# Tuning Support Vector Machines for Robust Phoneme Classification with Acoustic Waveforms

*Jibran Yousafzai*[1], *Zoran Cvetković*[1], *Peter Sollich*[2]

[1]Department of Electronic Engineering and [2]Department of Mathematics,
King's College London, WC2R 2LS, UK
[jibran.yousafzai, zoran.cvetkovic, peter.sollich] @kcl.ac.uk

## Abstract

This work focuses on the robustness of phoneme classification to additive noise in the acoustic waveform domain using support vector machines (SVMs). We address the issue of designing kernels for acoustic waveforms which imitate the state-of-the-art representations such as PLP and MFCC and are tuned to the physical properties of speech. For comparison, classification results in the PLP representation domain with cepstral mean-and-variance normalization (CMVN) using standard kernels are also reported. It is shown that our custom-designed kernels achieve better classification performance at high noise. Finally, we combine the PLP and acoustic waveform representations to attain better classification than either of the individual representations over the entire range of noise levels tested, from quiet condition up to $-18$dB SNR.

**Index Terms**: Kernels, Phoneme classification, Robustness, Support vector machines

## 1. Introduction

Automatic speech recognition (ASR) systems lack the level of robustness inherent to human speech recognition (HSR). This has a detrimental effect when these systems are operated in adverse acoustical environments, while humans can still recognize isolated speech units above the level of chance at $-18$dB SNR, and significantly above it at $-9$dB SNR [1]. No ASR system achieves performance close to that of human auditory system under such severe noise. While language and context modelling are essential for reducing many errors in speech recognition, accurate recognition of phonemes and the related problem of classification of isolated phonetic units is a major step towards achieving robust recognition of continuous speech.

State-of-the-art ASR front-ends are mostly some variant of Mel-frequency cepstral coefficients (MFCC) or Perceptual Linear Prediction (PLP) [2]. These representations are derived from the short term magnitude spectra followed by non-linear transformations to model the processing of the human auditory system. They remove variations from speech signals that are considered unnecessary for recognition while preserving the information content. This allows for more accurate modelling when the data is limited. However it is not known whether, by reducing the dimension significantly, one also discards some of the information that makes speech such a robust message representation. To make these state-of-the-art representations of speech less sensitive noise, several methods have been proposed to reduce explicitly the effect of noise on spectral representations [3] in order to approach the optimal performance which is achieved when training and testing conditions are matched [4].

The alternative approach investigated in this paper is the use of high-dimensional acoustic waveform representations for robust classification in the presence of additive white Gaussian noise. PLP/MFCC are designed in a way that removes non-lexical invariances (sign, time alignment); however, for classification in the acoustic waveform domain these invariances need to be taken into account by incorporating them in the kernel. By doing this, in combination with straightforward noise adaptation in the kernel, classification performance can be made rather robust to noise [5]. In a key refinement, we try to capture the idea of time derivatives of cepstral features, which measure the rate of change of features and hence represent the dynamics of a speech signal. This is done by embedding variations in signal energy across a phoneme into the kernel. Our experiments demonstrate the effectiveness of the kernels tuned for acoustic waveforms under adverse conditions. For comparison, classification results in the PLP representation domain using standard SVM kernels are also reported. We show further that a convex combination [5] of the decision functions of the PLP and acoustic waveform SVM classifiers results in superior performance across the entire range of SNRs. While this study is focused on phoneme classification for comparison of the acoustic waveform and PLP representations of speech, we believe the results also have implications for the construction of ASR systems.

The SVM approach to classification of phonemes using error-correcting output codes (ECOC) [6] is reviewed briefly in Section 2. Kernel design for the classification task in the acoustic waveform domain is addressed in Section 3, including the extension to time-variation of signal energies. Section 4 presents techniques for noise adaptation in both the PLP and acoustic waveform domains. The classification results in the PLP and acoustic waveform domains are reported in Section 5, where we also discuss the combination of the PLP and acoustic waveform representations for improved accuracy. It is shown that the acoustic waveforms perform better for all SNRs below a crossover point between 6dB and 12dB SNR. Finally, Section 6 has conclusions and an outlook towards future work.

## 2. Classification using SVMs

Support vector machines [7] estimate decision surfaces separating two classes of data. In the case of speech recognition, one typically requires nonlinear decision boundaries which are constructed using kernels that implicitly map data points to high-dimensional feature vectors. A kernel-based decision function which classifies an input vector $\mathbf{x}$ is expressed as

$$h(\mathbf{x}) = \sum_i \alpha_i y_i \langle \varphi(\mathbf{x}), \varphi(\mathbf{x}_i) \rangle + b = \sum_i \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \tag{1}$$

where $\varphi$ is a non-linear mapping function while $\mathbf{x}_i$, $y_i = \pm 1$ and $\alpha_i$, respectively, are the $i$-th training sample, its class label and its Lagrange multiplier. $K(\mathbf{x}, \mathbf{x}_i)$ is a kernel function that satisfies Mercer's theorem and $b$ is the classifier bias determined by the training algorithm. Two commonly used kernels are the polynomial kernel, $K_p(\mathbf{x},\mathbf{x}_i) = (1+\langle\mathbf{x},\mathbf{x}_i\rangle)^{\Theta}$, and radial basis function (RBF) kernel, $K_r(\mathbf{x},\mathbf{x}_i) = e^{-\Gamma\|\mathbf{x}-\mathbf{x}_i\|^2}$.

To obtain a multiclass classifier, binary SVM classifiers are combined via ECOC [6] methods. A standard approach is to use $K(K-1)/2$ pairwise classifiers, each trained to distinguish two of the $K$ classes. For a test point $\mathbf{x}$, we then predict the class $k$ for which $d_k(\mathbf{x}) = \sum_{l=1,l\neq k}^{K} \xi(h_{kl}(\mathbf{x}))$ is minimized, where $\xi$ is some loss function and $h_{kl}(\mathbf{x})$ is the output of the classifier trained to distinguish classes $k$ and $l$, with sign chosen so that a positive sign indicates class $k$. We compared a number of loss functions $\xi(h)$; the hinge loss $\xi(h) = \max(1-h,0)$ performed best and is used throughout this paper.

# 3. Kernels for Acoustic Waveforms

For a classification task using SVMs, the most important issue is the use of appropriate kernels that express prior knowledge about the physical properties of the data. For acoustic waveforms, key properties are: $(a)$ *sign-invariance* and $(b)$ *shift-invariance* - the fact that a speech waveform and a version that is inverted or shifted in time are perceived as being the same. We construct below kernels that incorporate these invariances, but to be meaningful these require normalized waveforms. We then need to account separately for the $(c)$ *energy distribution* as illustrated in Figure 1 (top) using log-energy distributions of 20ms waveform subsegments at the phoneme center for phoneme classes /aa/ and /v/. Comparing these distributions shows that the energy of isolated phoneme segments or subsegments (see below) can be very useful in distinguishing them.

Incorporating these properties of acoustic waveforms into a kernel $K(\mathbf{x},\mathbf{x}_i)$, using initially the energy of the entire phoneme segment, results in a kernel $K_n(\mathbf{x},\mathbf{x}_i)$ given by

$$K_n(\mathbf{x},\mathbf{x}_i) = e^{-\left(\log\|\mathbf{x}\|^2-\log\|\mathbf{x}_i\|^2\right)^2/2a^2} K_s(\mathbf{x},\mathbf{x}_i)\,, \tag{2}$$

where

$$K_s(\mathbf{x},\mathbf{x}_i) = \frac{1}{(2n+1)^2}\sum_{u,v=-n}^{n} K_e(\mathbf{x}^{u\Delta},\mathbf{x}_i^{v\Delta})\,, \tag{3}$$

$$K_e(\mathbf{x},\mathbf{x}_i) = K(\mathbf{x},\mathbf{x}_i)+K(\mathbf{x},-\mathbf{x}_i)+K(-\mathbf{x},\mathbf{x}_i)+K(-\mathbf{x},-\mathbf{x}_i)\,, \tag{4}$$

$\Delta$ is the shift increment, $[-n\Delta, n\Delta]$ is the shift range, $\mathbf{x}^{u\Delta}$ is a segment of the same length as the original waveform $\mathbf{x}^0$ but extracted from a position shifted by $u\Delta$ samples. In this paper, we use the polynomial kernel, $K_p$ for both representations. However, evaluating $K_p$ for the waveforms requires normalization of $\mathbf{x}$ and $\mathbf{x}_i$ to give a sensible estimate of their closeness i.e. $K_p(\mathbf{x},\mathbf{x}_i) = (1+\langle\mathbf{x}/\|\mathbf{x}\|,\mathbf{x}_i/\|\mathbf{x}_i\|\rangle)^{\Theta}$. This is used as a baseline kernel for the waveform representation whereas $K_p$ in its vanilla form is used for the PLP representation.

Next, we evaluate the kernel $K_n$ over $T$ non-overlapping subsegments of the phoneme as

$$K_{n,s}(\mathbf{x},\mathbf{x}_i) = \sum_{t=1}^{T} e^{-\left(\log\|\mathbf{x}_t\|^2-\log\|\mathbf{x}_{i,t}\|^2\right)^2/2a^2} K_s(\mathbf{x}_t,\mathbf{x}_{i,t})\,, \tag{5}$$

where $\mathbf{x}_t$ and $\mathbf{x}_{i,t}$ are the $t^{\text{th}}$ subsegments of the test waveform $\mathbf{x}$ and the $i^{\text{th}}$ training sample, $\mathbf{x}_i$ respectively. This is done in order to capture the dynamics of speech over relatively shorter time durations in a manner similar to the time derivatives and second order derivatives of the cepstral coefficients. This extended kernel is sensitive to correlation between the individual subsegments of the phonemes and gives information about the phoneme energy over a finer resolution, which can help to distinguish phoneme classes with similar energy distributions but different energy profiles as shown in Figure 1 (bottom).

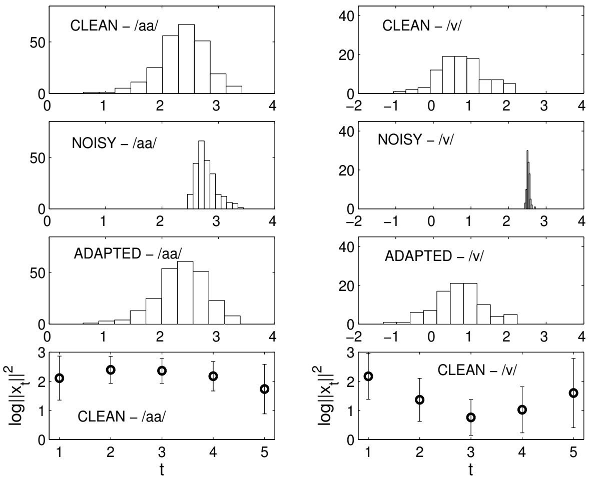

Figure 1: Histograms of log-energies of central subsegments of phoneme classes /aa/ and /v/ for clean waveforms (top), noisy waveforms at 0dB SNR ($2^{nd}$ row), noise adapted waveforms, i.e. $\log\left|\|\mathbf{x}\|^2-\sigma^2\right|$ (3rd row), and time profile (mean ± standard deviation) of log energy (bottom).

Since PLP, MFCC and other state-of-the-art representations are based on short-time magnitude spectra and contain information about the energy, using similar custom-designed kernels for classification in the PLP domain will not have any advantage over the standard (polynomial or RBF) kernels.

# 4. Noise Adaptation

Features extracted from the test data are adapted to noise to improve the robustness in both domains. Since the noise variance, $\sigma^2$ can be estimated during pause intervals (non-speech activity) between speech signals, we assume that its value is known. For the PLP representations, the features are standardized, i.e. scaled and shifted to have zero mean and unit variance per sentence. The optimal performance with PLP is obtained under matched training and test conditions [4]. However, this is an impractical target which could be achieved only if one had access to a large set of classifiers trained for different noise types and levels. Therefore, in order to have a fair comparison of PLP with acoustic waveforms, we use classifiers trained in quiet conditions, with feature vectors of the test data adapted to noise using cepstral mean and variance normalization (CMVN) [3], a noise compensation technique that modifies the cepstral coefficients in order to minimize the mismatch between the training and test data. Here, cepstral features are standardized on each noisy test sentence [3]. By attempting to 'decouple' the speech information from the noise, CMVN can significantly improve the performance of the PLP classifiers. Another common approach to reduce the mismatch between training and test data is multi-condition/multi-style training; however, CMVN and its variants generally perform better [8].

In the case of acoustic waveforms, the test data is normalized to $\sqrt{1+\sigma^2}$ whereas the training data is set to have a unit norm for computation of the inner product in the polynomial kernel. This is done to keep the norm of the speech *signal* roughly independent of the noise. Explicitly, let $\tilde{\mathbf{x}} = \mathbf{x}\sqrt{1+\sigma^2}/\|\mathbf{x}\|$ and $\tilde{\mathbf{x}}_i = \mathbf{x}_i/\|\mathbf{x}_i\|$ for a test waveform $\mathbf{x}$ and training waveform $\mathbf{x}_i$. Then the baseline polynomial kernel for the normalized waveforms is $K_p(\mathbf{x}, \mathbf{x}_i) = (1 + \langle\tilde{\mathbf{x}}, \tilde{\mathbf{x}}_i\rangle)^\Theta$ with $K_s$ and $K_e$ then defined as in $(3, 4)$.

Similar adaptation as in the polynomial kernel contribution is also required in $(5)$. The energy distributions of waveform subsegments change significantly with noise as illustrated in Figure 1 (2nd row) for an SNR of 0dB. Under the assumption that speech and noise are uncorrelated, subtracting the estimated noise variance ($\sigma^2$) from the energy of the subsegment of the noisy phoneme should result in distributions of the energies that are very similar to those of the clean subsegments as shown in Figure 1 (3rd row). We, therefore, use this adapted energy of the subsegments in evaluating $(5)$, giving

$$K_{n,s}(\mathbf{x}, \mathbf{x}_i) = \sum_{t=1}^{T} e^{\frac{-\left(\log\left|\|\mathbf{x}_t\|^2 - \sigma^2\right| - \log\left\|\mathbf{x}_{i,t}\right\|^2\right)^2}{2a^2}} K_s(\mathbf{x}_t, \mathbf{x}_{i,t}) \,. \tag{6}$$

As training for acoustic waveforms is performed in quiet conditions, noise adaption of the training data $\mathbf{x}_i$ is not required. The absolute value of the subtracted energy is used to catch the rare cases when speech and noise are anti-correlated. There are two important issues to be addressed when using $(6)$ in the presence of noise: $(a)$ *Normalization of the subsegments* - the use of $(6)$ requires normalization of the clean subsegments to unit norm and of the noisy ones to $\sqrt{1+\sigma^2}$. For short subsegments there can however 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. $(b)$ *Orthogonality* - using short (lower dimensional) subsegments makes fluctuations away from our assumed orthogonality of speech and noise more likely. To address these issues, we also consider a kernel $K_{n,u}$ which is obtained by replacing the last factor in $(6)$ by $K_s(\mathbf{x}, \mathbf{x}_i)$: this time-correlation part of the kernel is then left unsegmented, while the energies are still evaluated for subsegments of the phonemes. One might expect that $K_{n,s}$ outperforms $K_{n,u}$ in the absence of noise but it performs worse in high noise due to the two limitations discussed above.

In the next section, we show the performance of these kernels for the phoneme classification task in the acoustic waveform domain. Classification in the PLP domain is used as a benchmark for comparison with acoustic waveforms.

## 5. Results

Experiments are performed on the 'si' and 'sx' sentences of TIMIT database [9]. The training set consists of 3696 sentences from 168 different speakers. The core set is used for testing which consists of 192 sentences from 24 different speakers not included in the training set. We remove the glottal stops /q/ from the labels and fold certain allophones into their corresponding phonemes using the standard Kai-Fu Lee clustering [10], resulting in a total of 48 phoneme classes. Among these classes, there are 7 groups for which the contribution of within-group confusions towards multiclass error is not counted [10].

Regarding the SVM classifiers for acoustic waveform representation, results are reported for $K_{n,s}$ and $K_{n,u}$. For the PLP representation, comparable performance is obtained with polynomial and RBF kernels so we show results for the former.

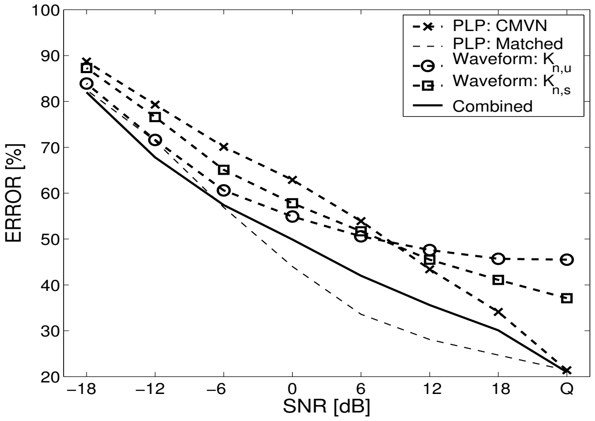

Figure 2: Classification results - PLP with CMVN using $K_p$ trained in quiet and matched conditions, waveforms using $K_{n,s}$ and $K_{n,u}$ kernels and the combination of PLP (CMVN) and waveforms ($K_{n,u}$).

Fixed hyperparameter values are used throughout for training binary SVMs: the degree of $K_p$, $\Theta = 6$ and the penalty parameter $C = 1$.

For the acoustic waveform representation, phoneme segments are extracted from the TIMIT sentences by applying a 100 ms rectangular window at the center of each phoneme waveform (of variable length), which at 16 kHz sampling frequency gives fixed length vectors in $\mathbb{R}^{1600}$. In the evaluation of $K_s$ defined in (3), we use a shift increment of $\Delta = 100$ samples ($\approx 6$ ms) over a shift range $\pm 100$ (so that $n = 1$), giving three shifted segments of length 1400 samples each. In evaluating $K_{n,s}$, each of these segments is broken into $T = 5$ subsegments of equal length whereas in $K_{n,u}$, the complete segments are used to compute the time-correlation. The value of $a$ is set to 0.5 for both kernels. For the PLP representation, we convert each waveform into a sequence of 13 dimensional feature vectors, their time derivatives and second order derivatives which are combined into a sequence of 39 dimensional feature vectors. Then, the 9 frames (with frame duration of 25ms and a frame rate of 100 frames/sec) closest to the center of a particular phoneme are concatenated to give a representation in $\mathbb{R}^{351}$. In this study, we focus on investigating robustness in the presence of additive white Gaussian noise. To test the classification performance of PLP 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.

Classification results using SVMs in the PLP and acoustic waveform domains are shown in Figure 2. For acoustic waveforms, classification results with kernels $K_{n,s}$ and $K_{n,u}$ are presented whereas $K_p$ is used for classification of PLP features. One observes that a PLP classifier trained on clean data gives very good performance when tested on clean data i.e. 21% error. (We achieve slightly better performance than [11] due to different cepstral representations.) But at 0dB SNR, we get an error of 63% even with CMVN. This can now be contrasted with the results of acoustic waveform classifiers. Classification with kernels, $K_{n,s}$ and $K_{n,u}$ exhibits a more robust behavior to noise and achieves improvements over PLP for noise levels above a crossover point between 12dB and 6dB SNR. The largest improvement over PLP, of 10% is achieved by $K_{n,u}$ at

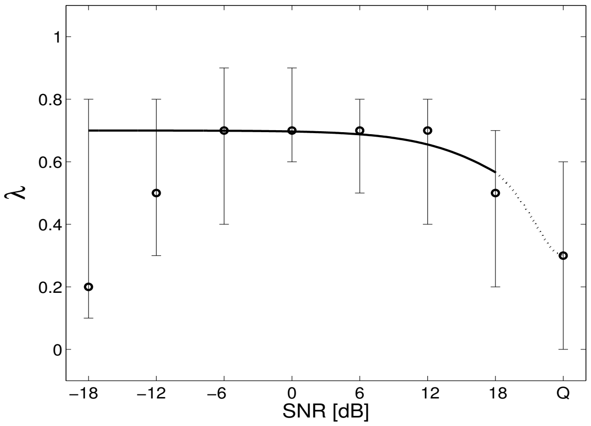

Figure 3: Optimal and approx. values of $\lambda$ for a range of test SNRs. $\lambda = 0$ corresponds to PLP classification with CMVN, $\lambda = 1$ is waveform classification with kernel $K_{n,u}$.

$-6$dB SNR. In a comparison of kernels for the acoustic waveforms, $K_{n,s}$ achieves an $8\%$ improvement in quiet conditions over $K_{n,u}$ because $K_{n,s}$ is sensitive to correlation between the individual phoneme subsegments however, $K_{n,u}$ performs better in high noise e.g. $K_{n,u}$ achieves a $4\%$ average improvement over $K_{n,s}$ between 6dB and $-18$dB SNRs. This is due to the limitations of $K_{n,s}$ in high noise as discussed in Section 4.

In our previous work [5], we established that a combination of the PLP and waveform classifiers attains better classification performance than either of the individual representations. As waveform classifiers with kernel $K_{n,u}$ achieve significantly better results in high noise, therefore we consider its convex combination with the decision values of the PLP classifiers with CMVN. For classifiers $h_p(\mathbf{x})$ and $h_w(\mathbf{x})$ in the PLP and waveform domains respectively, we define the combined classifier output as $h_c(\mathbf{x}) = \lambda(\sigma^2)h_w(\mathbf{x}) + \left[1 - \lambda(\sigma^2)\right] h_p(\mathbf{x})$, where $\lambda(\sigma^2)$ is a parameter which needs to be selected, depending on the noise variance, to achieve optimal performance. These binary classifiers are then combined for multiclass classification as described in Section 2. In Figure 3, the "optimal" $\lambda(\sigma^2)$ i.e. the values of $\lambda(\sigma^2)$ which give the minimum classification error for a given SNR of the test phoneme, are shown marked by 'o'. The error bars give a range of values of $\lambda(\sigma^2)$ for which the classification error is less than the minimum error $(\%) + 2\%$. We use an approximation of the optimal $\lambda(\sigma^2)$: $\lambda_{\text{app}}(\sigma^2) = \alpha + \beta/\left(1 + \left(\sigma_0^2/\sigma^2\right)\right)$, with $\alpha = 0.3$, $\beta = 0.4$ and $\sigma_0 = 0.09$ as shown in Figure 3 (solid line).

In Figure 2, we compare the classification performance in the PLP and acoustic waveform domains with the combined classifier for $\lambda_{\text{app}}(\sigma^2)$. One observes that the combined classifier often performs better or at least as well as the individual classifiers. Furthermore, the wide range of errorbars in Figure 3 indicates that the combined classifier is less sensitive to the values of $\lambda(\sigma^2)$. Due to this, we found no significant difference in the performance of the combined classifier for the optimal $\lambda(\sigma^2)$ and its approximation, $\lambda_{\text{app}}(\sigma^2)$. Moreover, the values of optimal $\lambda$ (between $0.2$ and $0.7$) for different SNRs suggest that the combined classifier is not simply a hard-switch between the two representations and a genuine improvement in performance is achieved when $0<\lambda(\sigma^2)<1$. Although the combined classifier does not achieve the impractical target of PLP classifier trained and tested in matched conditions for SNR$>-6$dB as shown in Figure 2, the gain in classification accuracy is significant compared to a standalone PLP classifier with CMVN. For instance, the combined classifier achieves an average of $11\%$ less error than the PLP classifier trained on clean data with CMVN for $-12$dB$\leq$SNR$\leq$ 12dB.

## 6. Conclusions

The robustness of phoneme classification to additive white Gaussian noise in the PLP and acoustic waveform domains was investigated using SVMs. We observe that embedding invariances and variations in the signal energy across a phoneme into the kernel can significantly improve the classification performance. While PLP representation allows very accurate classification of phonemes especially for clean data, its performance suffers severe degradation at high noise. On the other hand, the high-dimensional acoustic waveform representation, although not as accurate as PLP classification on clean data, is more robust in severe noise. Finally, we demonstrate that a convex combination of classifiers can achieve performance that is consistently better than both individual domains across the entire range of SNRs. Currently, our preliminary experiments with longer phoneme segments (2048 samples) has shown improved performance in high noise, indicating them to be a more suitable representation for acoustic waveforms. In future work, we plan to fine tune the segmentation of kernels by assigning weight factors to put more emphasis on the variations in energy and correlation of subsegments of phonemes. This would be done in order to be consistent with the time derivatives and second order derivatives of the PLP features.

## 7. References

[1] G. Miller and P. Nicely, "An Analysis of Perceptual Confusions among some English Consonants," *J. of the Acous. Soc. of America*, vol. 27, no. 2, pp. 338–352, 1955.

[2] H. Hermansky, "PLP Analysis of Speech," *J. of the Acous. Soc. of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[3] O. Viikki and K. Laurila, "Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition," *Speech Comm.*, vol. 25, pp. 133–147, 1998.

[4] M. Gales and S. Young, "Robust Continuous Speech Recognition using Parallel Model Combination," *IEEE Trans. on Speech and Audio Proc.*, pp. 352–359, 1996.

[5] J. Yousafzai, Z. Cvetković, and P. Sollich, "Custom-Designed SVM Kernels for Improved Robustness of Phoneme Classification," *In Proc. of EUSIPCO*, 2009.

[6] T. Dietterich and G. Bakiri, "Solving Multiclass Learning Problems via Error-Correcting Output Codes," *J. of AI Research*, vol. 2, pp. 263–286, 1995.

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

[8] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE Algorithm on Aurora2," *EuroSpeech*, 2001.

[9] W. Fisher, G. Doddington, and K. Goudie-Marshall, "DARPA Speech Recognition Research Database," *DARPA Sp. Recogn. Workshop*, pp. 93–99, 1986.

[10] K. F. Lee and H. W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models," *IEEE Trans. Ac. Speech Sig. Proc.*, vol. 37, no. 11, 1989.

[11] P. Clarkson and P. Moreno, "On the Use of Support Vector Machines for Phonetic Classification," *ICASSP*, 1999.