On the Potential for Robust ASR with Combined Subband-Waveform and Cepstral Features

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

***Abstract*— This work explores the potential for robust classification of phonemes in the presence of additive noise and linear filtering using high-dimensional features in the subbands of acoustic waveforms. The proposed technique is compared with state-of-the-art automatic speech recognition (ASR) front-ends on the TIMIT phoneme classification task using support vector machines (SVMs). The key issues of selecting the appropriate SVM kernels for classification in frequency subbands and the combination of individual subband classifiers using ensemble methods are addressed. Experiments demonstrate the benefits of the classification in the subbands of acoustic waveforms: it outperforms the standard cepstral front-end in the presence of noise and linear filtering for all signal-to-noise ratios (SNRs) below a crossover point between 12dB and 6dB. Combining the subband-waveform and cepstral classifiers achieves further performance improvements over both individual classifiers.**

***Index Terms*—Speech recognition, subbands, support vector machines, classification, robustness.**

I. Introduction

Automatic speech recognition (ASR) systems suffer severe performance degradation in the presence of environmental distortions, in particular additive noise and linear filtering. Humans, on the other hand, exhibit a very robust behavior in recognizing speech even in extremely adverse conditions. In particular, humans recognize isolated speech units above the level of chance already at -18dB SNR, and significantly above it at -9dB SNR [1]. Even in quiet conditions, the machine error rates for recognizing isolated nonsense syllables and phonemes are significantly higher than those of humans [2–5]. Although there are a number of factors preventing conventional ASR systems from reaching the human benchmark, several studies [4, 6–9] have attributed the marked difference between human and machine performance to the fundamental limitations of the ASR front-ends. These studies suggest that the large amount of redundancy in speech signals, which is removed in the process of the extraction of cepstral features such as Mel-Frequency Cepstral Coefficients (MFCC) [10], is in fact needed to cope with environmental distortions. Among these studies, work on human speech perception [4, 6, 8, 9] has shown explicitly that the information reduction that takes place in the conventional ASR front-ends leads to a severe degradation in human speech recognition performance and furthermore, that in noisy environments there is a high correlation between human and machine errors in recognition of speech with distortions introduced by typical ASR front-end processing. Over the years, techniques such as cepstral mean-and-variance normalization (CMVN) [11, 12] and vector Taylor series (VTS) compensation [13] have been developed that aim to explicitly reduce the effects of noise on the short-term spectra. However, the distortion of the cepstral features caused by additive noise and linear filtering depends on the speech signal, filter characteristics, noise type and noise level in a very complex fashion that makes feature compensation or adaptation very intricate and not sufficiently effective [11].

In our previous work we showed that using acoustic waveforms directly, without any compression or nonlinear transformation can improve the robustness of ASR front-ends to additive noise [14]. In this paper, we propose features for an ASR front-end which are derived from the decomposition of high-dimensional acoustic waveforms into frequency subbands, to achieve additional robustness to additive noise as well as robustness to linear filtering. This approach draws its motivation primarily from the experiments conducted by Fletcher [15], which suggest that the human decoding of linguistic messages is based on decisions within narrow frequency subbands that are processed quite independently of each other. This reasoning further implies that accurate recognition in any subband should result in accurate recognition overall, regardless of the errors in other subbands. While this theory has not been proved and some studies on the subband correlation of speech signals [16, 17] have even put its validity into question, there are some technical reasons for considering classification in frequency subbands. First of all, decomposing speech into its frequency subbands can be beneficial since it allows a better exploitation of the fact that certain subbands may inherently provide better separation of some phoneme classes than others. Secondly, the effect of wideband noise in sufficiently narrow subbands can be approximated as that of narrowband white noise and thus make the compensation of features be approximately independent of the spectral characteristics of the additive noise and linear filtering. Moreover, appropriate ensemble methods for aggregation of the decisions in individual frequency subbands can facilitate selective de-emphasis of unreliable information, particularly in the presence of narrowband noise.

The subband approach has also previously been used in [18–24] where it provided marginal improvements in recognition performance over its full band counterparts. Note that the front-end features employed in the previous works were the subband-based variants of cepstral features or multi-resolution cepstral features. By contrast, our proposed features are extracted from an ensemble of subband components of high-dimensional acoustic waveforms, and thus retain more information about speech that is potentially relevant to discrimination of phonetic units than the corresponding cepstral representations. Robustness of the proposed front-end features to additive noise and filtering is demonstrated by its comparison with the MFCC front-end on a phoneme classification task as it remains important in comparing different methods and representations [20, 25–32]. Standard feature compensation methods such as CMVN and VTS compensation are used throughout the experiments in order to reduce the mismatch between the training and test data. The results demonstrate the benefits of the subband classification in terms of its robustness to additive noise and linear filtering; for instance, in classifying noisy reverberant speech, it outperforms the MFCC classifier compensated using VTS for all SNRs below a crossover point between 12dB and 6dB. Finally, their convex combination yields further performance improvements over both individual classifiers.

II. Subband Classification Using Support Vector Machines

A. Support Vector Machines

Support vector machines (SVMs) are receiving increasing attention as a tool for speech recognition applications due to their good generalization properties [14, 26, 33–35]. Here we use them to compare the proposed subband-based representation with standard cepstral front-end features in terms of their robustness to noise and filtering on a TIMIT phoneme classification task. Their performance on this task can be expected to extend to continuous speech recognition [33, 36] using hybrid SVM - HMM frameworks [33], as well as, more directly, by means of frame-based architectures based on the token passing algorithm [35].

A binary SVM classifier estimates a decision surface that jointly maximizes the margin between the two classes and minimizes the misclassification error on the training set. For a given training set $(\mathbf{x}_1, \ldots, \mathbf{x}_p)$ with corresponding class labels $(y_1, \ldots, y_p)\,,\ y_i \in \{+1, -1\}$, an SVM classifies a test point $\mathbf{x}$ by computing a score function,

$$h(\mathbf{x}) = \sum_{i=1}^{p} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \tag{1}$$

where α_i is the Lagrange multiplier corresponding to the i^{th} training sample, $\mathbf{x}_i$, b is the classifier bias – these are optimized during training – and K is a kernel function. The class label of $\mathbf{x}$ is then predicted as $\operatorname{sgn}\left(h\left(\mathbf{x}\right)\right)$. While the simplest kernel $K(\mathbf{x}, \tilde{\mathbf{x}}) = \langle \mathbf{x}, \tilde{\mathbf{x}} \rangle$ produces linear decision boundaries, in most real classification tasks, the data is not linearly separable. Nonlinear kernel functions implicitly map data points to a high-dimensional feature space where the data could potentially be linearly separable. Kernel design is therefore effectively equivalent to feature-space selection, and using an appropriate kernel for a given classification task is crucial. Commonly used is the polynomial kernel, $K_p(\mathbf{x}, \tilde{\mathbf{x}}) = (1 + \langle \mathbf{x}, \tilde{\mathbf{x}} \rangle)^{\Theta}$, where the polynomial order Θ in K_p is a hyper-parameter that is tuned to a particular classification problem. More sophisticated kernels can be obtained by various combinations of basic SVM kernels. Here we use a polynomial kernel for classification with cepstral features (MFCC) whereas classification with acoustic waveforms in frequency subbands is performed using a custom-designed kernel described in the following.

For multiclass problems, binary SVMs are combined via error-correcting output codes (ECOC) methods [37]. In this work, for an M-class problem we train $N = M(M-1)/2$ binary pairwise classifiers, primarily to lower the computational complexity by training on only the relevant two classes of data. The training scheme can be captured in a coding matrix $w_{mn} \in \{0, 1, -1\}$, *i.e.* classifier n is trained only on data from the two classes m for which $w_{mn} \neq 0$, with $\operatorname{sgn}(w_{mn})$ as the class label. One then predicts for test input $\mathbf{x}$ the class that minimizes the loss $\sum_{n=1}^{N} \chi(w_{mn} f_n(\mathbf{x}))$ where $f_n(\mathbf{x})$ is the output of the n^{th} binary classifier and χ is a loss function. We experimented with a variety of loss functions, including hinge, Hamming, exponential and linear. The hinge loss function $\chi(z) = \max(1 - z, 0)$ performed best and is therefore used throughout.

B. Kernels for Subband Classification

For classification in frequency subbands, each waveform $\mathbf{x}$ is processed through an S-channel maximally-decimated perfect reconstruction cosine modulated filter bank (CMFB) [38] and decomposed into its subband components, $\mathbf{x}^s, s = 1, \ldots, S$. Several other decompositions such as discrete wavelet transform, wavelet packet decomposition and discrete cosine transform also achieved comparable, but somewhat inferior performance. The CMFB consists of a set of orthonormal analysis filters,

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

where $g[k] = \sqrt{2}\sin\left(\pi\left(k - 0.5\right)/2S\right)\,,\ k = 1, \ldots, 2S$, is a low-pass prototype filter. Such a filter bank implements an orthogonal transform, hence the collection of the subband components is a representation of the original waveform in a different coordinate system [38]. A maximally-decimated filter bank was chosen primarily because the sub-sampling operation avoids introducing additional unnecessary redundancies and thus limits the overall computational burden. However, we believe that redundant expansions of speech signals obtained using over-sampled filter banks could be advantageous to effectively account for the shift invariance of speech.

For classification in frequency subbands, an SVM kernel is constructed by partly following steps from our previous work [14], which attempted to capture known invariances or express explicitly the waveform qualities which are known to correlate with phoneme identity. First, an even kernel is constructed from a baseline polynomial kernel K_p to account for the sign-invariance of human speech perception as

$$K_e(\mathbf{x}^s, \mathbf{x}_i^s) = K_p'(\mathbf{x}^s, \mathbf{x}_i^s) + K_p'(\mathbf{x}^s, -\mathbf{x}_i^s) \tag{3}$$

where

$$K_p'(\mathbf{x}^s, \mathbf{x}_i^s) = K_p\left(\frac{\mathbf{x}^s}{\|\mathbf{x}^s\|}, \frac{\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right) = \left(1 + \left\langle \frac{\mathbf{x}^s}{\|\mathbf{x}^s\|}, \frac{\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|} \right\rangle\right)^{\Theta}, \tag{4}$$

is a modified polynomial kernel which acts on normalized input vectors and is used as a baseline kernel for classification in waveform subbands. On the other hand, the standard polynomial kernel K_p is used for classification with the cepstral representations, where feature standardization by CMVN [12] already ensures that feature vectors typically have unit norm.

Next, the temporal dynamics of speech are explicitly taken into account by means of features that capture the evolution of energy in individual subbands. To obtain these features, each subband component $\mathbf{x}^s$ is first divided into T frames, $\mathbf{x}^{t,s}, t = 1, \ldots, T$, and then a vector of their energies $\boldsymbol{\omega}^s$ is formed as,

$$\boldsymbol{\omega}^s = \left[\log\left\|\mathbf{x}^{1,s}\right\|^2, \ldots, \log\left\|\mathbf{x}^{T,s}\right\|^2\right] .$$

Finally, time differences [39] of $\boldsymbol{\omega}^s$ are evaluated to form the dynamic subband feature vector $\boldsymbol{\Omega}^s$ as $\boldsymbol{\Omega}^s = \begin{bmatrix}\boldsymbol{\omega}^s & \Delta\boldsymbol{\omega}^s & \Delta^2\boldsymbol{\omega}^s\end{bmatrix}$. This dynamic subband feature vector $\boldsymbol{\Omega}^s$ is then combined with the corresponding acoustic waveform subband component $\mathbf{x}^s$ forming kernel K_Ω given by

$$K_\Omega(\mathbf{x}^s, \mathbf{x}_i^s, \boldsymbol{\Omega}^s, \boldsymbol{\Omega}_i^s) = K_e(\mathbf{x}^s, \mathbf{x}_i^s) K_p(\boldsymbol{\Omega}^s, \boldsymbol{\Omega}_i^s), \tag{5}$$

where $\boldsymbol{\Omega}_i^s$ is the dynamic subband feature vector corresponding to the s^{th} subband component $\mathbf{x}_i^s$ of the i-th training point $\mathbf{x}_i$.

C. Stacked Generalization

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

$$f^s(\mathbf{x}^s, \boldsymbol{\Omega}^s) = \sum_i \alpha_i^s y_i K_\Omega(\mathbf{x}^s, \mathbf{x}_i^s, \boldsymbol{\Omega}^s, \boldsymbol{\Omega}_i^s) + b^s,\ \ s = 1, \ldots, S, \tag{6}$$

are then combined using stacked generalization [40] to obtain the binary classification decision for a test waveform $\mathbf{x}$. Our practical implementation of stacked generalization consists of a hierarchical two-layer SVM architecture, where the outputs of subband base-level SVMs are aggregated by a meta-level linear SVM. The decision function of the meta-level SVM classifier is of the form

$$h(\mathbf{x}) = \langle \mathbf{f}(\mathbf{x}), \mathbf{w} \rangle + v = \sum_{s} w^s f^s(\mathbf{x}^s, \mathbf{\Omega}^s) + v, \tag{7}$$

where $\mathbf{f}(\mathbf{x}) = \left[f^1(\mathbf{x}^1, \mathbf{\Omega}^1), \ldots, f^S(\mathbf{x}^S, \mathbf{\Omega}^S)\right]$ 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 each of the binary classifiers has its specific weight vector, optimized on an independent development/validation set $\{\tilde{\mathbf{x}}_j, \tilde{y}_j\}$, with a result of the form $\mathbf{w} = \sum_j \beta_j \tilde{y}_j \mathbf{f}(\tilde{\mathbf{x}}_j)$. Here $\mathbf{f}(\tilde{\mathbf{x}}_j) = \left[f^1(\tilde{\mathbf{x}}_j^1, \tilde{\mathbf{\Omega}}_j^1), \ldots, f^S(\tilde{\mathbf{x}}_j^S, \tilde{\mathbf{\Omega}}_j^S)\right]$ 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. While a base-level SVM assigns a weight to each supporting 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. Finally, ECOC methods are used to combine the meta-level binary classifiers for multiclass classification.

An obvious advantage of the subband approach for ASR is that with adequate normalization the effect of environmental distortions in sufficiently narrow subbands can be approximated as similar to that of a narrowband white noise. This facilitates the compensation of features to make them independent of the spectral characteristics of the additive noise and linear filtering. In a preceding paper [14], we proposed an ASR front-end based on the full-band acoustic waveform representation of speech where a spectral shape adaptation of the features was performed in order to account for the varying strength of contamination of the frequency components due to the presence of colored noise. On the other hand, compensation of the features in this work is performed solely using appropriate standardization. Furthermore, we found in our experiment that the weight vectors of the stacked classifiers can be tuned to classification in a particular environment by introducing similar distortions in its training data. To this end, a multi-style approach for training of the meta-level SVM classifiers is employed. Here, the meta-level classifiers are trained with the score feature vectors of a mixture of clean and noisy data to attain a reasonable compromise of classification performance over a wide range of test conditions. Note that the dimension of the score feature vectors that form the input to the stacked subband classifier (S) is very low compared to the typical MFCC or waveform feature vectors. Therefore only a limited amount of data is required to learn optimal weights of the meta-level classifiers. As such, stacked generalization offers flexibility and some coarse frequency selectivity for the individual binary classification problems, and can be useful in de-emphasizing information from unreliable subbands. The experiments presented in this paper show that the subband approach attains major gains in classification performance over state-of-the-art front-ends such as MFCC.

III. EXPERIMENTAL RESULTS

Experiments are performed using the 'si' (diverse) and 'sx' (compact) sentences of the TIMIT database [41]. The training set consists of 3696 sentences from 168 different speakers. Testing is performed using the core set which consists of 192 sentences from 24 different speakers not included in the training set. From the development set (1152 sentences by speakers not included in either the training or core test set), a small randomly selected subset comprising an eighth of its data points is used for the training of the meta-level subband classifiers. 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 [42], resulting in a total of $M = 48$ phoneme classes and $N = M(M-1)/2 = 1128$ classifiers. Furthermore, 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 [26, 42]. Hyperparameter values of the binary SVM classifiers are fixed as parameter optimization has 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 $C = 1$.

Each TIMIT test sentence is normalized to unit energy per sample and a noise sequence (pink or speech-babble noise from NOISEX-92) 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. Moreover, the noisy TIMIT sentences are convolved with a room impulse response with reverberation time $T_{60} = 0.2$sec and a -0.5dB spectral coloration (defined as the ratio of the geometric mean to the arithmetic mean of spectral magnitude), measured using an Earthworks QTC1 microphone in the ICSI conference room populated with people [43].

MFCC feature vectors are obtained by converting each sentence into a sequence of 13 dimensional feature vectors and combining them with their time differences and second order differences to give a sequence of 39 dimensional feature vectors. Then, the $T = 10$ 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 $\mathbb{R}^{390}$. Noise compensation of the MFCC features is performed via the vector Taylor series (VTS) method which has been extensively used in recent literature and is considered as state-of-the-art. In our experiments, Gaussian mixture models (GMM) with 64 mixture components were used to model the distributions of the MFCC features of clean training data. Additionally, CMVN [11, 12] is performed to standardize the MFCC features, fixing their range of variation for both training and test data.

In order to derive the subband features, acoustic waveforms segments $\mathbf{x}$ are extracted from the TIMIT sentences by applying a 100ms rectangular window at the centre of each phoneme, and then decomposed into subband components $\{\mathbf{x}^s\}_{s=1}^{S}$ using the cosine-modulated filter bank. For the results presented in this paper, the number of filter bank channels is limited to 16 in order to reduce the computational complexity. The dynamic subband feature vector, $\mathbf{\Omega}^s$ is computed by extracting $T = 10$ equal-length (25ms with an overlap of 10ms) frames around the centre of each phoneme, thus yielding a vector of dimension 30. These feature vectors are further standardized within each sentence of TIMIT for the evaluation of kernel K_Ω. The training of base-level SVM subband classifiers is always performed using clean data. On the other hand, the weight vectors of the stacked subband classifiers can be adapted for classification in a particular environment by introducing similar distortion to the relevant training data. We consider two particular scenarios for meta-level training: (a) training with the base-level SVM score vectors of clean data and (b) training with the base-level SVM score vectors of a mixture of clean and white noise (0dB SNR) contaminated data.

Results of the TIMIT phoneme classification experiments in the presence of additive noise and linear filtering and conducted under the setup detailed above are presented next. Figure 1 compares the classification performances of the subband and VTS-compensated

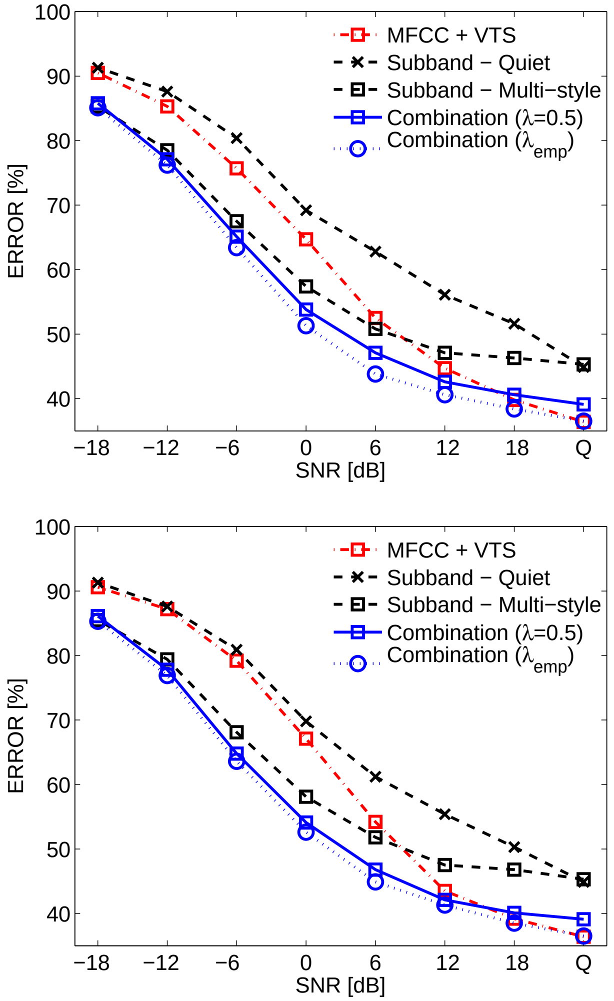

Fig. 1: *Classification with the subband and VTS-compensated MFCC classifiers, and their convex combination. Results are shown for the test data contaminated with pink noise (top) and speech-babble noise (bottom) as well as the ICSI conference room response. The convex combination curves correspond to the different settings of the paramater, λ.*

MFCC classifiers in the presence of linear filtering, and pink and speech-babble noise. In comparing the subband classifiers trained in quiet and multi-style training scenarios, one can observe that the performance of the stacked subband classifier trained in quiet conditions degrades relatively quickly even at low SNRs because its corresponding meta-level binary classifiers assign weights to different subbands that are tuned for classification only in quiet conditions. On the other hand, the multi-style trained subband classifier achieves consistent improvements over the one trained in quiet conditions and outperforms the MFCC classifiers for all SNRs below a crossover point between 12dB and 6dB, despite the mismatch of the noise level and type between training and testing conditions.

Since an obvious performance crossover between the MFCC and multi-style trained subband classifiers exists at moderate SNRs, we consider a convex combination of the scores of these classifiers using a combination parameter λ as discussed in [14]. Another factor motivating the convex combination approach was the stark difference in the error patterns of the two classifiers as observed from their confusion matrices (not shown here). This suggested that the errors of the subband and MFCC classifiers may be largely independent and a combination of the two may yield better performance than either of classifiers individually. Two different values of the combination parameter λ are considered, where λ=0 and 1 represent classification with, respectively, the VTS-compensated MFCC features and multi-style trained subband features. The first choice is $\lambda = 1/2$, which corresponds to the arithmetic mean of the MFCC and subband SVM classifier scores. In the second case, we set the combination parameter λ to a function $\lambda_{\text{emp}}(\sigma^2)$ which approximates the combination parameter values that are optimal on an independent development set. This approximated function was determined empirically in [14] for composite waveforms, rather than their subband components considered here, and it is given by

$$\lambda_{\text{emp}}(\sigma^2) = \eta + \zeta/[1 + (\sigma_0^2/\sigma^2)] \; ,$$

with $\eta = 0.2$, $\zeta = 0.5$ and $\sigma_0^2 = 0.03$. Note that $\lambda_{\text{emp}}(\sigma^2)$ also requires an estimate of the noise variance (σ^2) which was explicitly measured using the decision-directed estimation algorithm [44, 45]. Figure 1 compares the classification performances of the subband and MFCC classifiers with their convex combination. One observes that the combined classification with λ_{emp} consistently outperforms both of the individual classifiers across all SNRs; it attains a 5.4% and 7.1% average improvement over the subband and MFCC classifiers respectively, across all SNRs in both pink and speech-babble noise. Moreover, even the simple averaging of the subband and MFCC classifiers achieved by setting λ=$1/2$ provides a reasonable compromise between classification performance achieved in the two representation domains. While the performance of the combined classifier with λ=$1/2$ degrades only slightly (approximately 2%) as compared to the MFCC classifier for SNRs above a cross over point between 18dB and 12dB, it achieves relatively far greater improvements in high noise. For example, in pink noise the combined classifier with λ=$1/2$ attains a 10.9% and 3.6% improvement over the MFCC and subband classifiers at 0dB SNR, respectively. Quantitatively similar conclusions apply in the presence of speech-babble noise.

IV. Conclusions

This work investigated the potential of high-dimensional subband features for robust classification of phonemes, in comparison with the conventional MFCC front-end. The experiments demonstrated that classification with the subband features outperforms that with the cepstral features for SNRs below a crossover point between 12dB and 6dB. While the subband classifiers do not perform as well as the MFCC classifiers in low noise conditions, major gains across all noise levels can be attained by a convex combination [14]. In future work, we plan to investigate extensions of our approach to facilitate the recognition of continuous speech, by integrating it with other methods such as the hybrid phone-based HMM-SVM architecture [33] and the token-passing algorithm [35].

References

[1] G. Miller and P. Nicely, “An Analysis of Perceptual Confusions among some English Consonants,” *J. Acoust. Soc. Amer.*, vol. 27, no. 2, pp. 338–352, 1955.

[2] J. B. Allen, “How do humans process and recognize speech?,” *IEEE Trans. SAP*, vol. 2, no. 4, pp. 567–577, 1994.

[3] R. Lippmann, “Speech Recognition by Machines and Humans,” *Speech Comm.*, vol. 22, no. 1, pp. 1–15, 1997.

[4] B. Meyer, M. Wächter, T. Brand, and B. Kollmeier, “Phoneme Confusions in Human and Automatic Speech Recognition,” *Proc. INTERSPEECH*, pp. 2740–2743, 2007.

[5] J. Sroka and L. Braida, “Human and Machine Consonant Recognition,” *Speech Comm.*, vol. 45, no. 4, pp. 401–423, 2005.

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

[7] H. Bourlard, H. Hermansky, and N. Morgan, “Towards Increasing Speech Recognition Error Rates,” *Speech Comm.*, vol. 18, no. 3, pp. 205–231, 1996.

[8] K. K. Paliwal and L. D. Alsteris, “On the Usefulness of STFT Phase Spectrum in Human Listening Tests,” *Speech Comm.*, vol. 45, no. 2, pp. 153–170, 2005.

[9] S. D. Peters, P. Stubley, and J. Valin, “On the Limits of Speech Recognition in Noise,” *Proc. ICASSP*, pp. 365–368, 1999.

[10] S. B. Davis and P. Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” *IEEE Trans. ASSP*, vol. 28, pp. 357–366, 1980.

[11] C. Chen and J. Bilmes, “MVA Processing of Speech Features,” *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 257–270, 2007.

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

[13] P. J. Moreno, B. Raj, and R. M. Stern, “A Vector Taylor Series Approach for Environment-Independent Speech Recognition,” *Proc. ICASSP*, pp. 733–736, 1996.

[14] J. Yousafzai, Z. Cvetković, P. Sollich, and B. Yu, “Combined Features and Kernel Design for Noise Robust Phoneme Classification Using Support Vector Machines,” *To appear in the IEEE Trans. ASLP*, 2011.

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

[16] J. McAuley, J. Ming, D. Stewart, and P. Hanna, “Subband Correlation and Robust Speech Recognition,” *IEEE Trans. SAP*, vol. 13, no. 5, pp. 956 – 964, 2005.

[17] J. Ming, P. Jancovic, and F.J. Smith, “Robust Speech Recognition Using Probabilistic Union Models,” *IEEE Trans. SAP*, vol. 10, no. 6, pp. 403 – 414, Sept. 2002.

[18] H. Bourlard and S. Dupont, “Subband-based Speech Recognition,” *Proc. ICASSP*, pp. 1251–1254, 1997.

[19] C. Cerisara, J.-P. Haton, J.-F. Mari, and D. Fohr, “A Recombination Model for Multi-band Speech Recognition,” *ICASSP*, pp. 717 –720 vol.2, 1998.

[20] P. McCourt, N. Harte, and S. Vaseghi, “Discriminative Multiresolution Sub-band and Segmental Phonetic Model Combination,” *IET Elec. Letters*, vol. 36, no. 3, pp. 270 –271, 2000.

[21] P. McCourt, S. Vaseghi, and N. Harte, “Multi-Resolution Cepstral Features for Phoneme Recognition across Speech Sub-Bands,” *Proc. ICASSP*, pp. 557–560, 1998.

[22] S. Okawa, E. Bocchieri, and A. Potamianos, “Multi-band Speech Recognition in Noisy Environments,” *Proc. ICASSP*, pp. 641–644, 1998.

[23] S. Thomas, S. Ganapathy, and H. Hermansky, “Recognition Of Reverberant Speech Using Frequency Domain Linear Prediction,” *IEEE Sig. Proc. Letters*, vol. 15, pp. 681–684, 2008.

[24] S. Tibrewala and H. Hermansky, “Subband Based Recognition Of Noisy Speech,” *Proc. ICASSP*, pp. 1255–1258, 1997.

[25] H. Chang and J. Glass, “Hierarchical Large-Margin Gaussian Mixture Models for Phonetic Classification,” *Proc. ASRU*, pp. 272–275, 2007.

[26] P. Clarkson and P. J. Moreno, “On the Use of Support Vector Machines for Phonetic Classification,” *Proc. ICASSP*, pp. 585–588, 1999.

[27] S. Dusan, “On the Relevance of Some Spectral and Temporal Patterns for Vowel Classification,” *Speech Comm.*, vol. 49, pp. 71–82, 2007.

[28] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, “Hidden Conditional Random Fields for Phone Classification,” *Proc. INTERSPEECH*, pp. 1117–1120, 2005.

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

[30] K. M. Indrebo, R. J. Povinelli, and M. T. Johnson, “Sub-banded Reconstructed Phase Spaces for Speech Recognition,” *Speech Comm.*, vol. 48, no. 7, pp. 760–774, 2006.

[31] V. Pitsikalis and P. Maragos, “Analysis and Classification of Speech Signals by Generalized Fractal Dimension Features,” *Speech Comm.*, vol. 51, pp. 1206–1223, 2009.

[32] R. Rifkin, K. Schutte, M. Saad, J. Bouvrie, and J. Glass, “Noise Robust Phonetic Classification with Linear Regularized Least Squares and Second-Order Features,” *Proc. ICASSP*, pp. 881–884, 2007.

[33] A. Ganapathiraju, J. E. Hamaker, and J. Picone, “Applications of Support Vector Machines to Speech Recognition,” *IEEE Trans. Sig. Proc.*, vol. 52, no. 8, pp. 2348–2355, 2004.

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

[35] J. Padrell-Sendra, D. Martín-Iglesias, and F. Díaz-de María, “Support Vector Machines for Continuous Speech Recognition,” *Proc. EUSIPCO*, 2006.

[36] A. Halberstadt and J. Glass, “Heterogeneous Measurements and Multiple Classifiers for Speech Recognition,” *Proc. ICSLP*, pp. 995–998, 1998.

[37] T. Dietterich and G. Bakiri, “Solving Multiclass Learning Problems via Error-Correcting Output Codes,” *J. Artif. Intell. Res.*, vol. 2, pp. 263–286, 1995.

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

[39] S. Furui, “Speaker-Independent Isolated Word Recognition using Dynamic Features of Speech Spectrum,” *IEEE Trans. ASSP*, vol. 34, no. 1, pp. 52–59, 1986.

[40] D. Wolpert, “Stacked Generalization,” *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

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

[42] K. F. Lee and H. W. Hon, “Speaker-Independent Phone Recognition Using Hidden Markov Models,” *IEEE Trans. ASSP*, vol. 37, no. 11, pp. 1641–1648, 1989.

[43] “The ICSI Meeting Recorder Project - Room Responses,” Online Web Resource.

[44] Y. Ephraim and D. Malah, “Speech Enhancement Using a Minimum Mean-Square Error Short-time Spectral Amplitude Estimator,” *IEEE Trans. ASSP*, vol. ASSP-32, pp. 1109–1121, 1984.

[45] Y. Ephraim and D. Malah, “Speech Enhancement Using a Minimum Mean-Square Log-Spectral Amplitude Estimator,” *IEEE Trans. ASSP*, vol. ASSP-33, pp. 443–445, 1985.