An Implementation of Rational Wavelets and Filter Design for Phonetic Classification

Ghinwa F. Choueiter, *Student Member, IEEE*, and James R. Glass, *Senior Member, IEEE*

***Abstract*—Although wavelet analysis has been proposed for speech processing as an alternative to Fourier analysis, most approaches make use of off-the-shelf wavelets and dyadic tree-structured filter banks. In this paper, we extend previous wavelet-based frameworks in two ways. First, we increase the flexibility in wavelet selection by taking advantage of the relationship between wavelets and filter banks and by designing new wavelets using filter design methods. We adopt two filter design techniques that we refer to as filter matching and attenuation minimization. Second, we improve the flexibility in frequency partitioning by implementing rational as well as dyadic filter banks. Rational filter banks naturally incorporate the critical-band effect in the human auditory system. To test our extensions, we implement an energy-based measurement which we also compare in performance to the mel-frequency cepstral coefficients (MFCCs) in a phonetic classification task. We show that the designed wavelets outperform off-the-shelf wavelets as well as an MFCC baseline.**

***Index Terms*—Filter design, phonetic classification, rational wavelets.**

I. Introduction

THE MOST commonly used observations in automatic speech recognition (ASR) are based on a short-time spectral representation that assumes time-stationarity within fixed-sized time frames. For example, mel-frequency cepstral coefficients (MFCCs), which are the dominant representations for ASR, fall into this category [1]. Recently, wavelets and filter banks (FBs) have been introduced in ASR as potential speech processing tools to overcome the limitations of such spectral representations. The wavelet transform provides an improved signal representation with a tradeoff between time and frequency resolution. Moreover, the wavelet transform can be efficiently implemented using FBs that naturally take the critical-band effect into account, and it has also been shown to effectively emulate the cochlear transform [2]. We believe that these properties of the wavelet transform makes it an attractive representation for phonetic classification.

Much research on acoustic measurement extraction using wavelet analysis has been done [3]–[7]. Kim *et al.* proposed a modified octave-structured FB for speech recognition [8]. Their experiments showed better performance for the Daubechies wavelet over MFCCs on the task of Korean digit recognition. Tan *et al.* compared the discrete wavelet transform against the sampled continuous wavelet transform and MFCC as front-end processors in a speaker-independent HMM-based phoneme recognition system [9]. The experiments, performed over a subset of TIMIT [10], indicated marginal improvement of the wavelet-based front-end over MFCCs. Farooq and Datta used Daubechies wavelet packets to obtain a 24-band FB that mimics MFCCs [4]. The acoustic measurement was obtained by computing the log energy in each frequency band, and rotating the outputs with a discrete cosine transform. Phonetic classification was performed on the TIMIT corpus, over a limited subset of the data and the phonemes, and compared with MFCCs. Their results showed that the wavelet-based measurement outperformed MFCCs in the case of stops and unvoiced phonemes.

In this paper, we extend previous research on wavelet analysis for measurement extraction. Most of the wavelets used for speech analysis are generic and not particularly designed for the task at hand. For example, some wavelets might correspond to FBs that have a poor attenuation in the stopband causing energy leakage. Moreover, wavelet transforms are often implemented using dyadic FBs, which do not necessarily have a good frequency resolution. In this paper, we design new wavelets using two filter design techniques: filter matching and attenuation minimization. We also examine rational FBs as a potential extension to dyadic FBs [11], [12]. To the best of our knowledge, rational filter banks have not been previously implemented for acoustic measurement extraction in ASR.

We restrict our implementation of the wavelet-based acoustic measurement extraction to the task of context-independent phonetic classification. We believe this examination will give insight into the advantages and limitations of the proposed techniques. Although we do not perform speech recognition experiments, it has been previously shown that gains in phonetic classification tend to translate into gains in phonetic and word recognition [13], [14].

This paper is structured as follows. In Section II, we present a brief overview on rational wavelets and FBs, and in Section III, we elaborate on the filter design methods implemented for dyadic as well as rational FBs. In Section IV, we describe the experimental setup used to evaluate the framework and the proposed extensions. In Section V, we present the results for Daubechies wavelets and tree-structured dyadic FBs, and the designed wavelets and rational FBs. We summarize and propose future extensions to the current framework in Section VI.

II. Rational Wavelets and Filter Banks

The theory of wavelets and filter banks has been studied intensively over the past two decades [11], [15]–[17]. The range of

Manuscript received January 30, 2006; revised June 30, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mary P. Harper.

The authors are with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: ghinwa@mit.edu; glass@mit.edu).

Digital Object Identifier 10.1109/TASL.2006.889793

their application varies from image and signal processing to geophysics. Wavelets and FBs have evolved separately. A wavelet is a function with compact support capable of representing signals with good resolution both in the time and frequency domains. Wavelets, like sinusoids, are basis functions that span the square-integrable space, $L_2(\mathcal{R})$, and are used to develop series expansions of signals in that space. An FB, on the other hand, is an array of filters, which can be low-pass, band-pass, or high-pass, that decompose a signal into subbands over different regions of the spectrum. Within the multiresolution framework, continuous-time wavelets and discrete-time FBs are closely related. It has been shown that a wavelet transform can be efficiently implemented using FBs [11], [17]. It is this relation that is typically exploited and that we further stress in this research.

The following is a brief presentation of orthonormal rational wavelets and FBs. For readers who are interested in a broader overview on wavelets and FBs, we provide further theoretical background in Appendices I and II. More detail relevant to the research described in this paper can be found in [18].

As described in Appendix I, FBs are commonly implemented in a dyadic fashion, meaning that at each iteration of the FB, the spectrum is split in half. This results in a spectral decomposition that does not have a good resolution at the high frequencies. In this paper, we propose to use rational sampling to obtain a finer spectral resolution and naturally simulate the critical bandwidths. First, we define the Q-factor as the ratio of the bandwidth to the center frequency of a band. For the octave-band FB with a partitioning ratio of 1/2, the Q-factor is 2/3. We are interested in the more general partitioning ratio of $(M-1)/M$, where at each iteration of the FB, the spectrum is split into the ratios $1/M$ and $(M-1)/M$. To calculate the Q-factor corresponding to such a FB, we refer to the frequency partitioning after one iteration: the bandwidth of the highest frequency band is π/M, and the center frequency of that band is $\pi/2M + ((M-1)/M)\pi$. The expression for Q, in this case, becomes

$$Q = \frac{\text{bandwidth}}{\text{center frequency}} = \frac{\frac{\pi}{M}}{\frac{\pi}{2M} + \frac{M-1}{M}\pi} = \frac{1}{M - \frac{1}{2}}. \quad (1)$$

With this formula, we obtain the value of M that matches a desired Q-factor. For example, to closely approximate the Q-factor of the filters in the MFCC implementation, which is 0.1376, we need to set $M \approx 8$ and the resulting sampling ratio becomes 8/7. Fig. 1 illustrates the 40 MFSC filters used in the MFCC computation as well as those generated using a rational filter bank with a sampling factor of 8/7. Another interesting sampling ratio is 6/5 which closely approximates the Bark scale.

There has been research on perfect reconstruction FBs with rational sampling factors [19] and nonuniform multirate FBs [20]. We base the remainder of our work on rational FBs on the research of Blu [12], [21], [22], where we refer the reader for further detail and proofs.

A rational FB can only approximate a wavelet transform. In other words, rational sampling factors with FIR filters do not lead to a multiresolution analysis, and the iteration of a FB does not generate a unique limit function. This is because the *wavelet* function corresponding to a rational FB is not shift-invariant. The shift error, however, can be made arbitrarily small when the function regularity increases. In [12], Blu designed an algorithm for the rational case that takes into consideration the shift error. Fig. 2 illustrates a rational FB as proposed in [12], where G is the low-pass filter and H is the high-pass filter. Both FB branches are downsampled by a factor of p; however, the low-pass branch is upsampled by a factor of q, while the high-pass branch is upsampled by a factor of $p-q$. We briefly describe the FB design algorithm in Section III-B. We use the term *rational wavelets* as mentioned in [12], since the functions do not satisfy the shift-invariance property and effectively are not wavelets. Also, we concentrate on the rational sampling factor of the form $M/(M-1)$, although the references study rational FBs with general sampling factor p/q. For example, referring to Fig. 2, $p = M$ and $q = M-1$.

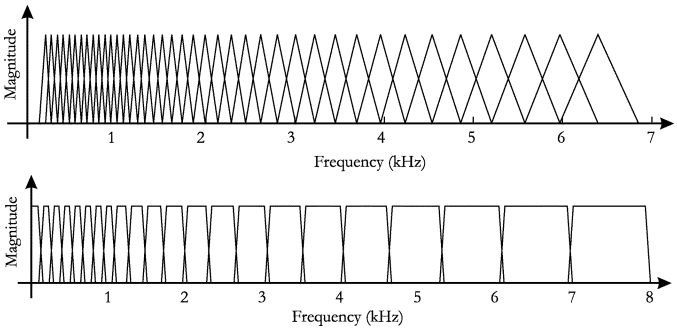

Fig. 1. Top figure illustrates the 40 MFSC filters used in the MFCC computation. Bottom figure shows the filters obtained using the rational filter bank with sampling factor 8/7.

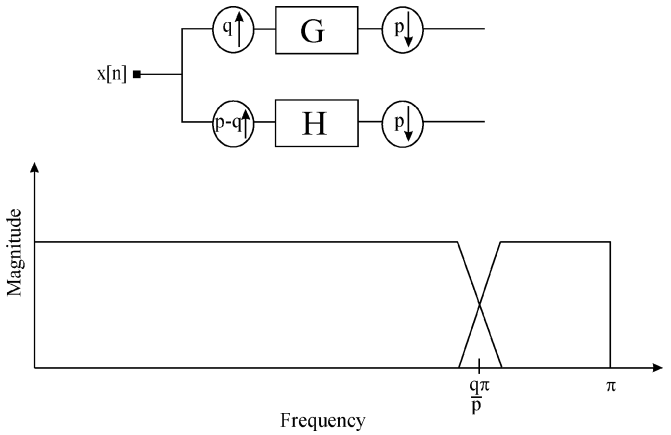

Fig. 2. Analysis channel of a rational FB of sampling factor p/q along with the corresponding frequency partitioning.

III. Filter Design

In this section, we describe two filter design methods implemented in this research. The first filter design method is referred to as **filter matching** and is implemented only for the dyadic case. The second method is referred to as **attenuation minimization** and is proposed by Blu [12] for the design of rational filter banks. We also implement it for the dyadic configuration, which is trivial in this case.

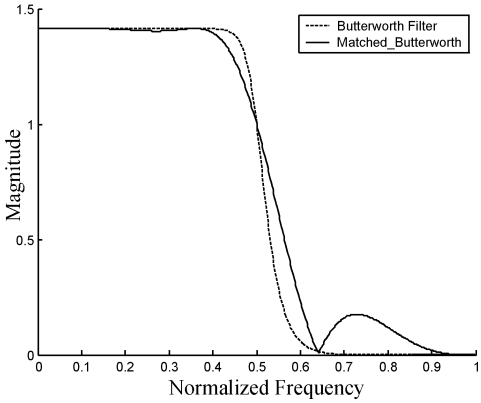

Fig. 3. Low-pass filter designed to match the order 10 Butterworth filter.

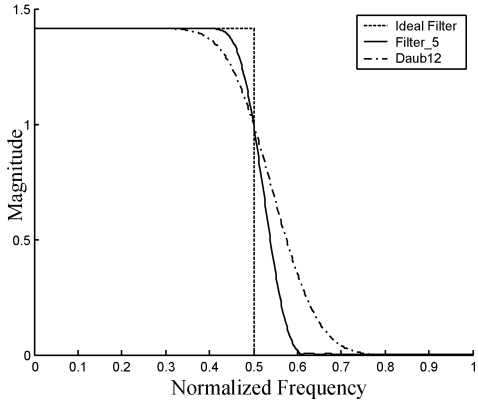

Fig. 4. Low-pass filter designed using attenuation minimization with the corresponding ideal filter it matches to and the Daub12 filter.

A. Filter Matching

The Filter Matching method refers to minimizing the difference in modulus between the designed and desired filter given some constraint. The minimization is formulated in the frequency domain. We are only concerned with orthonormal FBs where the analysis and synthesis systems can be modeled as paraunitary matrices. As described in Appendix I, a paraunitary matrix can be factored into blocks of delays and rotation matrices that are a function of $\underline{\theta}$. If we denote the desired filter by $H_d(\omega)$, and the paraunitary analysis filter by $H_p(\omega;\underline{\theta})$, the problem becomes that of minimizing

$$C(\underline{\theta},\lambda)=\int_{-\infty}^{\infty}||H_d(\omega)|-\lambda|H_p(\omega;\underline{\theta})||^2 d\omega \tag{2}$$

given the following constraint:

$$\frac{d^l H_0(\omega)}{d\omega^l}|_{\omega=\pi}=0 \quad l=0\ldots N-1 \tag{3}$$

where $H_0(\omega)$ is the frequency response of the analysis low-pass filter $h_0[n]$, and N is the number of desired zeros at π for $H_0(\omega)$ which is also the number of vanishing moments of the wavelet function [11]. Orthogonality of the FB is constrained by the lattice structure. The algorithm implementation is based on a constrained nonlinear optimization.

Fig. 3 shows a 30-tap filter designed to match the Butterworth filter given the constraints of orthogonality and 3 zeros at π.

B. Attenuation Minimization

1) Design of the Low-Pass Filter: The motivation behind the attenutation minimization algorithm is to find the best frequency-selective low-pass filter $G(z)$ given constraints of orthogonality and regularity of the FB. The degree of selectivity is defined as the difference in modulus between $G(z)$ and the ideal low-pass filter. It can be shown that by minimizing the attenuation band of $G(z)$, we minimize the difference between $G(z)$ and the ideal filter [12]. The problem is thus reduced to attenuation minimization, and can be formulated using the Lagrange multiplier method where we minimize

$$J(G)=\text{function}(G)-\lambda(\text{constraints}). \tag{4}$$

In our case, function $J(G)$ is the attenuation in the stopband, and the constraints are those of orthonormality and regularity of the FBs. In order to ensure perfect reconstruction of the FB, Blu devised a recursive implementation of the algorithm where the condition for convergence is minimal perfect reconstruction error [12]. By doing so, the algorithm itself focuses on the attenuation while the iterations minimize the reconstruction error.

2) Design of the High-Pass Filter: If the difference between the upsampling and downsampling factors is 1, as in our case, then there is a unique high-pass filter corresponding to the designed low-pass filter [12]. Referring to $\mathbf{G}$ as the polyphase matrix of $G(z)$ of size $(M-1)\times M$, we know that it is paraunitary. Furthermore, the rational FB is orthonormal and can also be represented by a paraunitary matrix $\mathbf{\Gamma_p}=[\mathbf{G}\ \mathbf{H}]^T$ where $\mathbf{H}$ is the polyphase representation of $H(z)$ and is of size $1\times M$. As shown in Appendix I, both $\mathbf{\Gamma_p}$ and $\mathbf{G}$ can be factored into Householder matrices

$$\mathbf{A_0}\prod_{i=1}^{M}\mathbf{V_i}=\mathbf{A_0}\prod_{i=1}^{M}(\mathbf{I}-(1-z^{-1})\mathbf{v_i}\mathbf{v_i^T}).$$

After obtaining $\mathbf{G}$, we factor it to get a rectangular constant matrix $\mathbf{A_0}$ of size $(M-1)\times M$. Next we complete $\mathbf{A_0}$ so that it becomes a square orthonormal matrix by adding a single row to it, in this case. Hence, $\mathbf{\Gamma_P}$ can be written as

$$\mathbf{\Gamma_p}=[\mathbf{A_0}\ \mathbf{H_{row}}]^T\cdot(\mathbf{I}-(1-z^{-1})\mathbf{v_1}\mathbf{v_1^T})\ldots(\mathbf{I}-(1-z^{-1})\mathbf{v_M}\mathbf{v_M^T}). \tag{5}$$

We can then compute $\mathbf{H}$ as

$$\mathbf{H}=\mathbf{H_{row}}\cdot(\mathbf{I}-(1-z^{-1})\mathbf{v_1}\mathbf{v_1^T})\ldots(\mathbf{I}-(1-z^{-1})\mathbf{v_M}\mathbf{v_M^T}). \tag{6}$$

We use this method to design dyadic as well as rational filter banks as illustrated in Figs. 4 and 5, respectively. Fig. 4 shows the magnitude response of a 30-tap low-pass filter designed by **attenuation minimization** and implemented in a dyadic FB. We refer to it here as DyadicAM30. For comparison we include the ideal filter and the low-pass filter corresponding to the Daubechies wavelet of order 12, Daub12. Unlike Daub12, the designed filter exhibits a very good attenuation in the stopband.

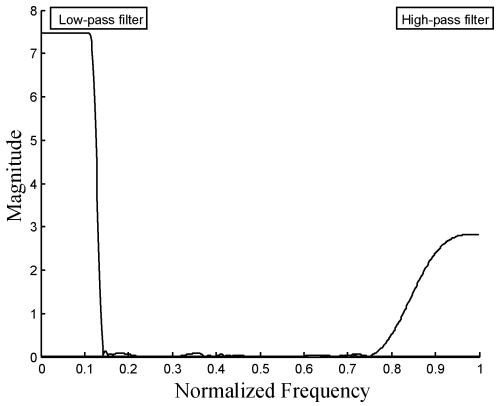

Fig. 5. Low-pass and high-pass filters corresponding to the rational FB of sampling factor 8/7.

TABLE I
DESCRIPTION OF THE STANDARD TRAIN, DEVELOPMENT, CORE TEST, AND FULL TEST DATA SETS OF THE TIMIT CORPUS

Set	# Speakers	# Utterances	# Hours
Train	462	3696	3.14
Development	50	400	0.34
Core Test	24	192	0.16
Full Test	118	944	0.81

Fig. 5 illustrates the designed low-pass and high-pass filters for a rational FB of sampling ratio 8/7.

IV. EXPERIMENTAL SETUP

To evaluate the wavelet-based framework and our proposed extensions, we set up phonetic classification experiments on the TIMIT corpus. We use a segment-based classifier and compare against a baseline that uses MFCCs. The following is a description of the setup.

A. TIMIT Corpus

TIMIT is a corpus of continuous read speech from 630 speakers, 438 males, and 192 females, representing eight major dialect groups of American English [10]. There are 61 phone labels used in the TIMIT phonetic transcriptions. Following common practice, the 61 phone labels are collapsed into 39 labels prior to scoring, and the glottal stops are ignored [23]. The data sets we use in the classification experiments are the standard Train, Development, Core Test, and Full Test sets, which omit all as dialect sentences. The 462-speaker Train set is used for training in all the experiments, the 50-speaker Development set for classification as well as significance scoring, the 24-speaker Core Test set for classification, and the 118-speaker Full Test set primarily for significance scoring. The sets are described in Table I. There is no overlap in speakers between any of Train, Development, and Full Test sets, and the sentences in the training set are different from those in the development and test sets. The Core Test set, on the other hand, is a subset of the Full Test set. It was designed to contain one female and two males from each of the eight different dialect regions, and has proven to be a challenging test set.

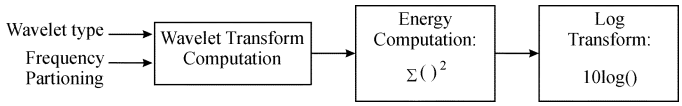

Fig. 6. Flowchart of the computational stages for the wavelet-based acoustic measurement.

B. Classifier

The classification experiments are performed using a segment-based classifier [24]. Each segment is represented by a fixed-size measurement vector. Since we are only dealing with phonetic classification, we obtain the segments from the phonetic transcriptions. In all the experiments, normalization and principal component analysis (PCA) are performed on the acoustic observations to whiten the feature space. The measurements are then modeled using diagonal Gaussian mixture models (GMMs). Maximum likelihood training is used to estimate the parameters of the Gaussian models. Classification is implemented using Maximum *a posteriori* decisions, i.e., phone priors are used.

C. Baseline

The speech waveform is first preemphasized by a factor of 0.97 prior to any processing. Next a Hamming window is applied to obtain speech frames and the 256-point short-time Fourier transform (STFT) is computed for the 25.6-ms frames at a rate of 5 ms. Forty MFSCs are computed, and 14 MFCCs are then obtained per frame [1]. A 76-dimensional observation vector is extracted for each segment in the TIMIT phonetic transcriptions. The segmental measurement is a concatenation of 3 MFCC and energy averages computed over the segment in a 3–4–3 proportion, 2 MFCC and energy derivatives computed using linear least-squared error regression over a time frame of 40 ms centered at the start and end of the segment, and a log duration [13].

Diagonal GMMs are used to model the acoustic measurements with a minimum of 61 datapoints per mixture component and a maximum of 96 mixture models per phone. With this baseline configuration, we obtain a classification error of approximately 23.9% on the Development set, 24.6% on the Core Test set, and 24.4% on the Full Test set.

D. Wavelet-Based Acoustic Measurement

To evaluate the suggested extensions to the wavelet and FB implementations, we propose an energy-based acoustic measurement. Fig. 6 illustrates the stages involved in its computation.

Stage 1) Compute the wavelet transform of the input speech frame. In all the experiments related to the wavelet-based acoustic measurement, the frame rate is 200 frames per second (5 ms per frame) with a frame size of 20 ms. In this stage, we specify the wavelet type and the frequency decomposition, whether we are using wavelet packets or rational sampling.

Stage 2) Compute the energy of each frequency band resulting in N coefficients where N is the number of frequency bands analyzed in Stage 1.

Stage 3) Compute the log of the energy coefficients.

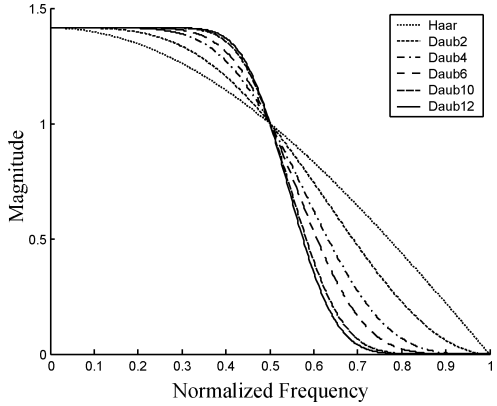

Fig. 7. Low-pass filters corresponding to the Haar, Daub2, Daub4, Daub6, Daub10, and Daub12 wavelets.

TABLE II
DESCRIPTION OF THE IMPLEMENTED DAUBECHIES WAVELETS AND THE CORRESPONDING ERROR RATE ON THE DEVELOPMENT SET. THE RESULTS ARE FOR THE 26-BAND TREE-STRUCTURED FB IMPLEMENTATION

Wavelet type	# Zeros at $\omega = \pi$	Filters length	% Error Rate
Daub4	4	8	25.4
Daub6	6	12	24.7
Daub10	10	20	24.3
Daub12	12	24	23.7

The frame sizes used for computing the wavelet-based and MFCC measurements are set to 20 and 25.6 ms, respectively. The values were optimized for each configuration. The resulting N-dimensional measurement is used to generate a segmental measurement of dimension $(5N + 6)$, which is extracted over given phonetic segments similarly to the acoustic measurement described in Section IV-C. N can range between 18 and 30 coefficients and the dimensionality of the segmental vector lies between 96 and 156. PCA is used to project the feature space onto 76 dimensions as well as whiten it for improved Gaussian mixture modeling.

V. RESULTS AND DISCUSSION

A. Daubechies Wavelets

First, we test the acoustic measurement using Daubechies wavelets. Tree-structured FBs are used to obtain the frequency partitions. Table II lists the implemented wavelets with a brief description and the corresponding error rates on the Development set. A wavelet of order n will have a corresponding low-pass filter with n zeros at π. From Fig. 7, we notice that the larger the number of zeros at π, the narrower the transition region and the sharper the filter cutoff. This also leads to a lower error rate as illustrated in the fourth column of Table II. Intuitively, sharp cutoff is a desired characteristic of filters since it implies good frequency selectivity. We adopt the 26-band tree described in Table III in the subsequent dyadic implementations and evaluations. This FB is actually reminiscent of the one proposed in [4], although the frequency bands are not identical. The frequency bands used here are selected to closely emulate the critical-band effect where eight filters divide the 0–1-kHz region into equal bands, and the rest of the filters approximate logarithmic partitions.

TABLE III
FREQUENCY BANDS CORRESPONDING TO THE 26-BAND FB

Filter #	Lower cutoff frequency (Hz)	Upper cutoff frequency (Hz)	Bandwidth (Hz)
1	0	125	125
2	125	250	125
3	250	375	125
4	375	500	125
5	500	625	125
6	625	750	125
7	750	875	125
8	875	1000	125
9	1000	1250	250
10	1250	1500	250
11	1500	1750	250
12	1750	2000	250
13	2000	2250	250
14	2250	2500	250
15	2500	2750	250
16	2750	3000	250
17	3000	3250	250
18	3250	3500	250
19	3500	3750	250
20	3750	4000	250
21	4000	4500	500
22	4500	5000	500
23	5000	5500	500
24	5500	6000	500
25	6000	7000	1000
26	7000	8000	1000

B. Designed Filters

1) Filter Matching: We test the filter matching algorithm by matching it to two desired signals: 1) the Butterworth filter of order 10 and cutoff frequency $\pi/2$ and 2) the ideal low-pass filter. The resulting filters are denoted DyadicFM_Butterworth and DyadicFM_Ideal, respectively, where FM stands for Filter Matching. They both have three zeros at π and are 30-tap filters. The designed filters are tested with the 26-band tree-structured FB, and their corresponding error rate on the Development set is 24.1% and 23.5%, respectively.

2) Attenuation Minimization: For the dyadic case, six filters denoted DyadicAM{tap #} are designed, where AM stands for attenuation minimization. The filters are 10, 16, 20, 26, 30, and 34-tap filters, respectively, and they all have a regularity order set to 1. We also design the rational FBs listed in Table IV. We refer to the rational filters as RationalAM{*sampling factor*}. The regularity order is also set to 1 for all of them. The rational FBs are iterated on the low-pass channel N times to generate N bands. We iterate until the lower cutoff of the last band-pass filter is at or close to 1 kHz. We then used DyadicAM30 designed in Section III-B to divide the 0–1-kHz region into eight equipartitions. This is done to obtain a frequency partition that models the critical-band spectral resolution. The fifth column in Table IV gives the overall number of filters in the FB. The length of the filters is large which is necessary to obtain narrow passbands and also good frequency selectivity as is the case here. Fig. 8 shows the error rates, which vary between 23.2% and 24.7%, on the Development set for the different filters designed using attenuation minimization.

TABLE IV
DESCRIPTION OF THE DESIGNED FILTERS FOR THE RATIONAL FB

Filter name	Regularity order	Low-pass filter length	High-pass filter length	# Filters
RationalAM6/5	1	194	44	18
RationalAM7/6	1	226	43	20
RationalAM8/7	1	226	41	22
RationalAM10/9	1	191	32	26

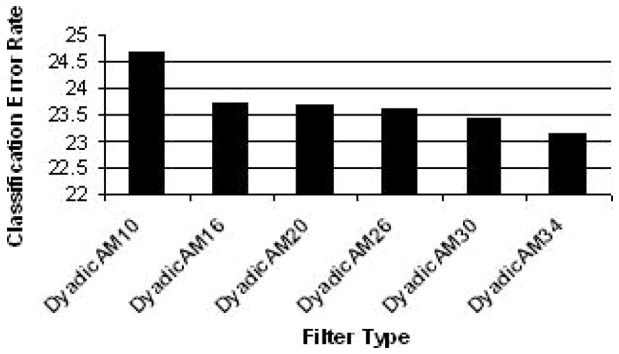

Fig. 8. Error rates on the Development set for the six filters designed using the attenuation minimization technique.

TABLE V
CLASSIFICATION PERFORMANCE (OVERALL AND PHONETIC SUBCLASSES) OF THE FIVE SELECTED ACOUSTIC MEASUREMENTS AND THE BASELINE (MFCC) ON THE DEVELOPMENT SET. VOW STANDS FOR VOWEL, NAS FOR NASAL, STP FOR STOP, WFR FOR WEAK FRICATIVE, SFR FOR STRONG FRICATIVE, AND CL FOR CLOSURE

Acoustic Measurement	(%) 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
Daub12	23.6	30.4	26.5	28.9	28.1	21.2	4.2
DyadicFM_Ideal	23.4	31.5	26.5	28.9	25.4	23.8	3.7
DyadicAM30	23.4	30.7	23.5	28.7	26.9	21.2	4.3
DyadicAM34	23.1	30.4	23.4	28.7	27.6	21.0	3.6
RationalAM8/7	23.2	30.5	25.5	26.4	27.7	22.7	3.3

C. Overview and Discussion of Classification Performance

For further evaluation of the results, we select five acoustic measurements listed in the first column of Table V which also gives the overall classification results as well as the error rates for the phonetic subclasses on the Development set. The results for the baseline classifier using MFCCs are included for comparison. Table VI also shows the evaluation of the five acoustic measurements on the Core Test and Full Test sets. Significance scoring is reported for the Development and the Full Test sets but not for the Core Test set since it has a small size. The McNemar significance test is used [25]. We make the following observations based on Tables V and VI. First, the overall classification error rates corresponding to all the acoustic measurements match or exceed that of MFCCs on the Development set. However, there is a difference in performance over the phonetic subclasses. For example, all the wavelet-based measurement outperform or roughly match the MFCCs in the vowel, weak fricative, and closure categories, but only RationalAM8/7 also shows improvement for the stops. These observations do not match the results reported by Farooq and Datta in [4]. It is worth noting that, as mentioned in Section V-A, the 26-band tree-structure is not identical to the one proposed in [4]. More important, the training and test sets for the two setups are different. One possible reason for our results might be the smoothness of all the implemented wavelets which renders them capable of capturing the harmonics of the vowels as well as the abrupt changes in the consonants. We also recall that stops are considered the most dynamic types of sounds. This could explain why RationalAM8/7, which exhibits a better overall frequency resolution, has the best result for the stop consonants. DyadicAM30 and DyadicAM34 show improvement over the MFCC for the nasals where most of the spectral information is concentrated below 1 kHz. This might explain why the high-frequency resolution introduced by RationalAM8/7 does not give any improvement. Furthermore, it does seem that the low-frequency selectivity provided by DyadicAM30 and DyadicAM34 improves the error rate of the nasal consonants. At this point, we do not claim to have a full intuition of the performance of the wavelet-based measurements over the different phonetic subclasses. We hope that further experiments will help shed the light on some of the results obtained. The results listed in Table V are also reminiscent of those obtained by Halberstadt [13], [26]. The difference in results over the phonetic subclasses suggests the possibility of implementing a hierarchical architecture where filters optimized to the different subclasses are designed. Another observation is that the Daub12 wavelet, performs the worst on the Core Test and Full Test sets, whereas the acoustic measurement corresponding to RationalAM8/7, consistently outperforms the rest of the measurements and the baseline. The difference in results over the baseline is also significant.

TABLE VI
CLASSIFICATION PERFORMANCE OF FIVE SELECTED MEASUREMENTS AND THE BASELINE (MFCC) ON THE CORE TEST AND FULL TEST SETS. MCNEMAR SIGNIFICANCE SCORES FOR THE DEVELOPMENT AND FULL TESTS ARE ALSO LISTED. (Y) OR (N) INDICATES WHETHER THE IMPROVEMENT IS STATISTICALLY SIGNIFICANT AT THE 0.05 LEVEL

Acoustic Measurement	(%) Error rate		McNemar significance	
	Core Test	Full Test	Dev	Full Test
MFCC	24.6	24.4	-	-
Daub12	25.9	24.7	0.39 (N)	-
DyadicFM_Ideal	25.2	24.5	0.15 (N)	-
DyadicAM30	24.9	24.3	0.15 (N)	0.56 (N)
DyadicAM34	24.6	24.1	0.019 (Y)	0.137 (N)
RationalAM8/7	24.0	23.8	0.045 (Y)	0.011 (Y)

Our best results compare favorably to those mentioned in the literature as well as those of the baseline classifier. Though the results mentioned here are for context-independent phonetic classification, they should not be used for direct comparison since the training and test conditions differ from one another.

The best reported result for context-independent phonetic classification is by Halberstadt [26]. He successfully experimented with heterogeneous measurements and multiple classifiers and obtained an error rate of 18.3% on the Core Test set. One of the issues addressed in Halberstadt's thesis is that of aggregating several acoustic models in order to boost the performance and robustness of the models [27]. To get an idea of the extent of expected improvement upon implementing aggregation on our acoustic measurements, four-fold aggregation was tested on the acoustic measurement corresponding to RationalAM8/7. Error rates of 21.8% on the Development

set, and 22.9% on the Core Test set are obtained. These results compare very well with the performance of the fourfold aggregated models tested for various segmental measurements in [27]. The error rates Halberstadt reported on the Development set ranged between 21.4% and 22.7%.

Other results are reported by Clarkson and Moreno who implemented support vector machines (SVMs), with various kernel functions, applied to phonetic classification. They obtained error rates that range between 22.9% and 23.7% on the Core Test set [28]. Chigier *et al.* experimented with several signal representations and reported 22.0% using perceptual linear predictive (PLP)-based measurements and a neural net classifier [29]. Chengalvarayan and Deng developed a new hidden Markov model that integrates generalized dynamic feature parameters into the model structure. The best result they reported is an error rate of 31.8% on a 20-speaker test set [30]. Zahorian *et al.* obtained 23.0% on the Core Test set using spectral/temporal features and binary-pair partitioned neural network classifier [31]. Recently, Gunawardana *et al.* implemented hidden conditional random fields for phone classification [32]. They reported a classification error rate of 21.7% on the Core Test set. Sha and Saul used large Gaussian mixture modeling for both phonetic classification and recognition [33]. Their best classification result on the Core Test set was 21.1%. It is interesting to note that this modeling method could be applied to any feature vector.

We believe our results are encouraging and portend further improvement upon the investigation of different wavelet-based acoustic measurements, the combination of several generated measurements, in addition to acoustic model aggregation.

VI. Conclusion and Future Work

We have presented a wavelet and FB framework for phonetic classification in which we have exploited two dimensions of the wavelet and FB theory: filter design and rational sampling. We have shown that off-the-shelf wavelets, particularly the Daubechies wavelets, do not always give the best results, and there is a need for wavelet design. We have also shown that a dyadic FB implementation is not optimal, and we have examined a method for rational FB 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 proved to be efficient in denoising tasks [34].

The framework is also limited to the task of phonetic classification. A natural extension would be phonetic recognition taking into account linguistic context-dependency such as coarticulation.

Given the results that we obtained, we feel that there is room for further experiments. In our implementation of this framework, we are mostly interested in the effect of filter design and rational FBs. However, given the trend in Table II, it would be interesting to see the effect of increasing the length of the Daubechies filters on the phonetic error rate. Furthermore, the 26-band tree-structured FB is one possible approach we adopt to mimic the critical bands and other designs can be proposed and implemented. It would also be worthwhile to implement different tree structures that are optimized for the different phonetic subclasses.

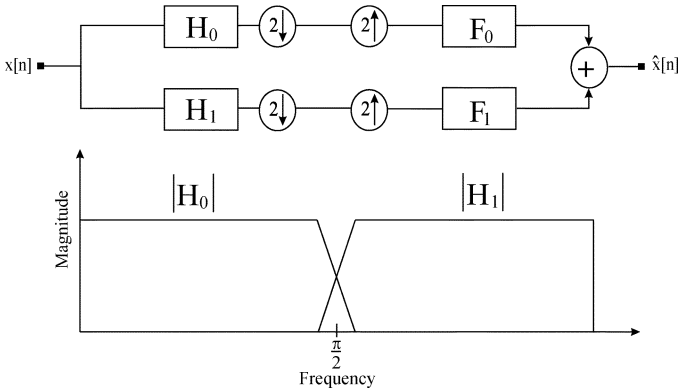

Fig. 9. Two-channel dyadic FB and the corresponding frequency spectrum partitioning. The analysis section consists of filtering followed by downsampling by 2 while the synthesis consists of upsampling by 2 followed by filtering.

A final, yet crucial point is that the current acoustic observation is a very simple energy-based measurement that takes little advantage of the multiresolution analysis provided by the wavelets and FBs. Hence, there is a need to design a different wavelet-based measurement that makes better use of the flexibility of the proposed framework.

Appendix I
Wavelets and Filter Banks

Filter Banks: Filter banks can be efficiently implemented using discrete finite-impulse response (FIR) filters, downsamplers, and upsamplers. An important requirement on the FB is perfect reconstruction, meaning that the input signal processed by some set of filters at one end of the channel should be perfectly reconstructed by another set of filters at the other end. Such filters are referred to as analysis and synthesis filters, respectively. Perfect reconstruction FBs can be used to implement series expansions of discrete-time signals in the $l_2(\mathcal{Z})$ space. Fig. 9 illustrates a perfect-reconstruction FB and the corresponding frequency bands for the two-channel case. In this case, $H_0(z)$ is the analysis low-pass, $H_1(z)$ is the analysis high-pass, $F_0(z)$ is the synthesis low-pass, and $F_1(z)$ is the synthesis high-pass. For perfect reconstruction, we would like the output to be at most a delayed version of the input, which, in the $\mathcal{Z}$-domain, is given by

$$\hat{X}(z) = z^{-L} X(z). \tag{7}$$

The output $\hat{X}(z)$ can be written as

$$\hat{X}(z) = \underbrace{\frac{1}{2}[H_0(z)F_0(z) + H_1(z)F_1(z)]}_{\text{Amplitude Distortion}} X(z) + \underbrace{\frac{1}{2}[H_0(-z)F_0(z) + H_1(-z)F_1(z)]}_{\text{Aliasing}} X(-z). \tag{8}$$

For (7) and (8) to be equal, we require that the Amplitude Distortion component be equal to a constant and the Aliasing component be equal to zero. In matrix notation, this is written as

$$\mathbf{H_m}(z)\mathbf{F_m}(z) = 2\mathbf{I} \tag{9}$$

where

$$\mathbf{H_m}(z) = \begin{bmatrix} H_0(z) & H_0(-z) \\ H_1(z) & H_1(-z) \end{bmatrix} \tag{10}$$

and

$$\mathbf{F_m}(z) = \begin{bmatrix} F_0(z) & F_1(z) \\ F_0(-z) & F_1(-z) \end{bmatrix}. \tag{11}$$

This is known as the perfect reconstruction criterion in the modulation domain, where $\mathbf{H_m}(z)$ is the **analysis modulation matrix** and $\mathbf{F_m}(z)$ the **synthesis modulation matrix**. An FB that satisfies (9) is referred to as *biorthogonal*. A perfect reconstruction FB is *orthonormal* if it satisfies

$$\mathbf{H_m}(z)\mathbf{H_m}^T(z^{-1}) = 2\mathbf{I}. \tag{12}$$

Based on (12), $\mathbf{H_m}(z)$ is referred to as a paraunitary matrix [16]. In this paper, we work only with orthonormal FBs. We implement the FBs in the polyphase rather than the modulation domain since it is more computationally efficient as illustrated in Fig. 10 [17]. In the polyphase domain, the condition for perfect reconstruction is

$$\mathbf{H_p}(z)\mathbf{F_p}(z) = \mathbf{I}. \tag{13}$$

Furthermore, the condition for orthonormality is

$$\mathbf{H_p}(z)\mathbf{H_p}^T(z^{-1}) = \mathbf{I} \tag{14}$$

where the **analysis polyphase matrix** is denoted

$$\mathbf{H_p}(z) = \begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H_{10}(z) & H_{11}(z) \end{bmatrix} \tag{15}$$

and the **synthesis polyphase matrix** is denoted

$$\mathbf{F_p}(z) = \begin{bmatrix} F_{00}(z) & F_{10}(z) \\ F_{01}(z) & F_{11}(z) \end{bmatrix}. \tag{16}$$

Note that $H_{ij}(z)$ is the jth polyphase component of the ith filter such that

$$H_i(z) = H_{i0}(z^2) + zH_{i1}(z^2). \tag{17}$$

In this paper, we are concerned with designing FBs characterized by paraunitary polyphase matrices. There are several methods available for factoring paraunitary matrices into smaller building blocks that are easy to manipulate. We proceed to describe the two methods that are implemented in this paper.

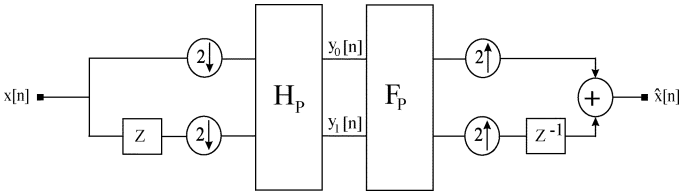

Fig. 10. Polyphase implementation of the two-channel FB. The upsampling and downsampling by 2 have been moved before and after the analysis and synthesis filters, respectively.

Lattice Factorization: A paraunitary matrix can be factored into building blocks consisting of delays and rotation matrices G_i [16]

$$\mathbf{G_i} = \begin{bmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{bmatrix}. \tag{18}$$

The lattice factorization of the paraunitary analysis filter $\mathbf{H_p}(z)$ can thus be written as

$$\mathbf{H_p}(z;\underline{\theta}) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \prod_{i=0}^{l-1} \left(\mathbf{G_i} \begin{bmatrix} 1 & 0 \\ 0 & z^{-1} \end{bmatrix} \right) \mathbf{G_l} \tag{19}$$

where $\underline{\theta} = [\theta_0, \theta_1, \ldots, \theta_l]$.

With l delay blocks and $l+1$ rotations matrices, the degree of every polyphase component of $\mathbf{H_p}(z)$ is l and, hence, that of $H_0(z)$ and $H_1(z)$ is $2l+1$. Such a structure imposes orthogonality on the filter bank [16]. One can now solve for $\underline{\theta}$ given some desired constraints, such as matching the frequency response of a filter and imposing a certain number of zeros at π on $H_0(z)$. The matrix $H_p(z;\underline{\theta})$ remains paraunitary for any value of $\underline{\theta}$. However, to impose at least one zero at π on $H_0(z)$, the following criteria should be satisfied [17]:

$$\sum_{i=0}^{l} \theta_i = \frac{\pi}{4}. \tag{20}$$

Householder Factorization: Another way of factoring a paraunitary matrix is using Householder matrices as the building blocks. The paraunitary analysis filter $\mathbf{H_p}(z)$ can then be written as

$$\mathbf{H_p}(z) = \mathbf{A_0} \prod_{i=1}^{M} \mathbf{V_i} \tag{21}$$

where $\mathbf{V_i}$, the Householder matrix is

$$\mathbf{V_i} = (\mathbf{I} - (1 - z^{-1}\mathbf{v_i}\mathbf{v_i^T})) \tag{22}$$

$\mathbf{v_i}$, $i = 1, \ldots, M$ are unitary vectors, and $\mathbf{A_0}$ is a constant unitary matrix [16].

Tree-Structured FBs: The FB that we have seen so far is a two-channel one. If one iterates on the low-pass channel, as shown in Fig. 11, we obtain a constant-Q octave band. In this simple case, the FB is said to have a dyadic structure meaning that at every iteration, the spectrum is split in half. The idea can

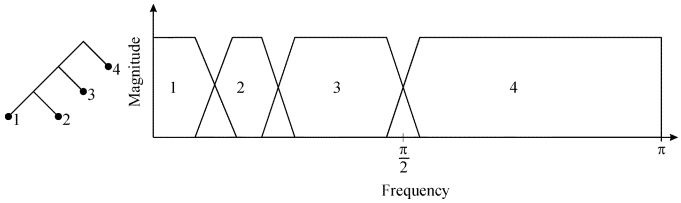

Fig. 11. FB iterated on the low-pass channel and the corresponding frequency partitioning.

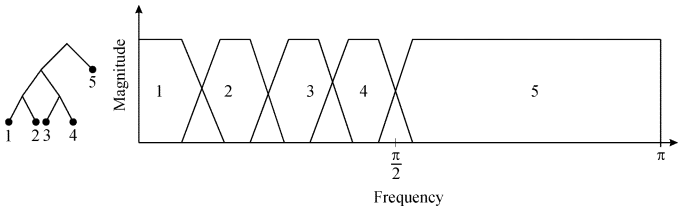

Fig. 12. Tree-structured implementation of a FB and the corresponding frequency partitioning. FB iterations occur at the high-pass as well as the low-pass channel.

be extended to arbitrary tree-structured FBs by also allowing iteration on the high-pass channel as illustrated in Fig. 12. Such structures are used to implement *wavelet packets*.

APPENDIX II
WAVELETS AND THE MULTIRESOLUTION FRAMEWORK

Wavelet Function: Suppose that $l_2(\mathcal{Z})$ is constructed as a collection of spaces W_m, which have the property of being spanned by functions of the form

$$\psi_{m,n}(t) = 2^{-m/2}\psi(2^{-m}t - n), \qquad m, n \in \mathcal{Z}. \tag{23}$$

These functions are known as wavelet functions and they form an orthonormal basis for the W_m spaces [11], [17]. They also satisfy a property, known as the two-scale or dilation equation

$$\psi(t) = \sqrt{2}\sum_n h_1[n]\varphi(2t - n) \tag{24}$$

which relates the wavelet function to another function known as the scaling function and is denoted by $\varphi_{m,n}(t)$. It can be shown that the set of scaling functions $\varphi_{m,n}(t)$ span the nested and complete spaces V_m such that [11]

$$\ldots V_1 \subset V_0 \subset V_{-1} \ldots. \tag{25}$$

The scaling functions are of the form

$$\varphi_{m,n}(t) = 2^{-m/2}\varphi(2^{-m}t - n) \quad m, n \in \mathcal{Z}. \tag{26}$$

They also satisfy a two-scale equation

$$\varphi(t) = \sqrt{2}\sum_n h_0[n]\varphi(2t - n). \tag{27}$$

The idea of a multiresolution framework manifests itself in (27) which relates the basis functions at one scale to those at a higher scale. Furthermore, (24) indicates that there is a relation between the wavelet and scaling functions and, hence, between the W_m and V_m spaces. It has been shown that W_m is the difference between V_{m-1} and V_m [11]. Fig. 13 illustrates the idea for $m = 0$ and shows that any signal belonging to V_{-1} can be represented in terms of a basis which spans a lower-resolution space V_0 plus some "detail" left out in W_0. This "detail" is, in turn, represented in terms of a basis which spans W_0. Hence, the original signal has been split into a coarse approximation and some detail. This representation can be easily extended to include more scales. Furthermore, by taking the Fourier transform of (24) and (27), we get

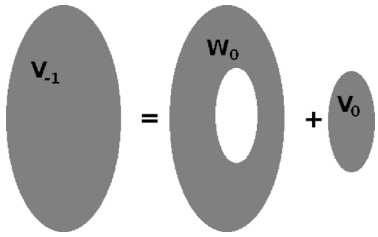

Fig. 13. Relation between space V_{-1} and spaces W_0 and V_0 where $V_{-1} = W_0 + V_0$.

$$\Phi(\omega) = \frac{1}{\sqrt{2}}H_0(e^{j\omega/2})\Phi\left(\frac{\omega}{2}\right)$$
$$\Psi(\omega) = \frac{1}{\sqrt{2}}H_1(e^{j\omega/2})\Phi\left(\frac{\omega}{2}\right) \tag{28}$$

where

$$H_0(e^{j\omega}) = \sum_{n\in\mathcal{Z}} h_0[n]e^{-j\omega n}$$
$$H_1(e^{j\omega}) = \sum_{n\in\mathcal{Z}} h_1[n]e^{-j\omega n}. \tag{29}$$

It can be shown that $h_0[n]$ and $h_1[n]$ are the low-pass and high-pass filters, respectively, of a two-channel FB, and the iterations of (28) converge to piecewise smooth scaling and wavelet functions if the corresponding filters satisfy certain conditions such as *regularity* [11]. At this point, it suffices to know that for a filter to be regular, it is necessary, but not sufficient, for it to have at least one zero at the aliasing frequency — π for the dyadic two-channel case. Furthermore, the wavelet transform can be efficiently implemented through an iteration of FBs.

REFERENCES

[1] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, pp. 357–366, 1980.

[2] I. Daubechies and S. Maes, "A non-linear squeezing of the continuous wavelet transform based on auditory nerve models," *Wavelets Med. Biol.*, pp. 527–546, 1996.

[3] I. Cohen, S. Raz, and D. Malah, "Shift invariant wavelet packet bases," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Detroit, MI, 1995, pp. 1081–1084.

[4] O. Farooq and S. Datta., "Mel filter-like admissible wavelet packet structure for speech recognition," *IEEE Signal Process. Lett.*, vol. 8, no. 7, pp. 196–198, Jul. 2001.

[5] Y. Hao and X. Zhu, "A new feature in speech recognition based on wavelet transform," in *Proc. Int. Conf. Spoken Lang. Process.*, Beijing, China, 2000, pp. 1526–1529.
[6] Y. Kaisheng and C. Zhigang, "A wavelet filter optimization algorithm for speech recognition," in *Proc. Int. Conf. Commun. Technol.*, Beijing, China, 1998, pp. 1–5.
[7] H. Wassner and G. Chollet, "New time-frequency derived cepstral coefficients for automatic speech recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, Philadelphia, PA, 1996, pp. 260–263.
[8] K. Kim, D. Youn, and C. Lee, "Evaluation of wavelet filters for speech recognition," in *Proc. IEEE Int. Conf. Syst. Man, Cybern.*, Nashville, TN, 2000, pp. 2891–2894.
[9] B. Tan, F. Minyue, A. Spray, and P. Dermody, "The use of wavelet transforms in phoneme recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, Philadelphia, PA, 1996, pp. 2431–2434.
[10] L. Lamel, R. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Proc. DARPA Speech Recognition Workshop*, 1986, pp. 100–109, Rep. SAIC-86/1546.
[11] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*. Englewood Cliffs, NJ: Prentice-Hall, 1995.
[12] T. Blu, "A new design algorithm for two-band orthonormal rational filter banks and orthonormal rational wavelets," *IEEE Trans. Signal Process.*, vol. 46, no. 6, pp. 1494–1504, Jun. 1998.
[13] A. Halberstadt, "Heterogeneous acoustic measurements and multiple classifiers for speech recognition," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Mass. Inst. Technol., Cambridge, MA, 1998.
[14] L. Lamel and J. Gauvian, "High-performance speaker-independent phone recognition using CDHMM," in *Proc. Eur. Conf. Speech Commun. Technol.*, 1993, pp. 121–124.
[15] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM, 1992.
[16] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[17] G. Strang and T. Nguyen, *Wavelets and Filter Banks*. Wellesley, MA: Wellesley-Cambridge Press, 1996.
[18] G. F. Choueiter, "A Wavelet and Filter Bank Framework for Phonetic Classification," M.S. thesis, Dept. Civil Environmental Eng., Mass. Inst. Technol., Cambridge, MA, 2004.
[19] J. Kovacevic and M. Vetterli, "Perfect reconstruction filter banks with rational sampling factors," *IEEE Trans. Signal Process.*, vol. 41, no. 6, pp. 2047–2066, Jun. 1993.
[20] P. Hoang and P. Vaidyanathan, "Non-uniform multirate filter banks: Theory and design," in *Proc. IEEE Int. Symp. Circuits Syst.*, Portland, OR, 1989, pp. 371–374.
[21] T. Blu, "Bancs de Filtres Iteres en Fraction d'Octave, Application au Codage de Son," (in French) Ph.D. dissertation, Ecole National Superieur des Telecommunications, Paris, France, 1996.
[22] ——, "Iterated filter banks with rational rate changes connection with discrete wavelet transforms," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3232–3244, Dec. 1993.
[23] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.
[24] J. Glass, J. Chang, and M. McCandless, "A probabilistic framework for feature-based speech recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, Philadelphia, PA, 1996, pp. 2277–2280.
[25] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Glasgow, U.K., 1989, pp. 532–535.
[26] A. K. Halberstadt and J. Glass, "Heterogeneous measurements and multiple classifiers for speech recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, Sydney, Australia, 1998, pp. 995–998.
[27] T. J. Hazen and A. K. Halberstadt, "Using aggregation to improve the performance of mixture Gaussian acoustic models," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Seattle, WA, 1998, pp. 653–656.
[28] P. Clarkson and P. Moreno, "On the use of support vector machines for phonetic classification," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Phoenix, AZ, 1999, pp. 585–588.
[29] H. Leung, B. Chigier, and J. Glass, "A comparative study of signal representations and classification techniques for speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Minneapolis, MN, 1993, pp. 680–683.
[30] R. Chengalvaryan and L. Deng, "Use of generalized dynamic feature parameters for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 232–242, May 1997.
[31] S. A. Zahorian, P. L. Silsbee, and X. Wang, "Phone classification with segmental features and a binary-pair partitioned neural network classifier," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Munich, Germany, 1997, pp. 1011–1014.
[32] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden conditional random fields for phone classification," in *Proc. Eur. Conf. Speech Commun. Technol.*, Lisbon, Portugal, 2005, pp. 1117–1120.
[33] F. Sha and L. K. Saul, "Large margin gaussian mixture modeling for phonetic classification and recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, 2006, pp. 265–268.
[34] M. Gupta and A. Gilbert, "Robust speech recognition using wavelet coefficient features," in *Proc. IEEE Automatic Speech Recognition Understanding Workshop*, 2001, pp. 445–448.

Ghinwa F. Choueiter (S'98) received the B.E. degree in computer and communications engineering from the American University of Beirut, Beirut, Lebanon, in 2002 and the S.M. degree in civil and environmental engineering in the Spoken Language Systems (SLS) Group, Massachusetts Institute of Technology, Cambridge, in 2004. Her Master's thesis, entitled "A wavelet and filter bank framework for phonetic classification," was supervised by Dr. J. R. Glass. Her work involved using wavelets and filter banks as a feature extraction tool for speech recognition systems. She is currently pursuing the Ph.D. degree in electrical engineering and computer science in the SLS Group.

Her research interests include signal processing, wavelet analysis and applications in speech technology, and machine learning.

James R. Glass (SM'78) received the S.M. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 1985, and 1988, respectively.

After starting in the Speech Communication Group at the MIT Research Laboratory of Electronics, he has worked since 1989 at the Laboratory for Computer Science, now the Computer Science and Artificial Intelligence Laboratory (CSAIL). Currently, he is a Principal Research Scientist at CSAIL, where he heads the Spoken Language Systems Group. He is also a Lecturer in the Harvard-MIT Division of Health, Sciences, and Technology. His primary research interests are in the area of speech communication and human–computer interaction, centered on automatic speech recognition and spoken language understanding. He has lectured, taught courses, supervised students, and published extensively in these areas.

Dr. Glass has previously been a member of the IEEE Signal Processing Society Speech Technical Committee and an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.