Practical Design of Filter Banks for Automatic Music Transcription

Filipe C. da C. B. Diniz, Luiz W. P. Biscainho, and Sergio L. Netto
Federal University of Rio de Janeiro
PEE-COPPE & DEL-Poli, POBox 68504, Rio de Janeiro, RJ, Brazil, 21945-972
{filiped, wagner, sergioln}@lps.ufrj.br

Abstract

This paper aims at exploring practical issues concerning the design of a bounded-Q fast filter bank (BQFFB) for automatic music transcription (AMT). Great effort is spent on avoiding hazardous effects in the spectral analysis stage of the AMT application. The result is a complete procedure for effectively designing the BQFFB tool. The analysis tool is then validated through computer experiments.

1. Introduction

Music transcription consists of writing down the score of a musical piece based on its execution. To achieve an effective result, the whole procedure is commonly performed by a well-trained music expert, most probably an experienced musician. Nowadays, great effort is being spent on attempting to automate such a task, forming the research field of automatic music transcription (AMT) [6].

In very general lines, the AMT can be divided into the following stages [10]:

1. Time-frequency analysis of the music signal;
2. Note identification;
3. Note onset (beginning) and offset (end) detection;
4. Timbre recognition;
5. Evaluation of results' coherence.

The foundation of such a system is the spectral analysis stage. One way of implementing it is based on filter banks [9], which separate the signal into narrow frequency bands. For AMT systems, an attractive category of filter bank is the bounded-Q fast filter bank (BQFFB) [2]. This tool presents very selective frequency bins, with low interchannel interference, and a non-uniform frequency distribution suitable for the organization of the Western musical scale.

The present paper explores the practical issues involved in the implementation of the BQFFB in order to ensure its effectiveness in AMT systems. The text is organized in the following way: Section 2 discusses the desired characteristics of spectral analysis tools for the envisaged application. Emphasis is given to the BQFFB tool, which combines the positive aspects of high channel selectivity, efficient channel distribution in the frequency domain, and reduced implementation cost. Section 3 describes some practical issues of the BQFFB design in an AMT environment, including a channel mapping to properly represent the equal tempered scale. Finally, Section 4 includes some numerical experiments that validate the main contributions of the paper.

2. Spectral Analysis

The most basic filter bank for spectral analysis is the sliding fast Fourier transform (sFFT), which consists of an FFT applied to consecutive windows of the input signal. The sFFT requires only one complex multiplication per output sample [4]. The main drawback of this tool is its inherent low selectivity. Their filters, transformed from low-order kernel filters [8], provide an attenuation of only 13 dB between adjacent channels.

In an attempt to surpass the FFT selectivity, the so-called fast filter bank (FFB) was proposed in [8]. In the FFB, higher-order kernel filters are employed in an FFT-like tree structure. As in the frequency-response masking (FRM) [7] technique, computational cost is kept low by the use of interpolated versions of the kernel filters, with most of their coefficients null. The overall result is a moderately complex but highly selective filter bank, with an attenuation of 56 dB between adjacent channels. This allows an improved distinction between components of the music signal, leading to a better note identification.

Frequency spacing of both sFFT and FFB is constant. This characteristic makes them somewhat inefficient in an AMT context. In fact, in the equal tempered scale, widely employed in Western music [10], the spacing between the fundamental frequencies of musical notes increases geometrically. As a consequence, linearly distributed channels capable of properly discriminate between contiguous notes in the low-frequency range of the spectrum tend to be too numerous in high frequencies. Conversely, a well matched high-frequency resolution corresponds to an insufficient resolution in the low-frequency range.

In an attempt to take advantage of the equal tempered scale, Brown [1] proposed the use of geometrically spaced channels in AMT systems. The result was the so-called constant-Q transform (CQT), which presents a high computational cost, since it is not able to take advantage of the FFT tree-like algorithm. The combination of the CQT with the

FFB resulted in the constant-Q fast filter bank (CQFFB) [3].

Another form of channel distribution uses a geometric spacing between different groups of channels (which may correspond to musical octaves) and a linear distribution inside each group. This approach is referred to as the bounded-Q transform (BQT), and was proposed in [5]. The BQT employs the FFT algorithm within each octave, greatly reducing the associated implementation cost.

2.1. Bounded-Q Fast Filter Bank

The high-selectivity counterpart of the BQT is denominated bounded-Q fast filter bank (BQFFB) [2]. The BQFFB implementation in the AMT context consists of submitting the input signal, at first, to a CQFFB presenting 10 channels geometrically spaced by a factor of 2. These channels can be associated to the 10 octaves perceivable by the human auditory system, between 20 Hz and 20 kHz. This first separation step presents high selectivity, but low computational cost due to the small number of channels. Each octave is then decimated by a proper factor, and the higher half of its complex spectrum (between π and 2π) is fed into 32 channels of an FFB, yielding a linear channel separation. The resulting resolution is enough to distinguish frequencies half-tone apart, which is generally the minimum spacing between notes in the equal tempered music scale. This entire procedure is represented in Figure 1.

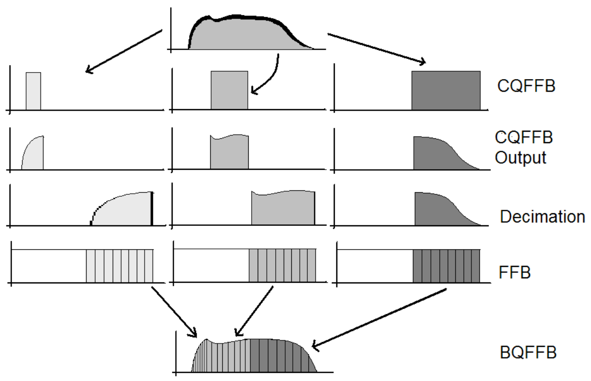


Figure 1. BQFFB implementation based on a CQFFB followed by an FFB.

Hence, the BQFFB combines the FFB high selectivity with the quasi-geometric spacing of the BQT. A comparative table with the properties of each spectral-analysis tool can be found in [2], along with detailed comparisons between their respective computational costs.

3. Practical Issues on the BQFFB Implementation

This section discusses a few points that should be taken into account in the BQFFB implementation to avoid artifacts in the time-frequency analysis.

3.1. End-of-Octave Gap

After the lower part (between 0 and π) of each octave spectrum is eliminated, it must be fed into the appropriate channels of the second-step FFB. However, picking one half of the channels creates a small gap at the end the octave. To illustrate this issue in a simple way, an N-channel FFB, with $N = 8$, is depicted in Figure 2. From this figure, one clearly observes that the normalized frequency range near 2π is in fact associated to the first channel. Therefore, the higher $N/2$ FFB channels do not entirely fill the desired π to 2π range.

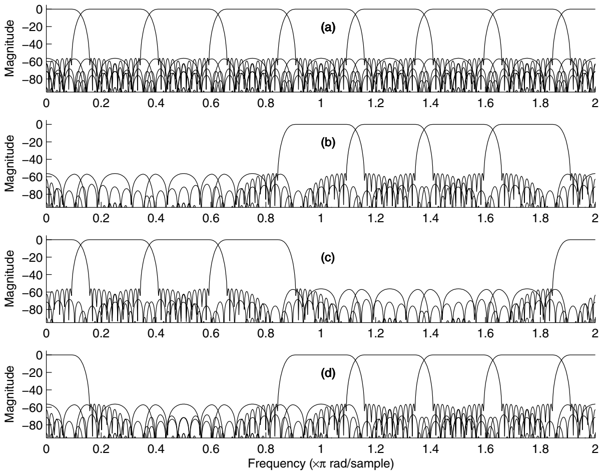


Figure 2. Magnitude responses of an 8-channel FFB: (a) Complete bank; (b) Higher $N/2$ channels; (c) Lower $N/2$ channels; (d) First channel along with higher $N/2$ channels.

Hence, in the BQFFB implementation, one must also take into consideration the first FFB channel, as depicted in Figure 2(d), to avoid the end-of-octave gap. The lower portion of this first channel does not impose any practical problem, since the CQFFB eliminates all components within this range in the first step of the BQFFB procedure, as seen in Figure 1.

3.2. Octave Filter Compensation

Obviously, the CQFFB that separates the spectrum in octaves is not ideal as depicted in Figure 1. The true magnitude responses of the individual filters can be observed in Figure 3.

This behavior can be compensated by incorporating a gain factor to each inner channel, thus equalizing the octave-separation procedure. The output of each FFB channel c inside a given octave o is multiplied by a gain given by

$$G_c = \frac{1}{H_o(\Omega_i)}, \qquad (1)$$

where $H_o(\Omega)$ is the frequency response of that octave filter and Ω_i is the center frequency of channel c. Naturally, within the stopband of $H_o(\Omega)$, the gains are set to zero.

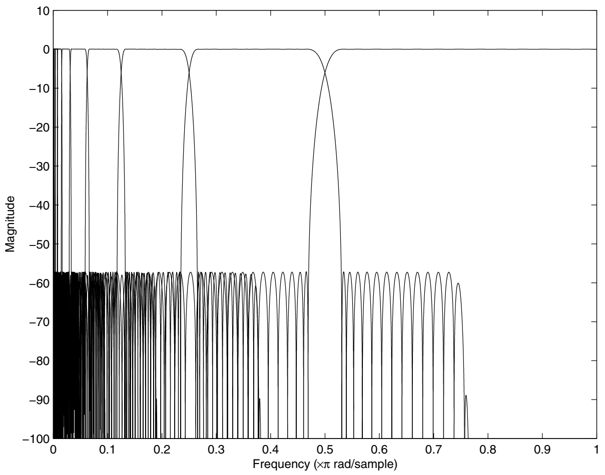


Figure 3. CQFFB presenting 10 channels geometrically spaced by a factor of 2. These channels correspond approximately to the 10 octaves perceived by the human auditory system, between 20 Hz and 20 kHz.

This simple procedure is illustrated in Figure 4. Figure 4(a) shows the individual channel magnitudes within the 7^{th} octave without gain compensation, whereas Figure 4(b) shows the compensating gain in each channel, yielding the flat result depicted in Figure 4(c).

3.3. Tempered-Scale Mapping

The previous steps allow one to separate the entire signal spectrum into octaves. These are defined as a function of the normalized sampling frequency 2π in the following way: The highest octave of interest lies within $\pi/2$ and π; the second highest octave lies between $\pi/4$ and $\pi/2$, and so on. After that, each of the M octaves is divided into N equally spaced channels, taking advantage of the FFT-like efficient algorithm.

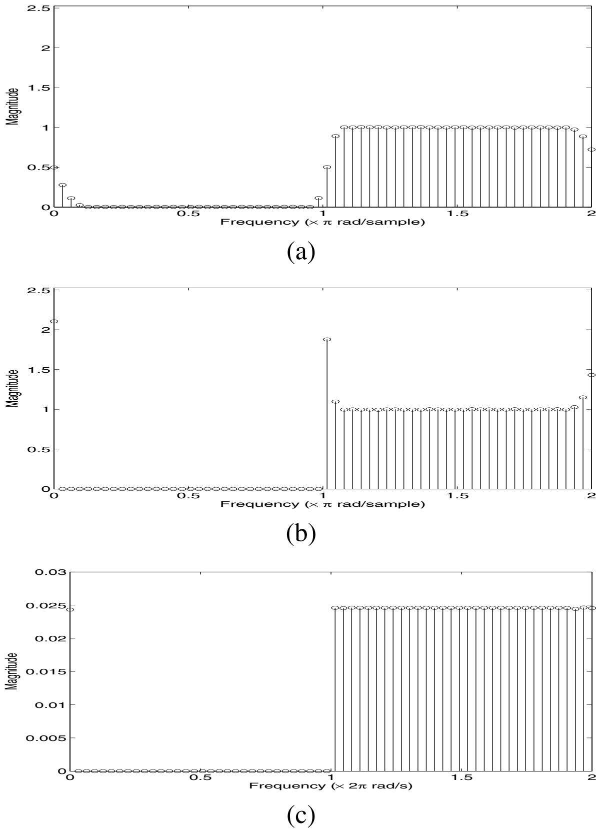
A simple strategy can be devised to map the BQFFB channels onto candidates to fundamental frequencies of the equal tempered scale:

1. Starting from a reference note (e.g. $A4 = 440$ Hz), compute the theoretical fundamental frequencies of musical notes spanning the complete spectrum, and define a half-tone band t_l around each note l.

2. Starting from the sampling frequency (e.g. $F_s =$ 44.1 kHz), once the number of octaves to be analyzed by the BQFFB (e.g. $M = 10$) and the number of channels within each octave (e.g. $N = 32$) have been chosen, define the lower and higher limits of each individual BQFFB channel.

3. Compose each candidate fundamental l by adding up the BQFFB channels contained in t_l weighted by their percentage of intersection.

For the values suggested above, one would consider the range associated to $A4$ from 427.5 Hz to 452.9 Hz. On the other hand, the three associated BQFFB channels should be: A from 425.2 Hz to 436.0 Hz; B from 436.0 Hz to 446.8 Hz; C from 446.8 Hz to 457.6 Hz. The weighted sum would then take 79% of channel A, 100% of channel B, and 56% of channel C to form the corresponding $A4$ channel. This process is illustrated in Figure 5.



(a)

(b)

(c)

Figure 4. Example of gain compensation after the octave separation stage: (a) Magnitude response without gain compensation; (b) Gain compensation; (c) Magnitude response with gain compensation.

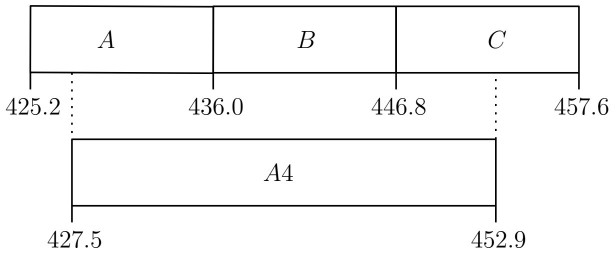


Figure 5. Mapping example between BQFFB channels and note channels.

4. Experiments

In order to illustrate the behavior of the BQFFB system described in this paper, two computer experiments are described.

In the first one, one performs the analysis of a sinusoidal signal whose frequency varies according to an ascending chromatic scale, i.e. where each note is one half-tone above its predecessor. This test verifies the BQFFB effectiveness along the frequency range, and shows the effects of the octave filter compensation and tone-mapping procedure described previously.

The second experiment tests the system under a practical AMT scenario. A real recording is submitted to the system to verify the percentage of notes correctly detected.

4.1. Experiment 1: Chromatic Scale

The chromatic scale used in this experiment extends from C5 (523 Hz) to E8 (5274 Hz). The respective input signal was submitted to a BQFFB with $M = 10$, $N = 32$, without any compensation procedure, generating the output seen in Figure 6.

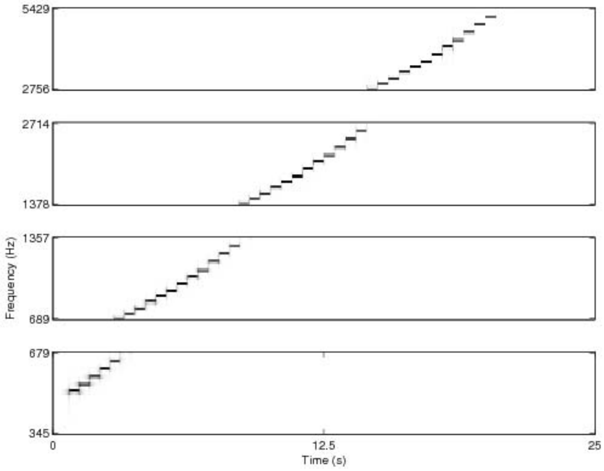


Figure 6. Experiment 1: BQFFB output for the signal containing the chromatic scale.

By applying the tone-mapping procedure, the BQFFB octaves with 32 filter channels are mapped onto musical octaves with 12 note channels each. The result is seen in Figure 7, where all the tones are properly distinguishable in frequency and time.

From the gray scale in Figure 7, one notices that distinct tones have been detected with different intensities, an effect of the non-ideal octave separation. The experiment was then repeated with the gain-compensating technique. Figure 8 shows the direct sum of the filter outputs in both cases without and with the gain compensation. Transients were purposely kept to serve as markers of note transitions. From the plots, one verifies that the gain compensation succeeded in equalizing the response of octave-separation filters.

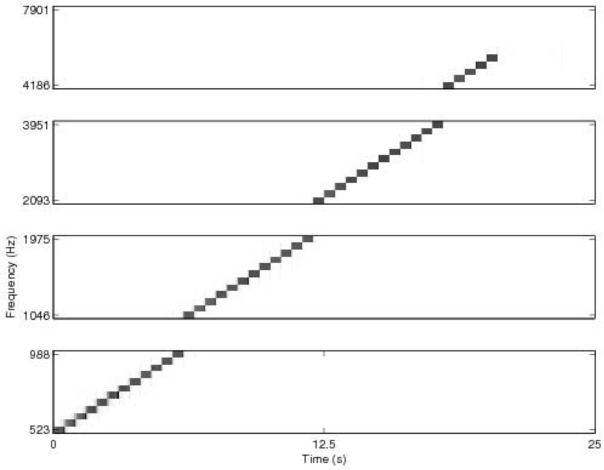


Figure 7. Experiment 1: BQFFB output for the signal containing the chromatic scale after the tempered scale mapping.

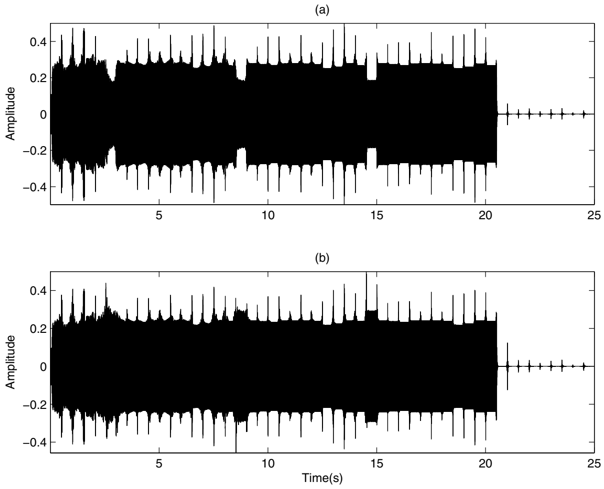


Figure 8. Experiment 1: BQFFB output for the signal containing the chromatic scale. (a) Before the octave filter compensation. (b) After the octave filter compensation.

4.2. Experiment 2: Real Recording

This experiment aims at evaluating the system performance over a real signal. This signal contains an execution of the piece “Flight of the Bumblebee”, by Rimski-Korsakoff, arranged for piano by Rachmaninoff. The main information to be retrieved concerns the detected notes. At this point, it is not important to estimate onsets and offsets. The task is to verify if the notes present in the musical score are indicated by the BQFFB output, which is shown in Figure 9. Note detection was based on the simple search for local maxima, at each time instant, at the outputs of the BQFFB channels. The results are shown in Figure 10.

Comparing the notes from the BQFFB output and from the musical score, one concludes that 80% of the notes were detected. It shows that the BQFFB can generate an output that can be presented to the next block in an AMT system in order to detect musical notes.

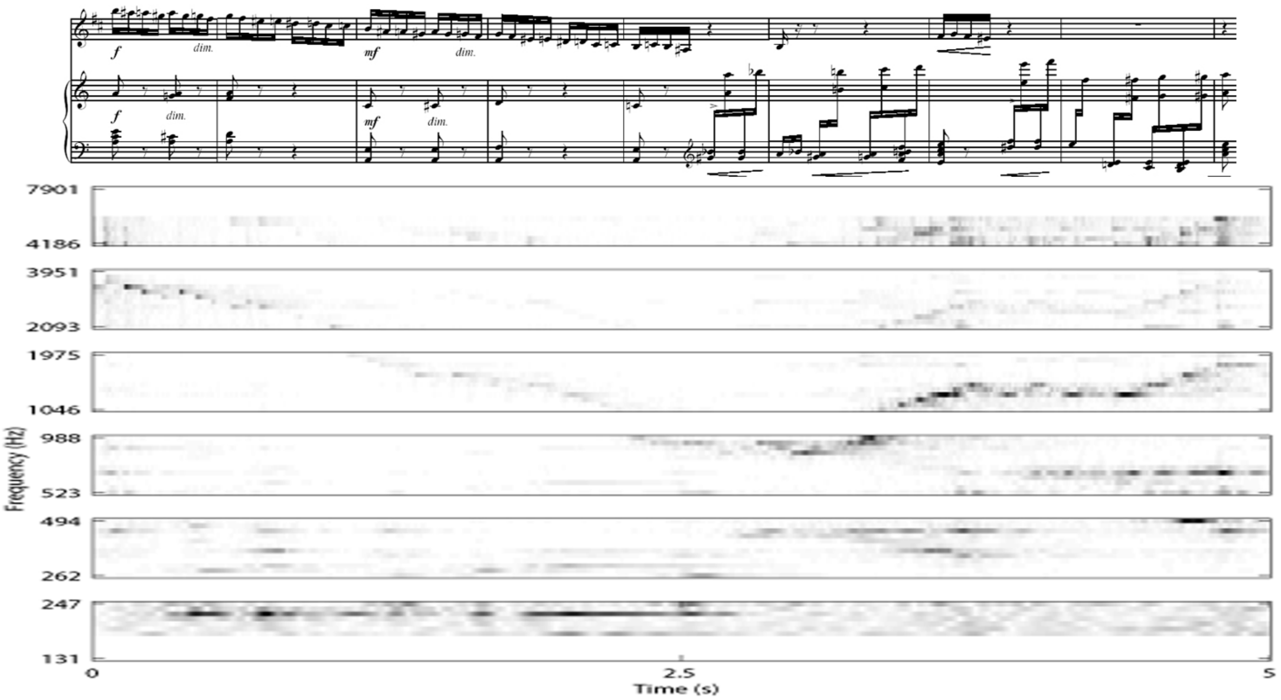


Figure 9. BQFFB output for the signal containing the "Flight of the Bumblebee".

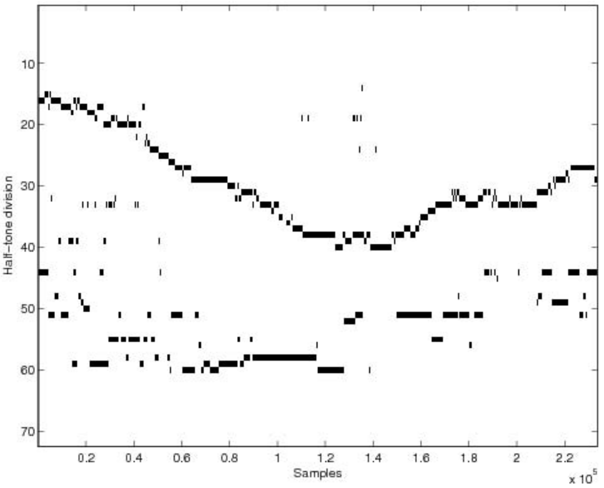


Figure 10. Simple note detection, of the "Flight of the Bumblebee extract", based on the BQFFB output after gain-compensation and channel-mapping procedures.

5. Conclusions

This paper explored practical issues concerning the bounded-Q fast filter bank (BQFFB) implementation for the automatic music transcription. The impact of some solutions were illustrated over some computer experiments using artificial and real signals. The results are quite promising when they are put into an AMT scenario, since it was possible to reach near 80% of detection rate after an extremely simple heuristics.

References

[1] J. C. Brown. Calculation of a constant Q spectral transform. *Journal of the Acoustical Society of America*, 89(1):425–434, 1991.

[2] F. C. C. B. Diniz, I. Kothe, S. L. Netto, and L. W. P. Biscainho. High-selectivity filter banks for spectral analysis of music signals. *EURASIP Journal Advances on Signal Processing*, 2007(PAPER ID 94704):1–12, 2007.

[3] C. N. dos Santos, S. L. Netto, L. W. P. Biscainho, and D. B. Graziosi. A modified constant-Q transform for audio signals. In *Proc. of ICASSP 2004 - Int. Conf. on Audio, Speech, and Signal Processing*, volume 2, pages 469–472, Montreal, Canada, May 2004. IEEE.

[4] B. Farhang-Boroujeny and Y. C. Lim. A comment on the computational complexity of sliding FFT. *IEEE Transactions on Circuits and Systems - II: Analog and Digital Signal Processing*, 39(12):875–876, 1992.

[5] K. L. Kashima and B. Mont-Reynaud. The bounded-Q approach to time-varying spectral analysis. Technical Report 28, Stanford University, Dep. of Music, Stanford, CA, USA, 1985.

[6] A. Klapuri and M. Davy. *Signal Processing Methods for Music Transcription*. Springer-Verlag, New York, USA, 2006.

[7] Y. C. Lim. Frequency-response masking approach for the synthesis of sharp linear phase digital filters. *IEEE Transactions on Circuits and Systems*, 33(4):357–364, 1986.

[8] Y. C. Lim and B. Farhang-Boroujeny. Fast filter bank (FFB). *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing*, 39(5):316–318, 1992.

[9] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice-Hall, Upper Saddle River, NJ, USA, 1992.

[10] D. William and E. Brown. *Theoretical Foundations of Music*. Wadsworth, Belmont, CA, USA, 1978.