ROBUST PHONEME CLASSIFICATION: EXPLOITING THE ADAPTABILITY OF ACOUSTIC WAVEFORM MODELS

Matthew Ager[1], Zoran Cvetković[2] and Peter Sollich[1]

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

ABSTRACT

The robustness of classification of isolated phoneme segments using generative classifiers is investigated for the acoustic waveform, MFCC and PLP speech representations. Gaussian mixture models with diagonal covariance matrices are used followed by maximum likelihood classification. The performance of noise adapted acoustic waveform models is compared with PLP and MFCC models that were adapted using noisy training set feature standardisation. In the presence of additive noise, acoustic waveforms have significantly lower classification error. Even for the unrealistic case where PLP and MFCC classifiers are trained and tested in exactly matched noise conditions acoustic waveform classifiers still outperform them. In both cases the acoustic waveform classifiers are trained explicitly only on quiet data and then modified by a simple transformation to account for the noise.

Index Terms— Speech Recognition, Robustness, Generative Classification, Phoneme, Acoustic Waveforms

1. INTRODUCTION

One of the key problems in automatic speech recognition (ASR) is to find phoneme classification methods that are robust 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 [11]. Clearly, 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 predicted input sequence is close to random guessing no useful information can be extracted at the later stages of recognition. Indeed, it has been observed that most of the inherent robustness of human speech recognition occurs early in the process, without access to context and language processing, e.g. already at -18dB SNR humans can still recognise isolated speech units above the level of chance [9]. The ultimate aim for an automatic speech classifier is to achieve performance close to that of the human auditory system in severe noise conditions. Developing methods for phoneme recognition, and the closely related problem of classification, that are robust to additive noise is a major step towards achieving that goal.

The current preferred speech representation is generally some variant of mel-frequency cepstral coefficients (MFCCs) [11] or perceptual linear prediction features (PLP) [6]. These representations are derived from the short term magnitude spectra followed by nonlinear transformations that model the processing of the human auditory system. They have the advantage of removing such variation from speech signals as is considered unnecessary for recognition and

have a much lower dimension than acoustic waveforms. This can allow for more accurate modelling when data is limited. However, it is not known whether this dimension reduction loses some information that gives communication by speech its inherent robustness. The alternative approach investigated in the paper is to use directly the acoustic waveform representation where the high dimensional space may give greater separation of the distributions of the different phonemes. If this is the case, classification from such representations should be more robust to additive noise. Additionally, noise modelling in the acoustic waveform domain is exact compared to approximations required for front-end representations which involve nonlinear processing.

In the following study Gaussian mixture models (GMMs) have been used to estimate the class-conditional densities of phonemes in all three representations considered: acoustic waveforms, PLP and MFCC. Classification in the presence of noise is performed using speech models adapted to account for particular noise conditions. Exact modelling of noisy data, given models trained in quiet and noise statistics, is straightforward in the acoustic waveform domain, and these exact models are used for classification of acoustic waveforms degraded by additive noise. Nonlinearities and dimension reduction involved in PLP and MFCC feature extraction make exact modelling of noisy speech in these two feature spaces very intricate. Hence, for classification in the PLP and MFCC domains, two approaches to noise adaptation are considered: a realistic one where PLP and MFCC models are trained in quiet conditions and then tested on features standardised [7] using the statistics of a training set that matches the testing conditions; and the unrealistic ideal approach where a separate model is trained for each noise condition and then the model that matches the testing condition in each case is used. This matched condition scenario is taken as an optimal baseline [12] for the MFCC and PLP representations. In order to give a fair comparison of this new approach to existing work we evaluate standard baseline classifiers with and without time derivatives.

The work presented in this paper significantly develops the experiments of our initial pilot study [1]. In that study we concentrated on a select subset of six phonemes, with the noise level fixed at the phoneme level. The data was extracted exactly from the centre of each segment and stops were exactly aligned at the release point. The PLP baseline used there did not include delta and delta-delta features. Additionally a PLP-waveform combined classifier was required to achieve adequate performance in low noise conditions.

The following experiments are more realistic, namely we have extended the ideas from [1] to the TIMIT core test set making our results directly comparable to existing benchmarks on the task. Here the signal to noise ratio (SNR) is specified at the sentence level, consequencely the local phoneme level SNRs will necessarily vary. Fur-

thermore the baseline is more stringent, i.e. results for MFCC and PLP with deltas and delta-deltas are considered. We also study three noise types, white, pink and speech weighted noise with model averaging to reduce the dependence on the number of model components. We see that the noise adaptation of the acoustic waveforms classifiers extends well to these more realistic conditions, even when the noise spectrum is estimated.

The results show the acoustic waveform representation to be significantly more robust to both white and pink noise than MFCC or PLP classifiers adapted using feature standardisation. The improvement is most significant at 12dB SNR when the absolute reduction in classification error rate compared to PLP is 17.6%. Even in the unrealistic case of matched condition training and testing for MFCC or PLP, acoustic waveform classifiers reduce the absolute error by 11.1% at 0dB SNR. Performance in pink noise is similar and although speech weighted noise is more challenging, acoustic waveforms still have lower error rates for all noise types below 18dB SNR.

2. GENERATIVE CLASSIFICATION IN THE PRESENCE OF ADDITIVE NOISE

Generative classification is particularly suited for robust speech classification as the estimated density models can capture the distribution of the noise corrupted phonemes. There are two approaches to derive the noisy class densities from data. The first method takes clean training data and corrupts it with generated noise followed by density estimation. Although this procedure will give good estimates with large datasets it has the obvious drawback that training must be repeated for each noise condition. The second approach combines density models of the clean training data and noise statistics that have been estimated independently. A clear advantage here is that the training data densities only need to be estimated once. However, such model combination methods can only be used where they are computationally tractable.

With that in mind we focus on acoustic waveforms, a speech representation where additive noise trivially acts additively. It immediately follows that the signal and noise distributions can be combined exactly by convolution. We considered white Gaussian noise, pink noise sampled from the Noisex-92 database and Gaussian noise generated to have the speech-weighted spectrum described in [10]. In all cases a range of noise levels were tested, parameterised by the global wideband SNR.

Without assuming any additional prior knowledge about the phoneme distributions we use Gaussian mixture models to estimate the class densities. We have chosen to train the models using maximum likelihood methods, in order to have a standard training platform on which the three different speech representations can be fairly compared. Other, discriminative, training objectives could be considered in due course, for example the large margin methods [13] that have recently shown promise.

In general a large number of parameters are required to specify GMMs, namely the mean, covariance matrix and weight of each component in the mixture. The number of parameters can be reduced by using diagonal covariance matrices. This will be a good approximation provided the data are presented in a basis where correlations between features are weak. For the acoustic waveform representation, this is clearly not the case on account of the strong temporal correlations in speech waveforms. We therefore systematically investigated candidate low-correlation bases derived from PCA, wavelet transforms and DCTs. Although the optimal basis for decorrelation is indeed the set of principal components, this initial investigation

showed that the lowest test error is in fact achieved with a DCT basis. The parameter count of the waveform models can be further reduced by observing the sign-invariance property of speech signals; it follows by symmetry that the phoneme distributions can be constrained to have zero mean. Hence the waveform classifiers use only diagonal covariance and mixture weight information.

One of the standard approaches used to select the number of optimal number of components in a mixture model is to minimise the classification error on a development set. As an alternative we investigate taking a model average (see e.g. [14]) over the number of components, i.e. to calculate the mean likelihood across models for a given data point. This gives uniformly better results in quiet conditions and can be interpreted as taking a uniform mixture of the mixtures. It is computationally helpful because it removes the need to optimise the number of components as a parameter during training. More importantly, the model averaging also gave better performance in all noise conditions, where it improved on any individual mixture.

With the models trained, classification is performed by predicting the class with the maximum likelihood weighte by the prior probabilities. The classification function $\mathcal{C}(x)$ that maps a test point x to a corresponding class label is defined as

$$\mathcal{C}(x) = \arg\max_{c=1,\ldots,k} \mathcal{L}^{(c)}(x) + \log(\pi_c) , \qquad (1)$$

where $\mathcal{L}^{(c)}(x)$ is the log likelihood of x given the model for class c and k is the total number of classes. π_c is the prior probability of class c, computed as the relative proportion of class c in the training dataset.

Local time alignment is an additional issue for acoustic waveforms. It would clearly be beneficial for the purpose of density modelling to align the data in a consistent manner; however, it is not straightforward to even define such an optimal alignment precisely. Rather than attempting to explicitly align the acoustic waveform data, a sliding window with a 1.6ms shift over a range of ±3.2ms was therefore used. This gives 5 shifted instances for each representative x. The log likelihood of the test point x is then taken as the log mean likelihood [1] taken over the shifts:

$$\mathcal{L}_s(x) = \log\Big(\frac{1}{2n+1}\sum_{p=-n}^{n}\exp(\mathcal{L}(x^{p\Delta}))\Big) \qquad (2)$$

where Δ (=1.6ms) is the shift increment, $[-n\Delta, n\Delta]$ is the shift range ($n = 2$ in our case), and $x^{p\Delta}$ denotes a time-shifted versions of x. Explicitly, $x^{p\Delta}$ is the segment of the same length and extracted from the same acoustic waveform as x but starting from a position shifted by $p\Delta$ in time. These modified log likelihoods are compared among the different classes to produce the classification. As MFCC and PLP use frames of magnitude spectra that are not sensitive to local time alignment there is no benefit in considering a similar averaging over shifts in these two domains; this was also experimentally verified.

We now consider the problem of noise adaptation for the different representations. One of the key advantages of the waveform representation is that the fitted density models can easily be modified to account for the presence of additive noise. Assuming that the noise power spectrum is known or can be estimated reliably, we simply need to perform a convolution with the appropriate Gaussian noise model. In this work we assume Gaussian noise of known variance σ^2; the resulting adapted density model has component covariance matrices $\tilde{\mathbf{C}}$ given by

$$\tilde{\mathbf{C}}(\sigma^2) = \frac{\mathbf{C} + \sigma^2\mathbf{N}}{1 + \sigma^2} , \qquad (3)$$

where $1 + \sigma^2$ is a sentence-level normalisation factor as explained in Section 3. $\mathbf{N}$ is the covariance matrix of the noise, normalised to have trace d. For white noise, $\mathbf{N}$ is the identity matrix, otherwise it has been estimated empirically from noise samples.

As MFCC and PLP features are highly non-linear transforms of the waveform data it is not possible to combine models of the training data and noise exactly. For MFCC an approximate combination [4] is possible. An alternative approach is to use models that are trained in quiet conditions and then tested on features standardised [7] using the statistics of a training set that matches the testing conditions. The other, less realistic, scenario we also consider is matched conditions. Here training and testing noise levels are the same, with a separate classifier trained for each noise condition.

3. DATASETS

Realisations of phonemes were extracted from the SI and SX sentences of the TIMIT database [5]. The training set consists of 3,696 sentences sampled at 16kHz. Each sentence is then normalised to have on average unit energy per sample. Noisy data is generated by applying noise samples additively at nine SNRs followed by the same average unit energy per sample normalisation. This ensures that the level of the input to the classifier is consistent and realistic. The SNRs were set at the sentence level and it is important to note that the local SNR of the individual phonemes may then differ significantly from the set value. In total ten testing and training conditions were run; -18dB to 30dB in 6dB increments and quiet (Q).

Following the extraction of the phonemes there are a total of 140,225 phoneme realisations. The glottal closures are removed and the remaining classes are then combined into 48 groups in accordance with [8, 13]. This is done to improve modelling by merging similar classes. Even after this combination some of the resulting groups have too few realisations. The smallest groups with fewer than 1,500 realisations were increased in size by the addition of shifted versions of the training data, as described – for the purpose of testing – in Section 2. These additions allow a greater number of mixture components to be reliably fitted and gave a small improvement for the acoustic waveform classifier of 0.6% when tested in quiet conditions.

MFCC and PLP features are obtained in the standard manner from frames of width 25ms, with an overlap of 15ms between neighbouring frames. A standard implementation [3] of MFCC and PLP with default parameter values is used to produce, from each frame, a 13-dimensional feature vector. (With the inclusion of first and second time derivatives this dimension increases to 39.)

Phonemes are extracted using the standard TIMIT segmentation. Our previous work gave successful classification using a 64ms window. For the MFCC and PLP representations, we therefore consider the five frames closest to the centre of each phoneme and concatenate their feature vectors to give 65-dimensional vector. In addition standard practice dictates that first and second time derivative features should be appended to the representation. We show results for both representations, those with and those without the time derivatives giving dimensions of 195 and 65 respectively.

For comparison, each sentence was divided into a sequence of 10ms non-overlapping frames to give the acoustic waveform representation, with the seven frames (70ms) closest to the centre of each phoneme resulting in a 1120-dimensional representation. The frames are individually processed using a 10ms DCT. A framewise DCT was used instead of a DCT over the full 70ms window as it gives a finer time resolution to capture features that are not stationary over that period. This DCT representation is nothing more than
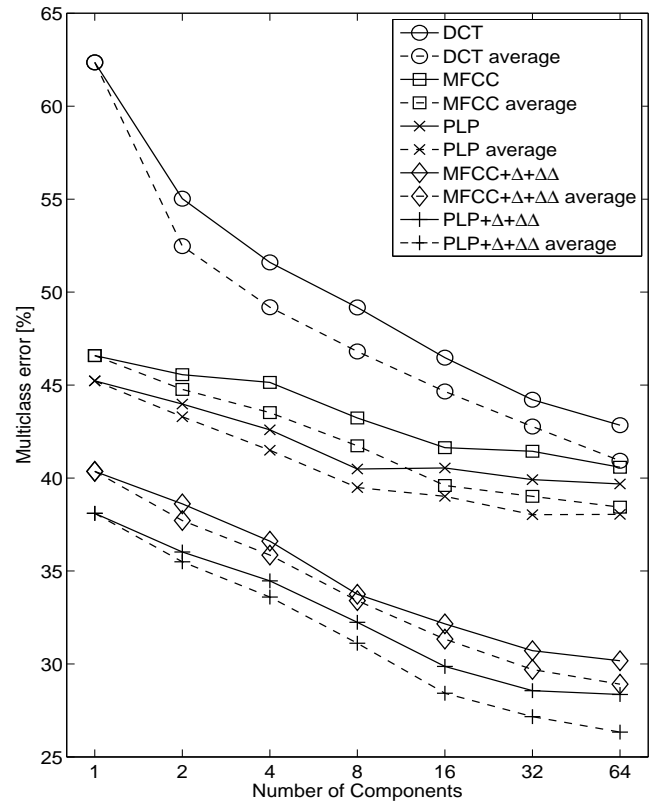

Fig. 1. *Model averaging for acoustic waveforms, PLP and MFCC models, all trained and tested in quiet conditions. Dashed: GMMs with number of components shown; solid: average over models up to number of components shown. The model average reduces the error rate in all cases.*

an orthogonal transformation of the original waveform segment and therefore the noise adaptation of (3) remains valid in the transformed domain for the case of white noise. The adaptation readily generalises to arbitrary coloured Gaussian noise; full covariances matrices would then in principle be required for the noise contribution on the right hand side of (3). We present results for two other noise types and see that this approximation using diagonal covariances in the DCT basis is sufficient to give good performance.

4. RESULTS

In the experiments Gaussian mixture models were tested with up to 64 components for all representations. We comment briefly on the results for individual mixtures, i.e. fixed number of components. Typically performance on quiet data improved with the number of components, although this has significant cost for both training and testing. The optimal number of components for MFCC and PLP models in quiet conditions was 64, i.e. the maximum considered here. However, in the presence of noise the lowest error rates were obtained with few components; typically the error rate stopped decreasing for mixtures with more than 4 components.

As explained above, rather than working with models with fixed numbers of components, we averaged over models, i.e. over the number of mixture components, in all the results reported below. Figure 1 shows that the improvement obtained by this in quiet condi-
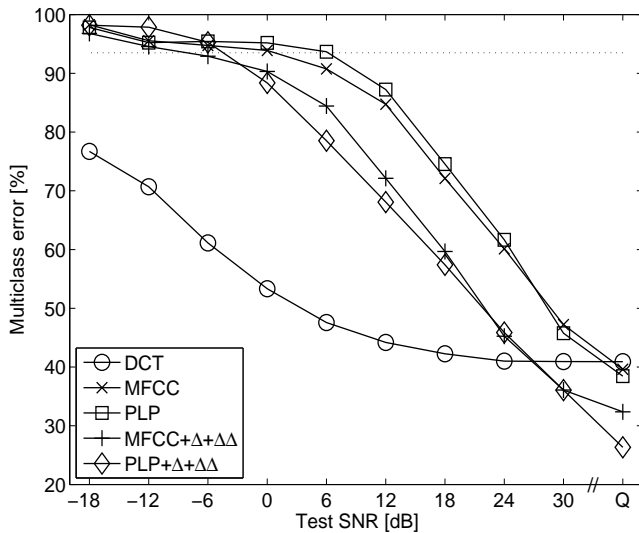
Fig. 2. *Comparison of adapted acoustic waveform classifiers in the DCT basis with MFCC and PLP classifiers trained in quiet conditions adapted by matched feature standardisation. All classifiers use the model average of mixtures up to 64 components. When the SNR is less that 24dB, acoustic waveforms are the significantly better representation; with an error rate below chance even at -18dB SNR. Dotted line indicates chance level at 93.5%.*

Fig. 3. *Comparison of training set standardisation with matched condition training for PLP and MFCC. Training set standardisation is more realistic as it only requires a set of transforms to be stored rather than retraining the classifiers at each condition. Matched condition training is expected to be optimal and accordingly improves accuracy for both representations, but would be difficult to achieve in practice.*

tions is approximately 2% for both acoustic waveforms and PLP with a small improvement seen for MFCC also. We checked (data not shown) that the model average likewise improved results in noise.

Our key results comparing the error rates for phoneme classification in the three domains are shown in Figure 2. The MFCC and PLP classifiers are adapted to noise using feature standardisation. This method is comparable with the adapted waveform models in so much as it only requires knowledge of the noise spectrum and the models trained in quiet conditions. The curve for acoustic waveforms is for models trained in quiet conditions and then adapted to the appropriate noise level using (3). We see that in quiet conditions the PLP representations gives the lowest error. The error rates for MFCC and PLP are significantly worse in the presence of noise, however, with acoustic waveforms giving an absolute reduction of 37.0% and 35.0% compared to MFCC and PLP respectively, both with delta and delta-deltas at 0dB SNR. These results strengthen the case that the adaptability of acoustic waveform models gives them a definite advantage in the presence of noise with the crossover point occuring above 30dB SNR. Curves are also shown for MFCC and PLP with deltas and delta-deltas. Again the same trend holds; performance is good in quiet conditions but quickly deteriorates as the SNR decreases. The cross-over point is around to 24dB for both representations. The chance-level error rate of 93.5% can be seen below 0dB SNR for the MFCC and PLP representations without deltas and below 6dB SNR when they are included, whereas the acoustic waveform classifier performs significantly better than chance with an error of 76.7% even at -18dB SNR.

The curves shown in Figure 3 compare the performance in white Gaussian noise of MFCC and PLP models adapted using feature standardisation with models that were trained in ideal matched conditions where the sentence level SNRs of the training data are exactly the same as those in testing. This is an unrealistic scenario that should give the best performance for the chosen representation and
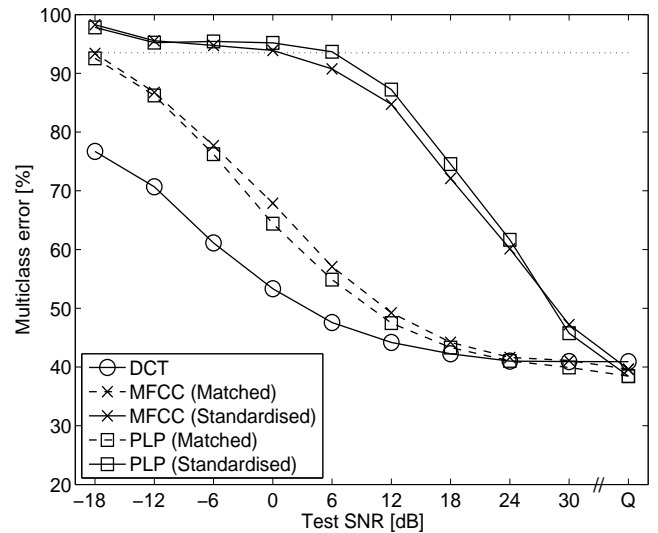
noise level [12]. Even relative to this baseline, however, acoustic waveforms perform better for all noise levels in our tests, with e.g. an absolute improvement in error rate over MFCC and PLP of 9.3% and 4.8% respectively at 0dB SNR. The cross-over occurs between 24dB and 30dB SNR.

The baseline results shown in Figure 3 do not include time derivatives as no analoguous derivative features are used in the acoustic waveform representation. Results for GMM classification on the TIMIT benchmark have previously been reported in [13, 2] with errors of 25.9% and 26.3% respectively. These studies use MFCC features and first and second time derivatives. To ensure that our baseline is valid we compared our experiment in quiet conditions for PLP with first and second derivatives included and obtained a comparable error rate of 26.3%. Beyond validating our implementation, this has implications for how fixed length representations are constructed from variable length phonemes. Previous work has concentrated on generating fixed length representations [2] via frame averaging across the entire phoneme. We have instead used representations derived from windows of fixed length: in line with our eventual goal of moving towards continuous speech recognition, these could be implemented directly in existing continuous ASR systems. It is then worth noting that the representation we have considered contains only 5 frames (or less in the case of very short utterances) from the centre of each phoneme, but performs essentially identically to the frame averaging method of [2] where information across the entire length of a phoneme is captured in the features.

Figure 4 shows a comparison of acoustic waveforms adapted to white, pink and speech-weighted noise using (3). Similar results are obtained in white and pink noise although the effect speech-weighted noise is most significant below 18dB SNR. The curves are compared to PLP with deltas and delta-deltas using training set statistics that match the test conditions. We see that the error rates for the PLP
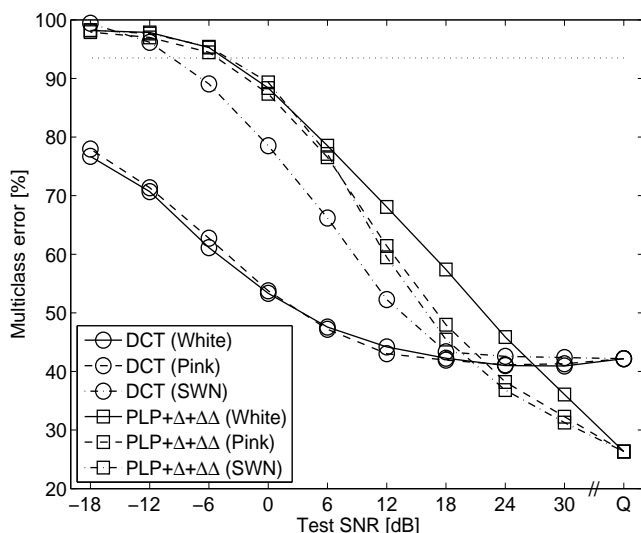
Fig. 4. *Comparison of performance in pink noise and speech-weighted noise of adapted acoustic waveform classifiers in the DCT basis with PLP classifiers trained in quiet conditions adapted by matched feature standardisation. All classifiers use the model average of mixtures up to 64 components.*

classifier is similar for speech-weighted noise, although the best performance crosses to acoustic waveforms below 18dB SNR.

In order to have a direct comparison with the existing delta and delta-delta results we must develop analoguous derivatives in the acoustic waveform domain. Alternatively we are currently investigating the use of additional consecutive frames to include the extra information used by the deltas.

5. CONCLUSIONS

This study has compared phoneme classification using generative classifiers, by considering the MFCC, PLP and waveform representations. Our results show that the waveform representation is more robust than PLP or MFCC in the presence of three types of noise. We emphasise that this performance was achieved with a waveform classifier trained exclusively on quiet data, with the noise being included via a simple transformation of the fitted class-conditional densities. For MFCC and PLP, we firstly took classifiers that were likewise trained on quiet data and adapted using features standardised on an appropriate noisy training set. Here waveforms perform significantly better in white Guassian noise improving by over 35% compared to both MFCC and PLP at 0dB. The performance of the acoustic waveform classifiers is consistently better for SNRs less than 24dB.

In a second scenario, under matched training we allowed the MFCC and PLP classifiers access to training data corrupted with exactly the same noise distribution as in testing. Such idealised conditions would clearly be difficult to achieve in practice, especially when dealing with general Gaussian rather than white noise and where in principle separate MFCC and PLP classifiers would need to be trained for a range of different noise power spectra. Even compared to this stringent baseline performance from the waveform classifiers is superior, with absolute improvements in white Gaussian noise of 14.6% and 11.1% at 0dB with the performance cross-over between 24dB and 30dB SNR.

Our study has shown that phoneme classification with improved robustness to additive noise can be achieved in the acoustic waveform domain. The results support the conclusions and extend our previous work [1] to more realistic and challenging conditions. There are a number of directions for further development of the methods demonstrated here. In particular the issue of finding an optimal basis transformation for the waveforms could be generalised to be class dependent, rather that using the same transform for all classes. It would also be interesting to compare explicitly with the phoneme sets used in experiments on human speech recognition [10]. We have had obtained some promising preliminary results using the DCT representation in conjunction with HMMs on such a consonant-vowel classification task. Our future work is focused on extending the results to continuous speech recognition in the presence of noise.

6. REFERENCES

[1] M. Ager, Z. Cvetković, P. Sollich, and B. Yu. Towards robust phoneme classification: Augmentation of PLP models with acoustic waveforms. In *Proceedings of EUSIPCO*, 2008.

[2] P. Clarkson and P. Moreno. On the use of support vector machines for phonetic classification. In *ICASSP*, volume 2, pages 585–588, 1999.

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

[4] M. Gales and S. Young. Robust continuous speech recognition using parallel model combination. *IEEE Transactions on Speech and Audio Processing*, 4:352–359, 1996.

[5] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, and D. Pallett. *The DARPA TIMIT acoustic-phonetic continous speech corpus. NIST*. Linguistic Data Consortium, Philiadelphia, 1993.

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

[7] P. Jain and H. Hermansky. Improved mean and variance normalization for robust speech recognition. In *Proceedings of ICASSP*, 2001.

[8] K.-F. Lee and H.-W. Hon. Speaker-independent phone recognition using hidden markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(11):1641–1648, 1989.

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

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

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

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

[13] F. Sha and L. Saul. Large margin gaussian mixture modeling for phonetic classification and recognition. In *Proceedings of ICASSP*, 2006.

[14] S. Srivastava, M. Gupta, and B. Frigyik. Bayesian quadratic discriminant analysis. *Journal of Machine Learning Research*, 8:1287–1314, 2007.