ROBUSTNESS OF PHONEME CLASSIFICATION USING GENERATIVE CLASSIFIERS: COMPARISON OF THE ACOUSTIC WAVEFORM AND PLP REPRESENTATIONS

Matthew Ager, Zoran Cvetković, Peter Sollich, Jibran Yousafzai

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

ABSTRACT

The robustness of classification of isolated phoneme segments using generative classifiers is investigated for the acoustic waveform and PLP speech representations. Probabilistic PCA is used to fit a density to each phoneme class followed by maximum likelihood classification. The results show that although PLP performs best in quiet conditions, as the SNR decreases below 0dB acoustic waveforms have a lower classification error. This is the case even though the waveform classifier is trained explicitly only on quiet data and is then modified by a simple transformation to account for the noise, whereas for PLP separate classifiers are trained for each noise condition. Even at -18dB SNR, multiclass performance of classification from waveforms is still significantly better than chance level. In addition the effect of time-alignment is tested and initial solution shown.

Index Terms— Speech Recognition, Robustness, Generative Classification

1. INTRODUCTION

One of the key problems in automatic speech recognition is robust phoneme classification. 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]. Clearly, though, this approach can only decode messages sent via speech signals if the input sequence of elementary speech units is sufficiantly 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. Indeed, it has been observed that the majority of inherent robustness of human hearing occurs early in the process [2]. Even at -18dB SNR humans can still recognise isolated speech units above the level of chance. 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. Developing methods of phoneme recognition that are robust to additive noise could be one step towards achieving that goal.

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 is to use higher dimensional representations where the distributions of the different phonemes may be better separated. If this is the case classification from such representations should be more robust to additive noise. The aim of this paper is to assess the separability of phonemes in the two domains and how it changes as a function of SNR.

In the following study Probabilistic PCA was used to estimate the class-conditional densities. This results in one fitted Gaussian per class. This model was chosen for its simplicity and clear interpretation. It is possible that more sophicated mixture models could give better performance. However the aim of our work was not to find the optimal classifier but to illustrate that acoustic waveforms are a viable representation for robust phoneme recognition. A representative subset of the TIMIT database was used to evaluate both binary and multiclass error rates on the task of isolated phoneme recognition.

2. GENERATIVE CLASSIFICATION

Realisations of six phonemes (/b/, /f/, /m/, /r/, /t/, /z/) were extracted from the TIMIT database. 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. This window is shorter than for typical human tests, although our experiments show it is sufficient to capture enough information for class separation. The natural space in which to perform classification for the waveforms is the

hypersphere $\mathbb{S}^{1023}$ as each sample has 1024 entries and is normalised to unit norm. As the mean value of each class was zero within sampling error, the class-conditional densities were constrained to have zero mean. This is natural as an inverted waveform is perceived as the same phoneme. For comparison the default 12th order PLP cepstra of the data were taken, leading to 4 frames of 13 coefficients [6]. The 4 frames were concatenated to give a PLP representation in $\mathbb{R}^{52}$. PLP representations were not normalised, and we allowed nonzero means for their class-conditional densities.

The tests were carried out in quiet conditions and also on data with additive white Gaussian noise. The SNR was decreased until the error approached that of chance level, i.e. 83.3% in the case of six classes. In total this gave six testing and training conditions; -18dB, -12dB, -6dB, 0dB, 6dB and quiet (Q).

Each class-conditional density model was derived using Probabilistic PCA [7] in the appropriate d-dimensional space ($d = 1024$ for waveforms, $d = 52$ for PLP). The method gives a maximum likelihood fitted Gaussian that is conditional on the dimension q of the principal subspace. It has the simple interpretation that the variance due to the smallest $d - q$ eigenvalues is redistributed isotropically. This method gave better results than first projecting the data onto the principal subspace, as no information from the data is discarded. In the case of waveforms the best results were found when the full empirical Gaussian model was taken, i.e. $q = 1024$. However for PLP the optimal q was dependent on the noise level and q was therefore optimised by cross-validation on the training set.

PPCA uses the eigenpairs (v_i, λ_i) of the empirical covariance matrix, with the eigenvalues ordered in decreasing order. Given the parameter q, the smoothed spectrum is defined as

$$\hat{\lambda}_i = \begin{cases} \lambda_i + \sigma^2 & i \leq q \\ \sigma^2 & i > q \end{cases}$$

where $\sigma^2 = \frac{1}{d}\sum_{i=q+1}^{d} \lambda_i$. The log likelihood of a test point x can then be written as, with μ as the mean of the Gaussian.

$$\begin{aligned} \mathcal{L}(x) &= -\frac{1}{2}\Big\{ d\ln(2\pi) + \sum_{i=1}^{d} \ln(\hat{\lambda}_i) \Big\} \\ &\quad -\frac{1}{2}\Big\{ \sum_{i=1}^{d} \frac{\langle v_i, (x-\mu)\rangle^2}{\hat{\lambda}_i} \Big\} \end{aligned}$$

Classification is then performed in the standard way, by predicting the class with the maximum likelihood (which implicitly assumes uniform prior probabilities over different classes). The classification function $\mathcal{C}(x)$ that maps a test point x to a corresponding class label is defined as

$$\mathcal{C}(x) = \arg \max_{c=1,\ldots,6} \mathcal{L}^{(c)}(x)$$

One of the key 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 SNR is ldB, the resulting density model for white noise is given by

$$\tilde{\lambda}_i(l) = \frac{\hat{\lambda}_i + k(l)}{1 + dk(l)} \quad \text{where } \; k(l) = \frac{10^{-0.1l}}{d}$$

For the PLP representations, on the other hand, there is no similarly obvious method for including noise in the density models. We therefore assume here that noisy data matched to the test conditions are available for training, and train one separate set of PLP density models for each test noise condition. (Other methods have been proposed to reduce explicitly the effect of noise on spectral representations [8] but are not explored here, for fairness of comparison with the waveform case.)

As PLP uses frames of magnitude spectra it is less sensitive to time alignment. In the case of waveforms though 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 align the data, a sliding window with a 10 sample shift over a range of ± 100 samples was used. This gives 21 shifted instances x_s for each representative x. The log likelihood of the test point x is then taken as the log mean likelihood taken over the shifts:

$$\mathcal{L}_s(x) = \ln\Big(\frac{1}{21} \sum_{k=-10}^{10} \exp(\mathcal{L}(x_{10k})) \Big)$$

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.

3. RESULTS

Figure 1 shows a direct comparison of the two representations considered. It displays the multiclass classification error, i.e. the probability of a test phoneme being classified incorrectly, as a function of SNR. The key observation is that, while PLP performes better up 0dB SNR, as the noise level increases beyond this point the waveforms representations lead to significantly lower errors and perform better than chance even down

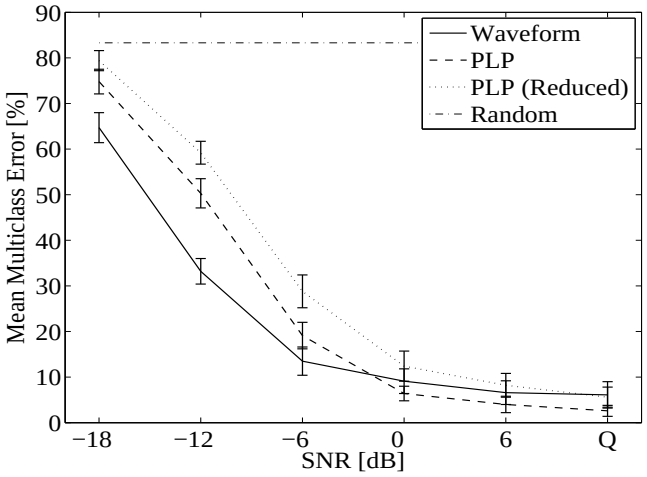

Fig. 1. Mean Multiclass Classification Error Rate of the six phonemes. The three curves are for waveform and PLP representations and PLP with a smaller training set. Dashed-dotted line: Error rate for random guessing.

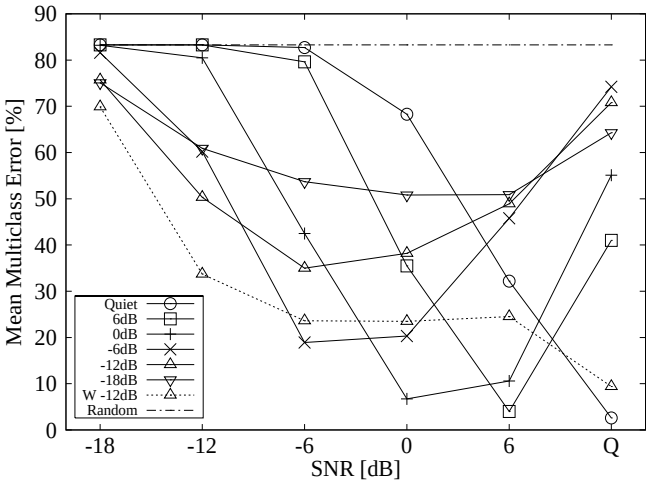

Fig. 2. Mean Multiclass Classification Error Rate for PLP trained on data corrupted by fixed levels of noise as given in the legend. Each curve shows one classifiers tested across all levels of test noise. The dotted curve shows the result of using the waveform classifier adjusted for data at -12dB SNR.

to -18dB SNR. This is in spite of the fact that for waveforms only a single classifier was trained in quiet conditions and then modified as above to allow for the inclusion of additive noise, while the PLP classifier for each test noise level was trained explicitly on training data corrupted with a matching level of noise.

Arguably, training the PLP density models should require less data as the representation is of significantly lower dimension than for waveforms. Hence for a given amount of data the model parameters can be better estimated. The dotted line in Figure 1 therefore shows additonal results where the PLP classifiers were trained on a smaller subset of the data. The size of this reduced subset was taken in proportion to the dimension of the PLP representation, i.e. $52/1024 \approx 5\%$ of the original training set. In this case the PLP and waveform representations perform comparably in quiet conditions with waveforms outperforming PLP at all higher noise levels. We would expect that this relation would persist when classifiers for both representations are trained on larger corpora, with the waveform classifiers improving on the results shown in Figure 1.

We further explored how the classifiers perform in the presence of a mismatch between training and test conditions. The PLP representation turns out be very sensitive to such mismatch. For example the classifier trained in quiet conditions performs well when the testing conditions are also quiet but is at chance level already by -6dB SNR. Figure 2 shows classification errors for the scenario where the noise level for testing is not known and models trained at fixed noise levels are considered; a key feature is that such classifiers can perform poorly even if the test noise level is *lower* than the one used for training. The dotted line shows, for comparison, the results for the waveform classifier "trained" at -12dB SNR (i.e. trained in quiet conditions and with noise of -12dB SNR included as explained above). Comparison with the PLP curves shows that the waveform model is considerably less sensitive to variation in test noise conditions when the full range is considered; in particular, performance does not deteriorate (and indeed improves) when test noise levels are lower than in training.

We also performed binary classification tests on all 15 pairs taken from our 6 phonemes, and observed similar trends as in the full multiclass classification. The mean error ranged from 1.9% in quiet to 33.3% at -18dB SNR for waveforms compared with 0.6% to 40.3% for PLP. The results for individual pairs had the same general pattern, with some additional variation arising from the fact that not all pairs are equally confusable.

Finally we show in Figure 3 the effect of including shifts in the evaluation of the likelihoods as explained in Sec. 2. As expected the inclusion of shifts gave no significant improvement for PLP, and so we do not show the relevant data. For waveforms, on the other hand, we see an improvement of at least one standard deviation at all noise levels.

4. CONCLUSIONS

In this study we have compared phoneme classification using generative classifiers, comparing the PLP and waveform representations. Our results show that the waveform representation is more robust than PLP in the presence of additive white noise. The point at which PLP started to perform worse was at an SNR of 0dB or lower. Even at -18dB SNR the mul-

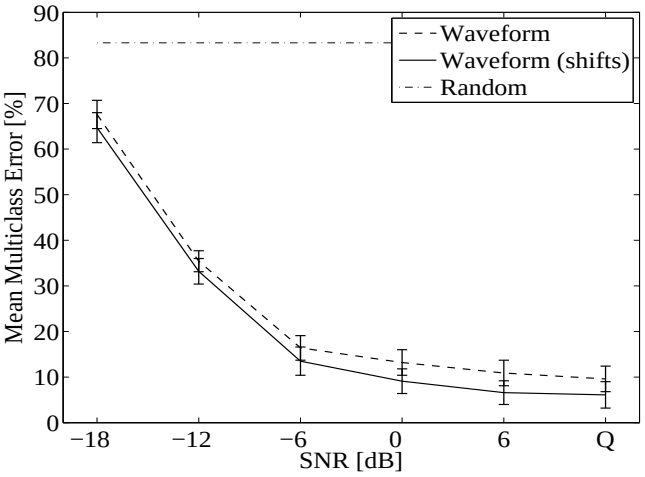

Fig. 3. Mean multiclass classification error showing an improvement when shifts are included at the testing stage. The inclusion of shifts in training also gave a small improvement.

ticlass classification error for the waveform representation is still significantly below chance level. 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 PLP, on the other hand, we allowed the classifier access to training data corrupted with exactly the same level of noise as in testing. Such an idealised scenario would clearly be difficult to achieve in practice, especially if one is dealing with coloured rather than white noise and where in principle one would then need to retrain separate PLP classifiers also for a range of different noise power spectra.

The waveform representation does in principle suffer from the difficulty of aligning waveforms appropriately, but we showed that this can be addressed in a simple manner by averaging the likelihoods over shifted versions of the test waveform.

Classifiers based on the PLP representation performed best for quiet conditions but where the SNR is below 0dB the acoustic waveform representation is superior. Adjusting the size of the respective training sets to be proportional to the dimension of the representation in each case, we found that difference in performance at high SNRs is reduced and waveforms perform comparably or better than PLP at all noise levels. This suggests that somewhat larger training sets than those considered here would be needed to accurately train the waveform classifiers, but that they would then perform competitively with PLP and rather better at low SNRs.

We also explored the effects of mismatch between training and noise conditions. Consistent with our other results, the waveform classifiers are rather less sensitive to such a mismatch. In fact, PLP classifiers can perform poorly even if test conditions are more benign than during training, whereas the waveform classifiers always improve their performance in such conditions.

Our proof-of-principle study has shown that phoneme classification robust to additive noise is possible in the acoustic waveform domain. There are, of course, a number of directions for further research to develop the methods demonstrated here. To make the scenario more realistic, tests should be performed on larger sets of phonemes; it would also be interesting to compare explicitly with the phoneme sets used in experiments on human speech recognition [9]. Evidently, more powerful density models will need to be explored, for example mixtures of Gaussians. It is worth emphasising that also such mixture models would preserve the desirable property of waveform representations that noise of known SNR (or more generally power spectrum) is trivial to include; as long as accurate estimates of the noise conditions are available, this could be done in real time. As an alternative, class-conditional densities could be modelled by isolating the most non-Gaussian components and representing their densities via empirical distributions. Finally, for both PLP and waveform classifiers, training under a combined range of noise conditions would be explored [8] to reduce sensitivity to any mismatch between training and test conditions.

5. REFERENCES

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

[2] G. A. Miller and P. E. 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] Daniel P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource.

[7] M. Tipping and C. Bishop, "Probabilistic principal component analysis," 1997.

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

[9] S. A. Phatak and J. B. Allen, "Syllable confusions in speech-weighted noise," *Acoustical Society of America Journal*, vol. 121, no. 4, pp. 2312–2326, Apr. 2007.