

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

Matthew Ager¹, Zoran Cvetković², Peter Sollich¹ and Bin Yu³

Department of Mathematics¹ and Division of Engineering²
King's College London, London, WC2R 2LS, UK

Department of Statistics³
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 show a method to 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. Issues of noise modelling and time-invariance of acoustic waveforms are also addressed with initial solutions shown. The resulting combined classifier has greater accuracy than PLP alone and is significantly more robust to the presence of additive noise during testing.

Index Terms— Speech Recognition, Robustness, Generative classification, Acoustic waveforms, PLP

1. INTRODUCTION

One of the key problems in automatic speech recognition 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 [1]. 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 the input 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; even at -18 dB SNR humans can still recognise isolated speech units above the level of chance[2]. The ultimate aim for an automatic speech classifier is to achieve performance close to that of the human auditory system in such severe noise conditions.

The current preferred speech representation is generally some variant of PLP[3], RASTA[4] or MFCC[5]. 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 have the advantage that they remove such variation from test signals as is considered 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 dimensional 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 acoustic waveforms in the presence of additive white Gaussian noise. This is achieved using regularised Gaussian mixture models in the form of mixture of probabilistic PCA[6]. For comparison and later combination, we also test classifiers on PLP representations. The main aim of this work is not to find optimal classifiers but to illustrate that acoustic waveforms can be a viable representation and improve the robustness phoneme classification.

Many noise compensation methods have been proposed to reduce explicitly the effect of noise on spectral representations [7]. However the proposed methods perform no better than the matched condition approach [8]. 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 quiet PLP models are used and the other where PLP models trained on matched noise conditions are available. In both cases we assume the noise level is known or can be estimated reliably. 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 the paper is to illustrate how acoustic waveforms can be used to improve the robustness of a PLP classifier in the presence of additive noise. This is achieved by taking a convex combination of the log-likelihoods of PLP density models with those of a waveform classifier. When the combination parameter is allowed to vary as a function of SNR, the performance of the derived classifier has greater accuracy than PLP alone and is significantly more robust to additive noise.

2. GENERATIVE CLASSIFICATION

Generative classification was performed using density estimates derived from mixtures of Probabilistic PCA (MPPCA) [6]. Probabilistic PCA (PPCA) uses the eigenpairs (v_i, λ_i) of the empirical covariance matrix, with the eigenvalues ordered in decreasing order. To achieve some dimensionality reduction while modelling data with a Gaussian distribution, the empirically estimated covariance matrix is replaced by a lower rank approximation of the form:

$$\mathbf{C} = r^2 \mathbf{I} + \mathbf{W}\mathbf{W}^T \quad (1)$$

where the q columns of \mathbf{W} are given by $\sqrt{\lambda_i} v_i$ for the corresponding index i . r^2 is then taken as the mean of the remaining $d - q$ eigenvalues (where $d_{\text{wave}} = 1024$ for the waveform

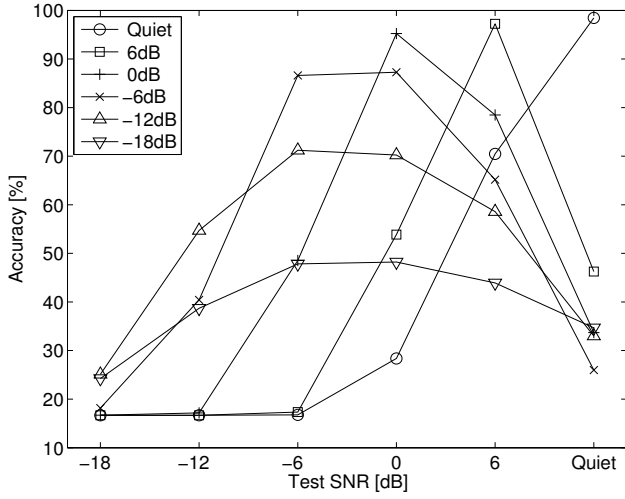


Figure 1: Multiclass accuracy of PLP classifier as a function of test SNR. Each curve shows the accuracy of the classifier trained at the corresponding SNR indicated by the curve marker. The training data was modelled using MPPCA with one component and a principal dimension of 40.

representation and $d_{\text{plp}} = 52$ for PLP):

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

MPPCA represents the class conditional distribution for each phoneme with a mixture of such regularised Gaussians; the model parameters are optimised using the EM algorithm[6], hence a maximum likelihood estimate of the density is achieved. Given a data point x , the log-likelihood function $\mathcal{L}(x)$ is defined in equation 3 as the logarithm of the density of the c -component mixture evaluated at x .

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

where \mathbf{C}_i , μ_i and w_i are the covariance matrix, mean and mixture weight of the i^{th} component. Classification is then performed in the standard way, by predicting the class with the maximum log-likelihood $\mathcal{L}^{(k)}(x)$ (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

$$H(x) = \arg \max_{k=1, \dots, K} \mathcal{L}^{(k)}(x) \quad (4)$$

The same type of modelling is used both for PLP and acoustic waveforms. One of the advantages of the waveform representations is that the fitted density models can easily be modified to allow for the presence of additive noise. 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. When the noise variance is σ^2 and $\tilde{\lambda}_i$ are the eigenvalues of \mathbf{C} , the spectrum of the resulting density model for waveforms corrupted by white noise is given by

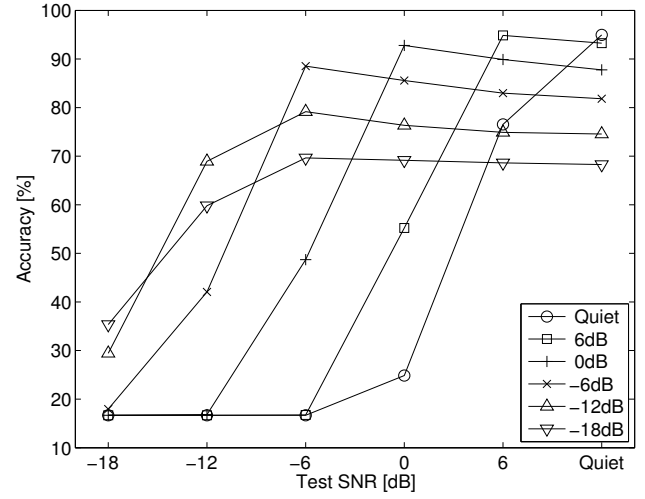


Figure 2: Multiclass accuracy of waveform classifier as a function of test SNR. Each curve shows the accuracy of the classifier when adapted to the SNR indicated by the curve marker. The adaption is achieved using equation 5. Data modelled using MPPCA with four components and principal dimension of 500.

$$\tilde{\lambda}_i(\sigma^2) = \frac{\hat{\lambda}_i + \frac{\sigma^2}{d}}{1 + \sigma^2} \quad (5)$$

For classification of noisy acoustic waveforms this type of noise modelling is used. Since PLP is a highly non-linear transformation it is not possible to model noise in a similar way. Noise compensated modelling of PLP distributions is currently an active area of research[7]. Since the main aim of this paper is to assess the sensitivity of classification in the PLP domain to additive 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.

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/. Rather than attempting to explicitly do this, a sliding window was used to give a number of shifted versions of the test point. The log-likelihood of the test point x is instead taken as $\mathcal{L}_s(x)$, the log-mean-likelihood taken over the shifts:

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

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

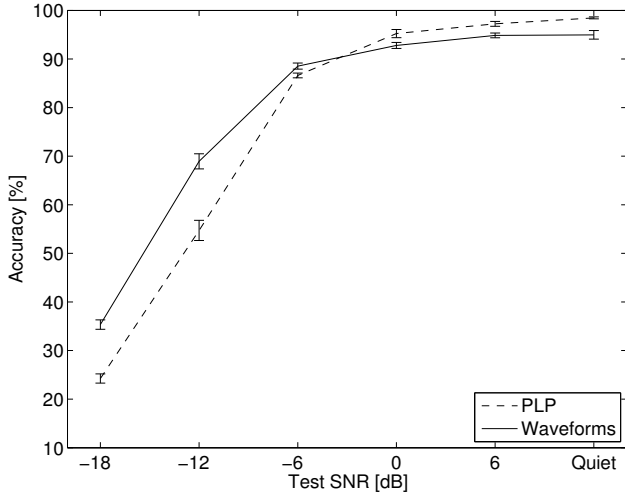


Figure 3: Multiclass accuracy of PLP and waveform classifiers as a function of test SNR. Here PLP performance is greater than acoustic waveforms where the SNR is above 0dB. Below that value however waveforms are significantly more accurate.

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.

2.1 RESULTS OF CLASSIFICATION IN PLP AND ACOUSTIC WAVEFORM DOMAINS

Realisations of six phonemes (/b/, /f/, /m/, /n/, /t/, /z/) were extracted from the TIMIT database[9]. 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 representatives, of which 80% were used for training and 20% for testing; error bars were derived by considering five different such splits. A single 64ms rectangular window was then applied to the data followed by normalisation. 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 were constrained to have zero mean.

For comparison the default 12th order PLP cepstra of the data were taken, leading to 4 frames of 13 coefficients[10]. The 4 frames were concatenated to give a PLP representation in \mathbb{R}^{52} . Finally the SNR range 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.

The PLP classes was 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 of these classifiers trained on data corrupted at the correspond-

ing noise levels. Each of the curves represents a different training SNR. It is clear that PLP is highly sensitive to mismatch between training and testing conditions. For example, when 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 classes were modelled with $c = 4$ and $q = 500$. The waveforms are less sensitive to condition mismatch. Taking the 6dB classifier as an example again we see that for matched conditions the accuracy is 94.9% and when tested in quiet it remains high at 91.6%. Although the matched performance is lower than that of PLP at this noise level, the decrease due to mismatch is much less. It should be stressed that the waveform models are only trained in quiet conditions and then adapted appropriately using equation 5.

Another interesting comparison is to consider the matched conditions which is equivalent to taking the upper envelopes of the two plots, these are shown in figure 3. In this case PLP gives greater accuracy than waveforms down to 0dB SNR where the situation reverses. We seek to combine the strengths of each representation, specifically the high accuracy of PLP classifiers at high SNR and the robustness of waveforms at all noise levels. Ideally this would result in a single classifier that only need be trained in quiet conditions and then easily adapted to the test environment.

3. REPRESENTATION COMBINATION

By considering the results obtained for PLP and waveforms that were trained and tested on matched noise conditions. i.e. the two accuracy curves shown in figure 3, it can be seen that a classifier may give better results if it can choose between the two, dependent on the SNR. We propose a method to achieve this. The following convex combination of the two log-likelihoods can be used with each term being standardised by the relevant representation dimension. Let $\mathcal{L}_{\text{plp}}^{(k)}(x)$ and $\mathcal{L}_{\text{wave}}^{(k)}(x)$ be the log-likelihoods of a point x for the k^{th} . Then the combined log-likelihood $\mathcal{L}_{\alpha}^{(k)}(x)$ parameterised by α is given as

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

where $d_{\text{plp}} = 52$ and $d_{\text{wave}} = 1024$ are the dimensions of the PLP and waveform representations respectively. In this form α should be almost zero for high SNR and close to one for low SNR in order to give the desired improvement in accuracy.

We investigated the effect of varying the combination parameter α on the classification accuracy. The ranges of α that gives good performance when using the combined log-likelihood, $\mathcal{L}_{\alpha}(x)$, are show in figures 4 and 5, where the errors bars indicate the range of α that gives classification accuracy within 1% of the maximum value. The first of the figures shows the results of combining only the quiet PLP model where in quiet conditions the range of suitable α is large, however when noise is present the accuracy is more sensitive to the choice of α . The second figure is a plot for the matched PLP models showing a large range for higher SNRs but again very narrow for low SNR.

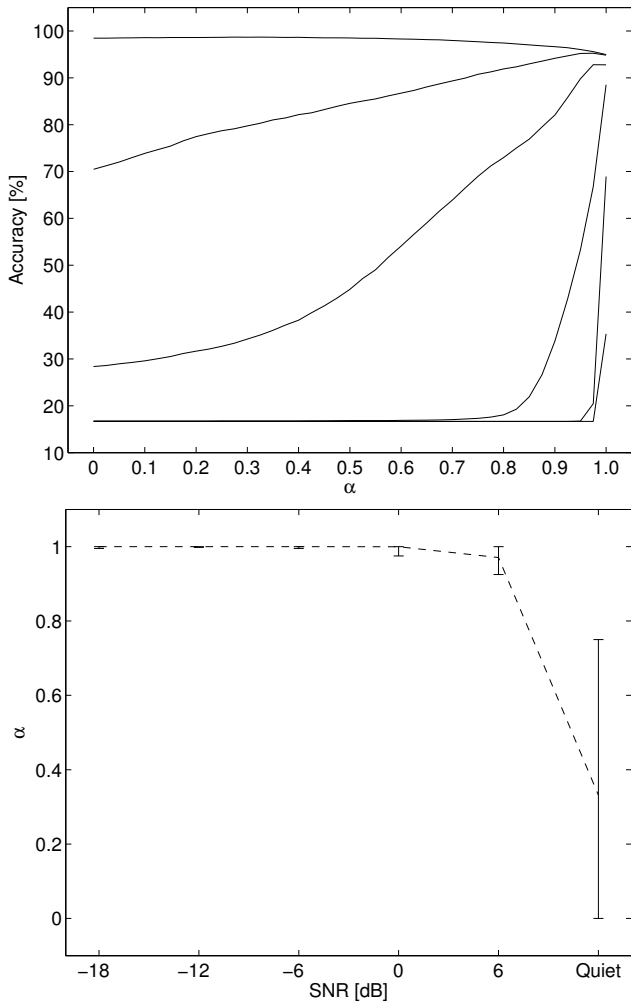


Figure 4: Top plot shows the result of combining a PLP classifier trained in quiet conditions with a noise adapted waveform model. Each of the six testing conditions has been plotted showing that the highest accuracy is obtained when $\alpha = 1$ for noisy test conditions. The bottom plot is the range of α that gives an accuracy within 1% of the maximum achievable. The dashed curve shows a possible function of the form in equation 8 to fit the range, with $s_0 = 11\text{dB}$ and $\beta = 0.7$

We use this information to fit a combination function for α that varies with SNR, denoted here as s . 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 s_0 and β .

$$\alpha(s) = \frac{1}{1 + e^{\beta(s-s_0)}} \quad (8)$$

Suitable curves fitted to the results are shown in figure 4 where $s_0 = 11\text{dB}$, $\beta = 0.7$ for quiet training conditions. The equivalent results for matched conditions shown in figure 5, $s_0 = 2\text{dB}$, $\beta = 0.3$ gives the good results.

3.1 COMBINED CLASSIFIER RESULTS

The accuracy of the combined classifier using quiet models is shown as the bold curve in figure 6, with results of PLP

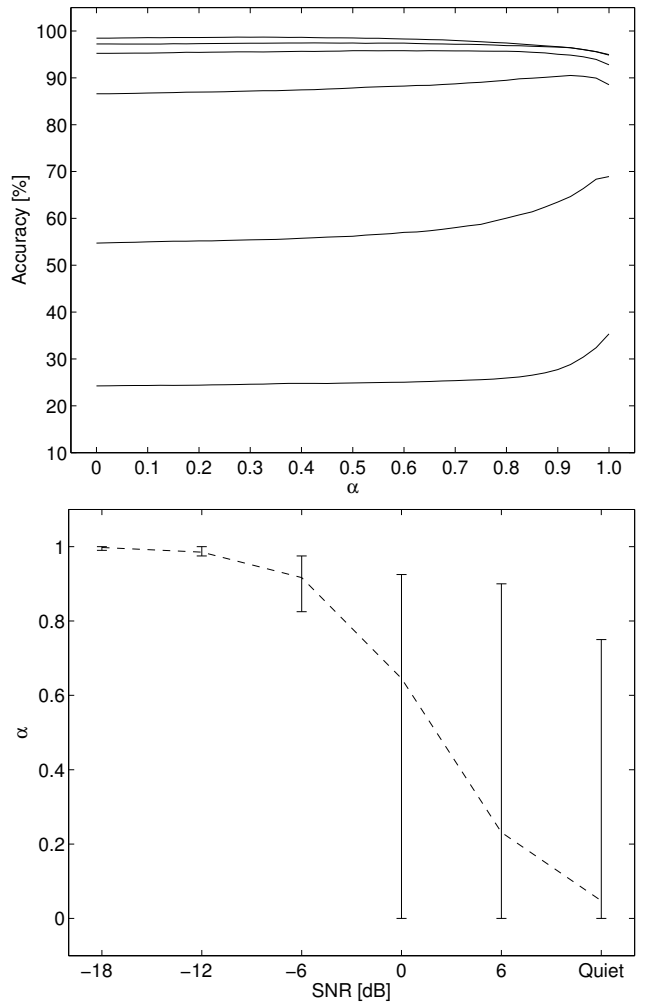


Figure 5: Top plot shows the result of combining a PLP classifier trained under matched conditions with a noise adapted waveform model. Each of the six testing conditions has been plotted. The bottom plot is the range of α that gives a classification accuracy within 1% of the maximum achievable. For high SNR the range of α that gives good accuracy is large but as the noise level increases the change in sensitivity become more sensitive to the value of α . The dashed curve shows a possible function of the form shown in equation 8 to fit the range, with $s_0 = 2\text{dB}$ and $\beta = 0.3$

and waveforms classifiers also plotted for comparison. In quiet conditions the combined classifier is as accurate as PLP, corresponding to $\alpha = 1$. When noise is present it is at least as accurate as for waveforms alone and in fact there is a slight improvement at 6dB SNR.

The other extreme to consider is when matched PLP models are used. This is shown in figure 7. Again the combined classifier performs as well as both PLP and waveforms with an 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 acous-

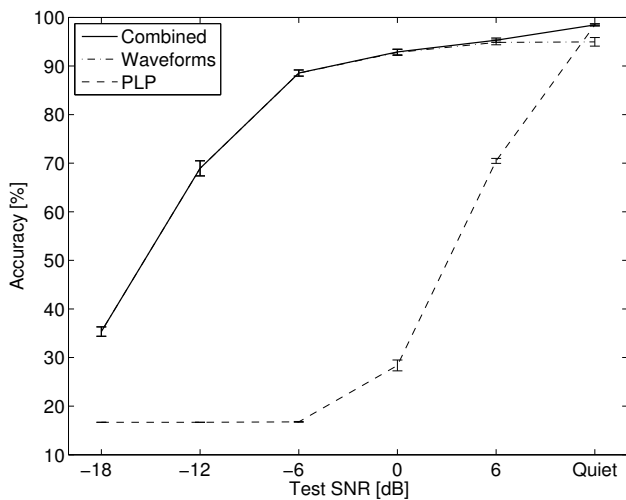


Figure 6: Performance of combined classifier for case where only quiet PLP models are used. Here accuracy in quiet test conditions is equivalent to using PLP. When noise is present accuracy is similar to that for the noise adapted waveform models alone, with a slight improvement at 6dB SNR.

tic waveforms has improved the classification accuracy of matched PLP alone significantly with SNR 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 matched conditions. The inset of figure 7 shows the expected range of accuracy for the combination of 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 results shown here, we will extend the experiments to a larger set of phonemes and also to larger databases containing more realisations of each phoneme. We would expect improvement for both PLP and waveform classifiers but especially so for acoustic waveforms due to their high dimensional representation where additional training data would improve density estimation. The results could also be generalised further to produce a classifier robust to noise other than white Gaussian noise. We expect the waveform classifier to be particularly suited to the task, since only an estimate of the noise covariance structure would be necessary to accurately adapt quiet models.

There is no reason to assume that the convex combination of the log-likelihoods gives an optimal classifier. It is possible that a 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.

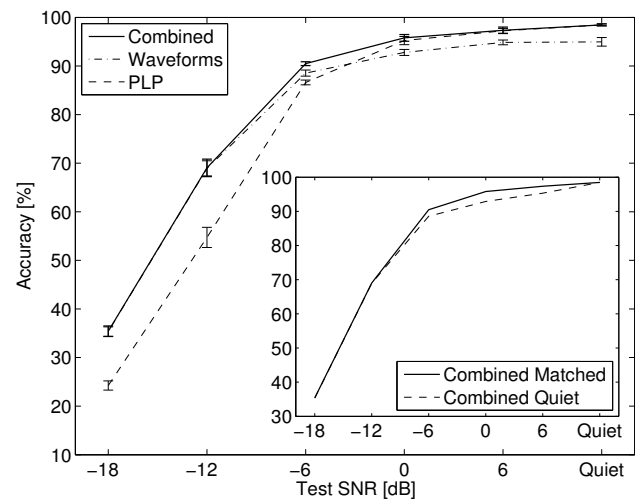


Figure 7: Performance of combined classifier when matched PLP models are used. The combined classifier is uniformly as accurate as those it is derived from and gives improvement at -6 dB SNR. Inset is a comparison of the combined quiet classifier from figure 6.

REFERENCES

- [1] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewoods Cliffs, 1993.
- [2] G. Miller and P. Nicely, "An analysis of perceptual confusions among some English consonants," *Acoustical Society of America Journal*, vol. 27, pp. 338–352, 1955.
- [3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Acoustical Society of America Journal*, vol. 87, pp. 1738–1752, Apr. 1990.
- [4] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct. 1994.
- [5] 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.
- [6] M. Tipping and C. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.
- [7] R. Rose, "Environmental robustness in automatic speech recognition," *Robust2004 - ISCA and COST278 Workshop on Robustness in Conversational Interaction*, Aug 2004.
- [8] M. Gales and S. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transactions on Speech and Audio Processing*, pp. 352–359, Sept. 1996.
- [9] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, and D. Pallett, "The DARPA TIMIT acoustic-phonetic continuous speech corpus. NIST," Feb. 1993.
- [10] D. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource.