

Improving the Robustness of Phoneme Classification Using Hybrid Features

Jibrán Yousafzai[†], Zoran Cvetković[†] and Peter Sollich[‡]

Department of Electronic Engineering[†] and Department of Mathematics[‡]
King's College London

Abstract—This work is concerned with improving the robustness of phoneme classification to additive noise with hybrid features using support vector machines (SVMs). In particular, the cepstral features are combined with local energy features of acoustic waveform segments to form a hybrid representation. The local energy features are taken into account separately in the SVM kernel, and a simple subtraction method allows them to be adapted effectively in noise. This hybrid representation with mean and variance normalization of the cepstral features contributes significantly to the robustness of phoneme classification and narrows the performance gap to the ideal baseline of classifiers trained under matched noise conditions. Further improvements are obtained by extending the multiclass prediction method from standard discrete error-correcting codes to adaptive continuous codes.

Index Terms—Hybrid features, Phoneme classification, Robustness, Support vector machines, Continuous codes

I. INTRODUCTION

Accuracy of automatic speech recognition (ASR) systems rapidly degrades when operated in adverse acoustical environments. While language and context modelling are essential for reducing many errors in speech recognition, accurate recognition of phonemes and the related problem of classification of isolated phonetic units is a major step towards achieving robust recognition of continuous speech [1, 2]. Indeed, phoneme classification has been the subject of several recent studies [3–6].

State-of-the-art ASR systems use cepstral features, normally some variant of Mel-frequency cepstral coefficients (MFCC) or Perceptual Linear Prediction (PLP) [7], as their front-end for processing of speech signals. These representations are derived from the short term magnitude spectra followed by non-linear transformations to model the processing of the human auditory system. This work is focused on the task of phoneme classification using these features in the presence of additive noise although we believe the results also have implications for the construction of continuous speech recognition systems. We propose that a set of hybrid features, formed by combining the standard cepstral features with the local energy features of acoustic waveform segments, can contribute to the robustness of phoneme classification in noise. The local energy features can then be adapted effectively in noise by taking into account the approximate orthogonality of clean speech and noise. A two-stage learning framework is used for classification. In the first stage, support vector machines (SVMs) are used as baseline binary classifiers. Error-correcting output code (ECOC) methods are then used to form a multiclass classifier. Since ECOC methods use predefined discrete codes which assign same weights to the learned binary SVM classifiers regardless of the underlying training data and classification performance, we use the outputs (scores) of the binary SVMs to learn adaptive continuous codes [8] in the second stage to improve the classification performance. Our experiments demonstrate the effectiveness of the hybrid features in improving the robustness of classification in noise. Moreover, it is shown that the continuous

relaxation of output codes further improves the classification performance and narrows the performance gap to the ideal classifiers trained under matched noise conditions [9] *e.g.* the hybrid features with continuous codes achieve an average improvement of around 8% over the PLP features with discrete ECOC in the presence of additive white Gaussian noise.

The SVM approach to classification of phonemes using error-correcting output codes (ECOC) [10] and continuous codes is reviewed briefly in Section II. Section III presents the proposed hybrid features and their adaptation in the presence of noise. Experimental setup is discussed in Section IV. The classification results in the presence of noise are reported in Section V. Finally, Section VI draws some conclusions.

II. CLASSIFICATION METHOD

An SVM [11] binary classifier estimates decision surfaces separating two classes of data. In the simplest case these are linear, but for most pattern recognition problems one requires nonlinear decision boundaries. These are constructed using kernels instead of dot products, implicitly mapping data points to high-dimensional feature vectors. A kernel-based decision function which classifies an input vector \mathbf{x} is expressed as

$$h(\mathbf{x}) = \sum_i \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b, \quad (1)$$

where K is a kernel function, \mathbf{x}_i , $y_i = \pm 1$ and α_i , respectively, are the i -th training sample, its class label and its Lagrange multiplier, and b is the classifier bias determined by the training algorithm. Two commonly used kernels are the polynomial and radial basis function (RBF) kernels given by

$$K_p(\mathbf{x}, \mathbf{x}_i) = (1 + \langle \mathbf{x}, \mathbf{x}_i \rangle)^\Theta. \quad (2)$$

$$K_r(\mathbf{x}, \tilde{\mathbf{x}}) = e^{-\Gamma \|\mathbf{x} - \tilde{\mathbf{x}}\|^2}. \quad (3)$$

Comparable performance is achieved with both kernels; results are reported for the polynomial kernel throughout this study.

SVMs are binary classifiers trained to distinguish between two groups of classes. For multiclass classification, they can be combined via predefined *discrete* error-correcting output codes (ECOC) [10]. To summarize the procedure briefly, N binary classifiers are trained to distinguish between M classes using the coding matrix $\mathbf{W}_{M \times N}$, with elements $w_{mn} \in \{0, 1, -1\}$. Classifier n is trained on data of classes m for which $w_{mn} \neq 0$ with $\text{sgn}(w_{mn})$ as the class label; it has no knowledge about classes $m = 1, \dots, M$ for which $w_{mn} = 0$. The class m that one predicts for test input \mathbf{x} is then the one that maximizes the confidence, $\rho_m(\mathbf{x}) = -\sum_{n=1}^N \chi(w_{mn} h_n(\mathbf{x}))$. Here χ is some loss function and $h_n(\mathbf{x})$ is the output of the n^{th} classifier. The error-correcting capability of a code is commensurate to the minimum Hamming distance between pairs of code words

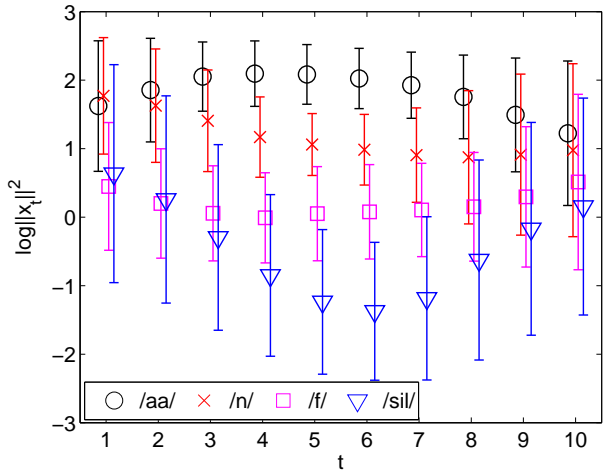


Fig. 1. Time profile (mean \pm standard deviation) of the local energy features τ (with $T = 10$ segments) of phoneme classes /aa/, /n/, /f/ and /sil/. The difference in the profiles indicates a correlation of these features with phoneme identity, which suggests that they should be useful for discriminating among phoneme classes.

[10]. Therefore, classification performance benefits from using error-correcting codes with larger Hamming distances between their rows. However one must also take into account the choice of accurate binary classifiers and the computational costs associated with such a code. In our previous work [12] on phoneme classification on a subset of the TIMIT database, a code formed by the combination of the *one-vs-one* (pairwise) and *one-vs-all* codes was used as this achieved better classification performance than either of the codes individually. A similar technique that implicitly combined the two different coding schemes to form an *all-and-one* coding strategy also improved classification performance in another study [13]. Since the construction of one-vs-all binary classifiers for a problem with large datasets is not computationally feasible, only one-vs-one ($N = M(M - 1)/2$) classifiers are used in the present study. A number of loss functions were compared; the hinge loss $[\chi(z) = (1 - z)_+ = \max(1 - z, 0)]$ performed best and is used throughout this paper. It should be noted that ECOC is a general method for solving multiclass problems by combining multiple binary classifiers, independently of the specific application and the learning algorithm used to construct the classifiers. Furthermore, the decoding scheme with discrete codes assigns the same weight to each learned binary classifier regardless of its performance.

In [8], Crammer & Singer proposed an approach developed from the theory of SVMs to address multiclass classification by solving a single optimization problem. Rather than using predefined discrete codes, this method improves the performance of the output codes by relaxing them to *continuous* codes. This relaxation procedure is cast as a constrained optimization problem. In this work, we use this procedure as a second stage of learning in the binary SVM *score* space to obtain a continuous code. Again, a polynomial kernel is used to learn continuous codes for the results presented in this paper.

III. HYBRID FEATURES

One of the reasons for which speech recognition in the cepstral domain is very sensitive to additive noise is the considerable distortion of decision boundaries caused by the noise. A standard noise adaptation technique for most large vocabulary ASR systems

using cepstral features is cepstral mean-and-variance normalization (CMVN) [14]. The algorithm computes the mean and variance of the feature vectors of a sentence and standardizes the cepstral vectors of that sentence so that each feature has zero mean and a fixed variance. We use a standardization that sets the variance of each feature to be equal to the inverse of the dimension of the cepstral feature vector so that, on average, the cepstral feature vectors have unit (squared) norm. CMVN contributes significantly to robustness by alleviating the effects of distortions caused by additive noise and linear filtering and limiting the range of deviation in the cepstral features. However, due to the non-linear transformations in the feature extraction process, the distortion in the cepstral features caused by additive noise is not merely an additive bias that can be fully characterized only by noise. Instead, this bias is jointly determined by speech, noise type and noise level in a complicated fashion, with the different components difficult to separate as detailed in [14].

In this paper, we propose that phoneme classification can be improved, in the presence of noise, by combining cepstral features with local energy features of acoustic waveform segments to form a hybrid representation. To this end, let $\mathbf{x} \in \mathbb{R}^D$ be a D -samples long acoustic waveform representation of a phoneme, and \mathbf{c} be the cepstral representation of the same phoneme. For speech recognition tasks, a *zero*-th order cepstral coefficient c_0 for each frame of speech is included in the cepstral representation, \mathbf{c} , due to its strong correlation to the energy level of that frame [14, 15]. Given that the evolution of energy in a phoneme is closely associated with phoneme identity as illustrated in Figure 1, c_0 is a useful cue for phoneme classification. However, in the presence of noise, the adapted c_0 cepstral feature that is obtained by applying CMVN to the distorted c_0 feature will still exhibit a significant level of contamination. To overcome this issue, we propose to embed the exact information about the local energies of the acoustic waveform segments. A straightforward adaptation of these energy features can then be performed by taking into account the approximate orthogonality of clean speech and noise. This adaptation results in the distributions of the local energy features of noisy speech to be close to those of the clean speech. To obtain these features, the fixed length acoustic waveform, \mathbf{x} , of a phoneme is divided into T non-overlapping segments, $\mathbf{x}_t \in \mathbb{R}^{D/T}$, $t = 1, \dots, T$ as illustrated in Figure 2. Let $\tau = [\tau_1, \dots, \tau_T]$ denote the local energy features of these segments such that $\tau_t = \log_{10} \|\mathbf{x}_t\|^2$. Then, the cepstral feature vector \mathbf{c} (including c_0) is augmented with the local energy feature vector τ for the evaluation of a hybrid kernel given by

$$K_n(\mathbf{c}, \tilde{\mathbf{c}}, \tau, \tilde{\tau}) = K_p(\mathbf{c}, \tilde{\mathbf{c}}) \sum_{t=1}^T K_\varepsilon(\tau_t, \tilde{\tau}_t), \quad (4)$$

where

$$K_\varepsilon(\tau_t, \tilde{\tau}_t) = e^{-(\tau_t - \tilde{\tau}_t)^2 / 2a^2}, \quad (5)$$

and a is a parameter that is tuned experimentally. Note that in K_n , we sum the exponential terms over T segments rather than using the standard RBF kernel from (3). This is done in order to avoid the local energy features of certain segments dominating the evaluation of the kernel. Alternatively, the local energy features can be standardized in a manner similar to the cepstral features and then evaluated using an RBF/polynomial kernel. In this paper, we report results using the former method as it avoids the additional step of feature standardization however similar classification performance is obtained using both strategies. Furthermore, non-overlapping segments of speech are used to extract the local energy features of phonemes in order to avoid the smoothing of the time-profiles of these features and to make their

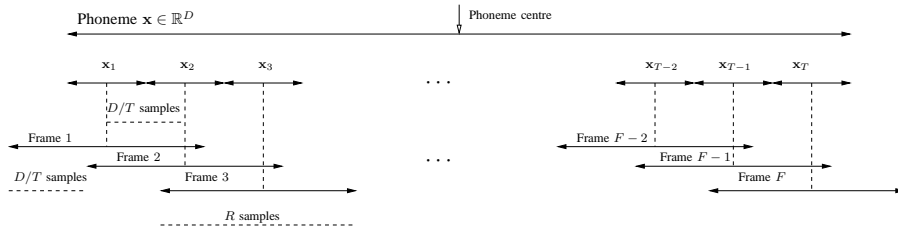


Fig. 2. Extraction of segments and frames from a waveform: an acoustic waveform, $\mathbf{x} \in \mathbb{R}^D$, is divided into T non-overlapping segments, each containing D/T samples. In addition overlapping frames, each containing R samples, are extracted to obtain the cepstral features, \mathbf{c} with an overlap of $R - D/T$ samples between two consecutive frames so that the frame rate equals the segment rate.

evolution more evident. It should also be noted that replacing the local energy feature vector, τ in (5) with a vector containing c_0 features of frames of a phoneme would lead to poor performance due to the significant contamination of the cepstral features.

In this work, we focus on classification of speech degraded by additive noise. For that purpose we will use classifiers that are trained in quiet conditions, with cepstral feature vectors of the test data adapted to noise using CMVN [14] and local energy features τ adapted as described next. Let $\mathbf{x} = \mathbf{s} + \mathbf{n}$, $\mathbf{x} \in \mathbb{R}^D$ be a noise corrupted waveform, where \mathbf{s} and \mathbf{n} represent the clean speech and the Gaussian noise vector, respectively. The energy of the clean speech can then be approximated as, $\|\mathbf{s}\|^2 \approx \|\mathbf{x}\|^2 - \|\mathbf{n}\|^2 \approx \|\mathbf{x}\|^2 - D\sigma^2$. The first approximation involved here is that, because speech and noise are uncorrelated, the vectors \mathbf{s} and \mathbf{n} are typically orthogonal. More precisely, $\langle \mathbf{s}, \mathbf{n} \rangle$ is of order $D^{-1/2}\|\mathbf{s}\|\|\mathbf{n}\|$ which can be neglected for large enough D . The second approximation then replaces the noise energy by its average value which is set by σ^2 , the noise variance per sample. We work here and throughout with a default normalization of waveforms to unit energy per sample, so that $1/\sigma^2$ is the SNR. Since σ^2 can be estimated during pause intervals (non-speech activity) between speech signals, we assume that its value is known. Applying these general arguments to the local energy features, we adapt these in the presence of noise by subtracting the estimated noise variance of a segment, $D\sigma^2/T$ from the energies of the noisy segments, *i.e.* $\tau_t = \log \|\mathbf{x}_t\|^2 - D\sigma^2/T$. This will provide an estimate of the local energies of the segments of clean speech. Following the reasoning above, using local energy features of shorter segments of acoustic waveform (lower D/T) would make fluctuations away from the orthogonality of speech and noise more likely, therefore K_ϵ should be evaluated on the energies of long enough segments of speech. It should be noted that the adaptation discussed here is performed only on the test features because training is performed in quiet conditions; adaptation of the local energy features of the training data is therefore not required.

IV. EXPERIMENTAL SETUP

Experiments are performed on the 'si' and 'sx' sentences of TIMIT. The training set consists of 3696 sentences from 168 different speakers. The core test set is used for testing which consists of 192 sentences from 24 different speakers not included in the training set. We remove the glottal stops /q/ from the labels and fold certain allophones into their corresponding phonemes using the standard Kai-Fu Lee clustering [16], resulting in a total of 48 classes. Among these classes, there are 7 groups for which the contribution of within-group confusions towards multiclass error is not counted [16].

In this study, we focus on investigating robustness in the presence of additive white Gaussian noise and pink noise from NOISEX-92 database. To test the classification performance in noise, each

sentence is normalized to unit energy per sample and then a noise sequence with variance σ^2 (per sample) is added to the entire sentence. It should be noted that SNR at the sentence level is fixed but SNR at the level of individual phonemes will vary widely.

Two separate forms of the cepstral representations, \mathbf{c} , are considered. First, a sequence of frames with frame duration of 25ms and a frame rate of 100 frames/sec is derived from each sentence. Then, 9 frames (105ms) closest to the center of a particular phoneme are concatenated to give a representation in \mathbb{R}^{351} where each frame is represented by 13 cepstral coefficients (including c_0), their time derivatives and second order derivatives. Second, all frames within a variable length phoneme segment and its transition regions are included in the representation as proposed by Clarkson *et al.* [17] to give a representation in \mathbb{R}^{196} . The latter achieves slightly better classification performance in both quiet and noise conditions due to the additional knowledge encoded in the representation. However the former representation may be more suitable for use with HMMs for continuous speech recognition. Cepstral representations, such as PLP, MFCC and RASTA are known to achieve comparable recognition accuracy [18]. For classification with SVMs, PLP achieved slightly better classification performance and therefore it is used as the default cepstral representation for all experiments reported in this paper

For cepstral features two train-test scenarios are considered: (i) training SVM classifiers using clean data and then performing CMVN of test data to adapt to noise, and (ii) training and testing under identical noise conditions. The latter scenario is an impractical target which could in be achieved in practice only if one had access to a large set of classifiers trained for different noise types and levels. Nevertheless, we present the classification results in matched training and testing conditions as a reference, since this setup is considered to give the optimal achievable performance with cepstral features [9].

In order to obtain the local energy features from the acoustic waveforms, phoneme segments are extracted from the phonetically hand labelled TIMIT sentences by applying a 100 ms rectangular window at the center of each phoneme waveform (of variable length), which at 16 kHz sampling frequency gives fixed length vectors in \mathbb{R}^{1600} . Each of these vectors is broken into $T = 10$ non-overlapping segments of equal length resulting in 10 local energy features per phoneme. As mentioned previously, the local energy features are extracted from non-overlapping segments of speech in order to capture the precise evolution of energy in time and to avoid smoothing of these features which can be caused by using overlapping segments. The hyperparameter of K_ϵ in (5) is set to $a = 0.5$. It should be noted that the noise adaptation in the kernel requires an estimate of σ^2 but assumes no knowledge about the noise type or the shape of the noise spectrum.

Regarding the binary SVM classifiers, results are reported for both kernels, K_n and K_p , for comparison. Fixed hyperparameter values

are used throughout for training binary SVMs: the degree of K_p is set to $\Theta = 6$ and the penalty parameter $C = 1$. To learn the continuous code, the development set of TIMIT is used for training. The scores of $N = 1128$ one-vs-one binary SVM classifiers are concatenated and normalized to give a representation in \mathbb{R}^{1128} . The regularization parameter and the degree of the kernel are set to $\beta = 100$ and $\Theta = 6$.

V. RESULTS

Classification results using SVMs for the 9-frame PLP representation (in \mathbb{R}^{351}) and its hybrid representation in the presence of additive white Gaussian noise are shown in Figure 3(a). For the PLP representation, classification results with kernels K_p are presented whereas K_n is used for classification with the hybrid features. First, the use of continuous codes with kernel K_p slightly improves the performance over the standard ECOCs. We also observe that the hybrid features perform better throughout all noise conditions. For instance, at 6dB SNR, using the hybrid features with continuous codes reduces the error by 12% compared to the standard PLP features with discrete ECOCs; most of this reduction in error, around 9%, can be attributed to using the proposed hybrid features.

The hybrid features (with kernel K_n) used with continuous codes perform better than any other classifier trained in quiet conditions *e.g.* the performance compared to that achieved using the PLP features with discrete ECOCs is improved from 21.4% to 20.1% in quiet conditions and on average by around 8% across all SNRs tested. Furthermore, classification with the hybrid features using kernel K_n performs better than the standard PLP features with kernel K_p in matched condition. Even though the improvement achieved by K_n in low noise conditions ($\text{SNR} \geq 6\text{dB}$) is not more than 1–2%, we achieve a significant gain in high noise *e.g.* a 6% performance gain at -6dB SNR. Similar results are obtained for classification in the presence of pink noise as shown in Figure 3(b). Our setup was also tested with other noise types, *e.g.* speech-weighted noise [19] (results not shown here) and similar conclusions apply. Even though the two-stage learning framework using hybrid features and continuous codes does not achieve the impractical target of the classification in matched conditions, the gain in classification accuracy is significant and the performance gap between the quiet and matched conditions classifiers is reduced.

Next, we compare the classification performance of the 9-frame PLP representation in Figure 3(a) with that of the representation proposed by Clarkson *et al.* in Figure 4 in the presence of additive white Gaussian noise. For brevity, we will denote these representations by P^9 and P^c respectively. It should be noted that the P^c representation uses all frames within a given phoneme and its transition regions and relies on exact sentence segmentation whereas the P^9 representation uses only 9 frames (105ms) around the middle of a phoneme. The local energy features are combined with each of these PLP representations individually to obtain hybrid P^9 and hybrid P^c representations. From the figures, the P^c representation performs slightly better than the 9-frame representation P^9 for all noise conditions *e.g.* using kernel K_n with continuous codes, the hybrid P^c representation achieves an error of 19% whereas 20.1% error is achieved by the hybrid P^9 representation in quiet conditions. Furthermore, an average increase in error rate by 2.5% for the hybrid P^9 representation over the hybrid P^c representation is observed across all SNRs tested. This shows that the performance degradation caused by the lack of information about exact phoneme segmentation in the hybrid P^9 representation is not considerable.

Finally, in Table I, results of some recent experiments on the TIMIT phoneme classification task in quiet condition are presented

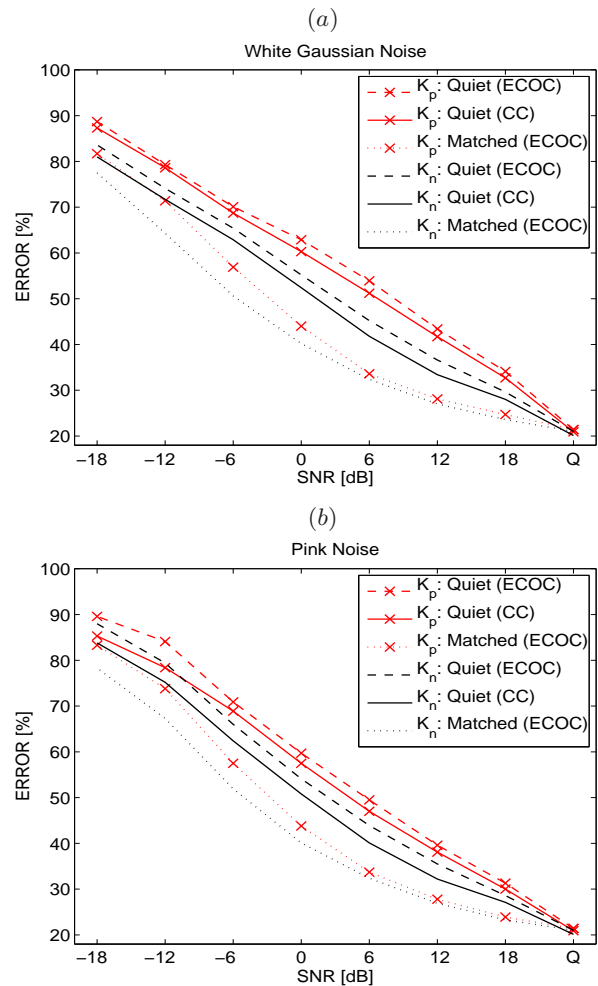


Fig. 3. SVM classification in the presence of (a) white Gaussian noise (b) pink noise from NOISEX-92 with the 9-frame PLP representation (\mathbb{R}^{351}) and its hybrid representation using K_p and K_n respectively, trained in both quiet and matched conditions. Classification results with the continuous code (CC) trained on binary SVM scores of clean development data using kernels K_p and K_n are also shown.

and compared with the results reported in this paper. Our classifiers improve over the best benchmark result in Table I in quiet conditions, 20.9% (obtained with RLS2 as described in [5]). It should be noted that these benchmarks use cepstral representations that encode information from the entire variable length phoneme and our result of 20.1% improves on these benchmarks even though we use a fixed length 9-frame cepstral representation. Further improvement can then be achieved by including all frames within a variable length phoneme and its transition regions, following the method considered by Clarkson *et al.* [17]. Furthermore, results presented in this paper significantly outperform the benchmarks in the presence of noise. For example, 77.8% classification error is reported by Rifkin *et al.* [5] at 0dB SNR in pink noise whereas our classifier using the hybrid representation achieves an error of 50.8% in same conditions as shown in Figure 3(b).

These experiments show that the hybrid representation, as proposed in this paper, contributes significantly to the robustness of phoneme classification to additive noise and therefore may be a suitable front-end for ASR systems. Although, we focus here on different flavors of

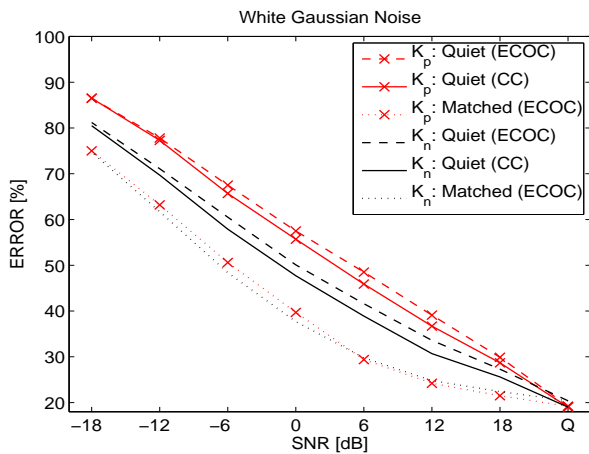


Fig. 4. SVM classification in the presence of white Gaussian noise with the P^c representation (\mathbb{R}^{196}) and its hybrid using kernels K_p and K_n respectively, trained in both quiet and matched conditions. Classification results with the continuous code (CC) using kernels K_p and K_n are also shown.

TABLE I

RESULTS OF SOME RECENTLY REPORTED EXPERIMENTS ON THE PHONEME CLASSIFICATION OF THE TIMIT CORE TEST SET IN QUIET CONDITION.

METHOD	ERROR
SVMs [17]	22.4%
Hidden CRF [3]	21.7%
Large Margin GMM [4]	21.1%
Hierarchical GMM [21]	21.0%
RLS2 [5]	20.9%
This Work (PLP - Fixed length, 9 Frames)	20.1%
This Work (PLP - Full variable phoneme length + Transitions)	19%

the PLP representations and their hybrids, hybrid MFCC features can be created in an analogous fashion. Similarly, improvements might be expected over the standard ETSI advanced front end (AFE) features [20] with the use of hybrid features and continuous codes. This might be an interesting area to explore in future work.

VI. CONCLUSIONS

The use of hybrid features is shown to contribute to the robustness of phoneme classification using SVMs. The approximate orthogonality of speech and noise is taken into account for an effective adaptation of the local energy features in noise of known SNR for evaluation in the SVM kernel. This significantly reduces the classification error in noise and narrows the performance gap to the classifiers trained under matched noise conditions. The multiclass classification method proposed by Singer *et al.* [8] further improves the results by continuous relaxation of the output codes. Finally, a comparison of the 9-frame PLP representation, P^9 with the P^c representation that uses the encoding scheme of Clarkson *et al.* [17] suggests that the performance degradation caused by the lack of information about the exact phoneme segmentation is not considerable.

REFERENCES

[1] A. Halberstadt and J. Glass, "Heterogeneous Measurements and Multiple Classifiers for Speech Recognition," *Proceedings of ICSLP*, pp. 995–998, 1998.
 [2] J. B. Allen, "Articulation and Intelligibility," *Synthesis Lectures on Speech and Audio Processing*, vol. 1, no. 1, pp. 1–124, 2005.

[3] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden Conditional Random Fields for Phone Classification," *Proceedings of Interspeech*, pp. 1117–1120, 2005.
 [4] F. Sha and L. K. Saul, "Large Margin Gaussian Mixture Modeling for Phonetic Classification and Recognition," *Proceedings of ICASSP*, pp. 265–268, 2006.
 [5] R. Rifkin, K. Schutte, M. Saad, J. Bouvrie, and J. Glass, "Noise Robust Phonetic Classification with Linear Regularized Least Squares and Second-Order Features," *Proceedings of ICASSP*, pp. 881–884, 2007.
 [6] M. Johnson *et al.*, "Time-domain Isolated Phoneme Classification Using Reconstructed Phase Spaces," *IEEE Transactions on Speech and Audio Proc.*, vol. 13, no. 4, pp. 458–466, 2005.
 [7] 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.
 [8] K. Crammer and Y. Singer, "On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2002.
 [9] M. Gales and S. Young, "Robust Continuous Speech Recognition using Parallel Model Combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 352–359, Sept. 1996.
 [10] T. Dietterich and G. Bakiri, "Solving Multiclass Learning Problems via Error-Correcting Output Codes," *Journal of AI Research*, vol. 2, pp. 263–286, 1995.
 [11] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.
 [12] J. Yousafzai, Z. Cvetković, P. Sollich, and B. Yu, "Combined PLP-Acoustic Waveform Classification for Robust Phoneme Recognition using SVMs," *Proceedings of EUSIPCO*, 2008.
 [13] N. Garcia-Pedrajas and D. 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.
 [14] C. Chen and J. Bilmes, "MVA Processing of Speech Features," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 15, no. 1, pp. 257–270, 2007.
 [15] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust Speech Recognition using the Modulation Spectrogram," *Speech Communication*, vol. 25, no. 1-3, pp. 117–132, 1998.
 [16] K. F. Lee and H. W. Hon, "Speaker-Independent Phone Recognition Using HMMs," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, 1989.
 [17] P. Clarkson and P. J. Moreno, "On the Use of Support Vector Machines for Phonetic Classification," *Proceedings of ICASSP*, vol. 2, pp. 585–588, 1999.
 [18] B. Milner, "A Comparison of Front-end Configurations for Robust Speech Recognition," *Proceedings of ICASSP*, vol. 1, pp. 797–800, 2002.
 [19] S. A. Phatak and J. B. Allen, "Consonant and Vowel Confusions in Speech-weighted Noise," *Journal of the Acoustical Society of America*, vol. 121, no. 4, pp. 2312–2326, April 2007.
 [20] ETSI standard doc., "Speech processing, Transmission and Quality aspects (STQ): Advanced front-end feature extraction," *ETSI ES 202 050*, 2002.
 [21] A. Halberstadt and J. Glass, "Heterogeneous Acoustic Measurements for Phonetic Classification," *Proceedings of EuroSpeech*, pp. 401–404, 1997.