16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

# COMBINED PLP – ACOUSTIC WAVEFORM CLASSIFICATION FOR ROBUST PHONEME RECOGNITION USING SUPPORT VECTOR MACHINES

*Jibran Yousafzai*[1], *Zoran Cvetković*[1], *Peter Sollich*[2] *and Bin Yu*[3]

Division of Engineering[1] and Department of Mathematics[2]
King's College London, WC2R 2LS, UK

Department of Statistics[3]
University of California, Berkeley, CA 94720, USA

## ABSTRACT

The robustness of phoneme classification to additive white Gaussian noise is investigated in acoustic waveform and PLP domains using support vector machines (SVMs). Classification in the PLP space gives excellent results at low noise level under matched training and testing conditions, but it is very sensitive to their mismatch. On the other hand, classification in the acoustic waveform domain is inferior at low noise levels, but exhibits a much more robust behaviour, and at high noise levels even with training on clean data significantly outperforms the classification in the PLP space with training under matched conditions. The two classifiers are then combined in a manner which attains the accuracy of PLP at low noise levels and significantly improves its robustness to additive noise.

## 1. INTRODUCTION

Language and context modelling have resulted in major breakthroughs that have made automatic speech recognition (ASR) possible. ASR systems, however, still lack the level of robustness inherent to human speech recognition [1, 2]. While language and context modelling are essential for reducing many errors in speech recognition, human speech recognition attains a major portion of its robustness early on in the process, before and independently of context information [3, 4]. In the extreme case, when phonemes or syllables are recognized at the level of chance (random guessing), no context and language modelling can retrieve any information from speech. In the other extreme, when all phonemes and syllables are recognized accurately, context and/or language modelling are not needed. Both ASR and human speech recognition operate between these two extreme conditions, therefore both sophisticated language-context modelling and accurate recognition of isolated phonetic units are needed to achieve a robust recognition of continuous speech. In recognizing syllables or isolated words, the human auditory systems performs above chance level already at -18dB SNR and significantly above it at -9dB SNR [4]. No ASR system is able to achieve performance close to that of human auditory systems in recognizing isolated words or phonemes under severe noisy conditions, as has been confirmed recently in an extensive study by Sroka and Braida [2]. Robust recognition of isolated phonemes and syllables is therefore a very important open problem of ASR.

Most of the state-of-the-art ASR front-ends are generally some variant of PLP[5], RASTA[6] or MFCC[7]. These representations are derived from the short term magnitude spectra followed by nonlinear 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 important speech information and have a much lower dimension than acoustic waveforms. Hence, they facilitate the estimation of probability distributions and significantly enhance the discrimination of different phonetic units. However it is not certain that in this process of peeling off speech components that are unnecessary for recognition one is not discarding part of the information that makes speech such a robust message representation.

In the representation domains which involve compression, different phonetic units although well separated may not be sufficiently apart and may start overlapping considerably at lower noise levels than they do in the original uncompressed domain of acoustic waveforms, consequently ending up with ASR systems which are very sensitive to noise and other forms of degradation. Several methods have been proposed to reduce explicitly the effect of noise on spectral representations [8] in order to approach the optimal performance which is achieved when the training and testing conditions are matched [9]. Our recent study indicates that classifiers in the high-dimensional acoustic waveform domain when trained in quiet conditions may outperform classifiers in the PLP domain trained under matched conditions in severe noise [10]. However, the classifiers in the PLP domain trained under matched conditions demonstrate superior performance when tested on phonemes corrupted by low levels of noise. In this paper, we consider combining SVM classifiers in the PLP and acoustic waveform domains to achieve the performance equivalent to the best of both domains across a wide range of SNRs. The method considered is convex combination of the decision functions of classifiers in the two domains. Preliminary experiments demonstrate the effectiveness of this method for robust phoneme recognition under adverse conditions. Furthermore, the combined classifier is desensitized to noise mismatch between training and testing conditions. It should be emphasized that this preliminary study is focused on the comparison of phoneme classification in acoustic waveform and PLP representation domains and their combination, rather than the design of a complete phoneme recognition system. Useful conclusions can be drawn by comparing this paper with [11], a similar approach using Gaussian mixture models (GMMs). SVM approach to classification of phonemes in the PLP and acoustic waveform domains is presented in Section 2. In Section 3, the results of classification in the individual feature spaces are reported. The method for combining classifiers of PLP and acoustic waveform domains is described in Section 4 where we also present results of the combined classifier and draw comparisons to classifiers in the individual domains. Finally, Section 5 draws some conclusions.

## 2. CLASSIFICATION METHOD

An SVM estimates decision surfaces separating two classes of data. In the simplest case these are linear but for speech recognition, one typically requires nonlinear decision boundaries. These are constructed using kernels instead of dot products, implicitly mapping data points to high-dimensional feature vectors [12]. A kernel-based decision function has the form

$$h(x) = \sum_i \alpha_i y_i K(x, x_i) + b \tag{1}$$

where $x_i$ are all training inputs, $y_i = \pm 1$ are class labels, while the bias term, $b$, and the $\alpha_i$ are parameters determined by SVM. Two commonly used kernels are polynomial and radial basis function (RBF) kernels given by (2) and (3), respectively,

$$K(x_i, x_j) = (1 + \langle x_i, x_j \rangle)^{\Theta} , \tag{2}$$

$$K(x_i, x_j) = e^{-\Gamma \|x_i - x_j\|^2} . \tag{3}$$

SVMs are binary classifiers that distinguish two classes or two groups of classes. To obtain a multiclass classifier, binary

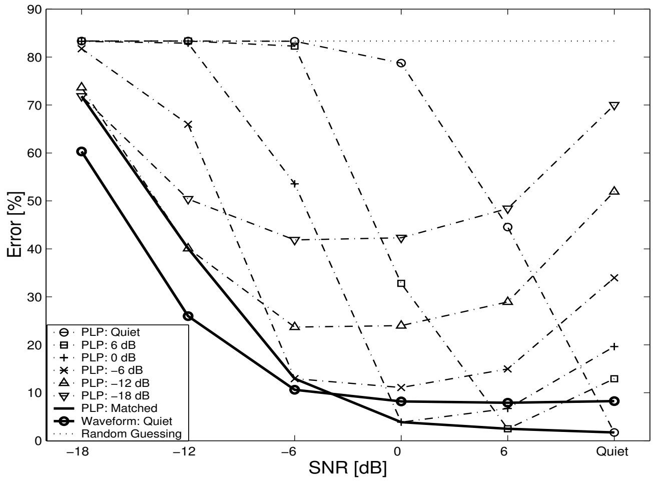

Figure 1: Multiclass error rate for SVM classifiers in the PLP and acoustic waveform domains. SVMs for acoustic waveforms are trained on clean data while, for PLP, training is done on noisy data sets with SNR indicated in the legend. Polynomial and shift-invariant even-polynomial kernels are used for PLP and acoustic waveform representations respectively. The bold lines represent the classification performance in the acoustic waveform domain and PLP domain (under matched conditions).

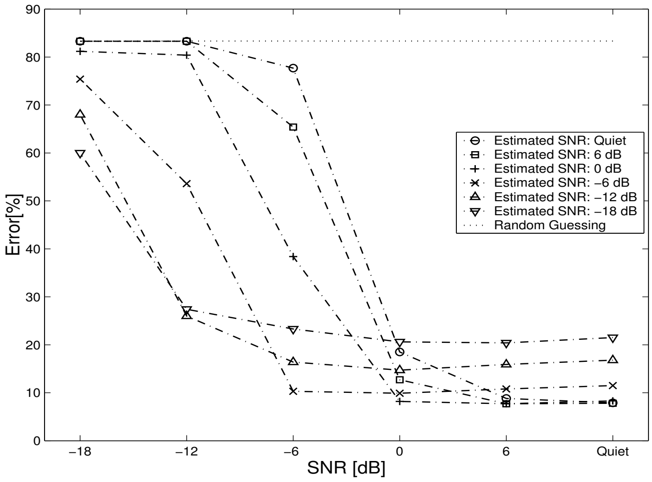

Figure 2: Multiclass error rates for SVM classifier in the acoustic waveform domain with mismatch between actual and estimated SNRs. Waveforms corrupted by noise with SNR as shown on the x-axis are normalized to $\sqrt{1+\tilde{\sigma}^2}$ where $\tilde{\sigma}^2$ is the noise variance corresponding to the estimated SNR as indicated in the legend.

SVM classifiers are combined via error-correcting code methods [13]. In particular, $L$ binary classifiers are trained to distinguish between $K$ classes using the coding matrix $\mathbf{M}_{K\times L}$, with elements $M_{kl} \in \{0,1,-1\}$. Classifier $l$ is trained on data of classes $k$ for which $M_{kl} \neq 0$ where $\mathrm{sgn}(M_{kl})$ is the class label; it has no knowledge about classes $k$ for which $M_{kl} = 0$. For example, in the case of one-vs-all classifiers ($L = K$), $M_{kl} = 1$, if $k = l$, otherwise $M_{kl} = -1$. For the one-vs-one classification strategy, on the other hand, $L = K(K-1)/2$, each classifier is trained on data from only two phoneme classes. Here all the elements of a column of the coding matrix $\mathbf{M}$ are set to 0 except for one $+1$ and one $-1$.

In order to form a multiclass classifier by combining the binary ones, given a test point $x$, the decision values, $h_l(x), l = 1,\cdots,L$, of the $L$ binary classifiers are computed to form a vector $\bar{h}(x) = [h_1(x),\cdots,h_L(x)]$. The class of $x$ is the predicted to be the index of the row, $\bar{M}_k = [M_{k1},\cdots,M_{kL}], k = 1,\cdots,K$ of the matrix $\mathbf{M}$ which is at the minimum distance from the vector $\bar{h}(x)$ according to some distance measure, $H(x) = \arg\min_k d(\bar{M}_k,\bar{h}(x))$. The distance measure is given as $d(\bar{M}_k,\bar{h}(x)) = \sum_{l=1}^{L}\xi(z_{kl})$ where $\xi$ is some loss function and $z_{kl} = M_{kl}h_l(x)$. Commonly used loss functions include hinge – $\xi(z) = (1-z)_+ = \max(1-z,0)$, Hamming – $\xi(z) = [1-\mathrm{sgn}(z)]/2$, exponential – $\xi(z) = e^{-z}$ and linear – $\xi(z) = -z$ loss functions.

The issues of primary importance in any multiclass classification task with SVMs are: $(a)$ the use/design of appropriate kernel and $(b)$ the choice of the coding matrix. A kernel function with prior knowledge about the physical properties of the data sets can significantly improve the performance of the individual binary classifiers. To this end, for classification using acoustic waveforms, we use *even kernels* [14] to take into account the fact that a speech waveform and its inverted version are perceived as being the same. An even version of a kernel $K$ can be obtained as

$$K_e(x_i,x_j) = K(x_i,x_j) + K(x_i,-x_j) + K(-x_i,x_j) + K(-x_i,-x_j)\ , \tag{4}$$

which is the approach used in this work. Furthermore, invariance of acoustic waveforms to time alignment can be incorporated into even kernel by defining a *shift-invariant even kernel* of the form

$$K_s(x_i,x_j) = \frac{1}{(2n+1)^2}\sum_{p=-n}^{n}\sum_{q=-n}^{n} K_e(x_i^{p\Delta},x_j^{q\Delta})\ , \tag{5}$$

where $\Delta$ is the shift increment, $[-n\Delta, n\Delta]$ is the shift range, and $x^{p\Delta}$ denotes a time-shifted version of $x$. In particular, $x^{p\Delta}$ is the segment of the same length and extracted from the same acoustic waveform as $x$ but starting from a position shifted by $p\Delta$ samples in time. Since PLP, MFCC and other state-of-the-art representations are based on the short-time magnitude spectra, using even kernel or shift-invariant kernel for classification in the PLP domain will not have any significant advantage over the standard (polynomial or RBF) kernels and that was confirmed in our experiments.

Regarding the choice of the matrix $\mathbf{M}$, since the error-correcting capability of a code is commensurate to the minimum Hamming distance, $\beta$, between pairs of code words, the classification task benefits from using matrices $\mathbf{M}$ with larger Hamming distances between their rows. However, depending on the data sets, one must balance the use of a matrix $\mathbf{M}$ having larger Hamming distance between its code words with a choice of accurate binary classifiers. For instance, our experiments showed that in the case of $K = 6$ classes, the multiclass classifier obtained from 3-vs-3 binary classifiers ($\beta = 6$ for the corresponding matrix) performed worse than the classifiers obtained from either one-vs-all ($\beta = 2$) or one-vs-one ($\beta = 1$) classifiers, because the individual binary 3-vs-3 classifiers were on average much less accurate than one-vs-one or one-vs-all classifiers. One possible choice for a coding matrix can be a complete dense code i.e. for $K$ classes, $L = 2^{K-1}-1$ and $\beta = 2^{K-2}$. However, this code suffers from the problem of scalability of the number of classifiers, $L$ with the number of classes, $K$. Since the goal is to extend this work to a complete set of phonemes, the complete dense code may not be an appropriate choice as our coding matrix. In this study, we report results using matrix $\mathbf{M}$ that combines both one-vs-all and one-vs-one classifiers as this combination performed better than either set of binary classifiers separately on its own. Moreover, the number of classifiers in this coding matrix scales well ($O(K^2)$) with the number of classes($K$) compared to a complete dense code.

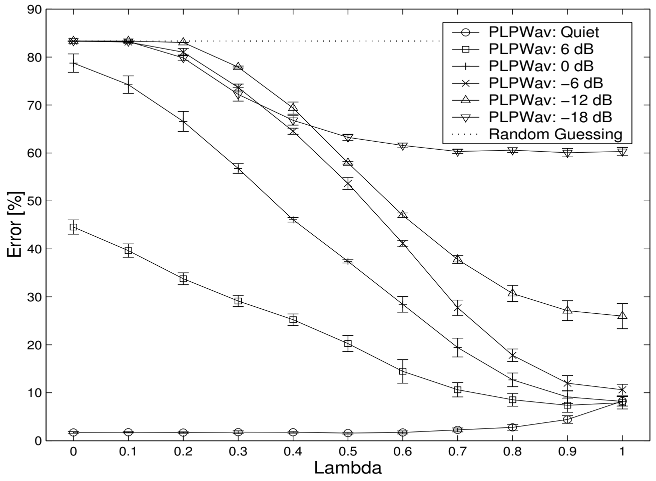

Figure 3: Multiclass classification error of combined classifier with training performed under quiet conditions. SNR of the test phonemes is indicated in the legend. $\lambda(\sigma^2) = 0$ represents the classification error in the PLP domain and $\lambda(\sigma^2) = 1$ corresponds to classification in the acoustic waveform domain.

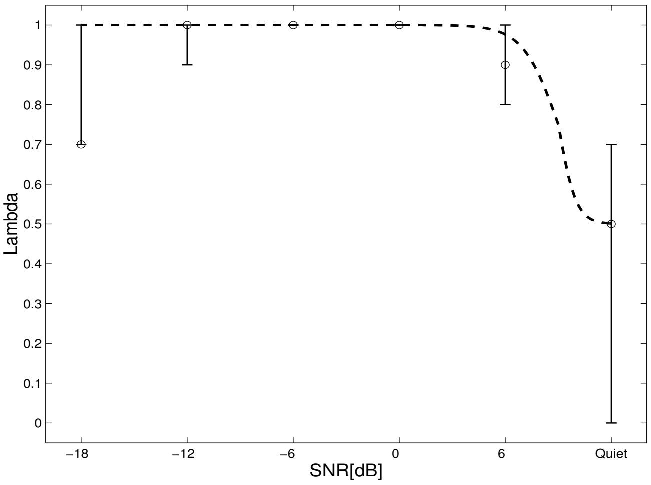

Figure 4: For the combined classifier trained in quiet condition, the optimal values of $\lambda(\sigma^2)$ for a range of test SNRs are shown as 'o'. The error bars give a range of values of $\lambda(\sigma^2)$ for which the classification error is less than the minimum error $(\%) + 2\%$. The dashed line is an approximation of the optimal $\lambda(\sigma^2)$ as given by (7).

## 3. CLASSIFICATION IN INDIVDUAL FEATURE SPACES

Experiments were performed on the realizations of six phonemes (/b/, /f/, /m/, /r/, /t/, /z/) extracted from the TIMIT database [15]. This set includes examples from fricatives, nasals, semivowels and voiced and unvoiced stops. In addition, this set of phonemes provides pairwise discrimination tasks of a varying level of difficulty. Each class consists of approximately 1000 representative acoustic waveforms, of which 80% were used for training and 20% for testing; error bars were derived by considering five different such splits. Phonetic segments used in this work were obtained by applying a 64ms rectangular window to each phoneme waveform (of variable length) at its center, which at 16kHz sampling frequency gives fixed-length vectors in $\mathbb{R}^{1024}$, followed by normalization to unit norm. For comparison, 12th order PLP representations of each 64ms phoneme segment were taken, leading to 4 frames of 13 coefficients [16]. These frames were then concatenated to give a representation in $\mathbb{R}^{52}$.

In the evaluation of shift-invariant kernel defined in (5), we use shift increment $\Delta = 25$ ($\approx 1.5$ ms) over a range of $\pm 100$ samples ($\approx \pm 6$ ms) to reduce computational effort. As pointed out before, PLP uses frames of magnitude spectra, it is less sensitive to time alignment. In this preliminary study, we investigated robustness to white Gaussian noise only. Since the noise variance, $\sigma^2$ can be estimated during pause intervals (non-speech activity) between speech signals, we assume for all classification approaches that the noise variance is known. Noise is added to the clean acoustic waveforms at phoneme level rather than sentence level in order to conduct the experiments in a more controlled environment. For data corrupted by noise, the acoustic waveforms representations of phonemes are normalized to $\sqrt{1+\sigma^2}$. This is done to keep the norm of the speech signal component roughly independent of noise. In the case of PLP, we experimented with both this normalization and normalization to unity independently of SNR, choosing the latter as it gave better performance. PLP features are standardized, i.e. scaled and shifted to have zero mean and unit variance on the training set.

Regarding the binary SVM classifiers, comparable performance is obtained with polynomial and RBF kernels for PLP representation so we show results for the former. For the waveform representation, the polynomial kernel performed better than the RBF kernel and the shift-invariant even-polynomial kernel outperformed both. Classification results using SVMs in the PLP and acoustic waveform domains are shown in Figure 1. The best results for both domains are compared here, i.e. shift-invariant even-polynomial kernel for waveforms and polynomial kernel for PLP. For both representations, a coding matrix that combines the one-vs-all and one-vs-one classifiers was used. Hinge loss function, which performed comparably or better than the Hamming, linear and exponential loss functions, is used to calculate the distance measure, $d$. One can observe that a PLP classifier trained on clean data gives excellent performance (less than 2% error) when tested on clean data, however at noise level as low as 6dB SNR, we get an error of 45%, while classification is at the level of chance for SNR smaller than 0dB. This observation is quite general: the PLP classifiers are highly sensitive to mismatch between the training and test conditions. For example, the PLP classifier trained at 6dB SNR does well when tested at the same SNR (3% error) but performs rather badly if the test noise level deviates in either direction (13% error for clean test data, 33% for 0dB SNR). The classifiers trained on very low SNRs ($-12$ and $-18$dB) give the best results for similarly noisy test conditions but perform very poorly in testing at low noise levels.

This can now be contrasted with the results for a classifier based on acoustic waveform data. One observes that although the performance of this classifier on clean data (7.5% error) is worse than that obtained by PLP classifier trained on clean data, it is significantly more robust to larger test noise levels as compared to the PLP classifier. For instance, there is no significant change in classification error (8%) up to a test noise level as high as 0dB SNR, whereas at the same SNR the corresponding PLP classifier trained on clean data has an error rate of 78%. It should be emphasized that best performance using acoustic waveform classifiers is obtained when training is performed on clean data; training on noisy data (results not shown) leads to poorer performance. This is a significant advantage: the acoustic waveform classifier can be trained once and for all on clean data and used with a broad range of test noise conditions; for the PLP classification, on the other hand, separate classifiers trained at various noise levels need to be constructed to give good performance.

In order to compare the best classification performance in both domains i.e. classification in the waveform domain with test points normalized to $\sqrt{1+\sigma^2}$ and in the PLP domain under matched training and testing conditions, we assumed to have the knowledge about the noise variance, $\sigma^2$. However, in practice, $\sigma^2$ may not be estimated accurately. This may result in a certain amount of mismatch

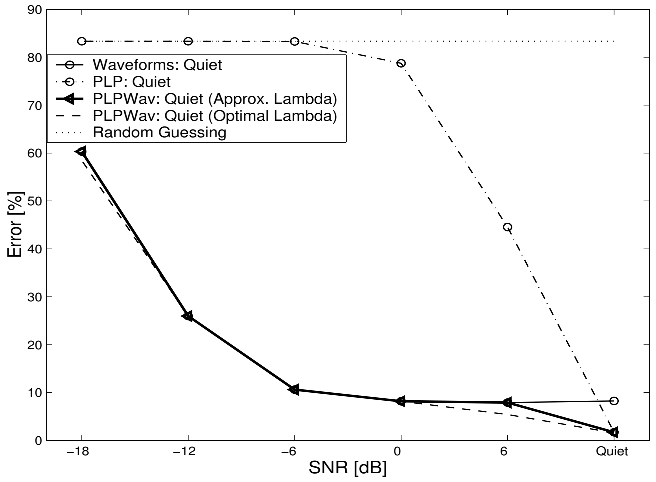

Figure 5: Comparison of classification performance in feature space of PLP, acoustic waveforms and their combined classifier under quiet training conditions. For the combined classifier, results for both optimal $\lambda$ and $\tilde{\lambda}$ are shown as indicated in the legend.

between the actual and estimated SNRs for waveform classification and correspondingly, a mismatch between training and testing conditions in the PLP domain. In Figure 2, waveform classification results are presented for different values of actual and estimated SNRs with possible mismatch between them. In order to analyze the sensitivity of both PLP and waveform classifiers to noise mismatch, we compare Figure 2 with plots from Figure 1 corresponding to PLP classifiers trained and tested under different conditions (dash-dotted curves). It is clear that waveform classification, although not as accurate as PLP classification under matched conditions, shows more robustness towards mismatch between estimated and actual SNRs at low noise levels (SNR $\geq$ 0dB). In severe noisy conditions (SNR $\leq -6$dB), the waveform classifiers outperform PLP classifiers both under matched and mismatched conditions.

In Figure 1, we finally compare classification in acoustic waveform domain with training done on clean data and classification in the PLP domain with training under matched conditions (the bold curves). We observe that classifiers in the waveform domain give better results for high noise (SNR $\leq -6$dB) whereas classification in the PLP domain under matched conditions performs excellently on clean data and at low noise levels (SNR $\geq$ 0dB). In the next section, we propose a method for combined classification in the PLP and acoustic waveform domains resulting in a classifier that is analogous to the best of both domains across all SNRs.

## 4. COMBINED PLP–ACOUSTIC WAVEFORM CLASSIFIER

Consider two sets of $L$ binary classifiers – one in acoustic waveform domain and other in the PLP domain. Let $\bar{h}_P(x_P) = [h_{P1}(x_P), \cdots, h_{PL}(x_P)]$ and $\bar{h}_W(x_W) = [h_{W1}(x_W), \cdots, h_{WL}(x_W)]$ be the vectors of decision values of classifiers in the PLP and acoustic waveform domain respectively where $x_P$ and $x_W$ are the PLP and acoustic waveform representations of phoneme $x$. We consider a convex combination of the decision values of the classifiers in the individual feature spaces, $\bar{h}_P(x_P)$ and $\bar{h}_W(x_W)$, to obtain the decision values of the combined binary classifiers, $\bar{h}_C(x) = [h_{C1}(x), \cdots, h_{CL}(x)]$ i.e.

$$\bar{h}_C(x) = \lambda(\sigma^2)\bar{h}_W(x_W) + \left[1-\lambda(\sigma^2)\right]\bar{h}_P(x_P)\,, \qquad (6)$$

where $\lambda(\sigma^2)$ is parameter which needs to be selected, depending on the noise variance, to achieve optimal performance. These binary classifiers are combined via ECOC methods described previously for multiclass classification. Two different scenarios for training are considered for this kind of combined classifier:

- Both acoustic waveforms and PLP classifiers trained on clean data,
- Acoustic waveform classifiers trained on clean data, PLP classifiers trained under matched conditions.

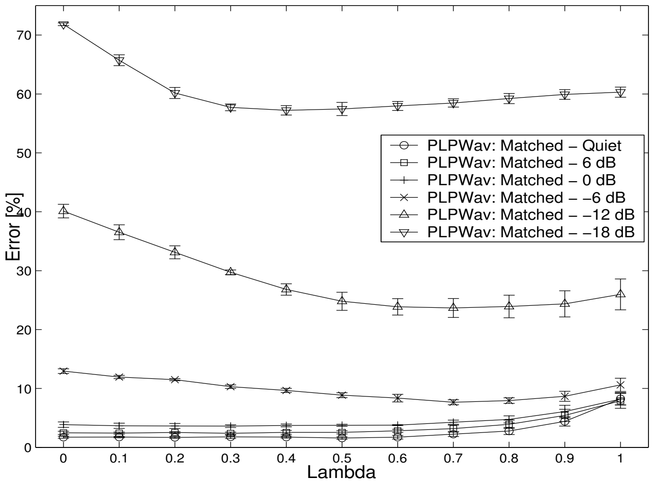

Figure 6: Multiclass classification error of the combined classifier: waveform classifiers trained on clean data combined with PLP classifiers trained under matched conditions.

Since training the acoustic waveform classifiers on noisy data does not give good results, that scenario is therefore not considered. For classification in the PLP domain in presence of noise, various noise compensation techniques have been developed. The two training approaches investigated here are extreme cases of noise compensation: training on clean data amounts to having no noise compensation at all whereas training under matched conditions is analogous to optimal noise compensation [9]. If some practical noise compensation method is used, the classification performance can be expected to fall between these two cases.

Figure 3 shows the performance of the combined classifier with training done in quiet conditions. Results are shown for different values of $\lambda(\sigma^2)$. When testing is done on clean data, the optimal performance is achieved with $0 \leq \lambda(\sigma^2) \leq 0.7$. However, $\lambda(\sigma^2) = 1$ gives the best results for test phonemes with SNR$\leq$ 6dB. This is due to the fact that the PLP classifiers are very sensitive to noise mismatch while waveform classifiers can tolerate a significant mismatch between training and testing conditions.

In Figure 4, 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 $(\%) + 2\%$. The dashed line is an approximation of the optimal $\lambda(\sigma^2)$ and is given by

$$\tilde{\lambda}(\sigma^2) = c + \frac{1-c}{1+e^{10log_{10}(\sigma_o^2/\sigma^2)}}\,, \qquad (7)$$

with $c = 0.5$ and $\sigma_o^2 = 1/8$.

In Figure 5, we compare the classification performance in the feature space of PLP and acoustic waveforms with the combined classifier under quiet training conditions. Two combinations are presented: for optimal $\lambda(\sigma^2)$ (the values of $\lambda(\sigma^2)$ which minimizes the test error) and for $\tilde{\lambda}(\sigma^2)$ selected according to (7). One can observe that for both choices of $\lambda$, the combined classifier performs as the better of the individual classifiers and the difference between optimal $\lambda$ and $\tilde{\lambda}$ is not significant.

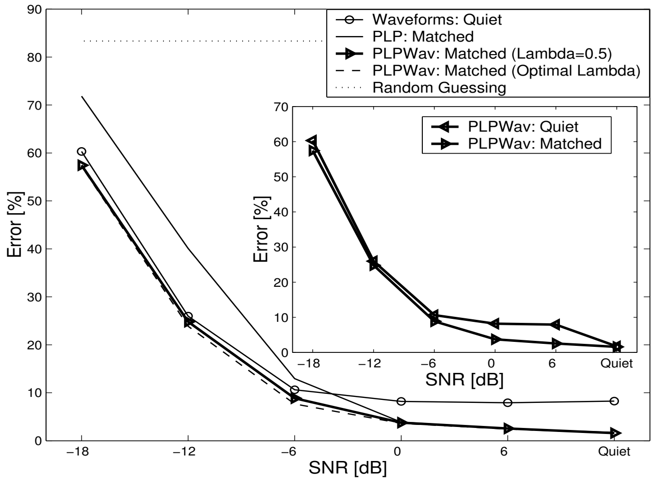

Figure 7: Comparison of classification performance in feature space of acoustic waveforms (trained in quiet conditions), PLP (trained under matched conditions) and their combined classifier. Inset: Comparison of combined classifiers under quiet and matched training conditions.

Now, we consider the scenario when acoustic waveform classifiers trained on clean data are combined with PLP classifiers trained under matched conditions. Figure 6 shows the performance of the combined classifier for different values of $\lambda(\sigma^2)$. It is evident that $\lambda(\sigma^2) = 0.5$ gives close to optimal performance for all test conditions. The comparison between individual and combined classifiers is shown in Figure 7. For clean data and low noise levels (SNR $\geq$ 0dB), the combined classifier exhibit similar performance to that of PLP classifier under matched conditions. However, under severe noisy conditions, the performance of the combined classifier is similar to that of the acoustic waveform classifier. Furthermore, no significant difference is observed between the optimal values of $\lambda(\sigma^2)$ and $\lambda(\sigma^2) = 0.5$. The inset of Figure 7 compares the performance of the combined classifiers under quiet and matched conditions. This clearly demonstrates that the combined classifiers are significantly desensitized to mismatch between training and testing conditions. It should be noted that the combined classifier consistently achieves superior performance than classifiers in either of the individual feature spaces for all SNRs.

## 5. CONCLUSIONS

The robustness of phoneme classification to additive white Gaussian noise in the PLP and acoustic waveform domains was investigated using SVMs. While PLP representation facilitates very accurate recognition of phonemes under matched conditions (especially for clean data), its performance suffers severe degradation with noise mismatch between training and testing conditions. On the other hand, the high-dimensional acoustic waveform representation, although not as accurate as PLP classification on clean data, is more robust to additive noise. Our results demonstrate that a convex combination of classifiers has a performance equivalent to the best of both domains. Moreover, the combined classifier can tolerate a significant mismatch between training and testing conditions. In future work, we plan to investigate the classification performance for a larger phoneme set and different types of noise. It will also be interesting to study different kernel functions which are finely tuned to the physical properties of speech data as that will play a crucial role in reducing the error significantly.

## REFERENCES

[1] R. Lippmann, "Speech Recognition by Machines and Humans," *Speech Comm.*, vol. 22, no. 1, pp. 1–15, 1997.

[2] J. Sroka and L. Braida, "Human and Machine Consonant Recognition," *Speech Comm.*, vol. 45, no. 4, pp. 401–423, 2005.

[3] G. Miller, G. Heise, and W. Lichten, "The Intelligibility of Speech as a Function of the Context of the Test Materials," *J. of Exp. Psychology*, vol. 41, pp. 329–335, 1951.

[4] 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.

[5] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech ," *J. of the Acous. Soc. of America*, vol. 87, pp. 1738–1752, Apr. 1990.

[6] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[7] F. Zheng, G. Zhang, and Z. Song, "Comparison of Different Implementations of MFCC," *Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, Sept. 2001.

[8] R. Rose, "Environmental Robustness in Automatic Speech Recognition," *Robust2004 - ISCA and COST278 Workshop on Robustness in Conversational Interaction*, Aug 2004.

[9] M. Gales and S. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. on Speech and Audio Processing*, pp. 352–359, Sept. 1996.

[10] J. Yousafzai, M. Ager, Z. Cvetković, and P. Sollich, "Discriminative and Generative Machine Learning Approaches towards Robust Phoneme Classification," *ITA - Workshop on Information Theory and Applications*, Jan. 2008.

[11] M. Ager, Z. Cvetković, P. Sollich, and B. Yu, "Toward Robust Phoneme Classification: Augmentation of PLP Models with Acoustic Waveforms," *EUSIPCO 2008*, Aug. 2008.

[12] A. Ganapathiraju, J. Hamaker, and J. Picone, "Applications of Support Vector Machines to Speech Recognition," *IEEE Trans. on Signal Proc.*, vol. 52, no. 8, pp. 2348–2355, 2004.

[13] T. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *J. of Artificial Intelligence Research*, vol. 2, pp. 263–286, 1995.

[14] M. Müller, R. Keiser, A. Nealen, M. Pauly, M. Gross, and M. Alexa, "'Point Based Animation of Elastic, Plastic and Melting Objects," in *Proc. of the ACM SIGGRAPH/EUROGRAPHICS Symposium on Computer Animation*, Aug 2004.

[15] W. Fisher, G. Doddington, and K. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," *Proc. DARPA Speech Recognition Workshop*, pp. 93–99, 1986.

[16] D. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, Online web resource.