ROBUSTNESS OF PHONEME CLASSIFICATION USING SUPPORT VECTOR MACHINES: A COMPARISON BETWEEN PLP AND ACOUSTIC WAVEFORM REPRESENTATIONS

Jibran Yousafzai, Zoran Cvetković, Peter Sollich and Matthew Ager

King's College London
Strand, London, WC2R 2LS, UK

ABSTRACT

Robustness of phoneme recognition to additive noise is investigated for PLP and acoustic waveform representations of speech using support vector machines (SVMs) combined via error-correcting code methods. While recognition in the PLP domain attains superb accuracy on clean data, it is significantly affected by mismatch between training and testing noise levels. The classification in the high-dimensional acoustic waveform domain, on the other hand, is more robust to additive noise. Moreover, these classifiers perform best when trained on clean data. We also show that the simpler structure of the waveform representation allows one to improve performance using custom-designed kernel functions.

***Index Terms*—** Speech recognition, robustness, support vector machines, error correction codes, 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, humans attain a major portion of their inherent 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. 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 basis hypothesis of our work is that compressed representations of speech such as PLP [5], because of the strong nonlinearities that link them to the original acoustic waveforms, lead to distributions of different speech units that can be harder to separate and may vary more with noise. In this study, we test this hypothesis by performing classification of phonemes in presence of noise using support vector machines in the acoustic waveform and PLP domains, with particular emphasis on exploring the mismatch between training and testing conditions. We review the classification approach in Section 2. The experiments, results of which are reported in Section 3, show that while classification using the PLP representation achieves considerably better results on clean data than the acoustic waveform representation, it is much more sensitive to additive noise not explicitly represented in the training data. A waveform classifier trained only on clean data, 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 speech recognition. Finally, Section 4 draws some conclusions.

2. METHODS

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 [6]. A kernel-based decision function has the form

$$h(x) = \sum_i \alpha_i y_i K(x, x_i) + b \tag{1}$$

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 (2) and (3), respectively,

$$K(x_i, x_j) = (1 + \langle x_i, x_j \rangle)^d \ , \tag{2}$$

$$K(x_i, x_j) = e^{-\Gamma\|x_i - x_j\|^2} . \quad (3)$$

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

We also introduce the *even-polynomial* kernel 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. This can be accounted for by using even-polynomial kernels of the form

$$K(x_i, x_j) = (1 + \langle x_i, x_j \rangle)^d + (1 - \langle x_i, x_j \rangle)^d . \quad (4)$$

In this work, SVMs are used as binary classifiers to distinguish two groups of phonemes, and these binary classifiers are then combined via error-correcting code methods to obtain multiclass classifiers [7, 8]. To summarize the procedure briefly, L binary classifiers are trained to distinguish between K phoneme classes using the coding matrix $\mathbf{M}_{K \times L}$, with elements $M_{kl} \in 0, \pm 1$. 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 $M_{kl} = 0$. 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, $L = K(K-1)/2$ and each classifier is trained on data from only two phoneme classes: All elements of each column of the coding matrix $\mathbf{M}$ are set to 0 except for one $+1$ and one -1. We also explore a recently proposed hybrid "all-and-one" method [9]. Broadly, this requires training both one-vs-one and one-vs-all classifiers; the latter are used to select the top two candidate classes and the final decision between these is made using the relevant one-vs-one classifier.

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, one chooses class k as 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 $\bar{h}(x)$,

$$H(x) = \arg\min_k d(\bar{M}_k, \bar{h}(x)) . \quad (5)$$

The distance measure is given by

$$d(\bar{M}_k, \bar{h}(x)) = \sum_{l=1}^{L} \xi(z_{kl}) \quad (6)$$

where ξ is some loss function and $z_{kl} = M_{kl} h_l(x)$. We show results below for the exponential loss function $\xi(z) = e^{-z}$, which performs comparably or better than the Hamming ($\xi(z) = [1 - \mathrm{sgn}(z)]/2$) and hinge ($\xi(z) = (1-z)_+ = \max(1-z, 0)$) losses.

For each of the classifier combination methods above we investigate the classification accuracy for both PLP and acoustic waveform representations of speech, and their robustness to additive noise. Note that no noise compensation methods are used in order to have a fair comparison between both representations. The results are reported in the next section.

3. RESULTS

Classification is performed on the following six phonemes from the TIMIT database: /b/, /f/, /m/, /r/, /t/ and /z/. Each phoneme set consists of 1000 examples. A rectangular window of 64 ms duration is applied to speech waveforms. All waveforms, $x_i \in \mathbf{R}^{1024}$, from the six phoneme classes are normalized to have unit norm. The PLP representation, $p_i \in \mathbf{R}^{52}$, is obtained by finding the 12^{th} order cepstral coefficients from each of four consecutive frames across a speech waveform. When the data is corrupted by additive 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 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. One-vs-one classifiers were trained on 800 examples and tested on 200 examples per class. For one-vs-all classifiers, the training set size for the single class was increased by a factor of five to balance the number of training examples from the other classes; this was done by adding waveforms shifted by $\pm 50 (\approx 3$ ms) and $\pm 100 (\approx 6$ ms) samples.

Regarding the binary SVM classifiers, comparable performance is obtained with polynomial and RBF kernels for the PLP representation so we show results for the former. For the waveform representation, the polynomial kernel performed better than the RBF kernel but as discussed below the even polynomial kernel outperformed both.

Robustness to Additive Noise

Classification results in the PLP and acoustic waveform domains are shown in Figure 1. The best results for both domains are compared here, i.e. even-polynomial kernel with all-and-one coding for waveforms and polynomial kernel with one-vs-all coding for PLP. 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 40%, while classification is at the level of chance for SNR less than 0dB. This observation is quite general: the PLP classifiers are highly sensitive to mismatch between the training and test conditions. In particular, 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, 30% 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.

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 (10% error) is worse than that obtained by PLP classifier trained on clean data, it is significantly more robust to larger test noise levels compared to the PLP classifier. For instance, we do not observe a significant change in classification error (12%) 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 quiet data and used with performance broadly comparable to PLP across a broad range of test noise conditions; for the PLP classifiers, on the other hand, separate classifiers need to be constructed for different noise levels to give good performance.

Multiclass Coding Methods

We compare different coding techniques in Figure 2 and 3, for classification in the acoustic waveform domain with polynomial and even-polynomial kernel respectively. For the polynomial kernel, we observe that one-vs-all coding performs better than one-vs-one for all levels of noise. For the even-polynomial kernel, on the other hand, one-vs-all gives better results than one-vs-one at low noise levels while the situation is reversed under extremely noisy conditions. The all-and-one coding strategy proves to be a useful compromise solution in this case that works close to optimal across all SNRs.

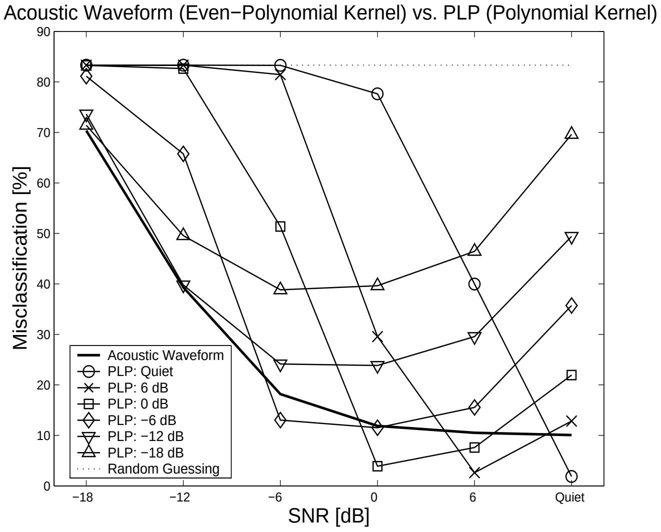

Fig. 1. *Classification results for PLP and acoustic waveform domains. SVMs for acoustic waveforms are trained on clean data and for PLP, training is done on noisy data sets with SNR as indicated by the legend.*

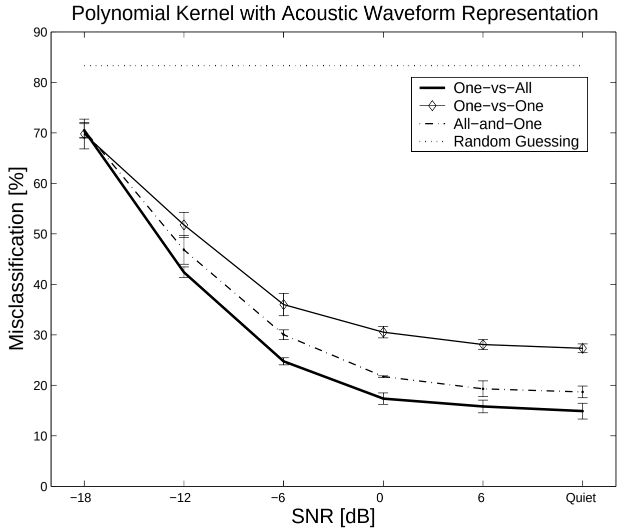

Fig. 2. *Classification results in the acoustic waveform domain using a polynomial kernel with one-vs-all, one-vs-one and all-and-one coding schemes*

Effect of Custom Kernel Design

In Section 2, we suggested the use of even-polynomial kernel and presented classification results using this kernel in this section. Now we discuss the quantitative effects of this kernel. Figure 4 shows the classification results for both polynomial and even-polynomial kernels in the acoustic waveform domain. We observe that the latter kernel choice leads to a reduction of around $5-10\%$ in the error rates across all levels of SNR except in extreme noisy conditions (i.e. -18dB SNR). This is a significant improvement given the fact that the even-polynomial kernel takes into account just one physical property of speech perception, and suggests that further improvements could be obtained by incorporating more prior knowledge into the kernel design. We are currently investigating the effects of time alignment on the recognition performance and different methods to embed this into custom designed kernels, with preliminary work showing promising results.

4. CONCLUSIONS

The robustness of phoneme classification to additive noise was investigated in numerical experiments using SVMs for acoustic waveform and PLP representations. The results, obtained using different coding schemes and loss functions, 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

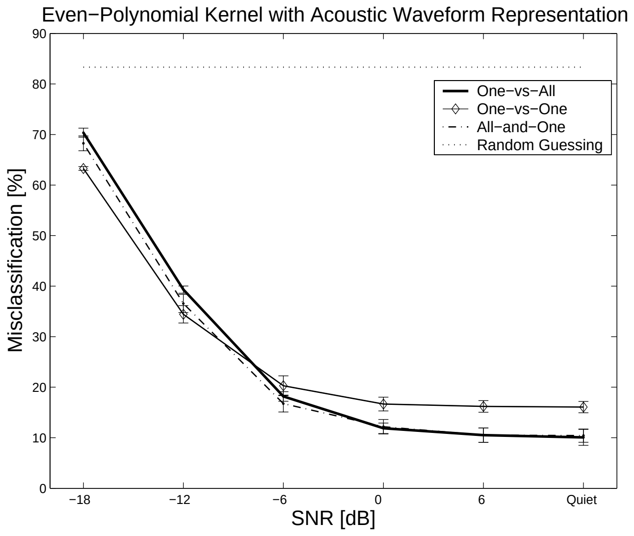

Fig. 3. *Classification results in the acoustic waveform domain using an even-polynomial kernel with one-vs-all, one-vs-one and all-and-one coding schemes*

classification on clean data, is more robust to additive noise and can tolerate significant mismatch between training and testing conditions. We showed further that the physically intuitive nature of the acoustic waveform representation allows one to custom design SVM kernels by incorporating prior knowledge, thus improving classification performance. In future work we plan to investigate larger phoneme sets and extend our work on custom-designed kernels to incorporate invariance to time alignment; it will also be interesting to study the effects of explicit noise-compensation techniques in both the PLP and waveform domains.

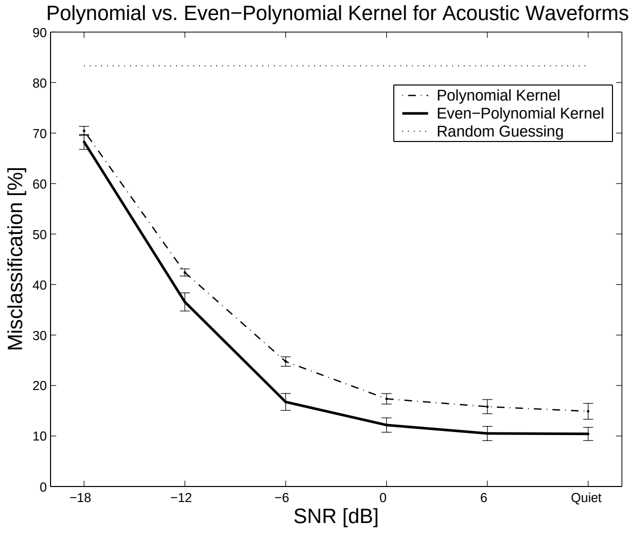

Fig. 4. *Even-polynomial vs. polynomial kernel for classification in acoustic waveform domain*

5. REFERENCES

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

[2] Jason J. Sroka and Louis D. Braida, "Human and machine consonant recognition," *Speech Communication*, 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] George A. Miller and Patricia E. 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," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, April 1990.

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

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

[8] Erin L. Allwein, Robert E. Schapire, and Yoram Singer, "Reducing multiclass to binary: a unifying approach for margin classifiers," *Journal of Machine Learning Research*, vol. 1, pp. 113–141, 2001.

[9] Nicolás García-Pedrajas and Domingo Ortiz-Boyer, "Improving multiclass pattern recognition by the combination of two strategies," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 1001–1006, 2006.