A Wavelet and Filter Bank Framework for Phonetic Classification.

Ghinwa Choueiter and James Glass

Introduction: Whereas there has been several approaches to feature extraction for Automatic Speech Recognition (ASR) systems, the most commonly used measurement remains the Mel-Frequency Cepstral Coefficient (MFCC) [1] despite the fact that it has several limitations. First, it is inherently a short-time spectral representation based on Fourier analysis which is limited in its time-frequency representation. Second, its computation is based on the inner product of the signal power spectrum with triangular band-pass filters where the selection of the triangular filter shape is quasi-arbitrary. Third, its performance is not robust under noisy conditions.

In this research, we propose a wavelet and filter bank framework for feature extraction to improve the signal representation. The solution we propose, mainly tackles the first two points mentioned above but could easily be extended to address the third issue.

The Wavelet and Filter Bank Framework: Wavelets are functions capable of representing signals with a good resolution in the time and frequency domains. The wavelet transform is well defined within the multiresolution framework which allows signal analysis at various scales.Wavelets are characterized by time locality which allows efficient capture of transient behaviour in a signal. Furthermore, the time-frequency resolution trade-off provided by the multiresolution analysis provides more flexibility and better signal representation over Fourier-based analysis.

A filter bank, on the other hand, is an array of filters used to decompose a signal into subbands over different regions of the frequency spectrum. Such an analysis is quite useful, especially when the signal has a non-uniform spectral content as in the case of speech.

Within the multiresolution framework, continuous-time wavelets are closely related to discrete-time filter banks where it has been proved that a wavelet transform can be implemented using filter banks [2, 3]. It is this relation that we study and exploit for the task of feature extraction for phonetic classification.

As reported in the literature, most of the filter banks implemented for speech analysis make use of off-the-shelf wavelets such as the Daubechies. While this is straightforward from a design point of view, it does not necessarily lead to corresponding adequate filters such as ones with sharp cutoff and low attenuation in the stopband. Given constraints such as orthonormality and desired filter features such as regularity we adopt two approaches for filter design. The first method referred to as **Filter Matching** attempts to match a desired filter shape while the second referred to as **Attenuation Minimization** minimizes the attenuation band. Despite their simplicity and limitations, the two methods give insight into the advantages of designing task-optimized filters. Figure 1 illustrates the result of designing a low-pass filter that attempts to match the ideal filter using the Attenuation Minimization method. We observe that the designed filter has a sharper transition band and lower attenuation in the stopband than the filter corresponding to the Daubechies of order 12. For the most part, the literature mentions

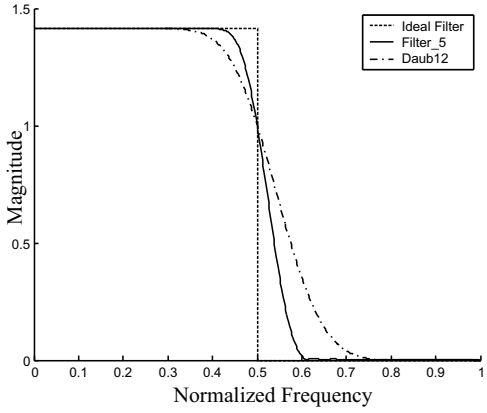

Figure 1: A designed low-pass filter (Filter_5) with the corresponding ideal filter it matches to and the Daub12 filter. Filter_5 was designed using the Attenuation Minimization method

filter banks implemented for speech analysis that generate octave bands by iteration of the low-pass

channel, or wavelet packets by iteration of both channels. The frequency partitions obtained in these cases, especially in the former, are not suitable for the task since octave-band filter banks do not have a good frequency resolution at the high frequency bands. Whereas we have replicated previously proposed solutions using wavelet packets and tree-structured filter banks, such solutions lead to a loss of the constant-Q characteristics.

This motivated our interest in filter banks customized for speech analysis. The objective is to develop filter banks that have a fine frequency resolution and can mimic the auditory filters. Fairly recently, there has been work done on filter banks with rational sampling factor [4] . Iterated rational filter banks give more flexibility in the frequency partitioning. More specifically, whereas the iterated dyadic filter banks are restricted to a single Q factor value, iterated rational filter banks can be designed to meet a range of Q values. Figure 2 illustrates a rational filter bank with the corresponding spectral partitioning obtained upon processing a signal with it. In our case, we restrict p to be q-1 so that our sampling factors are of the form q/(q-1). With such sampling factors we are able to obtain filter banks that mimic the human auditory system more naturally than octave-band filter banks.

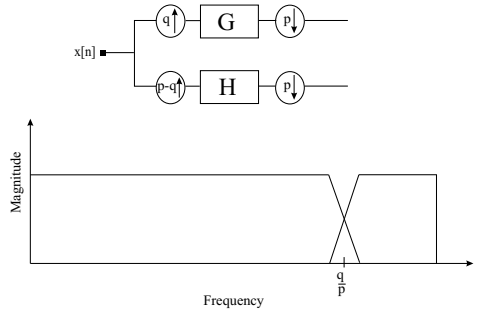

Figure 2: A rational filter bank with the corresponding rational frequency partitioning.

Results: Following the wavelet analysis, an acoustic feature was extracted as illustrated in Figure 3.

To evaluate the results, we select five acoustic measurements listed in Table 1 where the classification results for the phonetic subclasses are also reported. The baseline results are included for reference. The error rates corresponding to all the acoustic measurements match or exceed that of the MFCC on the Development set. The acoustic measurements are also evaluated on Test sets for significance level scoring. The McNemar significance test is used. A5 consistently outperforms measurements A1-A4. Compared to the baseline, A5 exhibits improvement that is statistically significant at the 0.05 level. We have also implemented 4-fold model aggregation for A5 obtaining 22.9% error rate on the 24-speaker Core Test set. We then combined this classifier with 8 other classifiers defined over 8 segmental features described in [5] obtaining an error rate of 18.5% on the same set, which is a slight improvement over the 18.7% obtained without the wavelet-based feature. Our results compare favorably to those mentioned in the

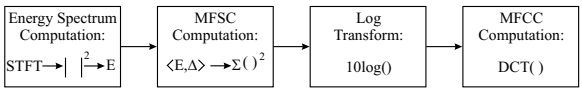

Figure 3: Flow Chart illustrating the extraction of the acoustic feature following wavelet analysis.

literature as well as those of the baseline classifier. The best error rate for context-independent phonetic classification is 18.3% on the Core Test set reported by Halberstadt who used hierarchical classifiers [5].

Future: The wavelet and filter bank framework for phonetic classification that we propose exploits two dimensions of the wavelet and filter bank theory: filter design and rational sampling. We show that off-the-shelf wavelets do not always give the best results, and there is a need for wavelet design. We also show that a dyadic filter bank implementation is not optimal, and we examine a method for rational filter bank design.

The framework, is however, still primitive in terms of design as well as implementation. First, the proposed energy-based acoustic measurement, though has some practical aspects from a computation point of view, does not take full advantage of the flexibility provided by the framework. A challenging task would be to design an acoustic measurement that would make full use of the proposed architecture. Furthermore, the framework is tested on the TIMIT corpus, which is a clean data set so the next step

Acoustic Meas.	(%) Error rate on the dev set						
	ALL	VOW	NAS	STP	WFR	SFR	CL
MFCC	23.9	31.6	25.3	27.4	28.5	21.5	4.2
A_1	23.6	30.4	26.5	28.9	28.1	21.2	4.2
A_2	23.4	31.5	26.5	28.9	25.4	23.8	3.7
A_3	23.4	30.7	23.5	28.7	26.9	21.2	4.3
A_4	23.1	30.4	23.4	28.7	27.6	21.0	3.6
A_5	23.2	30.5	25.5	26.4	27.7	22.7	3.3

Table 1: Classification performance (overall and phonetic subclasses) on the Development set. A_1 corresponds to the Daubechies wavelet of order 12, A_2 to a filter designed using Filter Matching to match the ideal filter, A_3 and A_4 are a 30-tap and 34-tap filters respectively designed using Attenuation Minimization, and A_5 corresponds to the rational filter bank of sampling factor $8/7$.

would be to implement it on a noisy data set, where wavelets have proven to be efficient in denoising tasks. The framework is also limited to the task of phonetic classification. A natural extension would be phonetic and word recognition. Finally, the filter design techniques that we have used in this thesis are simple and do not always give satisfactory results or even converge. It would be interesting to investigate other methods or even implement automatic filter optimization and generate filters that adapt to a task.

References:

[1] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. ASSP-28(4):357, 1980.

[2] G. Strang and T. Nguyen. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, Wellesley, MA, 1996.

[3] M. Vetterli and J. Kovacevic. *Wavelets and Subband coding*. Prentice-Hall, Englewood Cliffs, NJ, 1993.

[4] T. Blu. A new design algorithm for two-band orthonormal rational filter banks and orthonormal rational wavelets. *IEEE Transactions on Signal Processing*, 46(6):1494 – 1504, June 1998.

[5] A. K. Halberstadt and J. R. Glass. Heterogeneous measurements and multiple classifiers for speech recognition. In *Proc. ICSLP '98*, Sidney, Australia, November 1998.