ROBUST SPEECH RECOGNITION USING WAVELET COEFFICIENT FEATURES

Maya Gupta

Department of Electrical Engineering
Stanford University
Stanford, CA
guptama@stanford.edu

Anna Gilbert

AT&T Labs-Research, Room C215
180 Park Avenue
Florham Park, NJ
agilbert@research.att.com

ABSTRACT

We propose a new vein of feature vectors for robust speech recognition that use denoised wavelet coefficients. Greater robustness to unexpected additive noise or spectrum distortions begins with more robust acoustic features. The use of wavelet coefficients is motivated by human acoustic process modelling and by the ability of wavelet coefficients to capture important time and frequency features. Wavelet denoising accentuates the most salient information about the speech signal and adds robustness. We show encouraging results using denoised cosine packet features on small-scale experiments with the TIMIT database, its NTIMIT counterpart, and low-pass filter distortions.

1. INTRODUCTION

Current speech recognition systems perform well when tested on data similar to that used for training, however the lack of robustness of recognition systems continues to be a serious obstacle to practical speech recognition[1].

Speech recognition systems represent the speech waveform as feature vectors. A common set of feature vectors are some flavor of cepstral coefficients, such as Mel filter bank cepstral coefficients(MFCC), or LPC cepstral coefficients [2]. Acoustic and linguistic models are then used with the features to estimate what the speech waveform said.

Cepstral coefficients are a mature approach to feature vectors, but provide limited robustness, as evidenced by the difficulty of state-of-the-art systems to adapt to noise and distortions.

We propose a new vein of feature vectors, wavelet coefficients, to improve speech recognition robustness. Using wavelet coefficients is motivated by modelling of human acoustic processes and by the relationship of time-frequency coefficients to the Mel filterbank. Denoising theory and practice has shown that wavelet features can be robust to added noise and distortions.

This work was done while visiting AT&T Labs-Reearch.

The wavelet coefficients capture time and frequency localized information about the speech waveform that is impossible to obtain with a Fourier spectrum. Denoising the wavelet coefficients makes robustness part of the system. By including more localized time and frequency information, and by using wavelet denoising, we expect to be more robust to noise and spectrum distortions than cepstral coefficients. Encouraging results are shown in section 5 on a small-scale experiment with the TIMIT and NTIMIT database.

2. PREVIOUS WORK

Wavelets and time-frequency methods have been shown to be effective signal processing techniques over the last two decades for a variety of problems. In particular, wavelets have been successfully applied to denoising tasks and as robust features [3].

There has been recent interest in using the wavelet transform in speech recognition. One category of such papers, [4, 5, 6] uses a wavelet transform on the speech signal, computes the subband energies, and then uses these subband energies to replace Mel filterbank subband energies. This approach is slightly different from using the Mel filterbank in that the subband divisions induced by the wavelet transform are different from those in the Mel filterbank. The time information in the wavelet subbands, however, is lost into the subband energies. Sarikaya [5] uses a wavelet-packet tree that is a close approximation of the Mel-frequency division using Daubechies' 32-tap orthogonal filters. Our proposal differs in that we use the actual wavelet coefficients, and not the subband energies. This retains the time information. Furthermore, we denoise the wavelet coefficients to focus the features on the more salient information and improve robustness.

Another category of prior work in speech recognition that uses the wavelet transform is to apply it as an alternative to the cepstrum: Mel filterbank subband energies are computed, the log is taken, and then an inverse wavelet transform is performed [7, 8, 9].

3. WAVELETS AND WAVELET PACKETS

A wavelet expansion of a signal can be viewed as a tree expansion of recurrent low-pass and high-pass branches, with each filter followed by downsampling by a factor of two [10].

A wavelet transform expands only the low-pass branches of the tree, mapping N time samples into N wavelet coefficients. A wavelet packet transform expands the tree completely, mapping N time samples into $N log N$ wavelet coefficients. One can choose a wavelet packet tree pruning that results in an orthonormal basis of N coefficients that represent the signal in some optimal way, such as the minimum entropy representation.

The prototypical wavelet is the Haar wavelet, given by the low- and high-pass filters $\{1/\sqrt{2}, 1/\sqrt{2}\}$ and $\{1/\sqrt{2}, -1/\sqrt{2}\}$. Different filters (i.e., wavelets) may be used depending on the desired properties of the filterbank.

4. WAVELETS AND SPEECH

We propose to use an orthonormal set of the wavelet packet decomposition of the original time signal as features. We use the actual wavelet packet coefficients and not subband energies. In our experiments we used a local cosine packet [10] decomposition because the cosine packets form visually good matches to the speech signals and so we expect the cosine packet coefficients to represent the underlying information well.

There are several reasons why wavelet coefficients are a good approach to represent speech features for robust recognition. One physical model of the cochlea [11] suggests that it acts as a continuous wavelet transform in that different portions of the membrane respond to different frequency excitations logarithmically. Secondly, the Mel filterbank is a mature technology because it does work well. The subbands in the Mel filterbank are similar to those in wavelet decompositions in that both increase logarithmically in size as the frequency increases. Finally, wavelet (packet) decompositions are extremely successful in other scientific areas for denoising.

5. EXPERIMENT

We performed a small-scale experiment on the downsampled 8 kHz TIMIT and NTIMIT database, using only data from sentence 1 of region 1 (all speakers, mixed genders), yielding 1500 training phones and 436 test phones. We considered 40 phoneme classes, of which 26 appeared in this data set. We compared the adaptively denoised cosine-packet coefficients (CP) to standard mel-filterbank cepstral coefficients (MFCC).

The number of data points from each phone class in the test set is shown in Fig. 1.

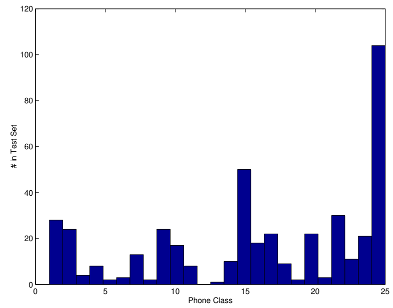

Fig. 1. Number of data points from each phone class in the test set

We designed the experiment to be as simple as possible to control the number of variables and effects. No information was used from outside the window given for a phone, thus there was no context-dependency nor were delta cepstral coefficients used. Both systems were trained only on the clean TIMIT data.

The feature vectors of the test phones were classified using the 1-Nearest Neighbor algorithm (1-NN) with Euclidean distance. 1-NN is not expected to be the optimal classification algorithm for phoneme recognition, but 1-NN is suitable for comparing the feature vectors without biasing the comparison by using a classifier known to work well with MFCC feature vectors. The 1-NN is known to perform well over a large class of problems and does not assume an underlying model (such as gaussianity) about the data.

Each phone was taken as a 32 ms (256 samples) time window centered around the center of the phone (we used the hand-segmented information available with the TIMIT database). If a phone was shorter than 32 ms, it was zeropadded. Then, we filtered each phone's time signal with the pre-emphasis filter $1 - .97z^{-1}$ and multiplied the result by a hamming window.

After that, for the wavelet CP analysis, we computed the 256 orthonormal cosine packet coefficients for each phone (using a basis experimentally optimized for discriminating silence). For each phone (training or test), we implemented standard wavelet denoising with a hard threshold [3]: we sorted the coefficients by magnitude and set to zero all but the top m coefficients, where m was a parameter we explored. Then we classified the test phones using the training phones and 1-NN.

For the MFCC analysis, after pulling out the centered 32 ms (256 time samples), zeropadding if necessary, preemphasis filtering and multiplying by the hamming window, we began the MFCC analysis by taking the magnitude of the fourier transform of each phone's time signal. Then we calculated the mel filterbank subband energies and computed the cepstrum. The cepstrum coefficients were normalized per sentence and then the mean was subtracted.

6. RESULTS

The results for the CP analysis depend on the number of coefficients not thresholded to zero. We experimented with keeping 80 to 248 coefficients (the rest of the coefficients are set to zero, and the classification is always done in the original 256 dimensions). In Fig. 2 we show the CP error rate as a function of the number of coefficients kept. The experimental results are noisy, but suggest a trade-off between thresholding out enough noise (keeping fewer coefficients) and retaining enough information (keeping more coefficients). Also, the classification was done in the original 256-D space (denoised coefficients are set to zero) and the small size of the training set undoubtedly had a worsening effect as the number of coefficients kept was decreased.

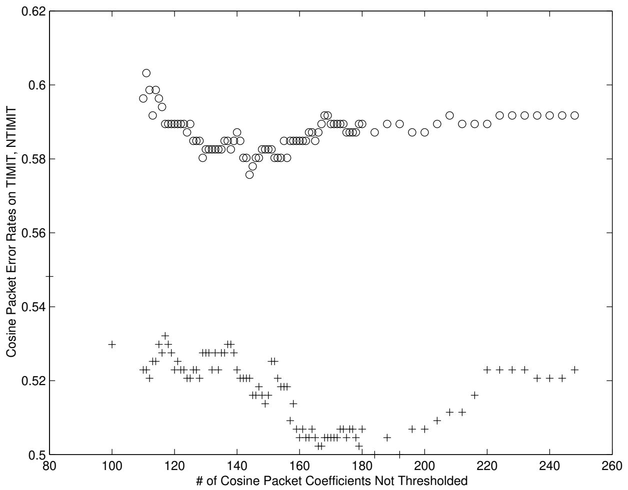

Fig. 2. CP error rate shown as a function of the number of coefficients not thresholded. Circles are CP error rates on NTIMIT data, crosses are CP error rates on TIMIT data

The best error rate on the NTIMIT data was 57.57% wrong when 144 coefficients were kept. The best error rate on the TIMIT data was 50% wrong when 184 coefficients were kept. Thus, as one would hypothesize, using more wavelet coefficients provides more information and is better in clean conditions, but more denoising (= fewer coefficients) is better for noisy environments.

For the other results in this section, 160 coefficients were kept in the denoising step (and the classification was done in the original 256 dimensions).

The MFCC results were 45% wrong on the clean TIMIT data and 62.39% wrong on the NTIMIT data. Thus the CP feature vectors were not providing as good clean performance but were, as theorized, able to degrade more gracefully in the presence of noise.

In Fig. 3 we show, for the NTIMIT data set, what percentage of each phone class was estimated correctly by CP

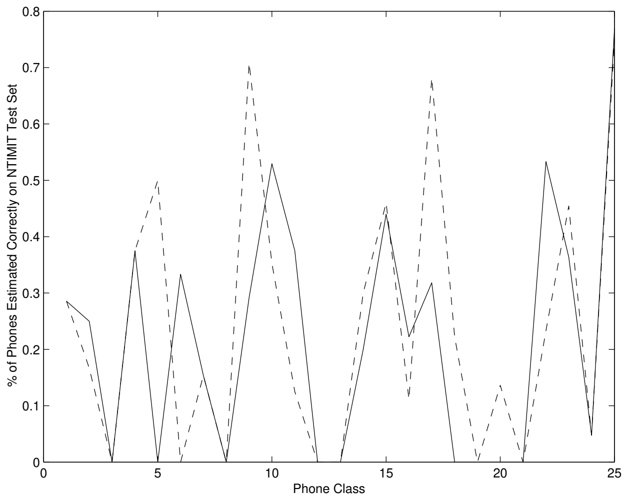

Fig. 3. Percentage correct for each phone class on the NTIMIT data, dotted line is CP, solid line is MFCC

and MFCC. Also, in Fig. 4 we plot the most likely class to be confused with each true class by MFCC and CP. These two figures show that the MFCC and CP methods perform differently over the classes and tend to confuse the classes differently, implying that the methods are thinking differently about the data.

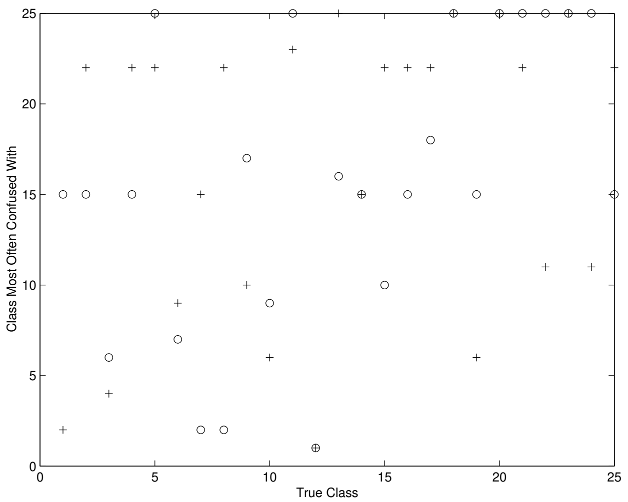

Fig. 4. Actual class plotted against the class most often confused with for the NTIMIT test set, circles represent cosine packet analysis and crosses represent MFCC features

We also experimented with lowpass filtering each test signal with a butterworth filter with cut-off frequency of 4 kHz. The results, shown in Table 1, show that the error increases less for the CP features than for the MFCC features.

	full spectrum	low-pass filtered speech
CP TIMIT	.5046	.6078
MFCC TIMIT	.4500	.5711
CP NTIMIT	.5849	.6055
MFCC NTIMIT	.6239	.6927

Table 1. Table with error rates for full spectrum and low-pass filtered speech

7. FUTURE WORK

This paper has proposed the use of wavelet coefficients for feature vectors and shown promising robustness results on a small experiment. Larger experiments need to be carefully designed to determine if wavelet coefficients can be a profitable and robust representation. There are also open theoretical and experimental questions of which wavelet and which basis are best for speech; best basis algorithms may be helpful [12].

We expect that a larger training set will have a positive effect on the ability to classify using denoised wavelet features, as such a high dimensional space was only very sparsely populated in our small-scale experiment.

8. CONCLUSIONS

Cepstral coefficient feature vectors are a mature technology that have not been shown to achieve good robustness to noise and distortion. In this paper we have provided theoretical and experimental reasons to investigate the use of wavelet coefficient feature vectors for robust speech recognition.

9. ACKNOWLEDGEMENTS

The authors would like to thank Bishnu Atal, Zoran Cvetkovic, Partha Parthasarathy, Mazin Rahim, Vinay Vaishampayan, and Vincent Vanhoucke for helpful discussions and mfcc code.

10. REFERENCES

[1] Louis Pols, "Flexible human speech recognition," *Proc. of the IEEE International Workshop on Automatic Speech Recognition and Understanding*, 1997.

[2] Lawrence Rabiner and Biing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[3] R. T. Ogden, *Essential Wavelets for Statistical Applications and Data Analysis*, Birkhauser, 1997.

[4] E. Erzin, A.E. Cetin, and Y. Yardimci, "Subband analysis for robust speech recognition in the presence of car noise," *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1995.

[5] Ruhi Sarikaya and John H. L. Hansen, "High resolution speech feature parameterization for monophone-based stressed speech recognition," *IEEE Signal Processing Letters*, vol. 7, no. 7, 2000.

[6] Kidae Kim, Dae Hee Youn, and Chulhee Lee, "Evaluation of wavelet filters for speech recognition," *Proc. of the IEEE International Conference on Systems, Man, and Cybernetics*, 2000.

[7] Kuansan Wang and David Goblirsch, "Extracting dynamic features using the stochastic matching pursuit algorithm for speech event detection," *Proc. of the IEEE International Workshop on Automatic Speech Recognition and Understanding*, 1997.

[8] P.M. McCourt, S.V. Vaseghi, and B. Doherty, "Multi-resolution sub-band features and models for hmm-based phonetic modelling," *Computer Speech and Language*, vol. 14, no. 3, 2000.

[9] J.N. Gowdy and Z. Tufekci, "Mel-scaled discrete wavelet coefficients for speech recognition," *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2000.

[10] Martin Vetterli and Jelena Kovacevic, *Wavelets and Subband Coding*, Prentice-Hall, Upper Saddle River, NJ, 1995.

[11] Ingrid Daubechies and Stephane Maes, "A nonlinear squeezing of the continuous wavelet transform based on auditory nerve models," *Wavelets in Medicine and Biology*, 1996.

[12] R. Coifman and M. Wickerhauser, "Entropy-based algorithms for best-basis selection," *IEEE Trans. on Information Theory*, vol. 38, no. 2, 1992.