A WAVELET AND FILTER BANK FRAMEWORK FOR PHONETIC CLASSIFICATION

Ghinwa F. Choueiter and James R. Glass

MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, MA 02139, USA
{ghinwa, glass}@mit.edu

ABSTRACT

In this paper, we present a wavelet and filter bank framework for context-independent phonetic classification with the aim of extending the work towards a full speech recognition system. The framework addresses the limitations of the Fourier analysis stage commonly used for short-time spectral representation of speech signals. Also, previous research pertaining to wavelet analysis for speech processing mostly makes use of off-the-shelf wavelets and dyadic-based signal decomposition. Our framework provides more flexibility by taking advantage of the relationship between wavelet transforms and filter banks, and using two filter design techniques as well as 'rational' wavelets. On the standard 39 phone TIMIT classification task, we achieve 22.9% error rate on the Core Test set using rational filter banks and 4-fold aggregation. This is improved to 18.5% when combined with multiple classifiers defined over non-wavelet acoustic measurements.

1. INTRODUCTION

The most commonly used measurement in automatic speech recognition (ASR) is the Mel-Frequency Cepstral Coefficient (MFCC) [1]. However, such a measurement is limited in its time-frequency representation since it is inherently a short-time spectral representation. Also, its computation is typically based on the inner product of the signal power spectrum with triangular band-pass filters where the filter shape selection is quasi-arbitrary. In this research, we propose a wavelet and filter bank framework for feature extraction to improve upon the signal representation as well as the filter selection.

Wavelets are functions with compact support capable of representing signals with good time and frequency resolution. The wavelet transform is well defined within the multiresolution framework allowing signal analysis at various scales. Filter banks have also emerged as signal processing tools that analyze and decompose signals into subbands over different regions of the spectrum. A wavelet transform can be implemented using an adequately designed filter bank [2, 3]. The filter bank features that we leverage are perfect reconstruction, regularity order, and orthogonality.

This research was supported in part by the SLS Affiliate program.

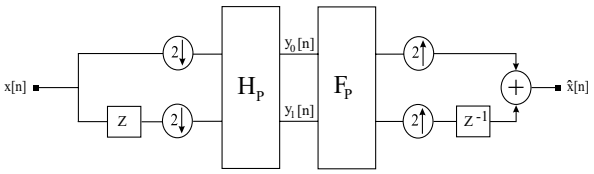

Fig. 1. A polyphase implementation of a 2-channel filter bank.

Much research has been done on wavelet analysis for speech recognition [4, 5, 6, 7]. However, most suffer from two main drawbacks which we attempt to address. First, the measurements are commonly extracted using off-the-shelf wavelets that are not optimized for speech processing. We examine two techniques, Filter Matching and Attenuation Minimization, to design a filter, and hence a wavelet that matches desired features. Second, filter banks implementing wavelet transforms are typically dyadic, splitting the spectrum in half. We examine tree-structured filter banks, which were previously proposed as potential solution, allowing iteration at both high and low channels of a 2-channel filter bank [8]. We then look into rational filter banks to obtain finer frequency resolution and naturally simulate the critical bands of the human auditory system [9].

In Section 2, we present a brief overview of the framework including the two filter design techniques and the rational filter banks. In Section 3, we describe the experimental setup, and in Section 4, we present our results. In Section 5, we summarize and propose future extensions to the current framework.

2. WAVELET AND FILTER BANK FRAMEWORK

Within the multiresolution framework, continuous-time wavelets and discrete-time filters are closely connected. Fil-

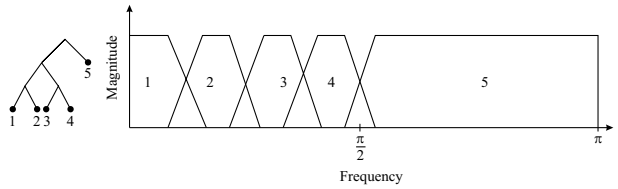

Fig. 2. A tree-structured implementation of a filter bank and the corresponding frequency partitions.

ter banks consisting of analysis and synthesis filters, downsamplers and upsamplers can be used to implement wavelet transforms efficiently. Fig.1 illustrates a 2-channel filter bank where the analysis and synthesis filters $\mathbf{H_p}(z)$ and $\mathbf{F_p}(z)$ are implemented in the polyphase domain using polyphase matrices [2, 3]. We are concerned with perfect reconstruction orthogonal filter banks characterized by paraunitary matrices that can be factorized using Lattice and Householder factorizations [3]. This will be useful when implementing the filter design methods.

2.1. Tree-Structured Filter Banks

A 2-channel dyadic filter bank splits the spectrum in half at each iteration, and results in a constant-Q octave band when iterated on the low-pass channel. Such a filter bank can be extended to an arbitrary tree-structure by allowing iteration on both channels. Special attention should be given to the mirroring of the spectrum on the high-pass channel [3]. Fig.2 illustrates an example of an iterated filter bank and the corresponding frequency partitions. We experiment with several frequency partitions, and adopt a tree-structure that consists of 8 filters uniformly distributed over 0-1 kHz, 4 over 1-2 kHz, 8 over 2-4 kHz, 4 over 4-6 kHz, and 2 over 6-8 kHz. The result is a 26-band filter bank structure that roughly incorporates the critical band effect.

2.2. Filter Design

We examine two filter design techniques with the aim of providing flexibility to the framework and overcoming the limitation of previous research pertaining to wavelet analysis for speech processing, as far as wavelet selection is concerned [5, 6, 8]. Instead of off-the-shelf wavelets, we are interested in those corresponding to filters with sharp cutoff and good stopband attenuation, and hence good frequency selectivity. In the two filter design methods, we take into consideration orthogonality and regularity order of the filter. The concept of regularity is related to the smoothness of the wavelet function which is, in turn, loosely related to the number of vanishing wavelet moments. Intuitively, the more the number of vanishing wavelet moments, the greater the number of zeros at the aliasing frequency and the more frequency selective the corresponding filter is.

2.2.1. Filter Matching

The Filter Matching method minimizes the difference in modulus between the designed and desired filter given that:

$$\frac{d^l H_0(\omega)}{d\omega^l}|_{\omega=\pi} = 0 \quad l = 0 \ldots N-1 \tag{1}$$

where $H_0(\omega)$ is the frequency response of the analysis low-pass filter being designed, and N is the number of desired vanishing moments. The designed filter is represented using lattice factorization which imposes the orthogonality of the filter bank [2, 3]. The algorithm implementation is based on a sequential quadratic programming method for constrained optimization. We test the algorithm by matching it to two desired filters: the Butterworth filter of order 10 and cutoff frequency $\pi/2$ and the ideal low-pass filter.

2.2.2. Attenuation Minimization

The Attenuation Minimization method is a special case of the rational filter bank design technique [9]. The idea is to match the filter to an ideal low-pass filter by minimizing the difference in modulus between the two filters. However, as it has been shown in [9, 10], the problem can be reduced to an attenuation minimization. All the filters are designed with a regularity order of 1 in order to guarantee convergence of the algorithm. Using this technique, we obtain filters that exhibit good attenuation.

Fig.3 illustrates low-pass filters designed using the two techniques. The ideal low-pass filter with normalized cutoff frequency 0.5 and the filter corresponding to the Daubechies wavelet of order 12 [11] are also included. *Match_Ideal* corresponds to a filter designed using the first technique to match the ideal filter. *Filter_5* and *Filter_6* are 30-tap and 34-tap filters designed using the second method.

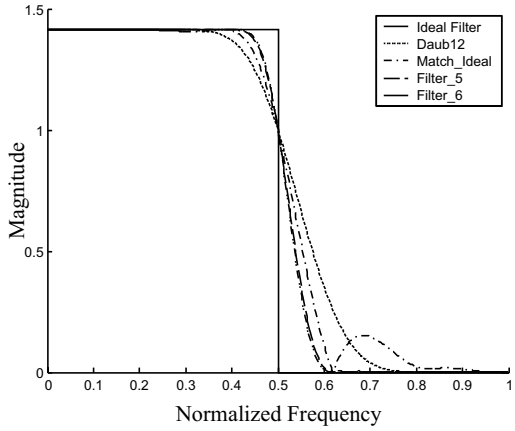

Fig. 3. The frequency responses of three designed filters. The ideal and Daub12 low-pass filters are included for reference.

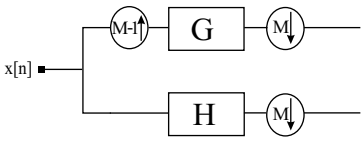

Fig. 4. A filter bank of rational sampling factor $M/(M-1)$.

2.3. Rational Wavelets and Filter banks

A dyadic filter bank splits the spectrum in half at each iteration, and has a sampling factor of 2. If the filter bank has a sampling factor of the general form $M/(M-1)$, the spectrum will be split into the bands: $[0, \frac{M-1}{M}\pi]$ and $[\frac{M-1}{M}\pi, \pi]$. We obtain a finer frequency resolution as well as a Q-factor tunable to the human auditory system. In fact, the Q-factor corresponding to such a filter bank is found to be:

$$Q = \frac{bandwidth}{center\ frequency} = \frac{1}{M - 1/2} \qquad (2)$$

This gives us M that matches a desired Q. Two interesting sampling factors are $8/7$ and $6/5$ that allow filter banks to mimic the Mel and Bark scale respectively.

It has been shown that rational filter banks can be put in the form illustrated in Fig.4 [12]. In [9, 10], Blu proposes an algorithm for designing such filter banks. Due to space constraints, we do not describe it in detail but give some insight behind it. The aim is to find the best frequency selective low-pass filter $G(z)$ given orthogonality and the regularity order of the filter bank. The problem is reduced to minimizing the attenuation band of $G(z)$ and is formulated using the Lagrange multiplier method for constrained minimization.

Blu devised a recursive implementation of the algorithm with condition for convergence being minimal perfect reconstruction error. After finding the low-pass filter, the high-pass filter is derived from it by taking advantage of orthogonality conditions and using Householder factorization [9].

We use this technique to examine rational filter banks. To our knowledge, there has been no implementation of rational filter banks to feature extraction tasks prior to this.

The regularity order is set to 1 for all the designed filters. The rational filter banks are iterated on the low-pass channel N times until the lower cutoff frequency of the last band-pass filter obtained is close to 1 kHz. A 30-tap filter designed using the Attenuation Minimization technique divides the 0-1 kHz region into 8 equipartitions. This gives us frequency partitions that mimic the critical bands. The length of the filters is large. This is necessary in order to obtain filters with narrow passbands and also good frequency selectivity.

3. EXPERIMENTS

The classification experiments are performed on the TIMIT corpus. Following common practice, the 61 phone labels in TIMIT are collapsed into 39 labels prior to scoring and glottal stops are ignored. We use the 462 speaker training set, 50 speaker development set, 24 speaker core test set for final testing, and 118 speaker full test set for significance level scoring. Segmental acoustic features described below are extracted over phonetic segments. Diagonal Gaussian mixture models are implemented with a minimum of 61 datapoints per mixture component and a maximum of 96 mixture models per phone.

The acoustic measurement of the baseline classifier consists of 14 MFCCs computed using short-time Fourier analysis at a rate of 5ms over 25.6ms frames. A 76-dimensional feature vector is obtained by concatenating 3 MFCC and energy averages over segment thirds, 2 MFCC and energy derivatives at segment boundaries, and a log duration [13].

For our framework, we first specify the wavelet type and the spectral partitions and then compute the wavelet transform over 20ms frames at a rate of 5ms. This is equivalent to processing the signal using an N-band filter bank. We then compute the energy for each of the N bands and perform their log transform. The result is an N-dimensional acoustic measurement used to generate a $(5N + 6)$-dimensi-onal segmental feature over a given segment derived similarly to the MFCC-based segmental feature. Principal component analysis is used to reduce the feature space dimensionality as well as whiten it. The feature space dimension used in the experiments is 76 similarly to the baseline. A 26-band tree-structured filter bank is used to extract the acoustic measurements (A_1-A_4 in Table 1) except for the rational filter bank (A_5), where 22 bands are extracted.

4. RESULTS

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. Also, the results listed in Table 1 are reminiscent of those obtained by Halberstadt [14]. Although the overall error rates are close to each other, there is a difference in performance over the phonetic subclasses. This suggests a hierarchical approach where filters optimized to different subclasses are designed.

The acoustic measurements are also evaluated on the Core Test and Full Test sets for significance level scoring. The McNemar significance test is used [15]. A_5 consistently outperforms measurements A_1-A_4. Compared to the baseline, A_5 exhibits improvement that is statistically sig-

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) of acoustic measurements A_1-A_5 and the baseline (MFCC) 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 $8/7$.

nificant at the 0.05 level.

We have also implemented 4-fold model aggregation [16] for A_5 obtaining 22.9% error rate on the Core Test set. We then combined this classifier with 8 other classifiers defined over 8 segmental features described in [14] obtaining an error rate of 18.5% on the same set, which is an improvement over the 18.7% obtained without the wavelet-based feature.

Our results compare favorably to those mentioned in the 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 [14]. Clarkson and Moreno obtained 22.9% on the same set using Support Vector Machines (SVM) [17]. Zahorian et al. obtained 23.0% on the Core Test set using spectral/temporal features and a neural network classifier [18].

5. DISCUSSION AND FUTURE WORK

We have presented a wavelet and filter bank framework for phonetic classification in which we have exploited two dimensions of the wavelet and filter bank theory: filter design and rational sampling.We have shown that off-the-shelf wavelets do not always give the best results, and there is a need for wavelet design. We have also shown that a dyadic filter bank implementation is not optimal, and we have examined a method for rational filter bank design.

The framework, is however, still primitive in terms of design as well as implementation. For example, it is tested on the TIMIT corpus, which is a clean data set. It would be challenging to implement it on a noisy data set, where wavelets have proven to be efficient in denoising tasks [4]. 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.

6. REFERENCES

[1] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-28, no. 4, pp. 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] M. Gupta and A. Gilbert, "Robust speech recognition using wavelet coefficient features," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Dec. 2001, pp. 445–448.

[5] K. Kim, D.H. Youn, and C. Lee, "Evaluation of wavelet filters for speech recognition," in *IEEE International Conference on Systems, Man, and Cybernetics*, Nashville, TN USA, Oct. 2000, vol. 4, pp. 2891–2894.

[6] B.T. Tan, R. Lang, H. Schroder, Minyue, A. Spray, and P. Dermody, "Applying wavelet analysis to speech segmentation and classification," in *Wavelet Applications, Proc. SPIE 2242*, 1994, pp. 750–761.

[7] H. Wassner and G. Chollet, "New time-frequency derived cepstral coefficients for automatic speech recognition," in *Proc. ICSLP '96*, Philadelphia, PA USA, Oct. 1996, vol. 4, pp. 260–263.

[8] O. Farooq and S. Datta., "Robust features for speech recognition based on admissible wavelet packets," in *Electronics Letters*, Dec. 2001, vol. 37, pp. 1554–1556.

[9] T. Blu, "A new design algorithm for two-band orthonormal rational filter banks and orthonormal rational wavelets," *IEEE Transactions on Signal Processing*, vol. 46, no. 6, pp. 1494 – 1504, June 1998.

[10] T. Blu, *Bancs de filtres iteres en fraction d'octave, Application au codage de son*, Ph.D. thesis, Ecole National Superieur des Telecommunications, Paris, France, april 1996, in French.

[11] I. Daubechies, *Ten Lectures on Wavelets*, vol. 61, SIAM Press, Philadelphia, PA USA, 1992.

[12] J. Kovacevic and M. Vetterli, "Perfect reconstruction filter banks with rational sampling factors," *IEEE Transactions on Signal Processing*, vol. 41, no. 6, pp. 2047 – 2066, June 1993.

[13] J. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, pp. 137–152, 2003.

[14] A. K. Halberstadt and J. Glass, "Heterogeneous measurements and multiple classifiers for speech recognition," in *Proc. ICSLP '98*, Sydney, Australia, Nov. 1998.

[15] L. Gillick and S.J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP '89*, Glasgow, UK, May 1989, vol. 1, pp. 532–535.

[16] T. J. Hazen and A. K. Halberstadt, "Using aggregation to improve the performance of mixture Gaussian acoustic models," in *Proc. ICASSP '98*, Seattle, WA USA, May 1998, pp. 653–656.

[17] P. Clarkson and P.J. Moreno, "On the use of support vector machines for phonetic classification," in *Proc. ICASSP '99*, Phoenix, AZ USA, Mar. 1999, vol. 2, pp. 585–588.

[18] S. A. Zahorian, P. L. Silsbee, and X. Wang, "Phone classification with segmental features and a binary-pair partitioned neural network classifier," in *Proc. ICASSP '97*, Munich Germany, Apr. 1997, pp. 1011–1014.