TOWARDS ROBUST PHONEME CLASSIFICATION: AUGMENTATION OF PLP MODELS WITH ACOUSTIC WAVEFORMS

Matthew Ager[1], *Zoran Cvetković*[2], *Peter Sollich*[1] *and Bin Yu*[3]

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

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

ABSTRACT

The robustness of classification of phoneme segments using generative classifiers is investigated for the PLP and acoustic waveform speech representations in the presence of white Gaussian noise. We combine the strengths of both representations, specifically the excellent classification accuracy of PLP in quiet conditions with the additional robustness of acoustic waveform classifiers. This is achieved using a convex combination of their respective log-likelihoods to produce a combined decision function. The resulting combined classifier is uniformly as accurate as PLP alone and is significantly more robust to the presence of additive noise during testing. Issues of noise modelling and time-invariant classification of acoustic waveforms are also considered with initial solutions used to improve accuracy.

1. INTRODUCTION

One of the key problems in automatic speech recognition (ASR) is robustness to additive noise. ASR systems can attribute much of their performance to language and context modelling, the principle being that classification errors made by the front-end can be remedied at a higher level [8]. However, this approach can only decode messages sent via speech signals if the input sequence of elementary speech units is sufficiently accurate. In the extreme case where that sequence is close to random guessing no useful information can be extracted at the later stages of recognition. Developing methods for robust classification of phonemes and isolated syllables is therefore essential for robust continuous speech recognition. Indeed, it has been observed that the majority of inherent robustness of human hearing occurs early in the process [7]; even at -18dB SNR humans can still recognise isolated speech units above the level of chance [6]. It is crucial for an automatic speech classifier to achieve performance close to that of the human auditory system even in such severe noise conditions.

The current preferred speech representation is generally some variant of Perceptual Linear Prediction (PLP) [4], Relative Spectral Transform - PLP (RASTA-PLP) [5] or mel-frequency cepstral coefficients (MFCC) [12] dependent on the particular task. These representations are derived from the short term magnitude spectra followed by non-linear transformations that model human auditory perception. They have the advantage that they remove such variation from test signals that is deemed unnecessary for recognition and have a much lower dimension than acoustic waveforms which can allow for more accurate modelling when data is limited. It is not known if this dimensionality reduction loses some information that gives speech additional robustness. An alternative approach that can be used to explore this possibility is to use higher dimensional representations, in particular acoustic waveforms.

In this paper we investigate the robustness of phomene classification in the acoustic waveform domain when additive white Gaussian noise is present. Regularised Gaussian mixture models in the form of mixture of probabilistic principial component analysers [10] are used to give preliminary classification results. For comparison and later combination, we also train and test classifiers on the lower dimensional PLP representation. The main aim of this work is not to find optimal classifiers but to instead illustrate that acoustic waveforms are a viable representation for speech and can be used to improve the overall robustness of phoneme classification.

Many noise compensation methods have been proposed to reduce the effect of noise on spectral representations [9]. However those methods perform no better than the matched condition approach [2]. i.e. training and testing in the same noise conditions. Throughout this paper we use no noise compensation of PLP feature vectors. Instead we consider the following two cases for the testing setup: one being where only PLP models that are trained in quiet conditions are used and the other where PLP models trained on noise conditions that match test conditions are available. In both cases we assume that the noise level is known or can be estimated reliably [11]. These two cases represent the extremes of the classifier performance and it is expected that the performance of a noise compensated PLP classifier would be between the two.

It would be very difficult to improve on the accuracy of PLP in quiet conditions, hence the focus of this work is to demostrate that acoustic waveforms can be used to improve the robustness of a PLP classifier in the presence of additive noise. This is achieved by taking the decision functions to be a convex combination of the log-likelihood functions of the respective PLP and acoustic waveform density models. This gives a one-parameter family of combined classifiers corresponding to PLP classification alone at one extreme and acoustic waveform classification alone at other. When the combination parameter varies as a function of SNR, the performance of the derived classifier is at least as accurate as a PLP classifier trained in conditions that match those in testing and is significantly more robust to additive noise.

2. GENERATIVE CLASSIFICATION

Generative classification was performed using density estimates derived from mixtures of probabilistic principal component analysis (MPPCA) [10]. This method gives a Gaussian mixture model where the covariance matrices of each component are regularised by replacement with a lower rank q approximation:

$$\mathbf{C} = r^2\mathbf{I} + \mathbf{W}\mathbf{W}^T \tag{1}$$

where the i^{th} column of the $d \times q$ matrix $\mathbf{W}$ is given as $\sqrt{\lambda_i}v_i$ corresponding to i^{th} eigenvalue λ_i and eigenvector v_i of the empirical covariance matrix with the eigenpairs in descending order. The regularisation parameter r^2 is then taken as the mean of the remaining $d-q$ eigenvalues:

$$r^2 = \frac{1}{d-q}\sum_{q+1}^{d}\lambda_i \tag{2}$$

This method is used to give maximum likelihood density estimates of the class conditional distribution for each phoneme. In order to use these density models for the classification of a data point x, a decision function is required, here it is given by the log-likelihood function $\mathscr{L}(x)$ is defined as the logarithm of the density of the c-component mixture evaluated at x:

$$\mathscr{L}(x) = \log\left(\sum_{i=1}^{c}\frac{w_i}{(2\pi)^{\frac{d}{2}}|\mathbf{C}_i|^{\frac{1}{2}}}e^{-\frac{\langle x-\mu_i,\mathbf{C}_i^{-1}(x-\mu_i)\rangle}{2}}\right) \tag{3}$$

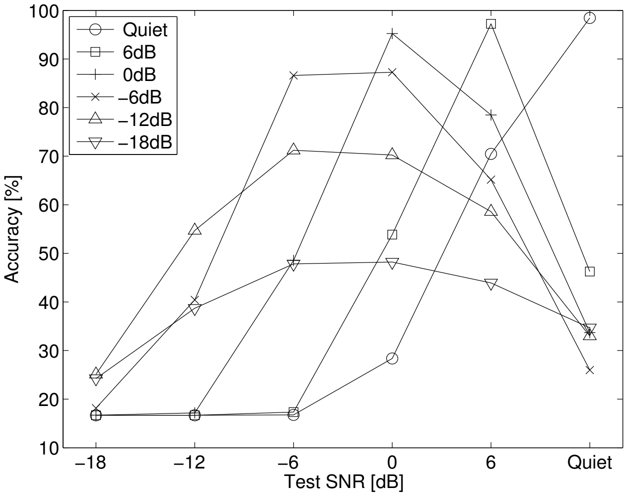


Figure 1: Multiclass accuracy of PLP classifiers as a function of test SNR. Each curve shows the accuracy of the classifier trained at the corresponding SNR indicated by the curve marker. The curves show the sensitivity of PLP classifiers when there is a mismatch between training and testing noise conditions. In particular the classifier trained at 0dB performs much worse when the test noise level is lower than the training level.

where $\mathbf{C}_i$, μ_i and w_i are the covariance matrix, mean and mixture weight of the i^{th} component respectively. Given this function, classification is then performed in the standard way, by predicting the class with the maximum log-likelihood $\mathscr{L}^{(k)}(x)$ (which implicitly assumes uniform prior probabilities over different classes). The classification function, $H(x)$, that maps a test point x to one of the corresponding K class labels is defined as:

$$H(x) = \arg\max_{k=1,\ldots,K} \mathscr{L}^{(k)}(x) \tag{4}$$

The same type of modelling is used both for PLP and acoustic waveforms. One of the advantages of the acoustic waveform representation is that the fitted density models can easily be modified to account for the presence of additive noise during testing. Assuming that the noise level (or more generally the noise power spectrum) is known or can be estimated reliably, we simply need to perform a convolution with the appropriate Gaussian noise model. For example, the covariance matrix of white Gaussian noise is a multiple of the identity matrix, $\frac{\sigma^2}{d}\mathbf{I}$, where σ^2 is the noise variance and d is the dimension of the data. Hence the noise-adapted model in that case is given by replacing each covariance matrix $\mathbf{C}_i$ with $\tilde{\mathbf{C}}_i(\sigma^2)$:

$$\tilde{\mathbf{C}}_i(\sigma^2) = \frac{\mathbf{C}_i + \frac{\sigma^2}{d}\mathbf{I}}{1+\sigma^2} \tag{5}$$

where $1+\sigma^2$ is the normalisation factor to ensure that the noise-adapted covariance matrix has unit trace. For classification of noisy acoustic waveforms this type of noise modelling is used. In contrast, PLP is a non-linear transformation and it is not possible to model noise directly in this manner. Noise-compensated modelling of PLP distributions is currently an active area of research [9]. Since the main aim of this paper is to assess the sensitivity of classification in the PLP domain to additive Gaussian noise, we do not perform any noise compensation but instead consider two extreme cases: training only on quiet data and training in matched noise conditions. These two scenarios cover the expected range of performance of PLP classifiers using optimal noise-compensation.

Another difference between the two domains is sensitivity to time alignment. PLP is not sensitive to small variations in time alignment as it uses frames of short-term magnitude spectra. In the case of waveforms however it would clearly be beneficial to align the data in a consistent manner. This is especially true in the case of stops such as /b/ and /t/. In general time alignment in the acoustic waveform domain is a not a well-defined problem. Therefore, rather than attempting to explicitly align the acoustic waveforms, a sliding window was used to give a number of shifted versions of the test point. These shifted versions are used during training and testing. During testing the log-likelihood of the test point x is instead given by $\mathscr{L}_s(x)$, the log-mean-likelihood when the mean is taken over the shifted versions of x:

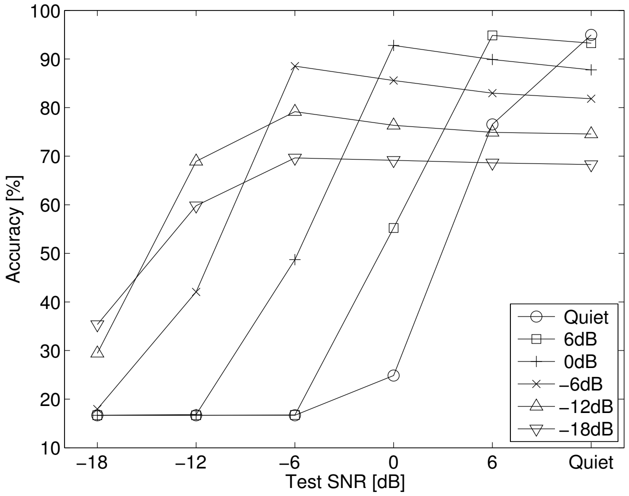


Figure 2: Multiclass accuracy of acoustic waveform classifiers as a function of test SNR. The curve marker indicates the adapted SNR of the classifier given by (5).

$$\mathscr{L}_s(x) = \log\left(\frac{1}{2n+1}\sum_{p=-n}^{n}\exp(\mathscr{L}(x^{p\Delta}))\right) \tag{6}$$

where Δ is the shift increment, $[-n\Delta, n\Delta]$ is the shift range, and $x^{p\Delta}$ denotes a time-shifted versions 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.

These modified log-likelihoods are compared among the different classes to produce the classification. The shift range was selected so that it would cover at least one fundamental period of a periodic waveform at the lower end of the typical frequency range of speech. We experimented with sample shifts of below 10 samples in the same shift range ±100, giving a greater number of shifted waveforms. Since this gave no noticeable improvement but increases computation time and memory requirements, all tests were carried out using the shifts in steps of 10 samples.

	Original Data	Shifted Data
/b/	-3.178 ± 0.029	-3.396 ± 0.010
/f/	-0.655 ± 0.016	-2.091 ± 0.007
/m/	-3.212 ± 0.009	-3.298 ± 0.009
/r/	-3.305 ± 0.004	-3.363 ± 0.007
/t/	-2.066 ± 0.044	-2.524 ± 0.021
/z/	-1.732 ± 0.048	-2.378 ± 0.021

Table 1: Standardised log-loss of the test data for acoustic waveform models in quiet conditions. The lower values in the right hand column indicates that the inclusion of shifted data results in more accurate density models.

The inclusion of these shifted versions of the acoustic waveforms has the additional benefit of increasing the size of the training dataset. Provided the shifted versions are sufficiently independent, this larger dataset should improve the model fit and hence classification accuracy. In particular the accuracy of the acoustic waveform classifier was improved from 89.2% to 95.1% when shifted versions are used during training and testing. In order to explain this improvement we consider the log-loss of the test data when shifted versions of the data is included and compare the values to when it is not present. The log-loss is a measure of how well a particular probability density fits the data with a lower value indicates a better model. The values given in Table 1 show that the inclusion of shifted versions of the data lead to a density estimates that more accurately model the phomene class distributions.

2.1 RESULTS OF CLASSIFICATION IN PLP AND ACOUSTIC WAVEFORM DOMAINS

In this preliminary study we consider only realisations of six phonemes (/b/, /f/, /m/, /r/, /t/, /z/) that were extracted from the TIMIT database [3]. This set includes examples from fricatives, nasals, semivowels and voiced and unvoiced stops. In addition, those classes provide pairwise discrimination tasks of a varying level of difficulty. A single 64ms rectangular window was then applied to the centre of each phoneme, except for /b/ and /t/ where the window was positioned to include the closure and release. The natural space in which to perform classification for the waveforms is the hypersphere $\mathbb{S}^{1023}$ as each sample has 1024 entries and unit norm. In addition the mean value of each class was zero within sampling error, the class-conditional densities for the acoustic waveform models were constrained to have zero mean.

Each phoneme class consists of approximately 1000 representatives, of which 80% were used for training and 20% for testing; error bars were derived by considering five different such splits and give an indication of the significance any differences in the accuracy of two classifiers. A range of SNRs was chosen to show classification accuracies that approached chance level, i.e. 16.7% in the case of six classes. In total this gave six testing and training conditions; -18dB, -12dB, -6dB, 0dB, 6dB and quiet.

For comparison the default 12th order PLP cepstra were computed for the 64ms segments. A sliding 25ms Hamming window was used with a overlap of 15ms leading to 4 frames of 13 coefficients [1]. These 4 frames were concatenated to give a PLP representation in $\mathbb{R}^{52}$. The data was then standardised prior to training so that each of features had zero mean and unit variance when the entire training set was considered.

The PLP phoneme distributions were modelled using a single component mixture with a principal dimension of 40, i.e. $c = 1$ and $q = 40$. We experimented with other parameters but only show the best results here. Figure 1 shows the test results for classifiers trained on data corrupted at the corresponding noise levels. Each of the curves represents a different training SNR. It is clear that PLP classifiers are highly sensitive to mismatch between training and testing noise conditions. For example, when training and testing conditions are matched at 6dB SNR accuracy is very high at 97.2% however if the same classifier is tested in quiet conditions this value falls to 46.3%. The analogous plot for waveform classifiers is shown in Figure 2, where the phoneme classes were modelled with $c = 4$ and $q = 500$. Acoustic waveform classifiers are less sensitive to differences between training and testing conditions. Taking the 6dB classifier as an example again we see that if training and testing conditions are matched the accuracy is 94.9% and when tested in quiet it remains high at 91.6%. Although the performance for matched conditions is lower than that of PLP at this noise level, the decrease due to mismatch between training and testing conditions is much less. It should be stressed that the waveform models are only trained in quiet conditions and then adapted appropriately according to (5).

The other scenario considered in this study was the case when training and testing conditions were matched which is equivalent to taking the upper envelopes of Figures 1 and 2, these are shown in Figure 3. In this case PLP gives greater accuracy than waveforms down to 0dB SNR where the situation reverses. Given these results we seek to combine the classification strengths of each representation, specifically the high accuracy of PLP classifiers at high SNRs and the robustness of acoustic waveform classifiers at all noise levels. Ideally this will result in a single combined classifier that only needs to be trained in quiet conditions and can be easily adapted to a range of noise conditions.

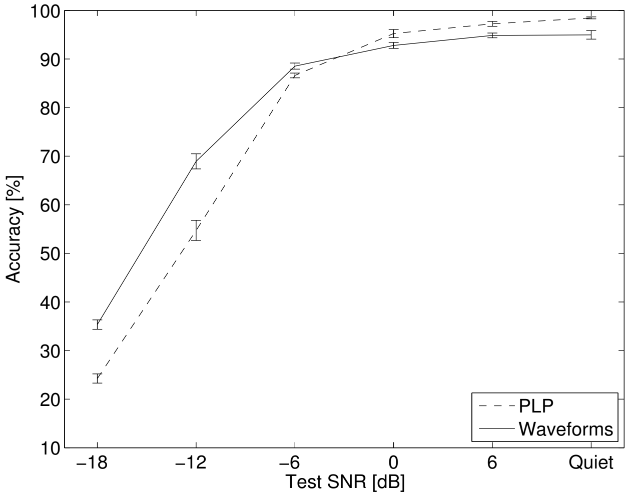


Figure 3: Multiclass accuracy of PLP and waveform classifiers as a function of test SNR. Here PLP performance is greater than acoustic waveforms where the SNR is above 0dB. Below that value however waveforms are significantly more accurate. These results suggest that a combined classifier could be more accurate.

3. COMBINATION OF CLASSIFIERS

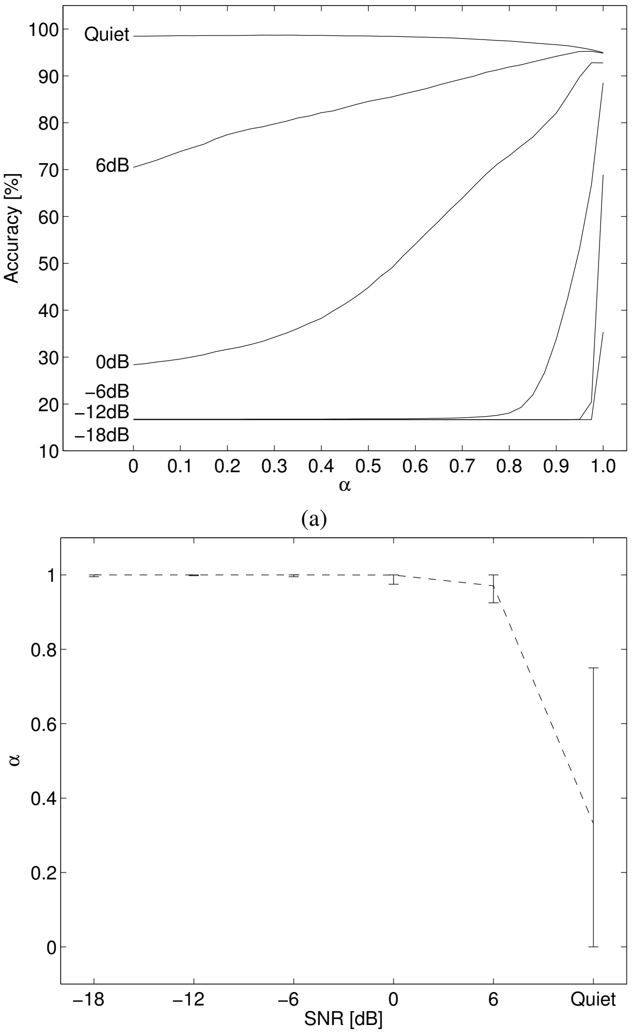
The results shown in Figure 3 suggest that it could be possible to construct a combined classifier that is at least as accurate as the better of those two across all SNRs. To investigate this concept we consider the following convex combination of the two log-likelihoods with each term being standardised by the relevant representation dimension. Let $\mathscr{L}^{(k)}_{\text{plp}}(x)$ and $\mathscr{L}^{(k)}_{\text{wave}}(x)$ be the log-likelihoods of a point x for the k^{th} class, then the combined log-likelihood $\mathscr{L}^{(k)}_{\alpha}(x)$ parameterised by α is given as

$$\mathscr{L}^{(k)}_{\alpha}(x) = \frac{(1-\alpha)}{d_{\text{plp}}}\mathscr{L}^{(k)}_{\text{plp}}(x) + \frac{\alpha}{d_{\text{wave}}}\mathscr{L}^{(k)}_{\text{wave}}(x) \qquad (7)$$

where $d_{\text{plp}} = 52$ and $d_{\text{wave}} = 1024$ are the dimensions of the PLP and acoustic waveform representations respectively. We would expect α to be almost zero for high SNRs and close to one for low SNRs in order to give the desired improvement in accuracy. The combined classifier function $H_\alpha(x)$ has the same form as (4) and is defined as:

$$H_\alpha(x) = \arg\max_{k=1,\ldots,K} \mathscr{L}_{\alpha}{}^{(k)}(x) \qquad (8)$$

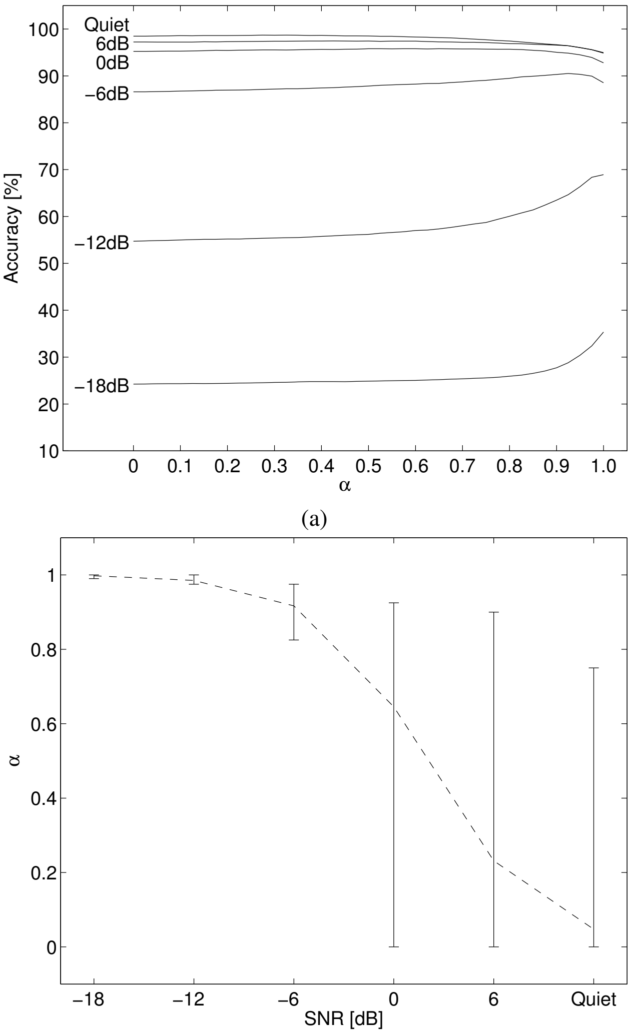
We then investigated the effect of varying the combination parameter α on the classification accuracy. The ranges of α that give good performance when using the combined log-likelihood, $\mathscr{L}_\alpha(x)$, are show in Figures 4(b) and 5(b), where the errors bars indicate the values of α that give classification accuracy within 1.0% of the maximum. Figure 4 shows the results of combining the PLP model trained in quiet conditions with noise-adapted acoustic waveform models. When this classifier is tested in quiet conditions the range of suitable α is large, however if noise is present the accuracy is



(a)

(b)

Figure 4: (a) shows the result of combining a PLP classifier trained in quiet conditions with the appropriate noise-adapted acoustic waveform classifier. The six test SNRs have been plotted, indicated by the curve labels. The highest accuracy is obtained for noisy test conditions when $\alpha = 1$. (b) gives the range of α that gives a classification accuracy within 1.0% of the maximum value. The dashed curve shows a function of the form (9) to fit the ranges indicated by the error bars. Here the parameter of the combination function are $\sigma_0^2 = 11$dB and $\beta = 0.7$.



(a)

(b)

Figure 5: (a) Accuracy of the classifier conbining the PLP classifiers that are trained in noise conditions that match test conditions with noise-adapted waveform classifiers. The six test SNRs have been plotted, indicated by the labels. For high SNRs, a value of α close to zero gives the best results. When SNR is low, α close to one is preferable. (b) shows the range of α that gives a combined classifier accuracy within 1.0% of the maximum. In particular the range of suitable α for high SNRs is large. The dashed curves show a possible function of the form (9) to fit that range, with $\sigma_0 = 2$dB and $\beta = 0.3$.

more sensitive to the choice of α, hence the fit of the combination function at low SNRs is important. Figure 5 is an analogous plot for the PLP models trained and tested on matched conditions showing a large range for higher SNRs but again very narrow for low SNRs.

We use this information to fit a combination function, $\alpha(\sigma^2)$. As the range of suitable α is large the particular form of this combination function is not critical, so we choose the following sigmoid function with two parameters σ_0^2 and β.

$$\alpha(\sigma^2) = \frac{1}{1 + e^{\beta(\sigma_0^2 - \sigma^2)}} \tag{9}$$

Suitable curves fitted to the results are shown in Figure 4(b) where $\sigma_0^2 = 11\text{dB}, \beta = 0.7$ for classifiers trained in quiet conditions. The equivalent results for PLP models trained in matched noise conditions that match the test conditions combined with noise-adapted acoustic waveform models are shown in Figure 5(b), with $\sigma_0^2 = 2\text{dB}, \beta = 0.3$ giving good results.

3.1 RESULTS OF COMBINED CLASSIFICATION

The accuracy of the combined classifier using models trained in quiet conditions is shown as the bold curve in Figure 6, with results of the individual PLP and acoustic waveform classifiers also plotted for comparison. In quiet conditions the combined classifier is as accurate as PLP alone, corresponding to $\alpha = 0$. When noise is present the combined classifier at least as accurate as the acoustic waveform classifier alone and significantly around −6dB SNR. The improvement observed in that range of SNRs justifies using the smooth combination function given by (9) rather than a function that simply switches from PLP classification to acoustic waveform classification when the SNR is below a given threshold.

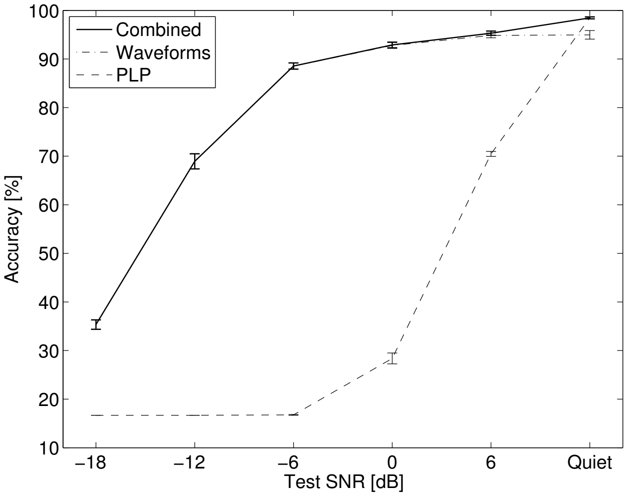


Figure 6: Performance of combined classifier for the case where only PLP models trained in quiet conditions are used. Here the accuracy in quiet test conditions is equivalent to using PLP. When noise is present the accuracy is similar to that for the noise adapted waveform models alone, with an improvement at 6dB SNR.

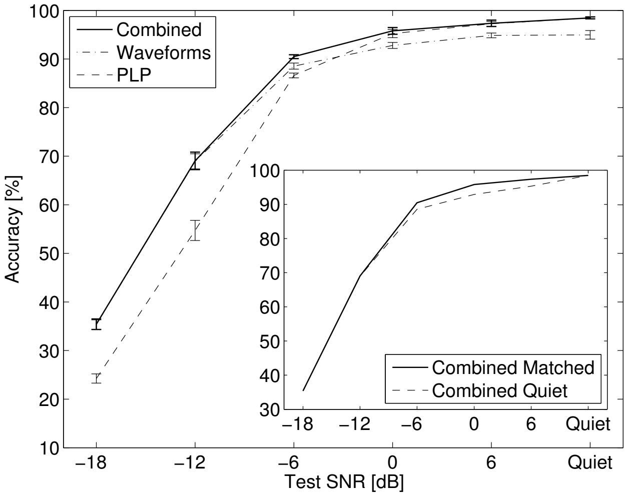


Figure 7: Performance of combined classifier when PLP models trained under matched conditions are used. The combined classifier is uniformly as accurate as those it is derived from and gives significant improvement at -6dB SNR. Inset is a comparison of the combined classifer trained only in quiet conditions from Figure 6.

The other scenario considered is when PLP models training on noise conditions that match those during testing are used for combination, as shown in Figure 7. Again the combined classifier is as least as accurate as the better of the two individual classifiers with a significant improvement at -6dB SNR. The combined classifier achieves 90.5% at -6dB SNR compared with 88.5% for waveforms and 86.6% for PLP alone at the same SNR. Even greater improvement over PLP alone can be seen at -12dB and -18dB with increases from 54.7% and 24.2% to 69.0% and 35.4% respectively.

The results have shown that the combination with acoustic waveforms has improved the classification accuracy of PLP classifiers alone significantly for SNRs below 0dB. We would expect the improvement to be even more significant if noise-compensated models are used for combination as typically they will be less accurate than for model trained on noise conditions that match those in testing. The inset of Figure 7 shows the expected range of accuracy for combination with noise-compensated PLP models.

4. CONCLUSIONS

In this work we have proposed a method to combine speech representations leading to a classifier that is more robust to mismatch between level of additive noise in training and testing, whilst retaining the excellent performance of PLP at high SNR. By considering the two contrasting scenarios of training only on quiet with the results obtained when using matched PLP models, we have been able to measure this improvement conditional on the combination with acoustic waveform classifiers.

To further validate the findings shown here, the experiments will be extended to a larger set of phonemes and larger databases containing more realisations of each phoneme class. We would expect improvement for both representations but especially so for acoustic waveforms due to their high dimensional representation where additional training data would improve density estimation. In addition the noise modelling described in (5) can be generalised to other noise types, we have carried out similar experiments using pink noise and speech-weighted noise with encouraging results.

The convex combination of the log-likelihoods demonstrated here may not be the optimal classifier. It is possible that more general combination functions could lead to an even greater improvement of accuracy. The results have however shown a practical method of combining existing phoneme classifiers in order to exploit their differing accuracy characteristics.

REFERENCES

[1] D. Ellis. PLP and RASTA (and MFCC, and inversion) in Matlab, 2005. online web resource, http://labrosa.ee.columbia.edu/matlab/rastamat/.

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

[3] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, and D. Pallett. The DARPA TIMIT acoustic-phonetic continous speech corpus. NIST. Feb. 1993.

[4] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech . *Acoustical Society of America Journal*, 87:1738–1752, Apr. 1990.

[5] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, Oct. 1994.

[6] G. Miller and P. Nicely. An analysis of perceptual confusions among some English consonants. *Acoustical Society of America Journal*, 27:338–352, 1955.

[7] S. Phatak and J. Allen. Syllable confusions in speech-weighted noise. *Acoustical Society of America Journal*, 121(4):2312–2326, Apr. 2007.

[8] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewoods Cliffs, 1993.

[9] R. Rose. Environmental robustness in automatic speech recognition. *Robust2004 - ISCA and COST278 Workshop on Robustness in Conversational Interaction*, Aug 2004.

[10] M. Tipping and C. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999.

[11] J. Yamauchi and T. Shimamura. Noise estimation using high frequency regions for speech enhancement in low SNR environments. In *Speech Coding, IEEE Workshop*, pages 56–61, Oct 2002.

[12] F. Zheng, G. Zhang, and Z. Song. Comparison of different implementations of MFCC. *Computer Science and Technology*, 16(6):582–589, Sept. 2001.