPCC predictive coefficients for digits 5–9 corresponding to three different speakers. The feature maps in columns (a) and (b) belong to the same speaker, whereas feature maps in columns (c) and (d) belong to different speakers.

For comparison, MFCC-based features [16] were chosen. After DCT, 12 coefficients (indices 2–13) were selected as features for the recognition system. Fig. 6 shows a sample comparison between KPCC features and corresponding MFCC features for digit *five* obtained before DCT operation for different signal-to-noise ratio (SNR) levels. As standard in MFCC [18], a window size of 25 ms with an overlap of 10 ms

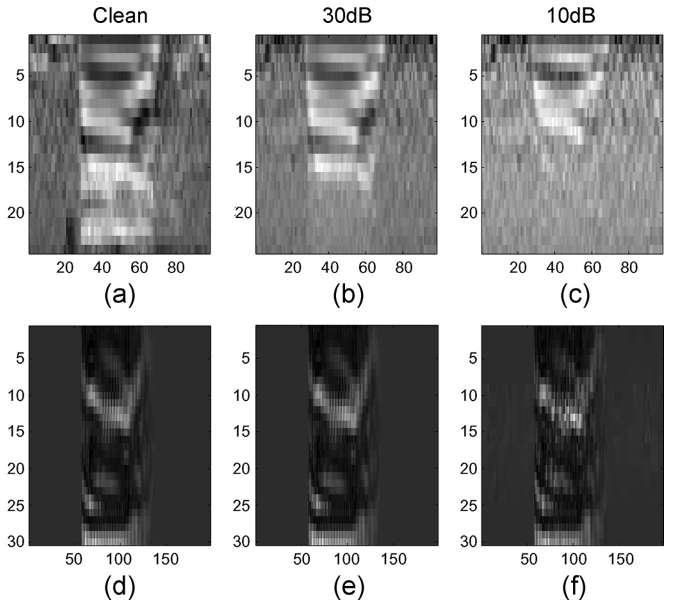

Fig. 6. (a)–(c) MFCC features and (d)–(f) KPCC features for digit *five* obtained before DCT, under different SNR conditions (clean, 30 dB, and 10 dB).

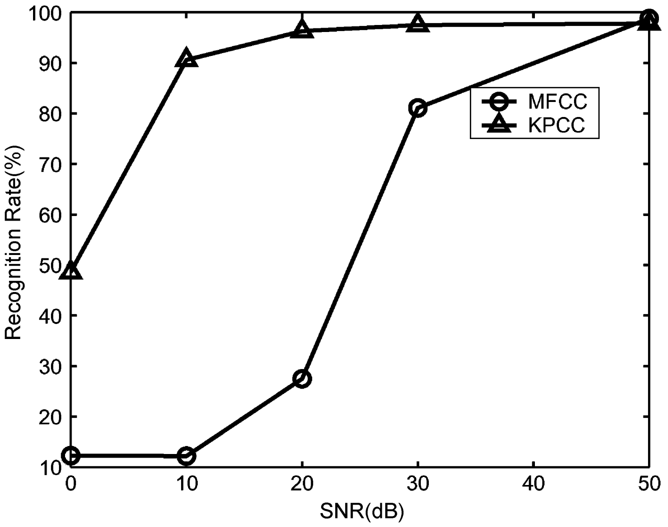

Fig. 7. Comparison of recognition rate for systems trained with MFCC and KPCC features for speech samples corrupted with white noise.

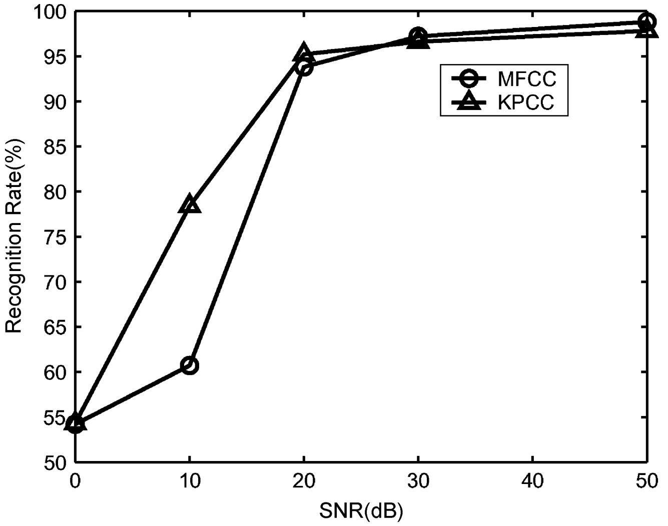

Fig. 8. Comparison of recognition rate for systems trained with MFCC and KPCC features for speech samples corrupted with babble noise.

was chosen, and cepstral features were obtained from DCT of log-energy over 24 Mel-scale filter banks. The degradation of spectral features for MFCC in the presence of white noise is evident, whereas KPCC features prevail at elevated noise levels.

For recognition experiments, we chose a simple isolated TI-DIGIT digit recognition task with a vocabulary size of 11 (zero to ten plus "O"). The training set contained two utterances of isolated digits each from 35 male speakers comprising a total of 770 utterances, and the test set contained isolated digits from 25 other male speakers for a total of 440 utterances. A recognition system was developed using the Hidden Markov Toolkit (HTK) [18], implementing a 14-state left-to-right transition model for each digit, where the probability distribution on each state was modeled as a four-mixture Gaussian. As a baseline, the same recognition system was developed using MFCC features comprising of 12 coefficients, without energy and delta features. The performance of MFCC and KPCC features were compared when speech signal was corrupted by additive noise during recognition conditions. It was assumed that the channel characteristics based on room impulse response remained fixed between training and testing conditions. In Section V, we discuss possible ways to embed robustness in KPCC features to channel variations.

Noise samples for the experiments were obtained from the *NOISEX* database [19] which are recorded sound clips in real-life environments (NOISE-ROM-0 from NATO:AC243/Panel3/RSG-10). The noise clippings were added to clean speech obtained from TIDIGIT database to generate test data. For these experiments, it was assumed that the channel characteristics based on room impulse response remained fixed between training and testing conditions. We considered four types of noise common in application environments: white noise (W), speech babble noise (B, cafeteria noise), factory noise (F, plate-cutting and electrical welding equipment), car interior noise (C, Volvo 340 at 75 mi/h under rainy conditions) and airplane cockpit noise (A, jet moving at 190 Kmi/h at 1000 feet and sound level 109 dBA). Figs. 7–11 summarizes the recognition rates obtained based on the two features under different noise statistics and SNR levels.

The following can be inferred from the plots.

1) For clean speech, the performance of both systems are comparable, with high recognition rates.
2) For white noise the recognition system with KPCC features demonstrates much better noise robustness than corresponding MFCC features. In fact, KPCC maintains acceptable (>90%) recognition performance for noise approaching signal levels (SNR reaching 10 dB).
3) KPCC features demonstrate significantly better performance in the presence of factory noise and slightly better performance in the presence of babble noise. An interesting observation can be made at this point by noting the trend in recognition rates for babble noise in comparison with other

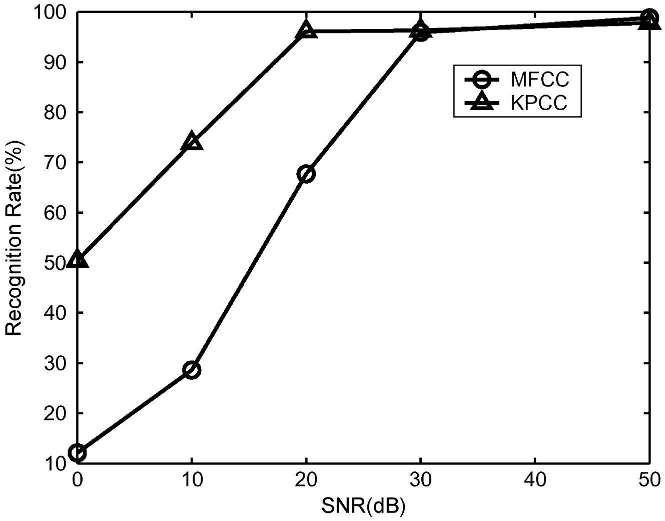

Fig. 9. Comparison of recognition rate for systems trained with MFCC and KPCC features for speech samples corrupted with factory noise.

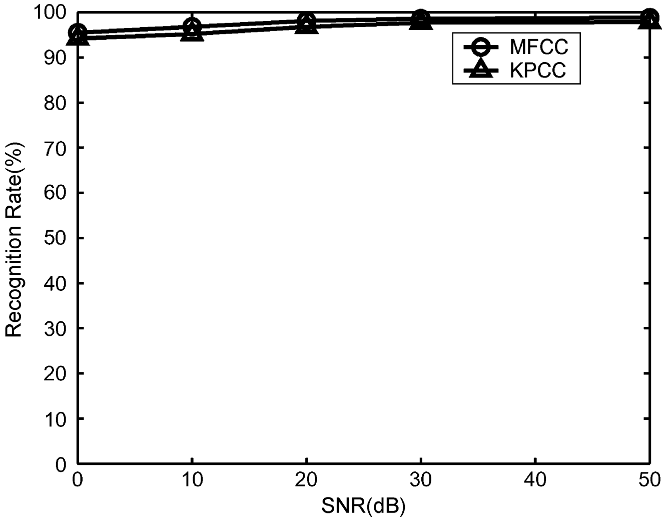

Fig. 10. Comparison of recognition rate for systems trained with MFCC and KPCC features for speech samples corrupted with car noise.

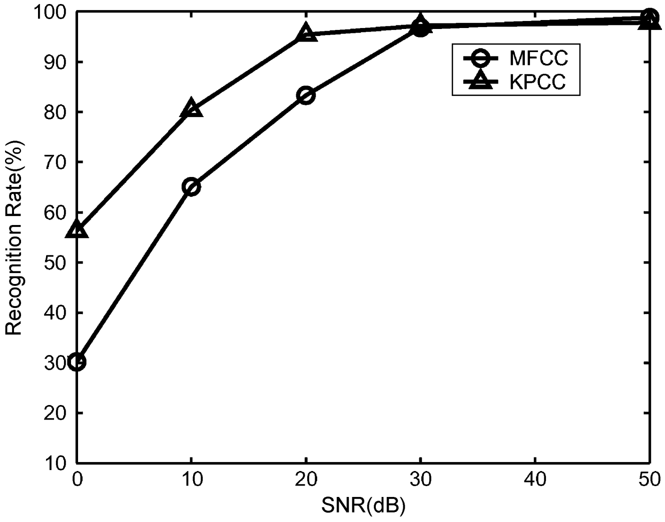

Fig. 11. Comparison of recognition rate for systems trained with MFCC and KPCC features for speech samples corrupted with airplane cockpit noise.

noise types. Babble noise primarily consists of speech signals produced by other humans and hence not only corrupts the entire information bearing frequency bands but also shares statistical properties of the reference signal. This attribute is reflected by a reduction in error rates even though KPCC features are more robust to MFCC features. For other sources of noise, the statistics are substantially different from reference statistics, which KPCC features utilize to extract noise robust features. This can be observed especially for white noise at very low SNR, for which KPCC features provide reasonable recognition performance.

4) The performance of both MFCC features and KPCC features do not degrade rapidly in the presence of car noise and yield similar relative decrease in recognition rates. This can be attributed to the very low-frequency nature of car noise, which keeps the higher frequency features intact for recognition purposes. Also, the baseline performance of the recognizer using MFCC features is better than the recognizer trained with KPCC features, which introduces a fixed offset in recognition performance as shown in Fig. 10.

V. Conclusion and Future Directions

In this paper, we presented a novel speech feature extraction procedure robust to noise with different statistics, for deployment with recognition systems operating under a wide variety of conditions. The approach is primarily data driven and effectively extracts nonlinear features of speech that are largely invariant to noise and interference with varying statistics.

An extension of the proposed work is to evaluate the features for a task of speaker verification. As can be seen in Figs. 4 and 5, the KPCC predictive coefficients bear more similarity for the same speakers as opposed to nonidentical speakers. The theory of KPCC features can be extended by linking growth transformation on regression functions with game theoretic principles in machine learning. The regression function estimation and growth transformation can be viewed as balancing criteria on optimizing the dual objective function (16). In this paper, a single iteration between the criterion was used to identify the robust features. One of the disadvantages of single iteration is the dominance of the stronger features over weak features. In principle, multiple iterations between the balancing criterions could be used to identify the weak features and could be used to improve the recognition accuracy of the system.

Another extension to this work is to design KPCC features that are robust to channel variations, a property which is naturally endowed in MFCC features due to cepstral-based signal processing [16]. Cepstral filtering techniques for extracting MFCC features assume that the channel variations can be modeled using a linear time-invariant filter that modifies the spectrum of the speech signal. It was shown using an example in Section II that KPCC regression procedure is equivalent to a filtering operation, where the shape of the filter is determined by a choice of the kernel function. Therefore, specific kernel functions can be used for constructing regression function which acts as a matched filter for reducing channel effects. In our future work, we will explore possibility of kernels with a bandpass filter response for eliminating slow-varying channel components, similar to RASTA-based methods [6].

ACKNOWLEDGMENT

The authors would like to thank S. Kumar and Y. Deng of CLSP at JHU for their help in developing the baseline MFCC system using HTK.

REFERENCES

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

[2] S. V. Vaseghi and B. P. Milner, "Noise-adaptive hidden Markov models based on Wiener filters," in *Proc. Eur. Conf. Speech Technol.*, Berlin, Germany, 1993, vol. II, pp. 1023–1026.

[3] A. Nadas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 37, no. 10, pp. 1495–1503, Oct. 1989.

[4] D. Mansour and B. H. Juang, "The short-time modified coherence representation and its application for noisy speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, New York, 1988, pp. 525–528.

[5] O. Ghitza, "Auditory nerve representation as a basis for speech processing," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1992, ch. 15, pp. 453–485.

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

[7] S. Chakrabartty, Y. Deng, and G. Cauwenberghs, "Robust speech feature extraction by growth transformation in reproducing kernel Hilbert space," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*, Montreal, QC, Canada, 2004, pp. 133–136.

[8] W. Rudin, *Functional Analysis*. New York: McGraw-Hill, 1973.

[9] S. Saitoh, "Theory of reproducing kernels and its applications," in *Longman Scientific and Technical*. Essex, U.K.: Harlow, 1988.

[10] T. Kailath and H. Weinert, "An RKHS approach to detection and estimation problems—Part II: Gaussian signal detection," *IEEE Trans. Inf. Theory*, vol. IT-21, no. 1, pp. 15–23, Jan. 1975.

[11] F. Girosi, M. Jones, and T. Poggio, "Regularization theory and neural network architectures," *Neural Comput.*, vol. 7, pp. 219–269, 1995.

[12] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[13] L. E. Baum and G. Sell, "Growth transformations for functions on manifolds," *Pacific J. Math.*, vol. 27, no. 2, pp. 211–227, 1968.

[14] P. S. Gopalakrishnan *et al.*, "An inequality for rational functions with applications to some statistical estimation problems," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 107–113, Jan. 1991.

[15] G. Wahba, *Spline models for observational data (CBMF-NSF Regional Conference Series in Applied Mathematics)*. Philadelphia, PA: SIAM, 1990, vol. 59.

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

[17] G. H. Golub and C. F. Van Loan, *Matrix Computation*. North Oxford, U.K.: Academic, 1983.

[18] The Hidden Markov Model Toolkit. [Online]. Available: http://www.htk.eng.cam.ac.uk

[19] The Noisex-92 Database. [Online]. Available: http://www.speech.cs.cmu.edu

[20] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, pp. 337–404, 1950.

[21] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations," *Philosoph. Trans. Roy. Soc.*, vol. 209, pp. 415–446, 1909, London, U.K.

Shantanu Chakrabartty (M'96) 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 The Johns Hopkins University, Baltimore, MD, in 2001 and 2004, respectively.

He is currently an Assistant Professor in the Department of Electrical and Computer Engineering, Michigan State University, 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, Tokyo, Japan. His current research interests include low-power analog and digital VLSI systems, hardware implementation of machine learning algorithms with application to biosensors and biomedical instrumentation.

Dr. Chakrabartty was a recipient of The Catalyst Foundation Fellowship from 1999 to 2004 and won the Best Undergraduate Thesis Award in 1996. He is currently a member of the IEEE BioCAS Technical Committee and IEEE Circuits and Systems Sensors Technical Committee.

Yunbin Deng (M'97) received the B.E. degree in control engineering from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 1997, the M.S. degree in electrical engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2000, and the M.S.E. and Ph.D. degrees in electrical and computer engineering from The Johns Hopkins University (JHU), Baltimore, MD, in 2002, and 2006, respectively.

He joined Intelligent Automation, Inc., Rockville, MD, as a Research Engineer in 2005. He is currently a Speech Scientist with LumenVox, LLC. His research interests include language and speech processing, robust and nonnative speech recognition, dialog systems, mixed-signal VLSI circuits and systems, and machine learning.

Dr. Deng won the Outstanding Overseas Chinese Student Award from the Chinese Scholarship Council in 2003. He is a member of SPIE. He is a Reviewer for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, the IEEE TRANSACTION ON CIRCUITS AND SYSTEMS, and the *Circuits, Systems, and Signal Processing Journal*.

Gert Cauwenberghs (S'89–M'94–SM'04) received the Ph.D. degree in electrical engineering from the California Institute of Technology, Pasadena, in 1994.

Previously, he was a Professor of Electrical and Computer Engineering at The Johns Hopkins University, Baltimore, MD. He joined the University of California at San Diego, La Jolla, as a Professor of Neurobiology in 2005. His research aims at advancing silicon adaptive microsystems to understanding of biological neural systems and to development of sensory and neural prostheses and brain–machine interfaces.

Prof. Cauwenberghs received the NSF Career Award in 1997, the ONR Young Investigator Award in 1999, and the Presidential Early Career Award for Scientists and Engineers in 2000. He is Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I, the IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING, and the IEEE SENSORS JOURNAL.