

# Combined Features and Kernel Design for Noise Robust Phoneme Classification Using Support Vector Machines

Jibrán Yousafzai<sup>†</sup>, *Student Member, IEEE*    Peter Sollich<sup>‡</sup>  
 Zoran Cvetković<sup>†</sup>, *Senior Member, IEEE*    Bin Yu<sup>\*</sup>, *Fellow, IEEE*

**Abstract**— This work proposes methods for combining cepstral and acoustic waveform representations for a front-end of support vector machine (SVM) based speech recognition systems that are robust to additive noise. The key issue of kernel design and noise adaptation for the acoustic waveform representation is addressed first. Cepstral and acoustic waveform representations are then compared on a phoneme classification task. Experiments show that the cepstral features achieve very good performance in low noise conditions, but suffer severe performance degradation already at moderate noise levels. Classification in the acoustic waveform domain, on the other hand, is less accurate in low noise but exhibits a more robust behavior in high noise conditions. A combination of the cepstral and acoustic waveform representations achieves better classification performance than either of the individual representations over the entire range of noise levels tested, down to  $-18\text{dB}$  SNR.

**Index Terms**—Robustness, Support vector machines, Acoustic waveforms, Kernels, Phoneme classification

## I. INTRODUCTION

State-of-the-art systems for automatic speech recognition (ASR) use cepstral features, normally some variant of Mel-Frequency Cepstral Coefficients (MFCC) [1] or Perceptual Linear Prediction (PLP) [2] as their front-end. These representations are derived from the short-term magnitude spectra followed by non-linear transformations to model the processing of the human auditory system. The aim is to compress the highly redundant speech signal in a manner which removes variations that are considered unnecessary for recognition, and thus facilitate accurate modelling of the information relevant for discrimination using limited data. However, due to the nonlinear processing involved in the feature extraction, even a moderate level of distortion may cause significant departures from feature distributions learned on clean data, making these distributions inadequate for recognition in the presence of environmental distortions such as additive noise and linear filtering. It turns out that recognition accuracy of state-of-the-art ASR systems is indeed far below human performance in adverse conditions [3, 4].

To make the cepstral representations of speech less sensitive to noise, several techniques [5–13] have been developed that

aim to reduce explicitly the effects of noise on spectral representations and thus approach the optimal performance which should be achieved when training and test conditions are matched [14]. State-of-the-art feature compensation methods include the ETSI advanced front-end (AFE) [11], which is based on MFCCs but includes denoising and voice activity detection, and the vector Taylor series (VTS) based feature compensation [7–10]. The latter estimates the distribution of noisy speech given the distribution of clean speech, a segment of noisy speech, and the Taylor series expansion that relates the noisy speech features to the clean ones, and then uses it to predict the unobserved clean cepstral feature vectors. Additionally, cepstral mean-and-variance normalization (CMVN) [12, 13] is performed to standardize the cepstral features, fixing their range of variation for both training and test data. CMVN computes the mean and variance of the feature vectors across a sentence and standardizes the features so that each has zero mean and a fixed variance. These methods contribute significantly to robustness by alleviating some of the effects of additive noise (as well as linear filtering). However, due to the non-linear transformations involved in extracting the cepstral features, the effect of additive noise is not merely an additive bias and multiplicative change of scale of the features, as would be required for CMVN to work perfectly [13].

Current ASR methods that are considered robust to environmental distortions are based on the assumption that the conventional cepstral features form a good enough representation, so that in combination with a suitable language and context modelling the performance of ASR systems can be brought close to human speech recognition. But for such modelling to be effective, the underlying sequence of elementary phonetic units must be predicted sufficiently accurately. This is, however, where there are still significant gaps between human performance and ASR. Humans recognize isolated speech units above the level of chance already at  $-18\text{dB}$  SNR, and significantly above it at  $-9\text{dB}$  SNR [15]. Even in quiet conditions, the machine phone error rates for nonsense syllables are higher than human error rates [3, 4, 16, 17].

Several studies [17–22] have attributed the marked difference between human and machine performance to the fundamental limitations of the feature extraction process. Among them, the studies on human speech perception [17, 19, 21, 22] have shown that the information reduction that takes place in the conventional front-ends leads to a severe drop in human speech recognition performance and that there is a high correlation between humans and machines in terms of recognition accuracy in noisy environment when both are exposed to speech with the kind of distortions introduced

Copyright (c) 2010 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

The authors are with the Department of Informatics<sup>†</sup> and the Department of Mathematics<sup>‡</sup> at King's College London, and the Department of Statistics<sup>\*</sup> at University of California, Berkeley (e-mail: {peter.sollich,zoran.cvetkovic}@kcl.ac.uk, binyu@stat.berkeley.edu). Financial support from EPSRC under grant EP/D053005/1 is gratefully acknowledged. Bin Yu thanks the National Science Foundation for grants NSF SES-0835531 (CDI) and NSF (60628102).

by typical ASR front-ends. This makes finding methods that would more effectively account for noise effects, or features whose probability distributions would not alter significantly in the presence of noise, of utmost importance for robust ASR.

In this paper, we propose combining cepstral features with high-dimensional acoustic waveform representations using SVMs [23–26] to improve the robustness of phoneme classification to additive noise (convolutional noise is not considered further in this paper). This is motivated by the fact that acoustic waveforms retain more information about speech than the corresponding cepstral representation. Furthermore, the linearity of the manner in which noise and speech are combined in the acoustic waveform domain allows for straightforward noise compensation. The same would of course be true for any linear transform, *e.g.* Fourier Transform (linear spectrum) or discrete cosine transform. The high-dimensional space of acoustic waveforms might also provide better separation of the phoneme classes in high noise conditions and hence make the classification more robust to additive noise. To effectively use acoustic waveforms with SVMs for phoneme classification, specially designed kernels that express the information relevant for recognition need to be designed, and this is one of the central themes of this paper. In addition, we explore the benefits of hybrid features that combine cepstral features with local energy features of acoustic waveform segments. These features can be compensated effectively, by exploiting the approximate orthogonality of clean speech and noise to subtract off the estimated noise energy before any non-linear transform is applied. The effectiveness of the hybrid features in improving robustness, when used with custom-designed kernels, is demonstrated in experiments. Acoustic waveforms and cepstral features are then compared on a phoneme classification task. Phoneme classification is a task of reasonable complexity, studied by other researchers [27–35] for the purpose of testing different methods and representations; the improvements achieved can be expected to extend to continuous speech recognition tasks [25, 36].

In broad terms, our experiments show that classification in the cepstral domain gives excellent results in low noise conditions but suffers severe degradation in high noise. Classification in the acoustic waveform domain is not as accurate as in the cepstral domain in low noise but exhibits a more robust behavior in severe noise. We therefore construct a convex combination of the cepstral and acoustic waveform classifiers. The combined classifier outperforms the individual ones across all noise levels and even outperforms the cepstral classifiers for which training and testing is performed under matched noise conditions.

Short communications of the early stages of this work have appeared in [37, 38]. Here we significantly extend our approach to account for the fine-scale and subband dynamics of speech. We also investigate in detail the issues of noise compensation and classifier combination, and perform experiments that integrate an SNR estimation algorithm for noise compensation of acoustic features in the presence of non-stationary noise. Basics of SVM classification are reviewed in Section II. In the same section we then describe our design of kernels for the classification task in the cepstral and the acoustic waveform

domains, along with the compensation of features corrupted by additive noise. Results on phoneme classification using the two representations and their combination are then reported in Section III. Finally, Section IV presents conclusions and an outlook toward future work.

## II. SVM KERNELS FOR SPEECH RECOGNITION

Support vector machines are receiving increasing attention as a tool for speech recognition applications [7, 24–26, 32, 39–41]. The main aim of the present work is to find representations along with corresponding kernels and effective noise compensation methods for noise robust speech recognition using SVMs. We focus on fixed-length,  $D$ -samples long, segments of acoustics waveforms, which we will denote by  $\mathbf{x}$ , and their corresponding cepstral representations  $\mathbf{c}$ , and compare them in a phoneme classification task. The classification in the acoustic waveform domain opens up a whole set of issues regarding the non-lexical invariances (sign, time alignment) and dynamics of speech that need to be taken into account by means of custom-designed kernels. Since this paper primarily focuses on comparison of different representations in terms of the robustness they provide, we conduct experiments using fixed-length representations which could potentially be used as front-ends for a continuous speech recognition system based on *e.g.* hidden Markov models (HMMs). Dealing with variable phoneme length lies beyond the scope of this paper; it has been addressed by means of dynamic kernels based on generative models such as Fisher kernels [41, 42], GMM supervector kernels [40], as well as generative kernels [39] that combine the strengths of generative statistical models and discriminative classifiers.

Our proposed approach can be used in a hybrid phone-based architecture that integrates SVMs with HMMs for continuous speech recognition [25, 26]. This is a two-stage process unlike the systems described in [39, 42] where HMMs and SVMs are tied closely together via dynamic kernels. It requires a baseline HMM system to perform a first pass through the test data. This generates for each utterance a set of possible segmentations into phonemes. The best segmentations are then re-scored by the discriminative classifier to predict the final phoneme sequence. This approach has provided improvements in recognition performance over HMM baselines on both small and large vocabulary recognition tasks, even though the SVM classifiers were constructed solely from the cepstral representations [25, 26]. The work presented in this paper can be integrated directly into this framework and would be expected to similarly improve the recognition performance over HMM baselines. This will be explored in future work, as will be extensions of our kernels for use with recently proposed frame-based architectures employing SVMs directly for continuous speech recognition using a token passing algorithm and dynamic time-warping kernel [43, 44].

### A. Support Vector Machines

Given a set of training data  $(\mathbf{x}_1, \dots, \mathbf{x}_p)$  with corresponding class labels  $(y_1, \dots, y_p)$ ,  $y_i \in \{+1, -1\}$ , an SVM attempts to find a decision surface which jointly maximizes the margin

between the two classes and minimizes the misclassification error on the training set. In the simplest case, these surfaces are linear but most pattern recognition problems require nonlinear decision boundaries, and these are constructed by means of nonlinear kernel functions. For the classification of a test point  $\mathbf{x}$ , an SVM trained to discriminate between two classes of data thus computes a score  $h(\mathbf{x}) = \sum_{i=1}^p \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b$  where  $K$  is a kernel function,  $\alpha_i$  is the Lagrange multiplier corresponding to the  $i^{\text{th}}$  training sample,  $\mathbf{x}_i$ , and  $b$  is the classifier bias. The class of  $\mathbf{x}$  is then predicted based on the sign of the score function,  $\text{sgn}(h(\mathbf{x}))$ . While  $b$  and the  $\alpha_i$  are optimized during training, the kernel function  $K$  has to be designed based on *a priori* knowledge about the specific task. The simplest kernel is the inner product function,  $K(\mathbf{x}, \tilde{\mathbf{x}}) = \langle \mathbf{x}, \tilde{\mathbf{x}} \rangle$ , which produces linear decision boundaries. Nonlinear kernel functions implicitly map data points to a high-dimensional feature space where decision boundaries are again linear. Kernel design is therefore effectively equivalent to feature-space selection, and using an appropriate kernel for a given classification task is crucial. Intuitively, the kernel should be designed so that  $K(\mathbf{x}, \tilde{\mathbf{x}})$  is high if  $\mathbf{x}$  and  $\tilde{\mathbf{x}}$  belong to the same class and low if  $\mathbf{x}$  and  $\tilde{\mathbf{x}}$  are in different classes.

Two commonly used kernels are the polynomial kernel

$$K_p(\mathbf{x}, \tilde{\mathbf{x}}) = (1 + \langle \mathbf{x}, \tilde{\mathbf{x}} \rangle)^\Theta, \quad (1)$$

and the radial basis function (RBF) kernel  $K_r(\mathbf{x}, \tilde{\mathbf{x}}) = e^{-\Gamma \|\mathbf{x} - \tilde{\mathbf{x}}\|^2}$ . The integer polynomial order  $\Theta$  in  $K_p$  and the width factor  $\Gamma$  are hyper-parameters which are tuned to a particular classification problem. More sophisticated kernels can be obtained by combining such basic kernels. In preliminary experiments we found that the standard polynomial and RBF kernels,  $K_p$  and  $K_r$ , lead to similar speech classification performance. Hence the polynomial kernel  $K_p$  in (1) will be used as the baseline for both the cepstral and acoustic waveform representations of speech. With cepstral representations having already been designed to extract the information relevant for the discrimination between phonemes, most of our effort will address kernel design for classification in the acoustic waveform domain. The approach will be to a large extent inspired by the principles used in cepstral feature extraction, considering the effectiveness of cepstral representations for recognition in low noise. The expected benefit of applying SVMs to acoustic waveforms directly is that owing to the absence of nonlinear dimension reduction, additive noise and acoustic waveforms are combined linearly. This leaves the nonlinear boundaries established on clean data less altered, and also makes noise compensation fairly straightforward.

For multiclass discrimination, binary SVM classifiers are combined via error-correcting output code methods [45, 46]. To summarize the procedure briefly,  $N$  binary classifiers are trained to distinguish between  $M$  classes using a coding matrix  $\mathbf{W}_{M \times N}$ , with elements  $w_{mn} \in \{0, 1, -1\}$ . Classifier  $n$  is trained only on data of classes  $m$  for which  $w_{mn} \neq 0$ , with  $\text{sgn}(w_{mn})$  as the class label. The class  $m$  that one predicts for test input  $\mathbf{x}$  is then the one that minimizes the loss  $\sum_{n=1}^N \chi(w_{mn} h_n(\mathbf{x}))$ , where  $h_n(\mathbf{x})$  is the output of the  $n^{\text{th}}$  classifier and  $\chi$  is some loss function.

The error-correcting capability of a code is commensurate with the minimum Hamming distance between the rows of a coding matrix [45]. However, one must also take into account the effect of the coding matrix on the accuracy of the resulting binary classifiers, and the computational costs associated with a given code. In previous work [47, 48], codes formed by the combination of the *one-vs-one* (pairwise) and *one-vs-all* codes achieved good classification performance. But since the construction of one-vs-all binary classifiers for a problem with large datasets is not computationally feasible, only one-vs-one ( $N = M(M - 1)/2$ ) classifiers are used in the present study. A number of loss functions were compared, including hinge:  $\chi(z) = \max(1 - z, 0)$ , Hamming:  $\chi(z) = [1 - \text{sgn}(z)]/2$ , exponential:  $\chi(z) = e^{-z}$ , and linear:  $\chi(z) = -z$ . The hinge loss function performed best and is therefore used throughout this paper. We also experimented with adaptive continuous codes for multiclass SVMs as developed by Crammer *et al.* [49]. We do not report the details here: although this approach resulted in slight reductions in classification error on the order of 1–2%, it did not change the relative performance of the various classification approaches discussed below.

## B. Kernels for Cepstral Representations

1) *Kernel Design:* The time evolution of energy in a phoneme strongly correlates with phoneme identity and should therefore be a useful cue for accurate phoneme classification. It is in principle encoded in the cepstral features, which are a linear transform of Mel-log powers, but difficult to retrieve from there in noise because of the residual noise contamination in the compensated cepstral features [13]. To improve robustness, we propose to embed the exact information about the short-term energies of the acoustic waveform segments, treating them as a separate set of features in the evaluation of the SVM kernel. A straightforward compensation of these features can then be performed as explained below, and we have previously shown that this works well in the sense that the compensated features have distributions close to those of features derived from clean speech [50]. To define the energy features, the fixed length acoustic waveform segment  $\mathbf{x} \in \mathbb{R}^D$  is divided into  $T$  non-overlapping subsegments,

$$\mathbf{x}_t \in \mathbb{R}^{D/T}, \quad t = 1, \dots, T, \quad (2)$$

such that the centres of frame  $t$  (as used for the calculation of MFCC features) and subsegment  $\mathbf{x}_t$  are aligned. Let  $\boldsymbol{\tau} = [\tau_1, \dots, \tau_T]$ , where  $\tau_t = \log \|\mathbf{x}_t\|^2$ , denote the local energy features of these subsegments<sup>1</sup>. Then, the cepstral feature vector  $\mathbf{c}$  is augmented with the local energy feature vector  $\boldsymbol{\tau}$  for the evaluation of a hybrid kernel given by

$$K_c(\mathbf{c}, \tilde{\mathbf{c}}, \boldsymbol{\tau}, \tilde{\boldsymbol{\tau}}) = K_p(\mathbf{c}, \tilde{\mathbf{c}}) \sum_{t=1}^T K_\varepsilon(\tau_t, \tilde{\tau}_t), \quad (3)$$

where  $K_p$  is as given in (1),  $K_\varepsilon(\tau_t, \tilde{\tau}_t) = e^{-(\tau_t - \tilde{\tau}_t)^2 / 2a^2}$ , and  $a$  is a parameter that is tuned experimentally. The vector  $\boldsymbol{\tau}$  is treated as a separate set of features in the hybrid SVM kernel

<sup>1</sup>We consider logarithms to base 10 throughout.

$K_c$  rather than fused with the cepstral feature vector  $\mathbf{c}$  on a frame-by-frame basis. We sum the exponential terms in (3) over  $T$  segments rather than use the standard polynomial or RBF kernels in order to avoid the local energy features of certain subsegments dominating the evaluation of the kernel. (Alternatively, the local energy features can be standardized using CMVN and then evaluated using an RBF or polynomial kernel; this yields similar classification performance.) Finally, local energy features are calculated using non-overlapping segments of speech in order to avoid any smoothing of the time-profiles.

2) *Noise Compensation*: To investigate the robustness of the hybrid features to additive noise, we train the classifiers in quiet conditions with cepstral feature vectors standardized using CMVN [13]. Applying CMVN also to the noisy test data provides some basic noise compensation. We standardize so that the variance of each feature is the inverse of the dimension of the cepstral feature vector. On average both training and test cepstral feature vectors then have unit norm. More sophisticated noise compensation methods, namely ETSI AFE and VTS, both followed by feature standardization using CMVN, are also compared below. We do not consider here multi-condition/multi-style classification methods [6] because a previous study [37] showed that they generally perform worse than AFE and VTS, due to high sensitivity to any mismatch between the type of noise contaminating the training and test data.

In using the hybrid kernel  $K_c$ , the local energy features  $\tau$  must also be compensated for noise in order for the classifiers to perform effectively. Given an actual or estimated SNR, this is done as follows. Let  $\mathbf{x} = \mathbf{s} + \mathbf{n}$ ,  $\mathbf{x} \in \mathbb{R}^D$  be a noise corrupted waveform, where  $\mathbf{s}$  and  $\mathbf{n}$  represent the clean speech and the noise vector, respectively. The energy of the clean speech can then be approximated as  $\|\mathbf{s}\|^2 \approx \|\mathbf{x}\|^2 - \|\mathbf{n}\|^2 \approx \|\mathbf{x}\|^2 - D\sigma^2$ . Two approximations are involved here. Firstly, because speech and noise are uncorrelated, the vectors  $\mathbf{s}$  and  $\mathbf{n}$  are typically orthogonal:  $\langle \mathbf{s}, \mathbf{n} \rangle$  is of order  $D^{-1/2}\|\mathbf{s}\|\|\mathbf{n}\|$  which can be neglected for large enough  $D$ . Secondly, we replace the noise energy by its average value  $\sigma^2$ , the noise variance per sample. We work throughout with a default normalization of clean waveforms to unit energy per sample, so that  $1/\sigma^2$  is the SNR. Applying these general arguments to the local energy features, we compensate these by subtracting the estimated noise variance of a subsegment,  $D\sigma^2/T$  from the energies of the noisy subsegments, *i.e.* we compute  $\tau_t = \log \|\mathbf{x}_t\|^2 - D\sigma^2/T$ . This provides an estimate of the local energies of the subsegments of clean speech. Following the reasoning above, using local energy features of shorter subsegments of acoustic waveform (lower  $D/T$ ) would make fluctuations away from the orthogonality of speech and noise more likely, therefore  $K_\varepsilon$  should be evaluated on the energies of long enough subsegments of speech. Note that the noise compensation discussed here is performed only on the test features because training of classifiers that use hybrid features is always performed in quiet conditions; compensation of the local energy features of the training data is therefore not required.

### C. Kernels for Acoustic Waveforms

The use of kernels that express prior knowledge about the physical properties of the data can also improve phoneme classification in the acoustic waveform domain significantly. We propose several modifications of baseline SVM kernels to take into account relevant physical properties of speech and speech perception.

1) *Kernel Design*: (a) *Sign invariance*. To account for the fact that a speech waveform and its inverted version are perceived as being the same, for any two waveforms  $\mathbf{x}$ ,  $\tilde{\mathbf{x}} \in \mathbb{R}^D$ , an *even* kernel can be defined from a baseline polynomial kernel  $K_p$  (or indeed any kernel) as

$$K_e(\mathbf{x}, \tilde{\mathbf{x}}) = K_p'(\mathbf{x}, \tilde{\mathbf{x}}) + K_p'(\mathbf{x}, -\tilde{\mathbf{x}}) + K_p'(-\mathbf{x}, \tilde{\mathbf{x}}) + K_p'(-\mathbf{x}, -\tilde{\mathbf{x}}) \quad (4)$$

where  $K_p'(\mathbf{x}, \tilde{\mathbf{x}})$  is a modified polynomial kernel given by  $K_p'(\mathbf{x}, \tilde{\mathbf{x}}) = K_p(\mathbf{x}/\|\mathbf{x}\|, \tilde{\mathbf{x}}/\|\tilde{\mathbf{x}}\|)$ . This kernel  $K_p'$ , which always normalizes its input vectors to unit length, will be used as a baseline kernel for the acoustic waveforms. On the other hand, the standard polynomial kernel  $K_p$  defined in (1) will be employed for the cepstral representations, where CMVN already ensures that feature vectors typically have unit norm.

(b) *Shift invariance*. A further invariance of acoustic waveforms, to time alignment, is incorporated by using a kernel of the form (with normalization constant  $c = 1/(2L + 1)^2$ )

$$K_s(\mathbf{x}, \tilde{\mathbf{x}}) = c \sum_{u,v=-L}^L K_e(\mathbf{x}^{u\delta}, \tilde{\mathbf{x}}^{v\delta}), \quad (5)$$

where  $\mathbf{x}^{u\delta}$  is a segment of the same length as the original waveform  $\mathbf{x}$  but extracted from a position shifted by  $u\delta$  samples,  $\delta$  is the shift increment, and  $[-L\delta, L\delta]$  is the shift range.

(c) *Phoneme energy*. As the energy of a phoneme correlates with phoneme identity, we embed this information into the kernel as

$$K_l(\mathbf{x}, \tilde{\mathbf{x}}) = c \sum_{u,v} K_\varepsilon(\log \|\mathbf{x}^{u\delta}\|^2, \log \|\tilde{\mathbf{x}}^{v\delta}\|^2) K_e(\mathbf{x}^{u\delta}, \tilde{\mathbf{x}}^{v\delta}),$$

where  $K_\varepsilon$  is as defined after (3).

(d) *Fine scale dynamics*. Further, the *dynamics* of speech over a finer timescale is captured by evaluating the kernel over  $T$  subsegments as

$$K_d(\mathbf{x}, \tilde{\mathbf{x}}) = c \sum_{u,v} \sum_{t=1}^T K_\varepsilon(\log \|\mathbf{x}_t^{u\delta}\|^2, \log \|\tilde{\mathbf{x}}_t^{v\delta}\|^2) K_e(\mathbf{x}_t^{u\delta}, \tilde{\mathbf{x}}_t^{v\delta}),$$

where  $\mathbf{x}_t$  and  $\tilde{\mathbf{x}}_t$  are the  $t^{\text{th}}$  subsegments of the waveforms  $\mathbf{x}$  and  $\tilde{\mathbf{x}}$ , respectively, and  $\mathbf{x}_t^{u\delta}$  is a subsegment of the same length as  $\mathbf{x}_t$  but extracted from a position shifted by  $u\delta$  samples. This kernel captures the information about the phoneme energy at a finer resolution, which can help to distinguish phoneme classes with different temporal dynamics and energy profiles.

(e) *Subband dynamics*. Features that capture the evolution of energy and the dynamics of speech in *frequency subbands* are also relevant for phoneme classification. To obtain these subband features, we divide the speech waveform into *frames* similar to those used to calculate MFCCs. The frames are centred at the same locations as the non-overlapping subsegments

in (2). But we choose the frames to have a larger length of  $R > D/T$  samples, to allow more accurate noise compensation and good frequency localization. To construct the desired subband energy features, let  $X^f[r]$ ,  $f = 1, \dots, F$ ,  $r = 1, \dots, R$  be the discrete cosine transform (DCT) of the  $f^{\text{th}}$  frame of phoneme  $\mathbf{x}$ . The DCT coefficients are grouped into  $B$  bands, each containing  $R/B$  coefficients, and the log-energy  $\omega_f^b$  of band  $b$  is computed as

$$\omega_f^b = \log \left( \sum_{r=1}^{R/B} X^f[(b-1)R/B + r]^2 \right), \quad (6)$$

$$f = 1, \dots, F, b = 1, \dots, B.$$

These subband features are then concatenated into a vector  $\boldsymbol{\omega} = [\omega_1^1, \dots, \omega_1^B, \dots, \omega_F^1, \dots, \omega_F^B]^T$  and its time derivatives [51, 52] are evaluated to form the dynamic subband features

$$\boldsymbol{\Omega} = [\boldsymbol{\omega}, \Delta\boldsymbol{\omega}, \Delta^2\boldsymbol{\omega}]^T. \quad (7)$$

The subband energy features  $\boldsymbol{\Omega}$  are then combined with the acoustic waveforms  $\mathbf{x}$  using a kernel  $K_{\boldsymbol{\Omega}}$  which is given by  $K_{\boldsymbol{\Omega}}(\mathbf{x}, \tilde{\mathbf{x}}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}}) = K_d(\mathbf{x}, \tilde{\mathbf{x}})K_p(\boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}})$ , where  $\tilde{\boldsymbol{\Omega}}$  is the subband feature vector corresponding to the waveform  $\tilde{\mathbf{x}}$ .

2) *Noise Compensation*: In order to make the waveform-based SVM classifiers effective in the presence of noise, feature compensation is again essential. In the presence of colored noise, the level of contamination in frequency components differs according to the noise strength at those frequencies and thus requires compensation based on the spectral shape of the noise. To compensate the features, let  $X[r]$ ,  $r = 1, \dots, D$  be the DCT of the noisy test waveform,  $\mathbf{x} \in \mathbb{R}^D$ . For the purposes of noise compensation, we consider the DCT of the whole phoneme segment rather than individual subsegments or frames. Given the estimated noise variance of the  $r^{\text{th}}$  frequency component  $\sigma_r^2$  (note that  $\sum_{r=1}^D \sigma_r^2 = \sigma^2$ ), the frequency component  $X[r]$  is scaled by  $1/\sqrt{1 + D\sigma_r^2}$  in order to depress the effect of DCT components with high noise variance on the kernel evaluation, giving  $\hat{X}[r] = X[r]/\sqrt{1 + D\sigma_r^2}$ . Note that we do not make specific assumptions on the spectrum of clean speech here, and simply take each DCT component to have the same average clean energy  $1/D$ . The spectrally adapted waveform  $\hat{\mathbf{x}}$ , obtained by inverse DCT of the scaled DCT coefficients  $\hat{X}[r]$ , is then used for the evaluation of time correlations (dot products) in the kernel  $K_p'$ ; note that the local and subband energy features are extracted from the unadapted waveform  $\mathbf{x}$  to obtain their exact values. Below we drop the hat subscript on  $\hat{\mathbf{x}}$ , to lighten the notation.

Let us consider now the overall normalization of acoustic waveforms. With clean speech, the kernel  $K_p'$  used in (4) effectively normalizes all waveforms to unit norm. However, the scaling of the norm of the acoustic waveforms in the presence of noise should be different, to keep the norm of the underlying clean speech signal approximately constant across different noise levels. Consider the inner product of the noise corrupted waveform  $\mathbf{x} = \mathbf{s} + \mathbf{n}$  (where  $\mathbf{s}$  is clean speech and  $\mathbf{n}$  is noise as before) with a waveform  $\mathbf{x}_i$  from the training set. Considering again that noise and clean speech are roughly uncorrelated, the contribution to this product from  $\mathbf{n}$  can be neglected, so that  $\langle \mathbf{x}, \mathbf{x}_i \rangle \approx \langle \mathbf{s}, \mathbf{x}_i \rangle$  where  $\mathbf{x}_i$  is a training point.

The clean speech signal appearing here should approximately have unit energy per sample,  $\|\mathbf{s}\|^2 \approx D$ . Arguing as in section II-B, we therefore need  $\|\mathbf{x}\|^2 \approx \|\mathbf{s}\|^2 + D\sigma^2 \approx D(1 + \sigma^2)$ . Thus, noisy waveforms have to be normalized to have a larger norm than clean waveforms, by a factor  $\sqrt{1 + \sigma^2}$ . So while the training waveform is normalized to unit norm in (4), the spectrally adapted test waveform  $\mathbf{x}$  is normalized to  $\sqrt{1 + \sigma^2}$  for the evaluation of the baseline polynomial kernel  $K_p'$ . To emphasize this, we write from now on generic kernels evaluated between a test waveform  $\mathbf{x}$  and a training waveform  $\mathbf{x}_i$  as  $K(\mathbf{x}, \mathbf{x}_i)$  rather than  $K(\mathbf{x}, \tilde{\mathbf{x}})$ .

A similar compensation as in the polynomial kernel can be used for  $K_{\epsilon}$  by subtracting the estimated subsegment noise variance,  $D\sigma^2/T$ , from the energy of each noisy subsegment  $\mathbf{x}_t$  to approximate the energies of the clean subsegments (see section II-B). The noise compensated kernel  $K_d$  is then

$$K_d(\mathbf{x}, \mathbf{x}_i) = c \sum_{u,v} \sum_{t=1}^T K_{\epsilon} \left( \log \left| \|\mathbf{x}_t^{u\delta}\|^2 - \frac{D\sigma^2}{T} \right|, \log \|\mathbf{x}_{i,t}^{v\delta}\|^2 \right) \times K_{\epsilon}(\mathbf{x}_t^{u\delta}, \mathbf{x}_{i,t}^{v\delta}). \quad (8)$$

As training for acoustic waveforms is performed in quiet conditions, local energy features of the training waveform  $\mathbf{x}_i$  are again not compensated. The spectral shape adaptation of the test waveform segment  $\mathbf{x} \in \mathbb{R}^D$  as discussed above is performed before the evaluation of  $K_{\epsilon}$  on subsegments in (8).

There are two potential drawbacks in using (8) in the presence of noise. Firstly, we need to normalize the clean subsegments to unit norm and the noisy ones to  $\sqrt{1 + \sigma^2}$ . However, for short subsegments there can be wide variation in local SNR in spite of the fixed global SNR, and so this normalization may not be in accordance with the local SNR. Secondly, using short (low dimensional) subsegments makes fluctuations away from the average orthogonality of speech and noise more pronounced. To avoid these problems, we also consider a modified kernel  $K_d'$  where we use  $K_{\epsilon}(\mathbf{x}^{u\delta}, \mathbf{x}_i^{v\delta})$  instead of  $K_{\epsilon}(\mathbf{x}_t^{u\delta}, \mathbf{x}_{i,t}^{v\delta})$ . This leaves the time-correlation part of the kernel unsegmented, while  $K_{\epsilon}$  is still evaluated over  $T$  subsegments of the phonemes. We will see that  $K_d$  gives significantly better performance than  $K_d'$  in less noisy conditions because of its sensitivity to the correlation of the individual subsegments. On the other hand,  $K_d$  performs worse than  $K_d'$  at high noise due to the two limitations discussed above.

Finally, if we want to use energies in the frequency subbands from (6) as features, as in  $K_{\boldsymbol{\Omega}}$ , then also these need to be compensated for noise. This is done in a manner similar to the local energy features in  $K_{\epsilon}$  as

$$\omega_f^b = \log \left| \sum_{r=1}^{R/B} X^f[(b-1)R/B + r]^2 - R \sum_{r=1}^{R/B} \sigma_r^2 \right|,$$

$$f = 1, \dots, F, b = 1, \dots, B,$$

prior to evaluating time derivatives to form the dynamic subband features  $\boldsymbol{\Omega}$ . The kernel  $K_{\boldsymbol{\Omega}}$  is then given by  $K_{\boldsymbol{\Omega}}(\mathbf{x}, \mathbf{x}_i, \boldsymbol{\Omega}, \boldsymbol{\Omega}_i) = K_d(\mathbf{x}, \mathbf{x}_i)K_p(\boldsymbol{\Omega}, \boldsymbol{\Omega}_i)$ , where  $\boldsymbol{\Omega}_i$  is the (uncompensated) subband feature vector of the training waveform  $\mathbf{x}_i$ . A modified kernel  $K_{\boldsymbol{\Omega}}'$  can be defined similarly, by replacing  $K_d$  by  $K_d'$ .

Methods for additive noise compensation of the cepstral and acoustic waveform features discussed in this section require an estimate of the noise variance ( $\sigma^2$ ) or the signal-to-noise ratio (SNR) of the noisy speech, a problem for which a number of approaches have been proposed [53–55]. The lowest classification error would be obtained in our approach from exact knowledge of the *local* SNR at the phoneme level. In most experiments, we assume that only the true *global* (per sentence) SNR is known, and approximate the local SNR by this global one. The intrinsic variability of speech energy across different phonemes means that this approximation will often be rather poor, and will not saturate the lower classification error bound that would result for known local SNR. In Section III-D, we then compare with classification results obtained by integrating the decision-direction SNR estimation algorithm [55, 56] into our proposed approach for compensation and normalization of acoustic waveform features. Because the SNR estimation is done frame by frame, this approach can track variations in local SNR, which should improve performance. On the other hand, the SNR estimates will deviate from the truth, also at the global level, and this will increase error rates. Our results show that these two effects essentially cancel in the resulting error rates.

### III. EXPERIMENTAL RESULTS

#### A. Experimental Setup

Experiments are performed on the ‘si’ (diverse) and ‘sx’ (compact) sentences of TIMIT database [57]. The training set consists of 3696 sentences from 462 different speakers. For testing we use the core test set which consists of 192 sentences from 24 different speakers not included in the training set. The development set consists of 1152 sentences uttered by 144 speakers not included in either the training or the core test set. The glottal stops /q/ are removed from the class labels and certain allophones are grouped into their corresponding phoneme classes using the standard Kai-Fu Lee clustering [58], resulting in a total of  $M = 48$  phoneme classes and  $N = M(M - 1)/2 = 1128$  binary classifiers. Among these classes, there are 7 groups for which the contribution of within-group confusions toward multiclass error is not counted, again following standard practice [32, 58].

Both artificial noise (white, pink) and recordings of real noise (speech-babble) from the NOISEX-92 database are used in our experiments. White noise was selected due to its attractive theoretical interpretation as probing in an isotropic manner the separation of phoneme classes in different representation domains. Pink noise was chosen because  $1/f$ -like noise patterns are found in music melodies, fan and cockpit noises, in nature etc. [59]. To test the classification performance of the cepstral features and acoustic waveforms in noise, each sentence is normalized to unit energy per sample and then a noise sequence with variance  $\sigma^2$  (per sample) is added to the entire sentence. The SNR at the level of individual phonemes can then still vary widely. For cepstral features, two training-test scenarios are considered: (i) training SVM classifiers using clean data, with standard noise compensation methods to clean the test features and (ii) training and testing

under identical noise conditions. The latter is an impractical target; nevertheless, we present the results as a reference for cepstral features, since this setup is considered to give the optimal achievable performance [14]. The cepstral features of both training and test data are standardized using CMVN in all scenarios. For classification using acoustic waveforms training is always performed with noiseless (clean) data, and then noisy test features are compensated as described in the previous section.

For the cepstral (MFCC) representation,  $\mathbf{c}$ , each sentence is converted into a sequence of 13 dimensional feature vectors, their time derivatives and second order derivatives. Then,  $F = 10$  frames (with frame duration of 25ms and a frame rate of 100 frames/sec) closest to the centre of a phoneme are concatenated to give a representation in  $\mathbb{R}^{390}$ . Along the same lines, each frame yields 14 AFE features (including log frame energy) and their time derivatives as defined by the ETSI standard, giving a representation in  $\mathbb{R}^{420}$ . For noise compensation with vector Taylor series (VTS) [7–9], several Gaussian mixture models (GMMs) were trained to learn the distribution of Mel log spectra of clean training data; we found all of them to yield very similar performance. For the results reported below, a GMM with 64 mixture components was used.

For the acoustic waveform representation, phoneme segments  $\mathbf{x}$  are extracted from the TIMIT sentences by applying a 100ms rectangular window at the centre of each phoneme, which at 16kHz sampling frequency gives fixed length vectors in  $\mathbb{R}^D$  with  $D = 1600$ . In the evaluation of the shift-invariant kernel  $K_s$  from (5), we use a shift increment of  $\delta = 100$  samples ( $\approx 6$  ms) over a shift range  $\pm 100$  (so that  $L = 1$ ), giving three shifted segments. Each of these segments is broken into  $T = 10$  subsegments of equal length for the evaluation of kernels  $K_d$  and  $K'_d$ .

For the subband features,  $\Omega$  (see (7)), the energy features and their time derivatives in  $B = 8$  frequency subbands of equal bandwidth are combined to form a 24-dimensional feature vector for each frame of speech. These subband features are standardized to zero mean and unit variance within each sentence of TIMIT. Then, the standardized subband features of  $F = 10$  frames with frame duration of 25ms ( $R = 400$  samples per frame) and a frame rate of 100 frames/sec, again closest to the centre of a particular phoneme, are concatenated to give a representation  $\Omega$  in  $\mathbb{R}^{240}$ . We did not use a larger number of subbands to avoid an excessive number of subband features, and also to keep enough frequencies,  $R/B$ , per subband to allow accurate noise compensation of the  $\omega_f^b$ .

The effect of custom-designed kernels on performance is investigated by comparing the different kernel functions defined above. The best classification performance with acoustic waveforms is achieved with  $K'_\Omega$ . For the cepstral representations, we compare the performance of the baseline kernel  $K_p$  with that of the hybrid kernel  $K_c$ . Initially, we experimented with different values of the hyperparameters for the binary SVM classifiers but decided to use fixed values for all classifiers as parameter optimization had a large computational overhead but only a small impact on the multiclass classification error. The degree of  $K_p$  is set to  $\Theta = 6$ , the penalty parameter (for slack

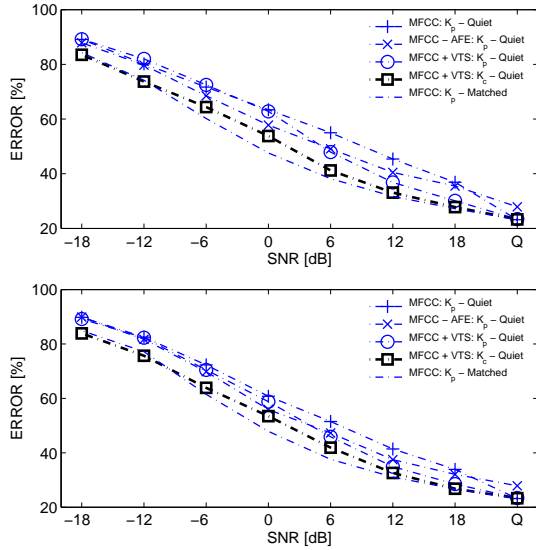


Fig. 1. Classification error versus SNR for SVM phoneme classification in the presence of (top) white noise, (bottom) pink noise, using the MFCC representation with standard kernel  $K_p$  and hybrid kernel  $K_c$ , for different training and test conditions and feature compensation methods.

variables in the SVM training algorithm) to  $C = 1$  and the value of  $a$  in  $K_e$  is tuned experimentally on the development data to give  $a = 0.5$ .

### B. Classification based on the Cepstral Representation

In Figure 1, the results of SVM phoneme classification with the polynomial kernel  $K_p$  in the presence of additive white and pink noise are shown for the standard MFCC cepstral representation, as well as for MFCC features compensated using VTS and AFE. For comparison, results for matched training and test conditions are presented as well. The plots demonstrate that the SVM classifier trained with the AFE representation outperforms the standard MFCC representation for SNRs below 18dB, but is worse in quiet conditions. The VTS-compensated MFCC features, on the other hand, perform comparably to standard MFCC in quiet, and thus in fact better than the more sophisticated AFE features. However, for SNR below 0dB, the classification performance of VTS-compensated MFCC features degrades relatively quickly as compared to the AFE features. Since the (log) frame energy is included in the AFE features as defined by the ETSI standard, we consider as a hybrid representation only the one formed by the combination of the local energy features and the VTS-compensated MFCC features, using kernel  $K_c$ . The results show that this hybrid representation performs better than both noise compensation methods (AFE and VTS) at all noise conditions and approaches the performance achieved under matched conditions. For instance, the hybrid representation achieves an average improvement of 5.5% and 5.8% over the standard VTS-compensated MFCC features and AFE features respectively, across all SNRs in the presence of white noise as shown in Figure 1(top), with similar conclusions in pink noise.

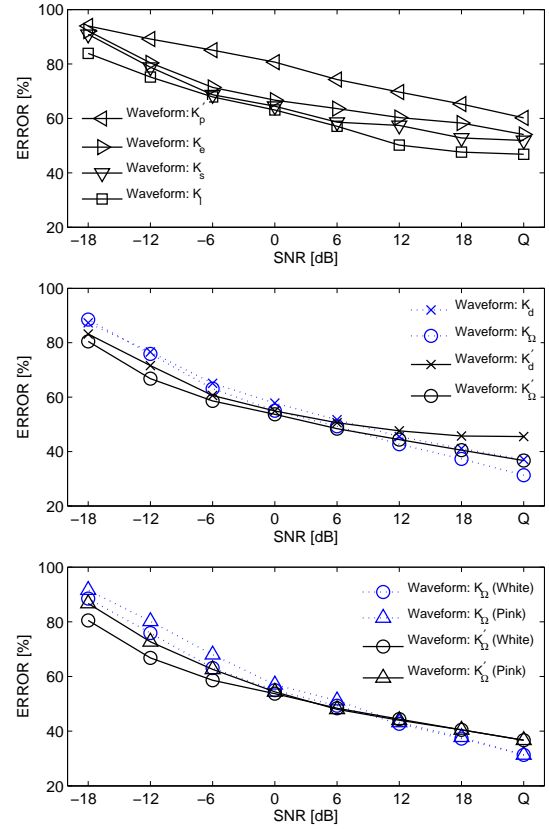


Fig. 2. Effects of custom-designed kernels on the classification performance of SVMs using acoustic waveform representations. (top) Results for classification with kernels  $K_p$ ,  $K_e$ ,  $K_s$ ,  $K_l$  in the presence of white noise. (middle) Classification with the more advanced kernels  $K_d$ ,  $K_\Omega$  and their unsegmented analogs  $K'_d$ ,  $K'_\Omega$  in white noise. (bottom) Comparison of classification with kernels  $K_\Omega$  and  $K'_\Omega$  in the presence of white and pink noise.

### C. Classification based on Acoustic Waveforms

Let us now consider classification using acoustic waveforms. First, Figure 2 illustrates the effects of our custom-designed kernels for acoustic waveforms on the classification performance in the presence of additive (white and pink) noise. As we had hoped, embedding more physical aspects of speech and speech perception into the SVM kernels does indeed reduce classification error. Classification results using acoustic waveforms with the standard SVM kernel  $K_p$  are shown in Figure 2(top); the resulting performance is clearly worse than MFCC classifiers (see Figure 1) for all noise levels. The even polynomial kernel  $K_e$  (see (4)), which is a sign-invariant kernel based on  $K_p$ , gives an 8% average improvement in classification performance. The largest improvement, 14%, is achieved at 0dB SNR in white noise. Adding shift-invariance and the noise-compensated local energy features to the kernel improves results further. The resulting kernels  $K_s$  and  $K_l$  reduce the classification error by approximately 3% and 4% respectively, on average across all noise levels. Overall, a reduction in classification error of approximately 18% is obtained by using kernel  $K_l$  over our baseline  $K_p$  kernel at 0dB SNR in white noise.

Next, the temporal dynamics of speech and information from frequency subbands are incorporated via the kernels  $K_d$

and  $K_\Omega$ . The results for these kernels are shown in Figure 2(middle). One can observe that these kernels give major improvements in low noise conditions because of their sensitivity to correlation of individual subsegments of phonemes, *e.g.*  $K_\Omega$  achieves 31.3% error in quiet condition, an improvement in classification performance of 15% over  $K_l$ . However,  $K_\Omega$  performs worse than  $K_l$  below a crossover point between  $-6$ dB and  $-12$ dB SNR, as anticipated in the discussion in Section II-C. In a comparison of  $K_\Omega$  with  $K'_\Omega$ , where the time-correlation part of the kernel is left unsegmented, we see that  $K_\Omega$  performs better than  $K'_\Omega$  in low noise conditions but the latter gives better results in high noise. Overall, incorporating invariances and additional information about the acoustic waveforms into the kernel results in major cumulative performance gains. For instance, in quiet conditions an absolute 30% reduction in error is achieved by  $K_\Omega$  over the standard polynomial kernel  $K'_p$ .

Figure 2(bottom) summarizes the classification results for the kernels that give the best classification performance with acoustic waveforms in white noise,  $K_\Omega$  and  $K'_\Omega$ , and compares with results for pink noise. It is clear that the noise type affects classification performance minimally in low noise conditions,  $\text{SNR} \geq 0$ dB. At extremely high noise,  $\text{SNR} \leq -6$ dB, pink noise has more severe effects than white noise. We also tested other noise types, *e.g.* speech-weighted noise (results not shown) and qualitatively similar conclusions apply.

#### D. Classifier Combination

In previous work [38, 50], we introduced an approach that combined the cepstral and acoustic waveform classifiers to attain better classification performance than either of the individual representations. Since waveform classifiers with kernel  $K'_\Omega$  achieve best results in high noise (see Figure 1 and Figure 2), we consider them in combination with the SVM classifiers trained on hybrid VTS-compensated MFCC features using kernel  $K_c$ . In particular, a convex combination of the scores of the classifiers in the individual feature spaces is considered, *i.e.* for binary classifiers  $h_w$  and  $h_m$  in the waveform and MFCC domains respectively, we define the combined classifier output as  $h_c = \lambda h_w + (1 - \lambda) h_m$ . Here  $\lambda = \lambda(\sigma^2)$  is a parameter which needs to be selected, depending on the noise variance, to achieve optimal performance. These combined binary classifiers are then in turn combined for multiclass classification as detailed in Section II-A.

Figure 3(top) shows the classification error on the core test set of TIMIT at various SNRs as a function of the combination parameter  $\lambda$ ; classification in the MFCC cepstral domain is obtained with  $\lambda = 0$ , whereas  $\lambda = 1$  corresponds to classification in the acoustic waveform domain. One can observe that the minimum error is achieved for  $0 < \lambda < 1$  for almost all noise conditions. To retain unbiased test errors on the core test set, we determine from the distinct *development* set the ‘‘optimal’’ values of  $\lambda(\sigma^2)$ ,  $\lambda_{\text{opt}}(\sigma^2)$ , *i.e.* the values of  $\lambda$  which give the minimum classification error for a given SNR. These are marked by ‘o’ in Figure 3(bottom); the error bars give the range of values of  $\lambda$  for which the classification error on the development set is less than the minimum error plus

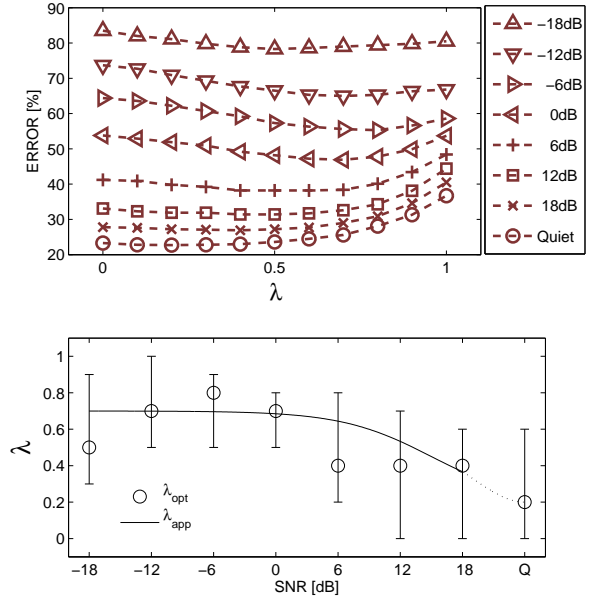


Fig. 3. (top) Classification error on the core test set over a range of SNRs in the presence of white noise as a function of  $\lambda$ ;  $\lambda = 0$  corresponds to classification with hybrid VTS-compensated MFCC features using kernel  $K_c$ ,  $\lambda = 1$  is waveform classification with kernel  $K'_\Omega$ ; (bottom) Optimal and approximate values of  $\lambda$  for a range of SNRs (in white noise). Error bars give the range of values of  $\lambda(\sigma^2)$  for which the classification error on the development set is less than the minimum error(%) + 2%.

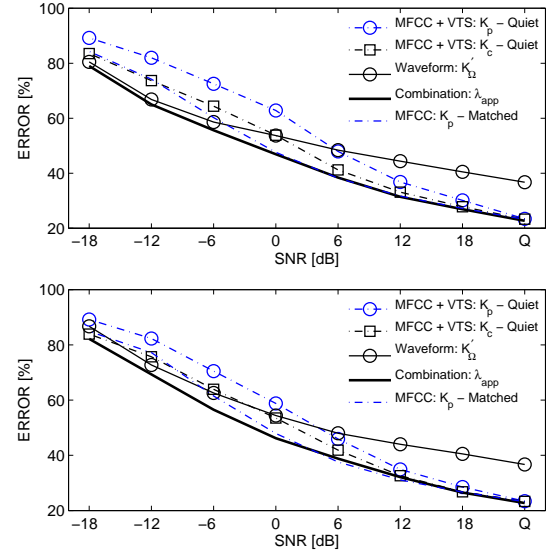


Fig. 4. Comparison of classification in the MFCC (with kernels  $K_p$  and  $K_c$ ) and acoustic waveform (with kernel  $K'_\Omega$ ) domains with the combined classifier, for  $\lambda_{\text{app}}(\sigma^2)$  given by (9) and (top) white and (bottom) pink noise. The combined classifier outperforms the MFCC classifier even under matched training and test conditions and in fact is more robust than individual MFCC and acoustic waveform classifiers.

2%. A reasonable approximation to  $\lambda_{\text{opt}}(\sigma^2)$  is given by

$$\lambda_{\text{app}}(\sigma^2) = \eta + \zeta / [1 + (\sigma_0^2 / \sigma^2)] , \quad (9)$$

with  $\eta = 0.2$ ,  $\zeta = 0.5$  and  $\sigma_0^2 = 0.03$ , and is shown in Figure 3(bottom) by the solid line.

Having determined  $\lambda_{\text{app}}(\sigma^2)$  from the development set, we can now go back to the core test set and compare (Figure 4)

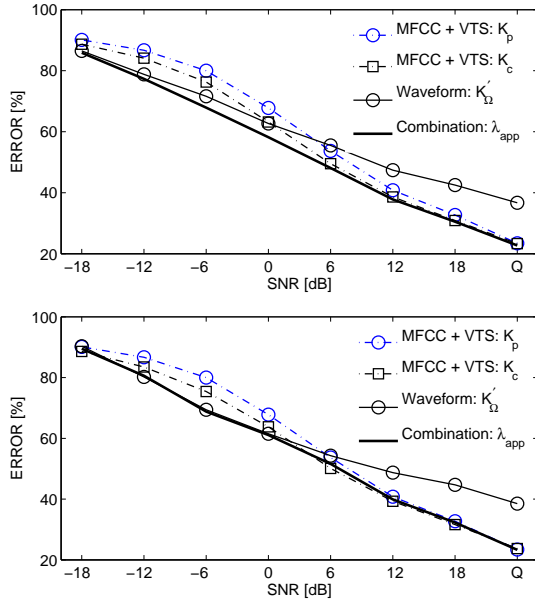


Fig. 5. Comparison of the classification error rates obtained with feature compensation and normalization using (top) the true global SNR and (bottom) the estimated local SNR [55, 56], in the presence of speech-babble noise from NOISEX-92. The difference in classification errors obtained is marginal.

the performance of the individual classifiers using hybrid VTS-compensated MFCC features (with kernel  $K_c$ ) and acoustic waveforms (with kernel  $K'_\Omega$ ) with their combination with  $\lambda = \lambda_{\text{app}}(\sigma^2)$ . One observes that the combined classifier always performs better or at least as well as the individual classifiers. Significantly, it even improves over the classifier trained with cepstral features in a matched environment; recall from Figure 1 that even the best cepstral classifiers that we found (with VTS noise compensation and hybrid features) never beat this matched scenario baseline. Similar results are obtained with  $\lambda = \lambda_{\text{opt}}(\sigma^2)$ : the wide error bars in Figure 3(bottom) show that the combined classifier is relatively insensitive to the precise value of  $\lambda$ . What is important is that  $\lambda$  stays away from the extreme values 0 and 1; for example  $0.2 < \lambda_{\text{app}}(\sigma^2) < 0.7$  so the combined classifier is not simply a hard switch between the two representations depending on the noise level. It should be noted that the gain in classification accuracy from the combination relative to standalone cepstral classifiers with kernels  $K_p$  or  $K_c$  is substantial. For instance, in white noise the combined classifier achieves an average of 12.3%, 10.1% and 5% reduction in error across all SNRs, when compared to classifiers with the VTS-compensated MFCC features, AFE features and hybrid VTS-compensated MFCC features respectively. Qualitatively similar behavior is observed in pink noise, as shown in Figure 4(bottom).

Up to this point in our experiments, the acoustic waveform features were normalized and compensated using the true global SNR as per our approximation in Section II-C2. As explained there, global and local SNR can differ significantly, so that we are not obtaining the theoretically achievable optimum performance that would result from known *local* SNRs. We now assess the effect of integrating an SNR estimation

TABLE I  
RESULTS FOR PHONEME CLASSIFICATION ON THE TIMIT CORE TEST SET IN QUIET CONDITION. FOR THE LAST LINE WE USED A VARIABLE LENGTH ENCODING AND CONTINUOUS ERROR-CORRECTING CODES.

METHOD	ERROR [%]
HMMs (MCE) [60]	31.4
GMMs [32]	26.3
HMMs (MMI) [35]	24.8
Multiresolution Subband HMMs (MCE) [34]	23.7
SVMs [32]	22.4
Large-Margin GMM (LMGMM) [29]	21.1
Hierarchical GMM [31]	21.0
RLS2 [30]	20.9
Hidden CRF [28]	20.8
Hierarchical LMGMM H(2,4) [27]	18.7
Committee Hierarchical LMGMM H(2,4) [27]	16.7
<b>SVMs - Hybrid Features (MFCC + VTS)</b>	<b>22.7</b>
<b>SVMs - Hybrid Features (PLP)</b>	<b>20.1</b>
<b>SVMs - PLP using [32, 49]</b>	<b>18.4</b>

algorithm into our approach. In particular, we use the decision-directed SNR estimation algorithm [55, 56] that provides a frame-by-frame estimate of SNR. We estimate from this the local phoneme SNR by averaging the SNR estimates for the  $T = 10$  frames closest to the phoneme centre. This value of the local SNR is then used throughout, *i.e.* for the normalization of test acoustic waveforms  $\mathbf{x}$ , the feature compensation of the local energy features  $\tau$  and subband features  $\Omega$  as described in Section II-C2 and for the evaluation of the combination parameter  $\lambda_{\text{app}}(\sigma^2)$ . For SNR estimation using this algorithm, it is assumed that the initial 100ms segment of each sentence contains no speech at all, to obtain an initial estimate of the noise statistics.

In Figure 5, we compare the classification performance achieved with noise compensation and feature normalization using the known global SNR and the estimated local SNR in the presence of speech-babble noise. The results show a marginal difference in the resulting classification errors, *e.g.* amounting to only 0.7% on average across all noise levels for the classification using acoustic waveforms. Quantitatively similar behavior is observed for the hybrid MFCC SVM classifier with kernel  $K_c$ . A slightly larger difference in the average classification error, *viz.* an increase of 2.3%, is observed for the combined classifier when the features are compensated and normalized according to the estimated local SNR. This is due to the use of the convex combination function  $\lambda_{\text{app}}(\sigma^2)$  which was determined using data normalized according to the true global SNR. Nonetheless, the combined classifier yields consistent improvements over the MFCC classifier with kernel  $K_p$  in both cases. This demonstrates that the combined classifier can easily tolerate the mismatch between the true global SNR and estimated local SNR; its performance remains superior to that of the classifiers trained with VTS-compensated MFCC features. Another set of experiments (results not reported here) showed that the combined classifier is also very robust to misestimation of the global SNR.

In Table I, results of some recent experiments on the TIMIT phoneme classification task in quiet condition are gathered (non-bold) and compared with the cepstral results reported in this paper (bold). We also show results obtained using SVM

classifiers trained with a hybrid cepstral representation with kernel  $K_c$  (Section II-B), but using PLP instead of MFCC features. This gives better performance in quiet conditions. Note that the benchmarks (non-bold entries in the table) use cepstral representations that encode information from the entire variable length phonemes and our result of 20.1% error improves on all benchmarks except [27] even though we use a fixed length cepstral representation. Further improvements can be expected by including all frames within a variable length phoneme and the transition regions [32] (see last entry in table), and by incorporating techniques such as committee classifiers [27, 61]. More importantly, our classifiers significantly outperform the benchmarks in the presence of noise. A classification error of 77.8% is reported by Rifkin *et al.* [30] at 0dB SNR in pink noise whereas our combined classifier achieves an error of 46.2% in the same conditions as shown in Figure 4(bottom).

One might be concerned about the computational complexity because for SVMs, training time scales approximately quadratically with the number of data points. However, this effect is independent of which front-end is used, and already a number of years ago it was possible to use SVMs for large vocabulary speech recognition tasks such as Switchboard recognition [25]. The only difference between front-ends arises in the time it takes to evaluate the required kernel elements  $K(\mathbf{x}_i, \mathbf{x}_j)$ . The evaluation of scalar products scales with the feature space dimension, leading to an increase in training time by a factor around five when using acoustic waveform rather than cepstral front-ends. The use of shift-invariant kernels leads to a similar factor, so that overall computation time is roughly an order of magnitude larger for waveforms than for cepstral features. This increase is modest and so our approach remains practical, particularly bearing in mind improvements in computing hardware since [25] was published: graphics processing units (GPUs) can provide approximate speedups of up to 100 times over standard SVM implementations such as LIBSVM [62].

#### IV. CONCLUSIONS

In this study, we proposed methods for combining cepstral and acoustic waveform representations to improve the robustness of phoneme classification with SVMs to additive noise. To this end, we developed kernels and showed that embedding invariances of speech and relevant dynamical information via custom-designed kernels can significantly improve classification performance. While the cepstral representation allows for very accurate classification of phonemes in low noise conditions, especially for clean data, its performance suffers degradation at high noise levels. The high-dimensional acoustic waveform representation, on the other hand, is less accurate on clean data but more robust in severe noise. We have also shown that a convex combination of the MFCC and acoustic waveform classifiers achieves performance that is consistently better than both classifiers in the individual domains across the entire range of noise levels.

The work reported in this paper could serve as a point of departure to address some key issues in the construction of

robust ASR systems. An important and necessary extension would be to investigate the robustness of the waveform-based representations to linear filtering. It would also be interesting to extend our work to handle continuous speech recognition tasks using SVM/HMM hybrids [25]. In future work, we would seek to further improve the results by incorporating techniques proposed by other authors, such as committee classifiers [27] that combine a number of representations with different parameters as well as hierarchical classification to reduce broad phoneme class confusions [61].

#### ACKNOWLEDGMENT

Zoran Cvetković would like to thank Jont Allen, Bishnu Atal, and Andreas Buja for encouragement and inspiration.

#### REFERENCES

- [1] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. ASSP*, vol. 28, pp. 357–366, 1980.
- [2] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, April 1990.
- [3] R. Lippmann, "Speech Recognition by Machines and Humans," *Speech Comm.*, vol. 22, no. 1, pp. 1–15, 1997.
- [4] J. Sroka and L. Braidia, "Human and Machine Consonant Recognition," *Speech Comm.*, vol. 45, no. 4, pp. 401–423, 2005.
- [5] M. Holmberg, D. Gelbart, and W. Hemmert, "Automatic Speech Recognition with an Adaptation Model Motivated by Auditory Processing," *IEEE Trans. ASLP*, vol. 14, no. 1, pp. 43–49, 2006.
- [6] R. Lippmann and E. A. Martin, "Multi-Style Training for Robust Isolated-Word Speech Recognition," *Proc. ICASSP*, pp. 705–708, 1987.
- [7] P. J. Moreno, B. Raj, and R. M. Stern, "A Vector Taylor Series Approach for Environment-Independent Speech Recognition," *Proc. ICASSP*, pp. 733–736, 1996.
- [8] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "High-Performance HMM Adaptation With Joint Compensation of Additive and Convolutional Distortions Via Vector Taylor Series," in *Automat. Speech Recogn. & Understanding*, pp. 65–70, 2007.
- [9] M. J. F. Gales and F. Flego, "Combining VTS Model Compensation and Support Vector Machines," *Proc. ICASSP*, pp. 3821–3824, 2009.
- [10] H. Liao, "Uncertainty Decoding For Noise Robust Speech Recognition," *Ph.D. Thesis, Cambridge University*, 2007.
- [11] ETSI standard doc., "Speech processing, Transmission and Quality aspects (STQ): Advanced front-end feature extraction," *ETSI ES 202 050*, 2002.
- [12] O. Viikki and K. Laurila, "Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition," *Speech Comm.*, vol. 25, pp. 133–147, 1998.
- [13] C. Chen and J. Bilmes, "MVA Processing of Speech Features," *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 257–270, 2007.
- [14] M. Gales and S. Young, "Robust Continuous Speech Recognition using Parallel Model Combination," *IEEE Trans. Speech Audio Process.*, vol. 4, pp. 352–359, Sept. 1996.
- [15] G. Miller and P. Nicely, "An Analysis of Perceptual Confusions among some English Consonants," *J. Acoust. Soc. Amer.*, vol. 27, no. 2, pp. 338–352, 1955.
- [16] J. B. Allen, "How do humans process and recognize speech?," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 567–577, 1994.
- [17] B. Meyer, M. Wächter, T. Brand, and B. Kollmeier, "Phoneme Confusions in Human and Automatic Speech Recognition," *Proc. INTERSPEECH*, pp. 2740–2743, 2007.

- [18] B.S. Atal, "Automatic Speech Recognition: a Communication Perspective," *Proc. ICASSP*, pp. 457–460, 1999.
- [19] S. D. Peters, P. Stubble, and J. Valin, "On the Limits of Speech Recognition in Noise," *Proc. ICASSP*, pp. 365–368, 1999.
- [20] Hervé Bourlard, Hynek Hermansky, and Nelson Morgan, "Towards Increasing Speech Recognition Error Rates," *Speech Comm.*, vol. 18, no. 3, pp. 205–231, 1996.
- [21] Kuldip K. Paliwal and Leigh D. Alsteris, "On the Usefulness of STFT Phase Spectrum in Human Listening Tests," *Speech Comm.*, vol. 45, no. 2, pp. 153–170, 2005.
- [22] Leigh D. Alsteris and Kuldip K. Paliwal, "Further Intelligibility Results from Human Listening Tests using the Short-Time Phase Spectrum," *Speech Comm.*, vol. 48, no. 6, pp. 727–736, 2006.
- [23] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.
- [24] A. Sloin and D. Burshtein, "Support Vector Machine Training for Improved Hidden Markov Modeling," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 172–188, 2008.
- [25] A. Ganapathiraju, J. E. Hamaker, and J. Picone, "Applications of Support Vector Machines to Speech Recognition," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2348–2355, 2004.
- [26] S. E. Krüger, M. Schaffner, M. Katz, E. Andelic, and A. Wendemuth, "Speech Recognition with Support Vector Machines in a Hybrid System," *Proc. INTERSPEECH*, pp. 993–996, 2005.
- [27] H. Chang and J. Glass, "Hierarchical Large-Margin Gaussian Mixture Models for Phonetic Classification," in *Automat. Speech Recogn. & Understanding*, pp. 272–275, 2007.
- [28] D. Yu, L. Deng, and A. Acero, "Hidden Conditional Random Fields with Distribution Constraints for Phone Classification," *Proc. INTERSPEECH*, pp. 676–679, 2009.
- [29] F. Sha and L. K. Saul, "Large Margin Gaussian Mixture Modeling for Phonetic Classification and Recognition," *Proc. ICASSP*, pp. 265–268, 2006.
- [30] R. Rifkin, K. Schutte, M. Saad, J. Bouvrie, and J. Glass, "Noise Robust Phonetic Classification with Linear Regularized Least Squares and Second-Order Features," *Proc. ICASSP*, pp. 881–884, 2007.
- [31] A. Halberstadt and J. Glass, "Heterogeneous Acoustic Measurements for Phonetic Classification," *Proc. EuroSpeech*, pp. 401–404, 1997.
- [32] P. Clarkson and P. J. Moreno, "On the Use of Support Vector Machines for Phonetic Classification," *Proc. ICASSP*, pp. 585–588, 1999.
- [33] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden Conditional Random Fields for Phone Classification," *Proc. INTERSPEECH*, pp. 1117–1120, 2005.
- [34] P. McCourt, N. Harte, and S. Vaseghi, "Discriminative Multi-resolution Sub-band and Segmental Phonetic Model Combination," *IET Electronics Letters*, vol. 36, no. 3, pp. 270–271, 2000.
- [35] M.I. Layton and M.J.F. Gales, "Augmented Statistical Models for Speech Recognition," *Proc. ICASSP*, pp. 129–132, 2006.
- [36] A. Halberstadt and J. Glass, "Heterogeneous Measurements and Multiple Classifiers for Speech Recognition," *Proc. ICSLP*, pp. 995–998, 1998.
- [37] J. Yousafzai, Z. Cvetković, and P. Sollich, "Towards Robust Phoneme Classification with Hybrid Features," *Proc. ISIT*, pp. 1643–1647, 2010.
- [38] J. Yousafzai, Z. Cvetković, and P. Sollich, "Tuning Support Vector Machines for Robust Phoneme Classification with Acoustic Waveforms," *Proc. INTERSPEECH*, pp. 2391–2395, 2009.
- [39] N. Smith and M. Gales, "Speech Recognition using SVMs," in *Adv. Neural Inf. Process. Syst.*, 2002, vol. 14, pp. 1197–1204.
- [40] W.M. Campbell, D. Sturim, and D.A. Reynolds, "Support Vector Machines using GMM Supervectors for Speaker Verification," *IEEE Signal Process. Letters*, vol. 13, no. 5, pp. 308–311, 2006.
- [41] J. Louradour, K. Daoudi, and F. Bach, "Feature Space Mahalanobis Sequence Kernels: Application to SVM Speaker Verification," *IEEE Trans. ASLP*, vol. 15, no. 8, pp. 2465–2475, 2007.
- [42] T. Jaakkola and D. Haussler, "Exploiting Generative Models in Discriminative Classifiers," in *Adv. Neural Inf. Process. Syst.*, 1999, vol. 11, pp. 487–493.
- [43] R. Solera-Urena, D. Martín-Iglesias, A. Gallardo-Antolín, C. Peláez-Moreno, and F. Díaz-de María, "Robust ASR using Support Vector Machines," *Speech Comm.*, vol. 49, no. 4, pp. 253–267, 2007.
- [44] J. Padrell-Sendra, D. Martín-Iglesias, and F. Díaz-de María, "Support Vector Machines for Continuous Speech Recognition," *Proc. EUSIPCO*, 2006.
- [45] T. Dietterich and G. Bakiri, "Solving Multiclass Learning Problems via Error-Correcting Output Codes," *J. Artif. Intell. Res.*, vol. 2, pp. 263–286, 1995.
- [46] R. Rifkin and A. Klautau, "In Defense of One-Vs-All Classification," *J. Mach. Learn. Res.*, vol. 5, pp. 101–141, 2004.
- [47] J. Yousafzai, M. Ager, Z. Cvetković, and P. Sollich, "Discriminative and Generative Machine Learning Approaches Towards Robust Phoneme Classification," *Proc. IEEE Workshop Inform. Theory Appl.*, pp. 471–475, 2008.
- [48] N. Garcia-Pedrajas and D. Ortiz-Boyer, "Improving Multiclass Pattern Recognition by the Combination of Two Strategies," *IEEE Trans. PAMI*, vol. 28, no. 6, pp. 1001–1006, 2006.
- [49] K. Crammer and Y. Singer, "On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines," *J. Mach. Learn. Res.*, vol. 2, pp. 265–292, 2002.
- [50] J. Yousafzai, Z. Cvetković, and P. Sollich, "Custom-Designed SVM Kernels for Improved Robustness of Phoneme Classification," *Proc. EUSIPCO*, pp. 1765–1769, 2009.
- [51] D. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, Online Web Resource.
- [52] S. Furui, "Speaker-Independent Isolated Word Recognition using Dynamic Features of Speech Spectrum," *IEEE Trans. ASSP*, vol. 34, no. 1, pp. 52–59, 1986.
- [53] J. Tchorz and B. Kollmeier, "Estimation of the Signal-to-Noise Ratio with Amplitude Modulation Spectrograms," *Speech Comm.*, vol. 38, no. 1, pp. 1–17, 2002.
- [54] E. Nemer, R. Goubran, and S. Mahmoud, "SNR Estimation of Speech Signals Using Subbands and Fourth-Order Statistics," *IEEE Signal Process. Letters*, vol. 6, no. 7, pp. 171–174, 1999.
- [55] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-time Spectral Amplitude Estimator," *IEEE Trans. ASSP*, vol. ASSP-32, pp. 1109–1121, 1984.
- [56] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Log-Spectral Amplitude Estimator," *IEEE Trans. ASSP*, vol. ASSP-33, pp. 443–445, 1985.
- [57] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallet, and N. Dahlgren, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," *Linguistic Data Consortium*, 1993.
- [58] K. F. Lee and H. W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models," *IEEE Trans. ASSP*, vol. 37, no. 11, pp. 1641–1648, 1989.
- [59] R. F. Voss and J. Clarke, "1/f Noise in Music: Music from 1/f Noise," *J. Acoust. Soc. Amer.*, vol. 63, no. 1, pp. 258–263, 1978.
- [60] C. Rathinavelu and L. Deng, "HMM-based Speech Recognition Using State-dependent, Discriminatively Derived Transforms on Mel-Warped DFT Features," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 243–256, 1997.
- [61] F. Pernkopf, T. V. Pham, and J. Bilmes, "Broad Phonetic Classification Using Discriminative Bayesian Networks," *Speech Comm.*, vol. 51, no. 2, pp. 151–166, 2009.
- [62] B. Catanzaro, N. Sundaram, and K. Keutzer, "Fast Support Vector Machine Training and Classification on Graphics Processors," *Proc. ICML*, pp. 104–111, 2008.



**Jibran Yousafzai** (S'08) received his B.S. degree in computer system engineering from GIK Institute, Pakistan in 2004 and the M.Sc. degree in signal processing from King's College London in 2006. In 2004-2005, he worked as a teaching assistant at GIK Institute. He is currently a Ph.D. candidate at the Department of Informatics at King's College London. His areas of interest include automatic speech

recognition, machine learning and audio processing for surround sound technology.

Committee of SAMSI and on the Board of Mathematical Sciences and Applications of the National Academy of Sciences in the US.



**Peter Sollich** is Professor of Statistical Mechanics at King's College London. He obtained an M.Phil. from Cambridge University in 1992 and a Ph.D. from the University of Edinburgh in 1995, and held a Royal Society Dorothy Hodgkin Research Fellowship until 1999. He works on statistical inference and applications of statistical mechanics to complex and disordered systems. He is a member of the Institute of Physics, a fellow of the Higher Education Academy, and

serves on the editorial boards of Europhysics Letters and Journal of Physics A.



**Zoran Cvetković** received his Dipl.Ing.El. and Mag.El. degrees from the University of Belgrade, Yugoslavia, in 1989 and 1992, respectively; the M.Phil. from Columbia University in 1993; and the Ph.D. in electrical engineering from the University of California, Berkeley, in 1995. He held research positions at EPFL, Lausanne, Switzerland (1996), and at Harvard University (2002-04). Between 1997 and 2002 he was a member of the technical staff of AT&T Shannon Laboratory. He

is now Reader in Signal Processing at Kings College London. His research interests are in the broad area of signal processing, ranging from theoretical aspects of signal analysis to applications in source coding, telecommunications, and audio and speech technology.



**Bin Yu** is Chancellor's Professor in the departments of Statistics and of Electrical Engineering & Computer Science at UC Berkeley. She is currently the chair of department of Statistics, and a founding co-director of the Microsoft Lab on Statistics and Information Technology at Peking University, China. She got her B.S. in mathematics from Peking University in 1984, M.S. and Ph.D. in Statistics from UC Berkeley in 1987 and 1990. She has published over 80

papers on a wide range of research areas including empirical process theory, information theory (MDL), MCMC methods, signal processing, machine learning, high dimensional data inference (boosting and Lasso and sparse modeling in general), bioinformatics, and remote sensing. She has been and is serving on many leading journals' editorial boards including Journal of Machine Learning Research, The Annals of Statistics, and Technometrics. Her current research interests include statistical machine learning for high dimensional data and solving data problems from remote sensing, neuroscience, and newspaper documents.

She was a 2006 Guggenheim Fellow, and is a Fellow of AAAS, IEEE, IMS (Institute of Mathematical Statistics) and ASA (American Statistical Association). She is a co-chair of National Scientific