SUBBAND ACOUSTIC WAVEFORM FRONT-END FOR ROBUST SPEECH RECOGNITION USING SUPPORT VECTOR MACHINES

Jibran Yousafzai[1], *Zoran Cvetković*[1], *Peter Sollich*[2]

Department of Electronic Engineering[1] & Department of Mathematics[2],
King's College London

ABSTRACT

A subband acoustic waveform front-end for robust speech recognition using support vector machines (SVMs) is developed. The primary issues of kernel design for subband components of acoustic waveforms and combination of the individual subband classifiers using stacked generalization are addressed. Experiments performed on the TIMIT phoneme classification task demonstrate the benefits of classification in frequency subbands: the subband classifier outperforms the cepstral classifiers in the presence of noise for signal-to-noise ratio (SNR) below 12dB.

Index Terms— Robustness, speech recognition, acoustic waveforms, subbands, support vector machines.

1 INTRODUCTION

Front-ends of state-of-the-art ASR systems are usually some variant of Mel-Frequency Cepstral Coefficients (MFCC) or Perceptual Linear Prediction (PLP) [1]. These representations are derived from the short term magnitude spectra followed by non-linear transformations to model the processing of the human auditory system. The main objective of these front-ends is to remove variations from speech signals that are considered unnecessary for recognition while preserving the relevant information content. However, most of the operations performed by the current ASR front-ends are not so accurate approximations of this objective. Compression of the highly redundant speech signals is primarily motivated by the need for accurate modelling of the information relevant for discrimination from limited data, thereby facilitating the development of commercial ASR systems. However, due to the nonlinear processing involved in the feature extraction, even a moderate level of distortion may cause significant departures from feature distributions learned on clean data, which makes these distributions inadequate for recognition in the presence of environmental distortions such as additive noise and linear filtering. As a result, recognition accuracy of state-of-the-art ASR systems is still far below the human performance in adverse conditions.

To make the cepstral representations of speech less sensitive to environmental distortions, several feature compensation methods [2–8] have been developed that aim to reduce explicitly the effects of additive noise and/or linear filtering on cepstral representations and thus approach the optimal performance which is achieved when training and testing conditions are matched [9]. These methods contribute significantly to robustness by alleviating some of the effects of noise. However, the resulting distortion in the cepstral features which most of these methods aim to correct is not merely an additive bias and multiplicative change of scale. Instead, this distortion is jointly determined by speech, noise type and noise level in a complicated fashion which makes the separation of noise from these features rather difficult [4].

Most of the current ASR methods considered as robust to environmental distortions are based on the assumption that the conventional cepstral features form a good enough representation to start with, so that in combination with a suitable language and context model, the performance of ASR systems can be brought close to human speech recognition. While essential for correcting many errors, context and language modelling are most effective when the underlying sequence of elementary phonetic units is predicted sufficiently accurately. Humans recognize isolated speech units above the level of chance at -18dB SNR, and significantly above it at -9dB SNR [10]. Even in quiet conditions, the current machine phone error rates for nonsense syllables are over an order of magnitude higher than human error rates [11–14]. This suggests that context and language modelling alone cannot bridge this performance gap. A number of studies [15–21] have attributed this marked difference between human and machine performance to the immense variability of speech as well as the fundamental limitations of the feature extraction process.

In this work we develop a novel front-end for ASR using SVMs that operates on an ensemble of high-dimensional subband components of acoustic waveforms, and investigate its robustness to additive noise on a phoneme classification task. This approach draws its motivation primarily from the experiments conducted by Fletcher [22] (a summary is presented in [23]), which suggests that the human decoding of linguistic message is based on decisions within narrow frequency subbands that are processed quite independently of each other. Furthermore, the high-dimensional subband components of acoustic waveforms retain more information about speech than the corresponding cepstral representations and thus facilite the construction of meaningful subband classifiers that may provide better separation of elementary phonetic units. Moreover, by constructing separate classifiers for each narrow subband, colored noise can be approximated as narrow-band white noise. We compare the noise robustness of high-dimensional subband representations with cepstral representations on a phoneme classification task; this is a problem of reasonable complexity frequently used for comparing different methods and representations [24–30]. At this stage we do not pursue a continuous speech recognition task because it depends both on the accuracy of labelling and segmentation, as well as how the two interact, which can blur the interpretation of the results on the comparison of different representations in terms of the robustness they provide. However, the improvements achieved on the classification task can be expected to extend to continuous speech recognition tasks [31, 32] given that SVMs can be used with hidden Markov models for continuous speech recognition as detailed in [32, 33].

For classification with acoustic waveforms in frequency subbands, custom-designed SVM kernels based on the physical properties of speech and speech perception, as proposed in our previous study [34, 35], are used. Decomposing an acoustic waveform into its subband components produces an ensemble of base-level binary subband classifiers. For each binary classification problem, the ensemble of base-level binary subband classifiers is combined using stacked generalization to form a meta-level binary classifier. The resulting meta-level binary classifiers are then further combined using error-correcting output code methods [36] for multiclass classification. For comparison, we also perform classification using a standard cepstral representation (MFCC) with state-of-the-art feature compensation such as vector Taylor series [5–8]. The experiments demonstrate the benefits of the acoustic waveform subband approach in providing robustness to noise. For example, the acoustic waveform classifier outperforms the cepstral classifier for signal-to-noise ratio (SNR) below 12dB.

The subband classification approach with acoustic waveforms using custom-designed kernels is reviewed in Section 2 where we also briefly discuss stacked generalization for combination of individual subband classifiers. The experimental setup is described in Section 3 and results are reported in Section 4. Finally, Section 5 draws some conclusions.

2 CLASSIFICATION METHOD

2.1 Support Vector Machines

An SVM [37] 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. For classification with SVMs, we consider fixed-length D-samples long acoustic waveform segments which we will denote by $\mathbf{x}$. Then a kernel-based decision function that classifies a test phoneme $\mathbf{x}$ is expressed as

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

where K is a kernel function, $\mathbf{x}_i$, $y_i = \pm 1$ and α_i, respectively, are the i-th training phoneme (clean speech), 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

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

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

In preliminary experiments, comparable performance was achieved with both kernels; in this work, a polynomial kernel is therefore used for classification with cepstral features (MFCC) whereas classification with acoustic waveforms in frequency subbands is performed using a custom-designed kernel developed from a baseline polynomial kernel.

2.2 Error Correcting Output Codes

SVMs are binary classifiers trained to distinguish between two classes. For multiclass tasks, they can be combined via predefined discrete error-correcting output codes [36]. To summarize the procedure briefly, N binary classifiers are trained to distinguish two groups of 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 minimizes the loss $\sum_{n=1}^{N} \chi(w_{mn} f_n(\mathbf{x}))$. Here χ is some loss function and $f_n(\mathbf{x})$ is the output of the n^{th} binary classifier. A number of coding strategies were considered. However, the construction of *one-vs-one* classifiers was computationally most feasible for a problem with large datasets as in our case; we therefore use $N = M(M-1)/2$ binary pairwise classifiers. 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 study.

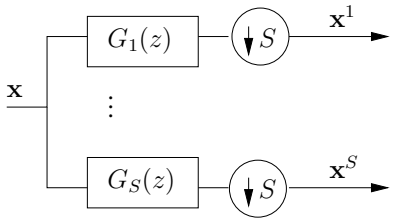


Fig. 1. *Decomposition of phonemes into its subband components using an S-channel cosine modulated filter bank.*

2.3 Kernels for Subband Classification

We propose some modifications to the baseline SVM kernels to take into account some physical properties of speech perception. To extract the relevant features for classification in frequency subbands, each waveform $\mathbf{x}$ is first decomposed into S subband components, $\mathbf{x}^s, s = 1, \ldots, S$, as shown in Figure 1 using a perfect reconstruction cosine modulated filter bank (CMFB) [38]. The filter bank consists of filters,

$$g_s[k] = \frac{1}{\sqrt{S}} g_p[k] \cos\left(\frac{2s-1}{4S}\,(2k - S - 1)\,\pi\right), \quad s = 1, \ldots, S,\ k = 1, \ldots, 2S \tag{4}$$

where the prototype filter $g_p[k]$ is a raised cosine function,

$$g_p[k] = \sqrt{2} \sin\left(\frac{\pi\,(k - 0.5)}{2S}\right),\ k = 1, \ldots, 2S. \tag{5}$$

a) Sign-invariance: The perception of acoustic waveforms by the human auditory system is invariant to sign and the same property extends to the subband components of acoustic waveforms as well. Hence we use an *even* kernel defined from a baseline polynomial kernel K_p, as proposed in our previous work [34, 35], to account for sign-invariance of the subband components:

$$\begin{aligned} K_e(\mathbf{x}^s, \mathbf{x}_i^s) = {} & K_p'(\mathbf{x}^s, \mathbf{x}_i^s) + K_p'(\mathbf{x}^s, -\mathbf{x}_i^s) \\ & + K_p'(-\mathbf{x}^s, \mathbf{x}_i^s) + K_p'(-\mathbf{x}^s, -\mathbf{x}_i^s) \end{aligned} \tag{6}$$

where K_p' is a modified polynomial kernel given by

$$\begin{aligned} K_p'(\mathbf{x}^s, \mathbf{x}_i^s) &= K_p(\mathbf{x}^s / \|\mathbf{x}^s\|, \mathbf{x}_i^s / \|\mathbf{x}_i^s\|) \\ &= (1 + \langle \mathbf{x}^s / \|\mathbf{x}^s\|, \mathbf{x}_i^s / \|\mathbf{x}_i^s\| \rangle)^{\Theta}. \end{aligned} \tag{7}$$

Kernel K_p', which acts on normalized input vectors, will be used as a baseline kernel for the acoustic waveforms. On the other hand,

the standard polynomial kernel K_p defined in (2) will be employed for the cepstral representations where feature standardization by cepstral mean-and-variance normalization (CMVN) already ensures that feature vectors typically have unit norm.

b) Subband dynamics: Features that capture the evolution of energy and the dynamics of speech in frequency subbands are important for discriminating among phonemes. To obtain the subband dynamic features, the speech waveform is divided into a sequence of overlapping *frames* similar to those used to calculate MFCCs (with the same frame duration and frame rate). Then the T frames closest to the phoneme center are used to construct the dynamic feature vector of that phoneme. Let $\mathbf{x}^{t,s}, t = 1, \ldots, T, s = 1, \ldots, S$ denote the s^{th} subband component of the t^{th} frame closest to the center of phoneme $\mathbf{x}$. Then the s^{th} subband energy vector is formed by concatenating the energies of the T frames in that subband as

$$\boldsymbol{\omega}^s = \left[\log \left\|\mathbf{x}^{1,s}\right\|^2, \ldots, \log \left\|\mathbf{x}^{T,s}\right\|^2\right], \; s = 1, \ldots, S \quad (8)$$

and its time derivatives [39, 40] are evaluated to form the dynamic subband feature vector $\boldsymbol{\Omega}^s$:

$$\boldsymbol{\Omega}^s = \left[\boldsymbol{\omega}^s \;\; \Delta\boldsymbol{\omega}^s \;\; \Delta^2\boldsymbol{\omega}^s\right], \;\; s = 1, \ldots, S. \quad (9)$$

This dynamic subband feature vector $\boldsymbol{\Omega}^s$ is then combined with the corresponding acoustic waveform subband component $\mathbf{x}^s$ using kernel K_Ω given by

$$K_\Omega(\mathbf{x}^s, \mathbf{x}_i^s, \boldsymbol{\Omega}^s, \boldsymbol{\Omega}_i^s) = K_e(\mathbf{x}^s, \mathbf{x}_i^s) K_p(\boldsymbol{\Omega}^s, \boldsymbol{\Omega}_i^s), \quad (10)$$

where $\boldsymbol{\Omega}_i^s$ is the dynamic subband feature vector corresponding to the s-th suband component $\mathbf{x}_i^s$ of the i-th training point $\mathbf{x}_i$. An additional invariance to time alignment can be incorporated by means of a *shift-invariant* kernel [35], which would likely improve the classification performance further [34, 35], but that approach is not pursued in the present study due to our modest computational resources.

2.4 Support Vector Machine Ensemble

For each binary classification problem, decomposing an acoustic waveform into its subband components produces an ensemble of S classifiers. The decision of the subband classifiers in the ensemble, given by

$$f^s(\mathbf{x}^s, \boldsymbol{\Omega}^s) = \sum_i \alpha_i^s y_i K_\Omega(\mathbf{x}^s, \mathbf{x}_i^s, \boldsymbol{\Omega}^s, \boldsymbol{\Omega}_i^s) + b^s \;, \; s = 1, \ldots, S \quad (11)$$

are then aggregated using ensemble methods to obtain the binary classification decision for a test waveform $\mathbf{x}$. Here α_i^s and b^s are the Lagrange multiplier corresponding to $\mathbf{x}_i^s$ and the bias of the s^{th} subband binary classifier. Under the assumption that the errors of the individual classifiers are independent with error rate $p < 1/2$, a simple combinatorial argument shows that even in the case of majority voting the probability that the result is incorrect can be bounded as

$$p_e < \frac{1}{2}(4p(1-p))^{\frac{S}{2}} \;,$$

for a large value of S. Therefore, the ensemble error decreases with an increase in the size of the ensemble S [41, 42]. While simple aggregation schemes like majority voting may yield some improvements in the classification performance, they do not exploit the importance of certain subbands in discriminating among a specific pair of phonemes because equal weights are assigned to all subband classifiers. Furthermore, the thresholding of scores into class labels (± 1) before voting may lose information that is useful for classification. In this light, we use stacked generalization [43] to aggregate the outputs of base-level SVMs in the ensemble.

Our practical implementation of stacked generalization [43] consists of a hierarchical two-layer SVM architecture, where the outputs of several base-level SVMs feed into a meta-level SVM implemented using a linear kernel. In our experiments, we found that the choice of a squashing function has little effect on the classification performance. Therefore, the raw base-level predictions are used in the construction of meta-level classifier for simplicity. The decision of the meta-level SVM classifier is of the form

$$h(\mathbf{x}) = \langle \mathbf{f}(\mathbf{x}), \mathbf{w} \rangle + v = \sum_s w^s f^s(\mathbf{x}^s, \boldsymbol{\Omega}^s) + v \;, \quad (12)$$

where $\mathbf{f}(\mathbf{x}) = \left[f^1(\mathbf{x}^1, \boldsymbol{\Omega}^1), \ldots, f^S(\mathbf{x}^S, \boldsymbol{\Omega}^S)\right]$ is the base-level SVM score vector of the test waveform $\mathbf{x}$, v is the classifier bias, and $\mathbf{w} = \left[w^1, \ldots, w^S\right]$ is the weight vector of the meta-level classifier. Note that this weight vector is determined from an independent development/validation set $\{\tilde{\mathbf{x}}_j, \tilde{y}_j\}$, and can be expressed as, $\mathbf{w} = \sum_j \beta_j \tilde{y}_j \mathbf{f}(\tilde{\mathbf{x}}_j)$, where $\mathbf{f}(\tilde{\mathbf{x}}_j) = \left[f^1(\tilde{\mathbf{x}}_j^1, \tilde{\boldsymbol{\Omega}}_j^1), \ldots, f^S(\tilde{\mathbf{x}}_j^S, \tilde{\boldsymbol{\Omega}}_j^S)\right]$ is the base-level SVM score vector of the training waveform $\tilde{\mathbf{x}}_j$, and β_j and $\tilde{y}_j$ are the Lagrange multiplier and class label corresponding to $\mathbf{f}(\tilde{\mathbf{x}}_j)$, respectively. Again, ECOC methods are used to combine the meta-level binary classifiers for multiclass classification.

While a base-level SVM assigns a weight to each support feature vector, stacked generalization effectively assigns an additional weight w^s to each subband based on the performance of the corresponding base-level subband classifier. Note that, unlike the results presented in previous chapter, we do not assume any knowledge of the noise statistics to perform feature compensation. Instead, we use stacked generalization to learn weight vectors that are tuned for classification in adverse conditions. For instance, the meta-level classifier can be trained using score feature vectors of noisy data or score feature vectors of a mixture of clean and noisy data. Moreover, since the dimension of the score feature vectors that form the input to the stacked subband classifier (S) is very small as compared to the typical MFCC or waveform feature vectors, only a very limited amount of data is required to learn the optimal weights of the meta-level classifier. As such, stacked generalization offers flexibility and some coarse frequency selectivity for the individual binary classification problems, and can be particularly useful in de-emphasizing information from unreliable subbands. The experiments show that major gains in the classification performance can be attained with this approach.

3 EXPERIMENTAL SETUP

Experiments are performed on the 'si' (diverse) and 'sx' (compact) sentences of the TIMIT database [44]. The training set consists of 3696 sentences from 168 different speakers. For testing we use the core test set which consists of 192 sentences from 24 different speakers not included in the training set. The development set consists of 1152 sentences uttered by 96 male and 48 female speakers not included in either the training or the core test set, with speakers from 8 different dialect regions. The glottal stops /q/ are removed from the class labels and certain allophones are grouped into their corresponding phoneme classes using the standard Kai-Fu Lee clustering [45], resulting in a total of $M = 48$ phoneme classes and $N = M(M-1)/2 = 1128$ classifiers. Among these classes, there are 7 groups for

which the contribution of within-group confusions toward multiclass error is not counted, again following standard practice [29, 45].

Experiments are performed with white and pink noises. This work is focused on the robustness of phoneme classification in order to get some assessment of the separation of phoneme classes in different representation domains and for that purpose, white (isotropic) noise was most appropriate. Robustness to pink noise was investigated because $1/f$-like noise patterns occur widely in nature and are also found in music, fan and cockpit noises [46–48]. To test the classification performance in noise, each TIMIT sentence is normalized to unit energy per sample and then a noise sequence is added to the entire sentence to set the sentence-level SNR. Hence for a given sentence-level SNR, signal-to-noise ratio at the level of individual phonemes will vary widely.

Initially, we experimented with different values of the hyperparameters for the binary SVM classifiers but decided to use fixed values for all classifiers as parameter optimization had a large computational overhead but only a small impact on the multiclass classification error: the degree of K_p is set to $\Theta = 6$ and the penalty parameter (for slack variables in the SVM training algorithm) to $C = 1$.

For cepstral features two training-test scenarios are considered: *(i)* training SVM classifiers using clean data with test features compensated via vector Taylor series (VTS) [5–8], and *(ii)* training and testing under identical noise conditions. The VTS algorithm aims to estimate the distribution of noisy speech given the distribution of clean speech, a segment of noisy speech, and the Taylor series expansion that relates the noisy speech features to the clean ones. After computing the distribution of the noisy speech, minimum mean-square estimation can be used to predict the unobserved clean cepstral feature vectors. The matched condition scenario, on the other hand, is an impractical target; nevertheless, we present the results as a reference, since this setup is considered to give the optimal achievable performance with cepstral features [9]. Additionally, the features of both training and test data are standardized using CMVN [4] in all scenarios. To obtain the cepstral (MFCC) representation, 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 $\mathbb{R}^{390}$. For noise compensation with VTS, a GMM with 64 mixture components was used to learn the distribution of Mel log spectra of clean training data.

To obtain the subband representation, phoneme segments $\mathbf{x}$ are extracted from the TIMIT sentences by applying a 100ms rectangular window at the center of each phoneme. At a 16kHz sampling frequency, this gives fixed length vectors in $\mathbb{R}^D$ with $D = 1600$ which are further decomposed into subband components $\{\mathbf{x}^s\}_{s=1}^{16}$ using a 16-channel cosine-modulated filter bank. For the dynamic subband feature vector, $\mathbf{\Omega}^s$ (see (9)), the log-energy and time derivatives subband components of the $T = 10$ frames closest to the center of a particular phoneme are combined to form a 30-dimensional feature vector. The dynamic subband feature vectors are further standardized within each sentence of TIMIT for the evaluation of kernel K_Ω (see 10).

For classification in subbands of acoustic waveforms, the training of base-level SVM subband classifiers is always performed with noiseless (clean) data. A random subset, one-eighth the size of the development data was selected for training of the meta-level SVMs in the stacked classifier as learning the optimal weights requires only very limited amounts of data. Again, two scenarios are considered for the training of stacked classifiers: *(i) multistyle training* - training with base-level SVM score vectors obtained from a small collection of clean and noise corrupted data, and *(ii)* training and testing under identical noise conditions.

4 RESULTS

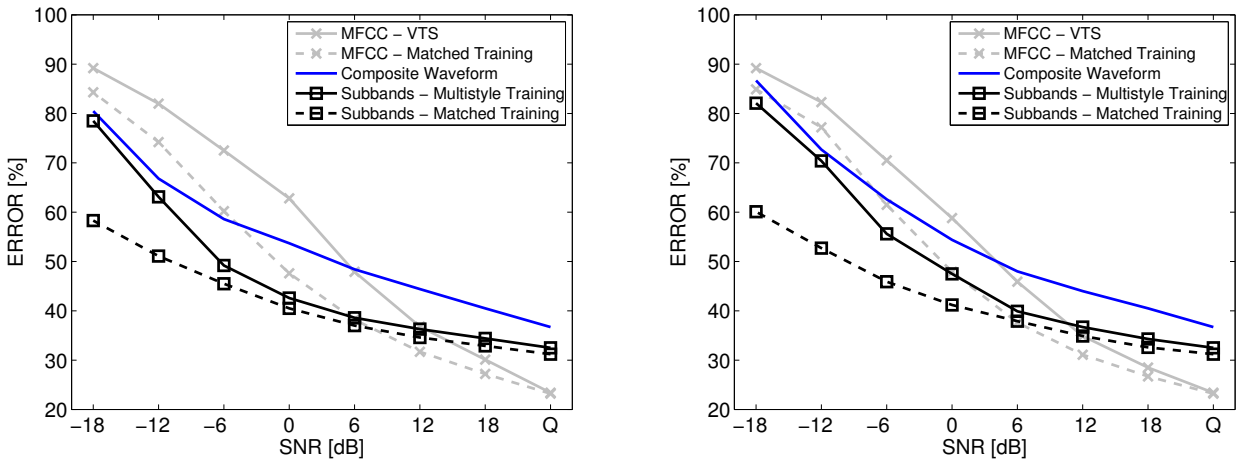
Figure 2 shows the results of SVM phoneme classification with the multistyle-trained stacked subband classifier, VTS-compensated MFCC classifier and composite acoustic waveform classifier [34] in the presence of additive white and pink noise. For comparison, results are presented for the stacked subband classifier and MFCC classifier in matched train-test conditions as well. The multistyle-trained meta-level acoustic waveform classifier is trained using base-level score vectors obtained from clean data and data corrupted by white noise at 0dB SNR only and then tested in white (matched) and pink (mismatched) noise. The amount of data used for training of the meta-level classifier was 5% of the data used for matched training of the MFCC classifier.

The results show that the stacked subband classifier exhibits better classification performance than the VTS-compensated MFCC classifier for SNR below 12dB whereas the performance crossover between MFCC and composite acoustic waveform classifiers is between 6dB and 0dB SNR. The stacked subband classifier achieves average improvements of 8.7% and 4.5% over the MFCC classifier across considered SNRs in the presence of white and pink noise, respectively. Moreover, the stacked subband classifier also significantly improves over the MFCC classifier trained and tested in matched conditions for SNRs below a crossover point between 0 and 6dB SNR, although the amount of the data used to learn the optimal weights of the meta-level classifier is a small fraction of the data set used for matched training of the MFCC classifier, and the training is done only using clean data and data corrupted by white noise at 0dB SNR. In particular, an average improvement of 6.5% in the phone error is achieved by the stacked subband classifier over the matched MFCC classifier for SNRs below 6dB in the presence of white noise.

Our previous work showed [34, 49] that the MFCC classifier suffers severe degradation in classification performance in the case of a mismatch in the noise type. On the other hand, the stacked subband classifier degrades gracefully in a mismatched environment as shown in 2(b). This is due to the decomposition of acoustic waveforms into a number of frequency subbands where, for each binary classification problem, colored noise can be approximated by narrow-band white noise.

It is worth noting that the proposed method achieves considerably better results than the errors reported in the literature on the same task. Rifkin *et al.* [27] report an error rate of 77.8% at 0dB SNR in pink noise whereas the subband classifier achieves an error of 48% in similar conditions as reported in Figure 2(b). Further improvements can be achieved by incorporating shift-invariance into SVM kernels, and with a convex combination of cepstral and acoustic waveform classifiers as proposed in [34, 35].

Figure 2 also shows a comparison of stacked subband classifier with MFCC classifier trained and tested in matched conditions. The matched stacked subband classifier significantly outperforms the matched MFCC classifiers for SNRs below 6dB. Around 13% average improvement is achieved by the subband classifier over the MFCC classifier for SNR $\leq$ 6dB in the presence of both white and pink noises. This suggests that the high-dimensional subband



(a) *White Noise* (b) *Pink Noise*

Fig. 2. *SVM classification in the subbands of acoustic waveforms and its comparison with MFCC and composite acoustic waveform representations* [34]. *The multistyle stacked subband classifier is trained with a small random subset consisting of clean and white-noise (0dB SNR) corrupted data which is one-eighth the size of the development set. It is tested on data corrupted with white noise (matched) and pink noise (mismatched). In the matched training case, noise levels as well as noise types of training and test data are identical for both MFCC and stacked subband classifiers.*

representation obtained from acoustic waveforms provides a better separation of phoneme classes compared to cepstral representation in high noise.

5 CONCLUSIONS

A novel subband acoustic waveform front-end for robust speech recognition using SVMs was proposed. We addressed the issues of kernel design for subband components of acoustic waveforms and the aggregation of the individual subband classifiers using stacked generalization. It is shown that an ensemble of classifiers trained on the subband components of the high-dimensional acoustic waveforms can contribute to the robustness of phoneme classification. The experiments show that the stacked subband classifier outperforms the cepstral classifier in the presence of noise for SNR below 12dB.

While the stacked subband classifier does not perform as well as the MFCC classifier in low noise conditions, their convex combination can achieve better performance than either of the individual classifiers as demonstrated in our previous studies [34, 35]. We are currently investigating the robustness of the proposed subband acoustic waveform front-end to linear filtering, with preliminary experiments showing encouraging results.

References

[1] 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.

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

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

[4] C. Chen and J. Bilmes, "MVA Processing of Speech Features," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 257–270, 2007.

[5] P. J. Moreno, B. Raj, and R. M. Stern, "A Vector Taylor Series Approach for Environment-Independent Speech Recognition," *Proceedings of ICASSP*, pp. 733–736, 1996.

[6] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "High-Performance HMM Adaptation With Joint Compensation of Additive and Convolutive Distortions Via Vector Taylor Series," *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 65–70, 2007.

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

[8] H. Liao, "Uncertainty Decoding For Noise Robust Speech Recognition," *Ph.D. Thesis, Cambridge University*, 2007.

[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] 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.

[11] J. Allen, "How Do Humans Process and Recognize Speech?," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, 1994.

[12] L. Deng and D. Sun, "Phonetic Classification and Recognition Using HMM Representation of Overlapping Articulatory Features for All Classes of English Sounds," *Proceedings of ICASSP*, pp. 45–48, 1994.

[13] S. Lee and J. Glass, "Real-Time Probabilistic Segmentation for Segment Based Speech Recognition," *Proceedings of ICSLP*, pp. 1803–1806, 1998.

[14] A. Robinson, M. Hochberg, and S. Renals, "IPA: Improved Modelling With Recurrent Neural Networks," *Proceedings of ICASSP*, pp. 37–40, 1994.

[15] B.S. Atal, "Automatic Speech Recognition: a Communication Perspective," *Proceedings of ICASSP*, pp. 457–460, 1999.

[16] S. D. Peters, P. Stubley, and J. Valin, "On the Limits of Speech Recognition in Noise," *Proceedings of ICASSP*, pp. 365–368, 1999.

[17] N. Morgan, Qifeng Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cretin, H. Bourlard, and M. Athineos, "Pushing the envelope - aside," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 81–88, 2005.

[18] Hervé Bourlard, Hynek Hermansky, and Nelson Morgan, "Towards Increasing Speech Recognition Error Rates," *Speech Communication*, vol. 18, no. 3, pp. 205–231, 1996.

[19] B. Meyer, M. Wächter, T. Brand, and B. Kollmeier, "Phoneme Confusions in Human and Automatic Speech Recognition," *Proceedings of INTERSPEECH*, pp. 2740–2743, 2007.

[20] K. Paliwal and L. Alsteris, "On the Usefulness of STFT Phase Spectrum in Human Listening Tests," *Speech Communication*, vol. 45, no. 2, pp. 153–170, 2005.

[21] L. Alsteris and K. Paliwal, "Further Intelligibility Results from Human Listening Tests using the Short-Time Phase Spectrum," *Speech Communication*, vol. 48, no. 6, pp. 727–736, 2006.

[22] H. Fletcher, *Speech and Hearing in Communication*, Van Nostrand, New York, 1953.

[23] J. B. Allen, "How do humans process and recognize speech?," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, 1994.

[24] 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.

[25] D. Yu, L. Deng, and A. Acero, "Hidden Conditional Random Fields with Distribution Constraints for Phone Classification," *Proceedings of INTERSPEECH*, pp. 676–679, 2009.

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

[27] 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.

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

[29] P. Clarkson and P. J. Moreno, "On the Use of Support Vector Machines for Phonetic Classification," *Proceedings of ICASSP*, pp. 585–588, 1999.

[30] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden Conditional Random Fields for Phone Classification," *Proceedings of INTERSPEECH*, pp. 1117–1120, 2005.

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

[32] A. Ganapathiraju, J. E. Hamaker, and J. Picone, "Applications of Support Vector Machines to Speech Recognition," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2348–2355, 2004.

[33] S. E. Krüger, M. Schaffñer, M. Katz, E. Andelic, and A. Wendemuth, "Speech recognition with support vector machines in a hybrid system," *Proceedings of INTERSPEECH*, pp. 993–996, 2005.

[34] J. Yousafzai, Z. Cvetković, P. Sollich, and B. Yu, "Combined Features and Kernel Design for Noise Robust Phoneme Classification Using Support Vector Machines," *Accepted for publication in the IEEE Transactions on Audio, Speech and Language Processing*.

[35] J. Yousafzai, Z. Cvetković, and P. Sollich, "Tuning Support Vector Machines for Robust Phoneme Classification with Acoustic Waveforms," *Proceedings of INTERSPEECH*, pp. 2391–2395, 2009.

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

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

[38] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

[39] D. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, Online Web Resource.

[40] S. Furui, "Speaker-Independent Isolated Word Recognition using Dynamic Features of Speech Spectrum," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[41] T. Dietterich, "Ensemble Methods in Machine Learning," *Lecture Notes in Computer Science: Multiple Classifier Systems*, pp. 1–15, 2000.

[42] L. Hansen and P. Salamon, "Neural Network Ensembles," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.

[43] David H. Wolpert, "Stacked Generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[44] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallet, and N. Dahlgren, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," *Linguistic Data Consortium*, 1993.

[45] K. F. Lee and H. W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[46] R. F. Voss and J. Clarke, "1/f Noise in Music: Music from 1/f Noise," *Journal of Acoustical Society of America*, vol. 63, no. 1, pp. 258–263, 1978.

[47] B. J. West and M. Shlesinger, "The Noise in Natural Phenomena," *American Scientist*, vol. 78, no. 1, pp. 40–45, 1990.

[48] P. Grigolini, G. Aquino, M. Bologna, M. Lukovic, and B. J. West, "A Theory of 1/f Noise in Human Cognition," *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 19, pp. 4192–4204, 2009.

[49] J. Yousafzai, Z. Cvetković, and P. Sollich, "Towards Robust Phoneme Classification with Hybrid Features ," *Proceedings of IEEE Symposium on Information Theory*, pp. 1643–1647, 2010.