DISCRIMINATIVE AND GENERATIVE MACHINE LEARNING APPROACHES TOWARDS ROBUST PHONEME CLASSIFICATION

Jibran Yousafzai, Matthew Ager, Zoran Cvetković and Peter Sollich

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

ABSTRACT

Robustness of classification of isolated phoneme segments using discriminative and generative classifiers is investigated for the acoustic waveform and PLP speech representations. The two approaches used are support vector machines (SVMs) and mixtures of probabilistic PCA (MPPCA). While recognition in the PLP domain attains superb accuracy on clean data, it is significantly affected by mismatch between training and test noise levels. Classification in the high-dimensional acoustic waveform domain, on the other hand, is more robust in the presence of additive white Gaussian noise. We also show some results on the effects of custom-designed kernel functions for SVM classification in the acoustic waveform domain.

Index Terms— Speech Recognition, Robustness, Discriminative Classification, Generative Classification, PLP

1. INTRODUCTION

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

The current preferred speech representation is generally some variant of PLP[5], RASTA[6] or MFCC[7]. These representations are derived from the short term magnitude spectra followed by nonlinear transformations to model the processing of the human auditory system. They have the advantage that they remove variations from speech signals that are considered unnecessary for recognition and have a much lower dimension than acoustic waveforms. However it is not certain that in this process of peeling off speech components that are unnecessary for recognition one is not discarding part of the information that makes speech such a robust message representation, consequently ending up with ASR systems which are very sensitive to noise and other forms of degradation.

The basic hypothesis of our work is that in representation domains which involve compression, different phonetic units although separated may not be sufficiently apart and may start overlapping considerably at lower noise levels than they do in the original uncompressed domain of acoustic waveforms. Moreover, compressed representations of speech, because of the strong nonlinearities that link them to the original acoustic waveforms, lead to distributions of different speech units that may vary significantly in the presence of noise. Hence, classification of speech units in acoustic waveform domain should be more robust to additive noise than in the domains of state-of-the-art representations all of which involve considerable nonlinear compression. In this study, we test this hypothesis by performing classification of phonemes in presence of noise using generative classifiers, in particular, MPCCA and discriminative classifiers i.e SVMs in the acoustic waveform and PLP domains, with particular emphasis on exploring the mismatch between training and testing conditions. Classification methods are presented in Section 2 and 3 and the test methodology is described in Section 4. The experiments, results of which are reported in Section 5, show that while classification using the PLP and MFCC representations achieve considerably better results on clean data than the acoustic waveform representation, it is much more sensitive to noise mismatch between training and test conditions. A waveform classifier, on the other hand, provides robust performance across a broad range of signal-to-noise ratios (SNRs). We also provide some insights into the importance of custom design of SVM kernels for improving the accuracy of phoneme classification. Finally, Section 6 draws some conclusions.

2. GENERATIVE CLASSIFICATION

Generative classification was performed using density estimates derived from mixtures of Probabilistic PCA (MPPCA) [8]. Probabilistic PCA (PPCA) uses the eigenpairs (v_i, λ_i) of the empirical covariance matrix, with the eigenvalues in descending 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:

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

where the s columns of $\mathbf{W}$ are given by $\sqrt{\lambda_i}v_i$ for the corresponding index i and r^2 is the mean of the remaining $t - s$ eigenvalues i.e.

$r^2 = \frac{1}{t-s}\sum_{s+1}^{t}\lambda_i$. MPPCA represents the class conditional distribution for each class with a mixture of such regularized Gaussians; the model parameters are optimized using the EM algorithm [8]. Given a data point x, the log likelihood function $\mathcal{L}(x)$ is defined in (2) as the log of the density of the mixture evaluated at x.

$$\mathcal{L}(x) = \ln\Big(\sum_{i=1}^{c}\frac{w_i}{(2\pi)^{\frac{t}{2}}|\mathbf{C}_i|^{\frac{1}{2}}}e^{-\frac{\langle x-\mu_i,\mathbf{C}_i^{-1}(x-\mu_i)\rangle}{2}}\Big) \qquad (2)$$

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 a corresponding class label is defined as

$$H(x) = \arg\max_{k=1,\ldots,K}\mathcal{L}^{(k)}(x) \qquad (3)$$

where $\mathcal{L}^{(k)}$ is the log likelihood function of class k. 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 $\hat{\lambda}_i$ are the eigenvalues of $\mathbf{C}$, the resulting density model for waveforms corrupted by white noise (and renormalized to unit length) is given by

$$\tilde{\lambda}_i(l) = \frac{\hat{\lambda}_i + \frac{\sigma^2}{t}}{1+\sigma^2} \qquad (4)$$

For the MFCC and PLP representations, there is no similarly explicit 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 MFCC/PLP density models for each test noise condition. (Other methods have been proposed to reduce explicitly the effect of noise on spectral representations [9] but are not explored here, for fairness of comparison with the waveform case. At any rate, the proposed methods perform no better than the matched condition approach [10].)

It is also beneficial to model the effects of time alignment of speech data. This is done by including shifted versions of the waveforms in the training set. It is expected that classification in the acoustic waveform domain will benefit more from the inclusion of the shifted data since PLP, MFCC and other state-of-the-art representations are based on short-time magnitude spectra and therefore shifted data would not carry significantly different information. For testing, we correspondingly use instead of the "bare" log likelihood $\mathcal{L}(x)$ its mean over shifts, i.e.

$$\mathcal{L}_s(x) = \ln\Big(\frac{1}{2n+1}\sum_{p=-n}^{n}\exp(\mathcal{L}(x^{p\Delta}))\Big) \qquad (5)$$

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.

3. DISCRIMINATIVE CLASSIFICATION

An SVM estimates decision surfaces separating two classes of data. In the simplest case these are linear but for speech recognition, one typically requires nonlinear decision boundaries. These are constructed using kernels instead of dot products, implicitly mapping data points to high-dimensional feature vectors [11]. A kernel-based decision function has the form

$$h(x) = \sum_i \alpha_i y_i K(x, x_i) + b \qquad (6)$$

where x_i are all training inputs, $y_i = \pm 1$ are class labels, the bias term, b and α_i are parameters determined by SVM. Two commonly used kernels are polynomial and radial basis function (RBF) kernels given by (7) and (8), respectively,

$$K(x_i, x_j) = (1 + \langle x_i, x_j\rangle)^{\Theta}\ , \qquad (7)$$

$$K(x_i, x_j) = e^{-\Gamma\|x_i - x_j\|^2}\ . \qquad (8)$$

As is commonly done, we choose the kernel parameters (Θ or Γ) and penalty parameter of SVM (C) by cross-validation.

SVMs are binary classifiers to distinguish two groups of classes, and these binary classifiers are then combined via error-correcting code methods to obtain multiclass classifiers [12]. To summarize the procedure briefly, L binary classifiers are trained to distinguish between K classes using the coding matrix $\mathbf{M}_{K\times L}$, with elements $M_{kl} \in \{0, 1, -1\}$. Classifier l is trained on data of classes k for which $M_{kl} \neq 0$ with $\mathrm{sgn}(M_{kl})$ as the class label; it has no knowledge about classes k for which $M_{kl} = 0$. For example, in the case of one-vs-all classifiers ($L = K$), $M_{kl} = 1$, if $k = l$, otherwise $M_{kl} = -1$. For the one-vs-one classification strategy, on the other hand, $L = K(K-1)/2$, each classifier is trained on data from only two phoneme classes. Here all the elements of a column of the coding matrix **M** are set to 0 except for one $+1$ and one -1.

To combine the binary classifiers into a multiclass classifier, for a given test point x, the decision values of the L binary classifiers $\bar{h}(x) = [h_1(x), \cdots, h_L(x)]$ are obtained. Then, class k is chosen to be the predicted class $H(x)$ if the k^{th} row of the coding matrix, $\bar{M}_k = [M_{k1}, \cdots, M_{kL}], k = 1, \cdots, K$ has the minimum distance from $\bar{h}(x)$, i.e. $H(x) = \arg\min_k d(\bar{M}_k, \bar{h}(x))$. The distance measure is given as $d(\bar{M}_k, \bar{h}(x)) = \sum_{l=1}^{L}\xi(z_{kl})$ where ξ is some loss function and $z_{kl} = M_{kl}h_l(x)$. Commonly used loss functions include hinge – $\xi(z) = (1-z)_+ = \max(1-z, 0)$, Hamming – $\xi(z) = [1 - \mathrm{sgn}(z)]/2$, exponential – $\xi(z) = e^{-z}$ and linear – $\xi(z) = -z$ loss functions.

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

$$K_e(x_i, x_j) = K(x_i, x_j) + K(x_i, -x_j) = K(x_i, x_j) + K(-x_i, x_j)\ , \qquad (9)$$

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

$$K_s(x_i, x_j) = \frac{1}{(2n+1)^2}\sum_{p=-n}^{n}\sum_{q=-n}^{n}K_e(x_i^{p\Delta}, x_j^{q\Delta})\ . \qquad (10)$$

As discussed previously, since PLP and MFCC are extracted from the short-time magnitude spectra, using *even kernel* or *shift-invariant* *kernel* for PLP and MFCC classification will not have any significant advantage over the standard (polynomial or RBF) kernels.

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

4. TEST METHODOLOGY

Experiments were performed on the realizations of six phonemes (/b/, /f/, /m/, /r/, /t/, /z/) extracted from the TIMIT database. This set includes examples from fricatives, nasals, semivowels and voiced and unvoiced stops. In addition, this set of phonemes provides pairwise discrimination tasks of a varying level of difficulty. Each class consists of approximately 1000 representative acoustic waveforms, of which 80% were used for training and 20% for testing; error bars were derived by considering five different such splits. Phonetic segments used in this work were obtained by applying to each waveform 64ms rectangular windows, which at 16kHz sampling frequency gives vectors in $\mathbb{R}^{1024}$, followed by normalization to unit norm. The natural space in which to perform classification of the waveforms is, therefore, the unit hypersphere $\mathbb{S}^{1023}$.

For comparison, 12th order MFCC and PLP representations of the data were taken, leading to 4 frames of 13 coefficients in both cases [13]. These frames were concatenated to give a representation in $\mathbb{R}^{52}$. Classification was performed using the two approaches described in Section 2 and Section 3. As noise statistics can be estimated during pause intervals (non-speech activity) between speech signals, we assume for all classification approaches that the noise statistics (i.e. the noise variance σ^2, for white noise) are known.

As pointed out before, PLP uses frames of magnitude spectra, it is less sensitive to time alignment. 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 align the data, a sliding window with $\Delta = 10$ sample shift (≈ 0.6 ms) over a range of ± 100 samples ($\approx \pm 6$ ms) was used. This gives 21 shifted instances for each representative x. 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 pitch range of speech. We experimented with other sample shifts Δ but in the generative models found that smaller values do not give noticeable performance improvements. For the discriminative (SVM) classifiers, we mainly used $\Delta = 25$ (≈ 1.5 ms) to reduce computational effort, in particular, in the evaluation of the shift-invariant kernel defined previously.

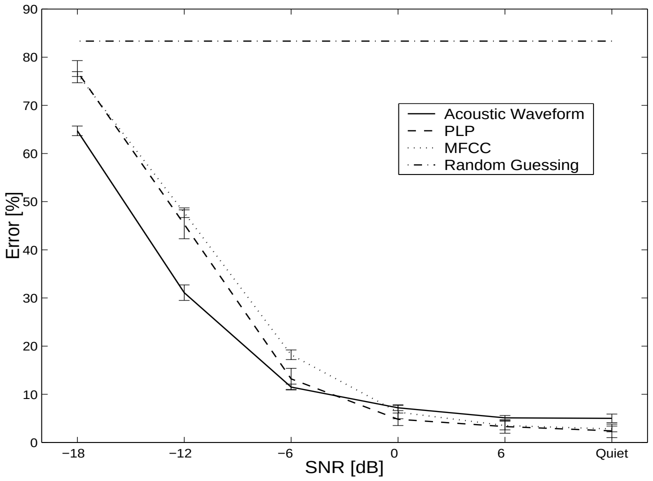

Fig. 1. Multiclass error rate for generative classification using MPPCA. The three curves are for waveform, PLP and MFCC representations. Dashed-dotted line represents the error rate for random guessing. In the case of PLP and MFCC, the classification is performed using density models estimated under matched noise conditions. The density model used for classification in the acoustic waveform domain is trained on clean data only and then adjusted according to (4) for noise modelling.

5. RESULTS

5.1. Results of Generative Methods

For MPPCA we tested systematically a variety of combinations of number of mixture components c and dimensions retained in each PPCA component, s. The best classification rate in high noise, for the waveform representation, is obtained for a mixture of $c = 4$ components with a PPCA dimension of $q = 500$ is used for waveforms. No improvement is observed when more components are used. This is likely because of the limited amount of data: the number of parameters in MPPCA scales as $\mathcal{O}(cst)$, which for $(c, s) = (4, 500)$ and $t = 1024$ is of order 2×10^6. This is already rather more than the number of data points (of order 40,000 per phoneme if for each example we include 21 shifts and all inverted images [$x \rightarrow -x$]), and increasing the number of components further is then likely to lead to significant overfitting. Reducing dimensionality by taking $s < t$ is beneficial here; without this, i.e. for $s = t$, the number of parameters is so large that essentially only a single Gaussian can be fitted reliably. The same trend was also observed with the PLP classifiers, with no consistent improvement seen when multiple component mixtures were used in place of a single Gaussian.

Figure 1 shows a comparison of the multiclass error rates for MPPCA applied to waveforms, PLP and MFCC representations. As expected PLP and MFCC outperform waveforms at high SNR with PLP being superior to MFCC. When the SNR is lower than 0dB, the waveform classification has higher accuracy and remains significantly above chance level even at -18dB.

The sensitivity of PLP to mismatch between training and testing conditions is illustrated in Figure 2. The curves show the performance of PLP trained at a fixed noise level and then tested on the

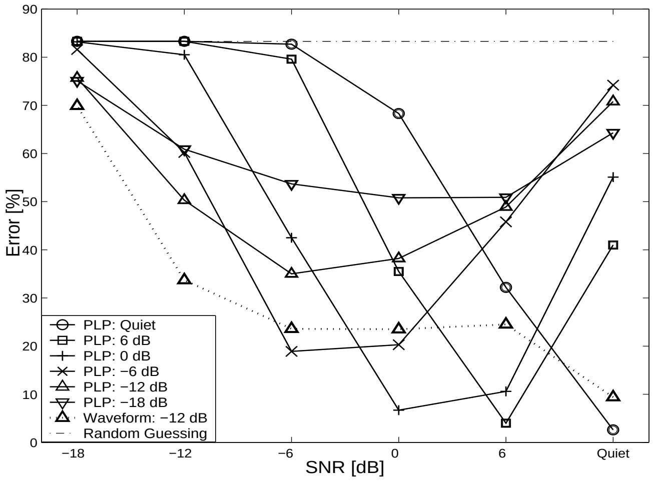

Fig. 2. Multiclass error rate for generative classifiers using the PLP representation. Density models are trained on data corrupted by fixed levels of noise as given in the legend. Each curve shows one classifiers tested across all levels of test noise. The dotted curve compares with the result of using the waveform classifier adjusted for data at -12dB SNR.

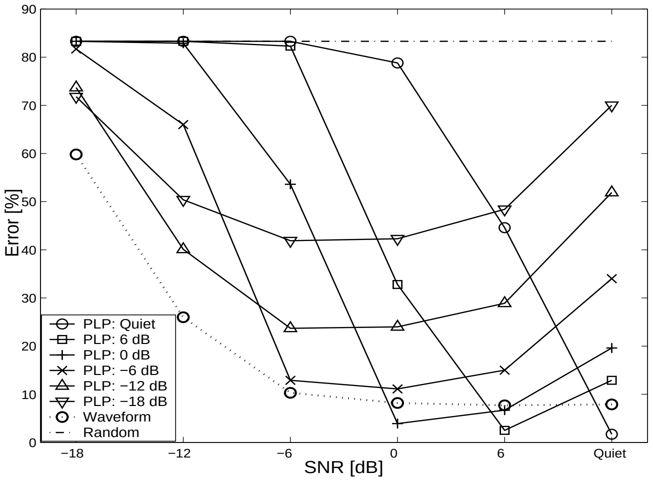

Fig. 3. Multiclass error rate for discriminative (SVM) classifiers in the PLP and acoustic waveform domains. SVMs for acoustic waveforms are trained on clean data while, for PLP, training is done on noisy data sets with SNR as indicated in the legend. Polynomial and shift-invariant even-polynomial kernels are used for PLP and acoustic waveform representations respectively.

range of noise levels. From the results it can be seen that PLP is very sensitive to mismatch between training and test conditions. For example PLP trained at 6dB has an error of 5% at matched test conditions but if the noise level differs by 6dB the error rate increases to 35%. For comparison a plot of the waveform classifier trained on quiet data and adjusted to -12dB is shown. The performance of the that classifier is less sensitive to mismatch and would make it favourable if there was an error in estimating σ^2.

5.2. Results of Discriminative Methods

Phoneme classification with MFCC using SVMs gives similar but slightly inferior performance to the PLP representation. Therefore, results of the classification accuracy for PLP and acoustic waveform representations of speech and their robustness to additive noise are reported here. For data corrupted by noise, the acoustic waveforms are normalized to $\sqrt{1+\sigma^2}$ where σ^2 is the noise variance. This is done to keep the norm of the signal component roughly independent of noise. In the case of PLP, we experimented with both this normalization and normalization to unity (as used throughout in the generative approach) independently of SNR, choosing the latter as it gave better performance. PLP features are standardized, i.e. scaled and shifted to have zero mean and unit variance on the training set.

Regarding the binary SVM classifiers, comparable performance is obtained with polynomial and RBF kernels for PLP representation so we show results for the former. For the waveform representation, the polynomial kernel performed better than the RBF kernel and the even polynomial kernel outperformed both. The shift-invariant even-polynomial kernel, finally, performed significantly better than all of the other kernels as discussed below. Here we use the shift range of ± 100 samples with $\Delta = 25$ samples. We also investigated the performance of the classifiers by adding time shifts to the test sets for the same range and Δ. In this particular case, the output of a binary classifier $h_l(x)$ for a test point x is given by the mean of the outputs of that binary classifier for time shifted versions of x, $h_l(x) = \frac{1}{2n+1} \sum_{p=-n}^{n} h_l(x^{p\Delta})$.

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

This can now be contrasted with the results for a classifier based on acoustic waveform data. One observes that although the performance of this classifier on clean data (7.5% error) is worse than that obtained by PLP classifer trained on clean data, it is significantly more robust to larger test noise levels as compared to the PLP classifier. For instance, there is no significant change in classification error (8%) up to a test noise level as high as 0dB SNR, whereas at the same SNR the corresponding PLP classifier trained on clean data has an error rate of 78%.

It should be emphasized that best performance using acoustic waveform classifiers is obtained when training is performed on clean data; training on noisy data (results not shown) leads to poorer performance. This is a significant advantage: the acoustic waveform classifier can be trained once and for all on clean data and used with a broad range of test noise conditions; for the PLP classifiers, on the other hand, separate classifiers need to be constructed for dif-

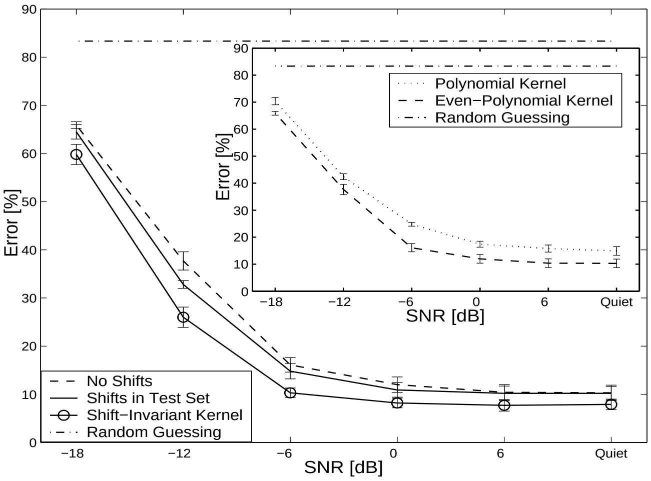

Fig. 4. Multiclass error rate for SVM classifiers with polynomial, even-polynomial and shift-invariant even-polynomial kernels in acoustic waveform domain. The inset figure compares the performance of waveform classifiers using polynomial and even-polynomial kernels. The main figure compares the performance of classifiers using even-polynomial kernel and shift-invariant even-polynomial kernel.

ferent noise levels to give good performance. Even then, optimal performance of the PLP classifiers which is achieved under matched conditions [10], is heavily dependent upon a fairly accurate estimate of the SNR (σ^2) and even a small error in this estimate can have a dramatic effect on classification performance.

In Section 3, the shift-invariant even-polynomial kernel was proposed. Figure 4 provides a quantitative assessment of the merits of designing a kernel to incorporate this shift and sign invariance. The inset figure in Figure 4 shows the classification results for polynomial and even polynomial kernels in the acoustic waveform domain with no shifted versions of acoustic waveforms included in training or testing. It is clear that the even kernel leads to a reduction of around 5% in the error rates across all levels of SNR. This is a significant improvement given the fact that the even-polynomial kernel takes into account just one physical property of speech perception, *i.e.* sign invariance, and suggests that further improvements could be obtained by incorporating additional prior knowledge into the kernel design. To this end, comparison of results of (a) even-polynomial kernel with no shifts in either training or test sets (b) even-polynomial kernel with shifts in only test set and (c) shift-invariant even-polynomial kernel is provided in the main plot of Figure 4. Almost no improvement is achieved by adding shifts of acoustic waveforms to the test set for SNRs above 0dB, however, there is approximately $1 - 3\%$ improvemener for SNRs of around 0dB and below. The figure further shows that the improvements obtained using the shift-invariant kernel are much more significant: approximately 2.5% for SNRs around 0dB and above, and $5 - 11\%$ for lower SNRs.

6. CONCLUSIONS

The robustness of phoneme classification to additive white Gaussian noise was investigated in experiments using generative and discriminative classifiers. Our results show that while PLP representation facilitates very accurate recognition of phonemes under matched conditions (especially for clean data), its performance suffers severe degradation with noise mismatch between training and testing conditions. On the other hand, the high-dimensional acoustic waveform representation, even though not as accurate as PLP classification on clean data, is more robust to additive noise and can tolerate significant mismatch between training and testing conditions. This observation holds for both generative and discriminative classifiers, and suggests that high-dimensional speech representations, such as acoustic waveforms, may be a more suitable front end for robust ASR than representations which involve non-linear dimension reduction. Optimal classification algorithms in the space of acoustic waveforms are still an open question. It is worth noting that density models in the domain of acoustic waveforms, lend themselves for very simple noise compensation. Furthermore, in the case of discriminative classifiers, major gains can be made by even straightforward custom kernel designs. There is still a considerable space for improvements in that direction, and kernel functions which are finely tuned to the physical properties of speech data will play a crucial role in further error reduction.

7. REFERENCES

[1] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Comm.*, vol. 22, no. 1, pp. 1–15, 1997.

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

[3] G. Miller, G. Heise, and W. Lichten, "The intelligibility of speech as a function of the context of the test materials," *Journal of Experimental Psychology*, vol. 41, pp. 329–335, 1951.

[4] G. Miller and P. Nicely, "An analysis of perceptual confusions among some english consonants," *Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, 1955.

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

[6] 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.

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

[8] Michael E. Tipping and Christopher M. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

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

[10] 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.

[11] J. Hamaker A. Ganapathiraju and J. Picone, "Applications of support vector machines to speech recognition," *IEEE Trans. on Signal Processing*, vol. 52, no. 8, pp. 2348–2355, 2004.

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

[13] Daniel P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource.