# Towards Robust Phoneme Classification With Hybrid Features

Jibran Yousafzai[†], Zoran Cvetković[†] and Peter Sollich[‡]
Department of Electronic Engineering[†] and Department of Mathematics[‡]
King's College London

***Abstract*—In this paper, we investigate the robustness of phoneme classification to additive noise with hybrid features using support vector machines (SVMs). In particular, the cepstral features are combined with short term energy features of acoustic waveform segments to form a hybrid representation. The energy features are then taken into account separately in the SVM kernel, and a simple subtraction method allows them to be adapted effectively in noise. This hybrid representation contributes significantly to the robustness of phoneme classification and narrows the performance gap to the ideal baseline of classifiers trained under matched noise conditions.**

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

## 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 and allow for more accurate modelling when data is limited. However, due to the nonlinear processing involved in the feature extraction, even small amounts of additive noise may cause significant departures from the distributions learned on noiseless data. Large amount of training data is required to retrain the system to a new environment. To make the cepstral representations of speech less sensitive to noise, several techniques such as cepstral mean and variance normalization (CMVN) [8] and multi-condition/multi-style training [9, 10] have been proposed to reduce explicitly the effects of noise on spectral representations with the aim of approaching the optimal performance which is achieved when training and testing conditions are matched [11]. State-of-the-art feature compensation methods for the cepstral representation of speech include the ETSI advanced front end (AFE) [12] and vector Taylor series (VTS) [13, 14]. In this work, we propose that a set of hybrid features, formed by combining the standard cepstral features (MFCC) with the short term/local energy features of acoustic waveform segments, can contribute to the robustness of phoneme classification in noise. This is motivated by the fact that the local energy features can then be adapted effectively in noise by taking into account the approximate orthogonality of clean speech and noise. Note that this work is focused on the task of phoneme classification using the hybrid features in the presence of additive noise although we believe the results also have implications for the construction of continuous speech recognition systems.

The SVM approach to classification of phonemes using error-correcting output codes (ECOC) [15] 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 and classification results in the presence of noise are reported in Section V. Finally, Section VI draws some conclusions.

## II. Classification Method

An SVM [16] 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

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

where $K$ is a kernel function, $\mathbf{x}_i$, $y_i = \pm 1$ and $\alpha_i$, 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

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

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

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) [15]. 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 $\mathrm{sgn}(w_{mn})$ as the class label; it has no knowledge about classes $m = 1, \ldots, M$ 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 $\chi$ is some loss function and $h_n(\mathbf{x})$ is the output of the $n^{\text{th}}$ classifier. The error-correcting capability of a code is commensurate to the minimum Hamming distance between pairs of code words [15]. 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 [17] 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 [18]. The construction of one-vs-all binary classifiers for a problem with large datasets is not computationally feasible. For instance, in the simplest case of equal number of training points per class, the training time for one-vs-all classifiers scales cubically ($\mathcal{O}(M^3)$) whereas for one-vs-one classifiers, it scales quadratically ($\mathcal{O}(M^2)$) with the number of classes, $M$. Therefore, 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.

## III. Hybrid Features

One of the reasons for which speech recognition in the cepstral domains is very sensitive to additive noise is the considerable distortion of decision boundaries caused by the noise. State-of-the-art feature compensation methods for most large vocabulary ASR systems using the cepstral representation as their front-end for processing speech include the ETSI AFE [12] and vector Taylor series (VTS) [13, 14]. Additionally, cepstral mean-and-variance normalization (CMVN) [19] is used to standardize the cepstral feature by limiting the range of deviation in the cepstral features (of both train and test data). These feature compensation methods contribute significantly to robustness by alleviating the effects of distortions caused by additive noise and linear filtering. 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 especially in severe noise as detailed in [19].

The evolution of energy in a phoneme strongly correlates with phoneme identity and is encoded in the cepstral features which is a linear transform of Mel log powers. It is therefore a useful cue for accurate phoneme classification however the compensated cepstral features will still exhibit a significant level of contamination [19]. To improve robustness, we propose to embed the exact information about the short term energies of the acoustic waveform segments and treat them as separate set of features in the evaluation of the SVM kernel. A straightforward adaptation of these 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 [20]. 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. The fixed length acoustic $\mathbf{x}$ is divided into $T$ non-overlapping segments, $\mathbf{x}_t \in \mathbb{R}^{D/T},\ t = 1, \ldots, T$ , such that the centres of frame $t$ and segment $\mathbf{x}_t$ are aligned as illustrated in Figure 1. Let $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_T]$ denote the local energy features of these subsegments such that[1] $\tau_t = \log \|\mathbf{x}_t\|^2, t = 1, \ldots, T$. Then, the cepstral feature vector $\mathbf{c}$ is augmented with the local energy feature vector $\boldsymbol{\tau}$ for the evaluation of a hybrid kernel given by

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

where

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

and $a$ is a parameter that is tuned experimentally. Note that the local energy feature vector $\tau$ is treated as a separate set of features in the hybrid SVM kernel $K_c$ (which is a product of two valid kernels) as defined in (4) rather than fusing the local energy feature vector $\tau$ with the cepstral feature vector $\mathbf{c}$ on a frame-by-frame basis. Furthermore, we sum the exponential terms over $T$ segments rather than using the standard polynomial or RBF kernels in order to to avoid the local energy features of certain subsegments dominating the evaluation of the kernel. Alternatively, the local energy features can be standardized using CMVN in a manner similar to the cepstral features and then evaluated using an RBF or polynomial kernel. In this paper, we adopt 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 evolution more evident.

To investigate the robustness of the hybrid features to additive noise, we train the classifiers in quiet conditions with cepstral feature vectors standardized using CMVN [19]. Several noise compensation methods such as ETSI AFE and VTS followed by feature standardization using CMVN, are also compared in this study. Furthermore, the classification performance of the hybrid features is also compared with a multi-condition/multi-style classifier [9, 10] trained with standard cepstral features. It will be shown that the multi-style training with cepstral features significantly improves the robustness however it is highly sensitive to the mismatch between the noise type contaminating the training and test data.

It is essential that the local energy features are compensated for environmental distortions in order for the classifiers to perform effectively. The local energy features $\boldsymbol{\tau}$ are compensated for noise 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 $\sigma^2$, 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 $\sigma^2$ can be estimated during pause intervals (non-speech activity) between speech signals, we assume that its value is known. A number of approaches [21–23] have been proposed over the past years for robust estimation of SNR. Applying these general arguments to the local energy features, we compensate these in the presence of noise by subtracting the estimated noise variance of a subsegment, $D\sigma^2/T$ from the energies of the noisy subsegments, *i.e.* $\tau_t = \log \left| \|\mathbf{x}_t\|^2 - D\sigma^2/T \right|$. This will provide an estimate of the local energies of the subsegments of clean speech. Following the reasoning above, using local energy features of shorter subsegments of acoustic waveform (lower $D/T$) would make fluctuations away from the orthogonality of speech and noise more likely, therefore $K_\varepsilon$ should be evaluated on the energies of long enough subsegments of speech. It should be noted that the noise compensation discussed here is performed only on the test features because training is performed in quiet conditions; compensation 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

[1] We consider logarithms to base 10 throughout.

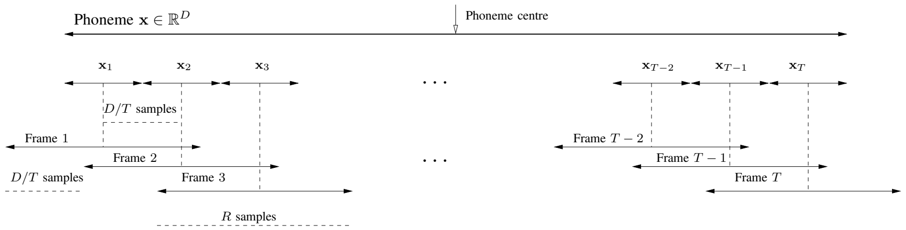


Fig. 1. Extraction of segments and frames from a waveform: an acoustic waveform, $\mathbf{x} \in \mathbb{R}^D$, is divided into $T$ non-overlapping segments, each containing $D/T$ samples. In addition overlapping frames, each containing $R$ samples, are extracted to obtain the cepstral features, $\boldsymbol{c}$ with an overlap of $R - D/T$ samples between two consecutive frames so that the frame rate equals the segment rate.

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 [24], 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 [24].

In this study, the robustness of phoneme classification is investigated in the presence of additive noise. Experiments are performed with white, pink, speech-weighted [25], factory floor and tank noises from the NOISEX-92 database. To test the classification performance of the cepstral features and acoustic waveforms in noise, each sentence is normalized to unit energy per sample and then a noise sequence with variance $\sigma^2$ (per sample) is added to the entire sentence. Three train-test scenarios for classification with cepstral features are considered: *i*) training SVM classifiers using clean data with standard noise compensation methods to clean the test features, *ii*) training on both clean and noisy data (with different noise levels and noise types) *i.e. multi-style* training, and *iii*) training and testing under identical noise conditions. The matched condition scenario is an impractical target; nevertheless, we present the results in matched training and testing conditions as a reference, since this setup is considered to give the optimal achievable performance with cepstral features [11]. It should be noted that the features of both training and test data are standardized using CMVN for all above-mentioned scenarios.

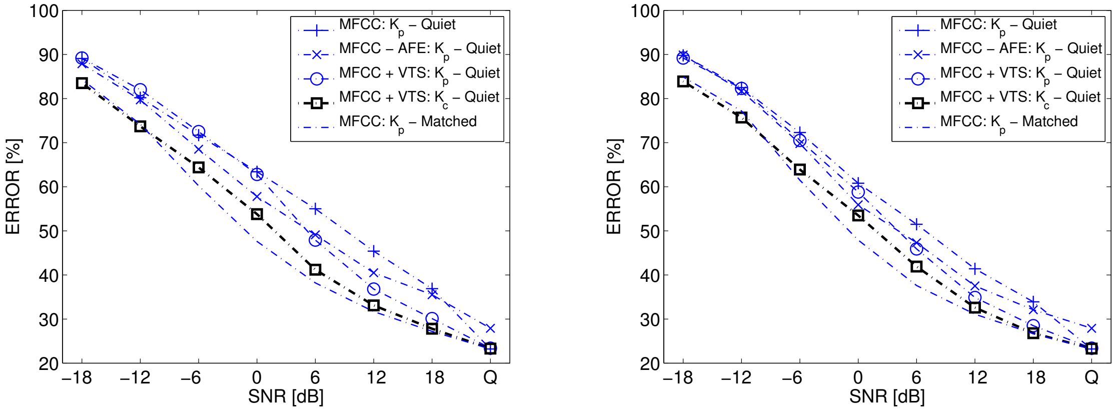
For the cepstral (MFCC) representation, $\mathbf{c}$, each sentence is converted into a sequence of 13 dimensional feature vectors, their time derivatives and second order derivatives which are combined into a sequence of 39 dimensional feature vectors. Then, $T = 10$ frames (with frame duration of 25ms and a frame rate of 100 frames/sec) closest to the center of a phoneme are concatenated to give a representation in $\mathbb{R}^{390}$. Along the same lines, each frame yields 14 AFE features (including log frame energy) and their time derivatives as defined by the ETSI standard giving a representation in $\mathbb{R}^{420}$ corresponding to 10 frames closest to the phoneme center. For noise compensation with vector Taylor series (VTS) [13, 14], a Gaussian mixture model (GMM) with 64 components was used to learn the distribution of the clean training data. 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 $T = 10$ local energy features per phoneme.

Regarding the SVM classifiers, comparable performance is obtained with $K_p$ and $K_r$ so we use the former as a baseline kernel and compare its performance with $K_c$. Initially, we experimented with different values of the hyperparameters to train the binary SVM classifiers but decided to use fixed values for all classifiers as they had a very little impact on the multiclass classification error: the degree of $K_p$ is set to $\Theta = 6$, the penalty parameter (for slack variables in the SVM training algorithm) to $C = 1$ and the value of $a$ in $K_\varepsilon$ from (5) is tuned experimentally and set to $0.5$. Using this setup, the results for SVM classification in the cepstral and acoustic waveform domains with custom-designed kernels, as detailed in section III, are reported in the next section.

## V. Results

In Figure 2, results of SVM phoneme classification with polynomial kernel $K_p$ in the presence of additive white and pink noise is shown for the MFCC cepstral representation using features compensation methods, VTS and AFE. For comparison, results are presented for the matched train and test conditions as well. The results demonstrate that the SVM classifier trained with the AFE representation outperforms MFCC representation for SNR below 18dB. On the other hand, the VTS-compensated MFCC features perform even better than the AFE in low noise conditions. However, for SNR below 0dB, the classification performance of VTS-compensated MFCC features degrades relatively quickly as compared to the AFE features. Since the (log) frame energy is included in the AFE features as defined by the ETSI standard, we only consider a hybrid representation formed by the combination of the local energy features and the VTS-compensated MFCC features with kernel $K_c$. The results show that this hybrid representation performs better than both noise compensation methods through all noise conditions and approaches the performance achieved under matched conditions. For instance, the hybrid representation achieves an average improvement of $5.5\%$ and $5.8\%$ over the standard VTS-compensated MFCC features and AFE features respectively, across all SNRs in the presence of white noise as shown in Figure 2(a). Similar conclusions are drawn when the test data is corrupted by pink noise as shown in Figure 2(b).

Another well known approach to make the cepstral features robust to additive noise is the multi-condition training setup. Here, the classifiers are trained on clean data as well as data corrupted by noise of different types and strengths. This style of training the classifiers contributes significantly towards the robustness of the cepstral features. In Figure 3(a), the multi-condition classifier is trained on clean data as well as data corrupted by white Gaussian noise *i.e.* 3 random noise contaminated versions of each feature vector in the training set, each corrupted with a different noise level ranging from 18dB to -18dB SNR, are added to the training set so that the size of the training set is 4 times larger than the original training set. When tested with data corrupted with the same noise type, the classification performance approaches that obtained



(a)

(b)

Fig. 2. SVM classification in the presence of (a) white noise (b) pink noise, using the MFCC representation with standard kernel $K_p$ and the hybrid kernel $K_c$. Curves correspond to different training and testing conditions and feature compensation methods for kernels $K_p$ and $K_c$.

with matched train/test conditions. However, the performance of this classifier significantly degrades when the test data is corrupted with pink noise from NOISEX-92 as shown in Figure 3(b). In this case, the classifier trained with hybrid VTS-compensated MFCC features using kernel $K_c$ in quiet condition achieves significantly better performance that the multi-condition trained classifier. Analyzing the results presented in Figures 3(a) and 3(b) together provides results for another interesting case of a partial mismatch between train and test noise types. In this case, the multi-style classifier is trained on data corrupted by white noise but tested under both pink and white noise conditions. It is evident that the hybrid VTS-compensated MFCC features outperform the multi-condition classifier for this case of partial mismatch.

Next, we present results of the multi-style classification with a complete mistmatch of noise types of training and test data. In Figure 3(c), results are reported using a multi-condition classifier that is trained on clean data as well as data corrupted by white, speech-weighted and pink noise from NOISEX-92 database *i.e.* 3 random noise contaminated versions of each feature vector in the training set, each corrupted with a different noise type and level, are added in the clean training set. Again, the size of the training set is 4 times larger than the original training set. Unlike the previous multi-condition training scenario, the training set contains the clean data as well as data corrupted by a mixture of 3 different noise types. The test data is corrupted by tank and factory floor noises from NOISEX-92. In this case, a clear improvement over the multi-condition trained classifier is obtained by classifier trained in quiet condition with the hybrid features *e.g.* 7.4% and 6.3% average improvements over the multi-condition trained classifier are achieved in the presence of factory floor and tank noise, respectively.

In Table I, results of some recent experiments on the TIMIT phoneme classification task in quiet condition are presented and compared with the results reported in this paper. We also present results obtained using SVM classifier trained with hybrid PLP cepstral representation with kernel $K_c$ as described in Section III which resulted in better classification performance in quiet conditions. Note that these benchmarks use cepstral representations that encode information from the entire variable length phoneme and our result of 20.1% improves on all benchmarks except [26] even though we use a fixed length cepstral representation. Further improvements can be achieved by including all frames within a variable length phoneme and its the transition regions, following the encoding method considered by Clarkson *et al.* [27]. Moreover, results presented in this paper significantly outperform the error reported by Rifkin *et al.* [6] (77.8%) at 0dB SNR in pink noise. In the same conditions, the hybrid classifier proposed in this paper achieves an error of 53.5% as reported in Figure 2(b).

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

| METHOD | ERROR [%] |
|---|---|
| HMMs (THMM-2) [28] | 30.4 |
| SVMs (MFCC) [27] | 22.4 |
| Large Margin GMM (LMGMM) [5] | 21.1 |
| Hierarchical GMM [29] | 21.0 |
| RLS2 [6] | 20.9 |
| Hidden CRF [30] | 20.8 |
| Hierarchical LMGMM H(2,4) [26] | 18.7 |
| Committee Hierarchical LMGMM H(2,4) [26] | 16.7 |
| **SVMs - Hybrid Features (MFCC + VTS)** | **22.7** |
| **SVMs - Hybrid Features (PLP)** | **20.1** |

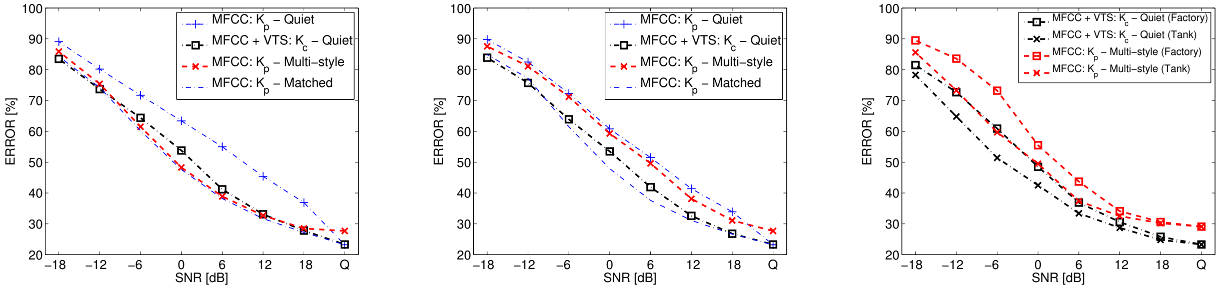
## VI. CONCLUSIONS

Hybrid representations that combine the cepstral features with the local energy features are shown to contribute to the robustness of phoneme classification with SVMs. The approximate orthogonality of speech and noise is taken into account for an effective compensation of the local energy features which are taken into account separately in the evaluation of the hybrid SVM kernel. The proposed method significantly reduces the classification error in noise and narrows the performance gap to the classifiers trained under matched noise conditions.

## REFERENCES

[1] A. Halberstadt and J. Glass, "Heterogeneous Measurements and Multiple Classifiers for Speech Recognition," *Proceedings of ICSLP*, pp. 995–998, 1998.

[2] J. Allen, "Articulation and Intelligibility," *Synthesis Lectures on Speech and Audio Proc.*, vol. 1, no. 1, pp. 1–124, 2005.



(a) (b) (c)

Fig. 3. Comparison of the hybrid noise-compensated MFCC features via VTS with kernel $K_c$ and the multi-style trained classifier with kernel $K_p$ with different training and test scenarios: (a) both training and test data corrupted by white noise (b) training and test data corrupted by white and pink noise, respectively (c) training data contaminated with white, pink and speech-weighted noises and tested under factory floor and tank noise from NOISEX-92.

[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] M. Johnson *et al.*, “Time-domain Isolated Phoneme Classification Using Reconstructed Phase Spaces,” *IEEE Trans. on Speech and Audio Proc.*, vol. 13, no. 4, pp. 458–466, 2005.

[5] F. Sha and L. K. Saul, “Large Margin Gaussian Mixture Modeling for Phonetic Classification and Recognition,” *Proceedings of ICASSP*, pp. 265–268, 2006.

[6] R. Rifkin *et. al.*, “Noise Robust Phonetic Classification with Linear Regularized Least Squares and Second-Order Features,” *Proceedings of ICASSP*, pp. 881–884, 2007.

[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] O. Viikki and K. Laurila, “Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition,” *Speech Communication*, vol. 25, pp. 133–147, 1998.

[9] M. Holmberg, D. Gelbart, and W. Hemmert, “Automatic Speech Recognition with an Adaptation Model Motivated by Auditory Processing,” *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 1, pp. 43–49, 2006.

[10] R. Lippmann and E. A. Martin, “Multi-Style Training for Robust Isolated-Word Speech Recognition,” *Proceedings of ICASSP*, pp. 705–708, 1987.

[11] 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.

[12] ETSI standard doc., “Speech processing, Transmission and Quality aspects (STQ): Advanced front-end feature extraction,” *ETSI ES 202 050*, 2002.

[13] M. J. F. Gales and F. Flego, “Combining VTS Model Compensation and Support Vector Machines,” in *Proceedings of ICASSP*, 2009, pp. 3821–3824.

[14] J. Li *et. al.*, “High-Performance HMM Adaptation With Joint Compensation of Additive and Convolutive Distortions Via Vector Taylor Series,” *IEEE Workshop on Automatic Speech Recogn. and Understanding*, 2007.

[15] T. Dietterich and G. Bakiri, “Solving Multiclass Learning Problems via Error-Correcting Output Codes,” *Journal of AI Research*, vol. 2, pp. 263–286, 1995.

[16] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[17] J. Yousafzai, Z. Cvetković, P. Sollich, and B. Yu, “Combined PLP-Acoustic Waveform Classification for Robust Phoneme Recognition using SVMs,” *Proceedings of EUSIPCO*, 2008.

[18] 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.

[19] 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.

[20] J. Yousafzai, Z. Cvetković, and P. Sollich, “Custom-Designed SVM Kernels for Improved Robustness of Phoneme Classification,” *Proceedings of EUSIPCO*, 2009.

[21] J. Tchorz and B. Kollmeier, “Estimation of the Signal-to-Noise Ratio with Amplitude Modulation Spectrograms,” *Speech Communication*, vol. 38, no. 1, pp. 1–17, 2002.

[22] S. Chandra Sekhar and T. V. Sreenivas, “Signal-to-Noise Ratio Estimation Using Higher-Order Moments,” *Signal Processing*, vol. 86, no. 4, pp. 716–732, 2006.

[23] R. Goubran E. Nemer and S. Mahmoud, “SNR Estimation of Speech Signals Using Subbands and Fourth-Order Statistics,” *IEEE Signal Proc. Letters*, vol. 6, no. 7, pp. 171–174, 1999.

[24] K. F. Lee and H. W. Hon, “Speaker-Independent Phone Recognition Using Hidden Markov Models,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, 1989.

[25] 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.

[26] H. Chang and J. Glass, “Hierarchical Large-Margin Gaussian Mixture Models for Phonetic Classification,” *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 272–275, 2007.

[27] 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.

[28] C. Rathinavelu and L. Deng, “HMM-based Speech Recognition Using State-dependent, Discriminatively Derived Transforms on Mel-Warped DFT Features,,” *IEEE Transactions on Speech and Audio Proc.*, vol. 5, no. 3, pp. 243–256, 1997.

[29] A. Halberstadt and J. Glass, “Heterogeneous Acoustic Measurements for Phonetic Classification,” *Proceedings of EuroSpeech*, pp. 401–404, 1997.

[30] D. Yu, L. Deng, and A. Acero, “Hidden Conditional Random Fields with Distribution Constraints for Phone Classification,” *Interspeech*, pp. 676–679, 2009.