PHONEME CLASSIFICATION IN FREQUENCY SUBBANDS USING ENSEMBLE METHODS

Nicholas Betteridge[1], *Zoran Cvetković*[2], *Peter Sollich*[1]

King's College London
[1]Department of Mathematics and [2]Division of Engineering
Strand, London, WC2R 2LS, UK
{nicholas.betteridge,zoran.cvetkovic,peter.sollich}@kcl.ac.uk

ABSTRACT

Phoneme classification in frequency bands of acoustic waveforms is studied. The goal is to investigate whether separate classifications across a number of subband signals, combined using appropriate machine learning algorithms, can provide performance similar to classification performed directly on the original acoustic waveforms. If this is the case, then combining subband classifications might lead to speech recognition algorithms that are robust to linear filtering and narrow-band noise. We perform proof-of-concept experiments on three binary phoneme classification tasks of varying difficulty, using Support Vector Machine subband classifiers which are combined by simple and weighted voting techniques as well as stacked generalization methods. We find that combining subband classifiers improves performance and that the improvement becomes more marked as the number of subbands increases.

Index Terms— Speech recognition, robustness, Support Vector Machines, ensemble methods, subban decompositions.

1. INTRODUCTION AND MOTIVATION

Substantial research efforts over the past decades devoted to the higher levels of speech recognition systems, *i.e.* language and context modeling, have resulted in major breakthroughs that have made automatic speech recognition (ASR) possible. There are currently many commercially available speech recognition systems covering applications which range from dictation and medical transcription to various customer service applications. However, state-of-the-art ASR systems still lack the level of robustness inherent to human speech recognition, which is manifested as a considerable degradation of performance in the presence of additive noise and/or linear filtering [1, 2]. The latter occurs, for instance, when the microphone used in the training process is different from the microphone used when the particular ASR task is actually performed, or if voice signal signal is passed to the system through a communication channel different from the channel under which the system was trained. This work is motivated by the need to improve the robustness of ASR to linear filtering and narrow-band additive noise. To this end we investigate phoneme classification in frequency bands of acoustic waveforms of speech.

To gain some intuition for the idea that speech recognition based on classification of subband components of acoustic waveforms followed by appropriate combination of these classifiers could indeed enhance robustness to linear filtering and narrow-band noise, assume that we have managed to construct such a classifier. Assume further that the subband components used at the first level of this classifier are contained within very narrow frequency bands. The effect of a reasonably smooth linear filter on a narrow-band component of a speech waveform would be approximately just amplitude scaling and a delay (see Figure 1), that is, the shape of the subband component of the acoustic waveform would not change much. Therefore, the performance of the subband classifiers, and consequently of the combined classifier, should not degrade considerably as a result of linear filtering. In the case of narrow-band noise, provided that the frequency support of the noise is known, the affected subband components can be excluded from the combining method, and hence the final classification result would again not be significantly affected. This subband approach may, however, exhibit inferior performance to classification on the original composite waveforms in the absence of linear filtering or narrow-band noise, because it imposes a very specific structure on the overall classification which may not be optimal. Our aim in this paper is to investigate whether this is indeed the case, or whether combining subband classifiers can give performance competitive with or perhaps superior to classification of the original waveforms.

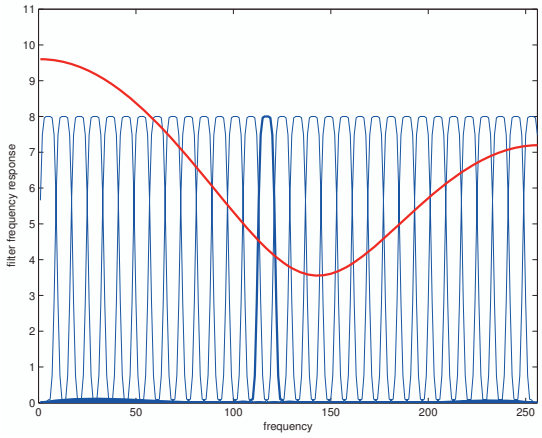

Fig. 1. *In narrow freqeuncy bands a reasonably smooth linear filter can be well approximated by a linear-phase constant-magnitude response. Therefore, the effect of a linear filter on a narrow-band signal will approximately amount to just amplitude scaling and a delay.*

Robust speech recognition has proved to be an extremely difficult problem of major importance to industry and of significant interest to acedeme. The method described above for approaching the specific issue of robustness to linear filtering obviously requires val-

idation. We focus here on studying the performance of classification in subbands in comparison to classification of composite waveforms and investigate also the impact of the number of subbands. Support vector machines (SVMs) will be used as subband classifiers and combined using popular ensemble methods [3, 4], such as majority voting, weighted majority voting, and stacked generalization. To enable a like-for-like comparison, SVMs will be also used as baseline classifiers, *i.e.* the classifiers that will be applied to composite waveforms.

2. METHODOLOGY

The data set used in this study consists of 64ms segments (1024 samples, at 16kHz sampling frequency) of phonemes from the TIMIT data base. For the purpose of our explorative study we considered the following three binary classification tasks:

- /iy/ (vowel) versus /sh/ (fricative) - easy classification task
- /iy/ (vowel) versus /n/ (nasal) - reasonably easy task
- /n/ (nasal) versus /m/ (nasal) - difficult task.

As indicated, from the structure of the phonemes involved, these tasks are expected to be increasingly difficult in the order listed.

2.1. Subband Decompositions

Each phoneme realization is decomposed into N subband components using a perfect reconstruction filter bank. For this purpose a cosine modulated filter bank is used consisting of filters

$$h_k[n] = \frac{1}{\sqrt{N}} h_{\rm pr}[n] \cos\left(\frac{2k-1}{4N}(2n-N-1)\pi\right), \quad k = 1,\ldots,N,\ n = 1,\ldots,2N \qquad (1)$$

where the prototype window $h_{\rm pr}[n]$ is a raised cosine function

$$h_{\rm pr}[n] = \sqrt{2}\sin\left(\frac{\pi(n-0.5)}{2N}\right),\ n = 1,2,\ldots,2N\,. \qquad (2)$$

This prototype filter satisfies the following two conditions

$$h_{\rm pr}[2N-n] = h_{\rm pr}[n],\ n = 1,\ldots,N$$

$$h_{\rm pr}[n]^2 + h_{\rm pr}[N-n-1]^2 = 2,\ n = 1,\ldots,N\,,$$

These imply that given an acoustic waveform $x[n]$, the ensemble of its subband components $\{x_1[n],\ldots,x_N[n]\}$, where $x_k[n]$ is the result of the convolution between $x[n]$ and $h_k[n]$, is a *tight frame* representation of $x[n]$ [5, 6]. The meaning of the tight frame property is that the distance between any two waveforms is magnified by the same amount in the subband component domain, and hence that the geometric configurations of data sets are not changed by means of this transform.

2.2. SVM classification and ensemble methods

For each of the N subband components we trained a standard SVM classifier with RBF kernel $K(x_k, x_k') = \exp(-\gamma\|x_k - x_k'\|^2)$ [7], on training data sets containing 1,500 examples of each of the two phonemes. The parameter γ and the standard SVM misclassification penalty parameter C were set separately in each case by a grid search over

$$\gamma \in \{0.05, 0.1, 1, 5, 20\} \text{ and } C \in \{0.1, 1, 10, 100, 1000\}\,.$$

More specifically, for every combination (γ, C) the test (misclassification) error was estimated by 5-fold cross-validation; the combination giving the lowest error was adopted as optimal and the SVM then retrained on the full training set. In the same fashion we trained SVM baseline classifiers on the original composite waveforms. This entire process was carried out separately for each of the three classification tasks at hand. We consider initially a set of $N = 8$ subbands; the effect of increasing N is studied later.

The splitting into subbands produces an ensemble of N classifiers. For a new (test) waveform x, one calculates the subband components x_k ($k = 1,\ldots,N$) and feeds each into its corresponding subband classifier to obtain a prediction $f_k(x_k)$. The key question is then how to combine these N "level-1" predictions optimally into an overall "level-2" prediction $f(x)$ of the phoneme class. We consider several choices. Firstly, the output $f_k(x_k)$ from each subband SVM is a real number and we have to decide whether to feed this directly into the level-2 classification. Conventionally, one thresholds SVM outputs to $\text{sgn}(f_k(x_k))$; we follow here the usual SVM convention of taking the class labels as ± 1 (rather than, say, 0 and 1). But such thresholding may lose information that is useful for classification at level 2. We therefore consider four possible "squashing functions"

$$\begin{aligned} F_1(f_k) &= \text{sgn}(f_k) \\ F_2(f_k) &= \begin{cases} f_k & \text{if} \quad |f_k| < 1 \\ \text{sgn}(f_k) & \text{else} \end{cases} \\ F_3(f_k) &= f_k \\ F_4(f_k) &= 1/[1+\exp(-f_k)] \end{aligned}$$

F_1 is hard thresholding; F_3 leaves the "raw" prediction of the level-1 SVMs intact. F_2 and F_4 provide a compromise, the former by thresholding only when $|f_k(x_k)| > 1$, and the latter by applying a sigmoid nonlinearity.

The subband classifications, transformed via squashing functions as above, still have to be combined into an overall classification $f(x)$. The simplest option is majority voting,

$$f(x) = \text{sgn}\left(\sum_{k=1}^{N} F_\alpha(f_k(x_k))\right)$$

where $\alpha \in \{1,2,3,4\}$ numbers the possible squashing functions that can be used; should the sum equate to zero, we have a draw and randomly predict class $+1$ or -1 with equal probability. As a generalization we also study weighted majority voting, where each $F_\alpha(f_k(x_k))$ is multiplied by a weight w_k. To set the weight we adopt the heuristic $w_k = 1 - \epsilon_k$, where ϵ_k is the cross-validation error obtained during the optimization of γ and C; this attributes less weight to less reliable subband classifiers as it should be. Finally we consider stacked generalization [8]. Here the subband classifiers are considered as preprocessors which transform a waveform x into an N-dimensional vector

$$\hat{x} = (F_\alpha(f_1(x_1)),\ldots,F_\alpha(f_N(x_N)))\,.$$

One can then use the training data preprocessed in this way to learn the optimal level-2 classifier. We implement the latter as an SVM with a dot-product kernel, so that the final classification is of the form

$$f(x) = \text{sgn}\left(\sum_{k=1}^{N} w_k F_\alpha(f_k(x_k)) + b\right)\,.$$

Compared to weighted majority voting this allows for an offset parameter b. More importantly, the $\{w_k\}$ and b are not set by hand

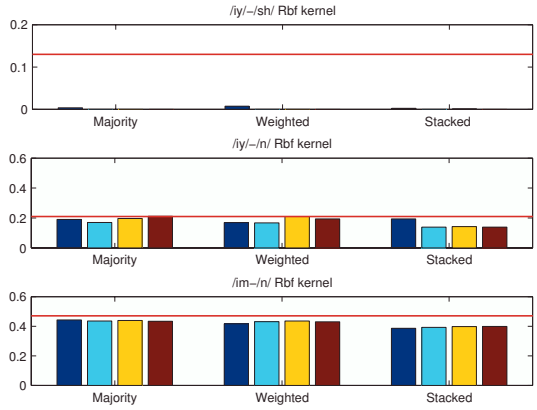

Fig. 2. *Test errors for the three classifi cation tasks considered (top to bottom). Within each graph, the three groups refer to the combination methods considered as indicated; within each group, the results for the squashing functions $F_1, \ldots, F_4$ are shown from left to right. Horizonal lines indicate the test error of the baseline classifi er trained (and tested) on the composite waveforms.*

but determined in a data-dependent way to give the best classification performance (again estimated by cross-validation, and used to optimize C).

3. CLASSIFICATION RESULTS

We show in Fig. 2 the test errors obtained in our experiments. The combination of the four possible squashing functions F_α with the three combination methods (majority voting, weighted voting and stacked generalization) gives us 12 results for each classification task. The test errors are obtained from 400 test examples per phoneme for each task; note that these examples are entirely unseen during the training of the level-1 (and, for stacked generalization, level-2) classifiers. We see from Fig. 2 that test errors across the three classification tasks increase in the order expected from the intuitive assessment of the difficulty. Much more importantly, however, the combination of subband classifiers yields *better* classification performance than the baseline classifiers trained on the composite waveforms: the availability of an ensemble of classifiers has helped rather than hindered classification ability. Given our motivation for this study, this result is very encouraging, suggesting that robustness to linear filtering and narrow-band noise might be achievable, while at the same time increasing recognition accuracy in the absence of these distortions.

Looking in more detail at Fig. 2, we see that stacked generalization performs consistently better than the other two combination methods. The increased flexibility in combining subband predictions that this method affords is therefore clearly worth the small computational overhead of having to train a level-2 classifier. Stacked generalization also performs best when one considers the test error separately for each of the two phonemes to be distinguished, giving similar errors for both rather than performing well on one and poorly on the other as happened in a few cases with the other methods.

As regards the effect of the squashing function on classification performance, Fig. 2 shows that this is relatively modest. Hard thresholding (F_1) is clearly suboptimal for /iy/-/n/ with stacked generalization but otherwise reasonable; the piecewise linear squashing function F_2 is seen to be a good compromise which performs close to optimally in all cases.

It should be stressed that the enhancement in classification performance that arises from combining subband classifiers is not due to the fact that the individual subband classifiers themselves are very reliable. In fact, as one might intuitively expect, the subband classifications are more error prone than those on the basis of the composite waveforms. For the task /iy/-/n/, for example, the average test error of the subband classifiers was 30%, which is considerably worse than the baseline classifier's error of 21%. However, after combining the classifications (using, say, hard thresholding followed by weighted majority voting) the test error or the subband ensemble was reduced to 17%, improving on the baseline result. This improvement is possible because the subband classifiers are sufficiently diverse to make up for the fact that each one of them individually is not very accurate.

Looking in more detail at the performance of individual subband classifiers is also revealing. Fig. 3 shows the histograms of decision values $f_k(x_k)$ for each of the $N = 8$ bands, separately for phonemes from each of the two classes in the /iy/-/n/ task. One notices that bands 6 to 8 separate the classes only very poorly; this is because at these higher frequencies distinguishing between the two phonemes is a substantially more difficult task.

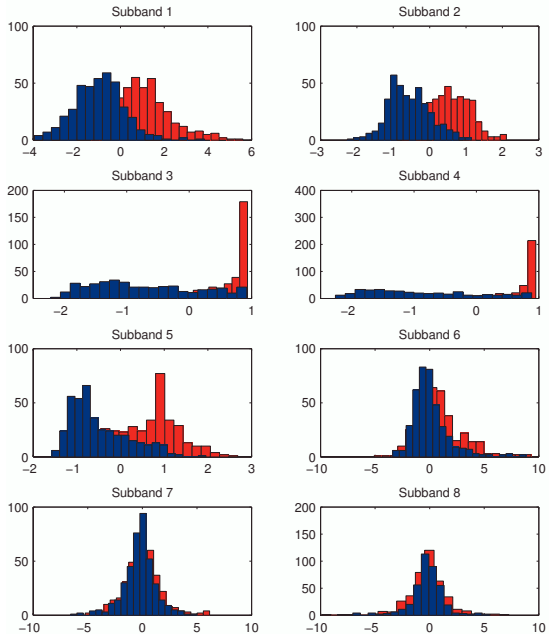

Fig. 3. *Decision values for test set by subband for /iy/-/n/. Blue: Test examples from class +1 (/iy/); red: class -1 (/n/).*

We finally consider the effect of increasing the number of subbands from $N = 8$ to $N = 16$. To keep computational effort manageable we reduced the number of training examples to 600 per phoneme in this case; to allow a fair comparison we then also re-ran the $N = 8$ experiments with this smaller training set. As Fig. 4 illustrates, the larger number of subbands tends to further enhance classification performance, thus continuing the trend we saw above

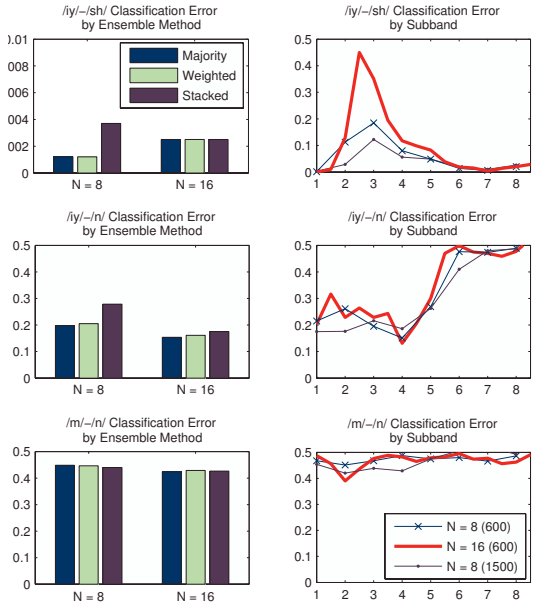

Fig. 4. *Left: Comparison of test errors for $N = 8$ and $N = 16$ subbands and 600 training examples per phoneme. Only the results for squashing function F_2 are shown. Right: Test errors according to subband; the x-axis is labelled by k for $N = 8$ and by $k/2$ for $N = 16$ so that equal values correspond to similar band frequencies. Test errors for the larger training set (1500 examples per phoneme) used initially for $N = 8$ are also shown.*

in going from $N = 1$ (baseline) to $N = 8$. The exception is the easy /iy/-/sh/ task, but here so few test examples are misclassified that the apparent increase of the test error for $N = 16$ is not statistically significant. On the right of Fig. 4 we show test errors for the individual subband classifiers, including also the $N = 8$ results for the original larger training set. The performance variation across subbands is similar between $N = 8$ and $N = 16$, being dependent mainly on the frequency of the subband considered. For /iy/-/n/ we have the performance degradation at high frequencies as observed previously. For /iy/-/sh/ this does not occur because of the noise-like nature of /sh/; conversely, performance here is worse in some of the lower frequency subbands, presumably because at least one of the two phoneme classes contains insufficient energy in this range.

4. CONCLUSION

We have investigated classification of phonemes by combination of predictions derived from subband components of the original composite waveform. This was motivated by the idea that such an approach might convey robustness both to linear filtering and narrow-band noise. While one might naively expect that a price may need to be paid for this gain in robustness, our key result is that classification performance is systematically *enhanced* by combining subband classifications. This effect arises because, even though the individual subband classifiers are rather poor, they are sufficiently diverse that their errors can, to a large extent, cancel out.

There are a number of possible ways in which the present proof-of-concept study could be extended. For simplicity we considered only binary classification tasks where two phonemes need to be distinguished; it would be interesting to see how combinations of subband classifiers fare in more realistic multiclass contexts. We employed SVMs here for their ease of use, placing less emphasis on the actual classification performance they yield. Particularly attractive as a conceptual alternative would be methods which predict class probabilities at level 1, such as Gaussian process classifiers [9]. These could then be combined into similarly probabilistic level-2 predictions by making appropriate (in)dependence assumptions between subband classifications and the final phoneme class. Other variations of the level-1 classifiers could be considered. For example, our SVM classifiers could presumably be further improved by applying ensemble methods already at level 1, e.g. by training several classifiers using different training subsets and combining them appropriately. As the sophistication of the classifiers in the method increases it will be particularly interesting whether the performance enhancements from subband combination persist or eventually decrease to zero.

Finally the potential of the proposed concept for attaining robustness to linear filtering and narrow-band noise needs to be investigated by performing classification on test data subject to these forms of degradation.

5. REFERENCES

[1] R. P. Lippmann, "Speech Recognition by Humans and Machines," *Speech Communication,* **22**: 1–15, 1997.

[2] J. J. Sroka and L. D. Braida, "Human and machine consonant recognition," *Speech Communication*, **45**:401–423, 2005.

[3] T. G. Dietterich, "Ensemble methods in machine learning", *Lecture Notes in Computer Science*, **1857**:1–15, 2000.

[4] J. V. Hansen, "Combining predictors: Comparison of five meta machine learning methods", *Inf. Sci.*, **119**:91–105, 1999.

[5] M. Vetterli and J. Kovačević,*Wavelets and Subband Coding*, Prentice Hall PTR, Englewood Cliffs, NJ, 1995.

[6] I. Daubechies,*Ten Lectures on Wavelets,* SIAM, Philadelphia, PA, 1992.

[7] N. Cristianini and J. Shawe-Taylor: *An introduction to Support Vector Machines*, Cambridge University Press, 2000.

[8] D. H. Wolpert, "Stacked generalization", *Neural Networks*, **5**:241–259, 1992.

[9] D. Barber and C. K. I. Williams, "Gaussian processes for Bayesian classification via hybrid Monte Carlo", *Advances in Neural Information Processing Systems*, **9**: 340–346, 1997.