

AN AUDITORY-BASED TRANSFORM FOR AUDIO SIGNAL PROCESSING

Qi (Peter) Li

Li Creative Technologies, Inc.
Florham Park, New Jersey 07932, USA
qli@ieee.org

ABSTRACT

An auditory-based transform is presented in this paper. Through an analysis process, the transform converts time-domain signals into a set of filter bank output. The frequency responses and distributions of the filter bank are similar to those in the basilar membrane of the cochlea. Signal processing can be conducted in the decomposed signal domain. Through a synthesis process, the decomposed signals can be synthesized back to the original signal through a simple computation. Also, fast algorithms for discrete-time signals are presented for both the forward and inverse transforms. The transform has been approved in theory and validated in experiments. An example on noise reduction application is presented. The proposed transform is robust to background and computational noises and is free from pitch harmonics. The derived fast algorithm can also be used to compute continuous wavelet transform.

Index Terms— Noise reduction, robust audio signal processing, fast transform, cochlea

1. INTRODUCTION

The Fourier transform (FT) is the most popularly used transform to convert signals from the time domain to frequency domain; however, it has fixed time-frequency resolution and the frequency distribution is restricted to be linear. These limitations generate problems in audio and speech processing such as the pitch harmonics, computational noise, and sensitivity to background noise. On the other hand, the wavelet transform (WT) provides flexible time-frequency resolution, but also has notable problems. First, no existing wavelet is capable of mimicking the impulse responses of the basilar membrane closely, so it cannot be directly used to model the cochlea or carry out related computation. Additionally, even though forward and inverse continuous wavelets transforms are defined for continuous variables, to the best of our knowledge, there is no numerical computational formula for real inverse continuous wavelet transforms (ICWT). No such function exists even in a commercial wavelet package. Discrete wavelet transform has been applied in speech processing, but the frequency distribution is limited to the dyadic scale which is different from the scale in the cochlea.

Motivated by the fact that the human auditory system outperforms current machine-based systems for acoustic signal processing, we developed an auditory-based transform to facilitate our future research in developing high performance systems. The traveling waves of the basilar membrane in the cochlea and its impose response have been measured and reported in the literature, such as

[1]. The basilar membrane tuning and auditory filter shapes have also been studied in the literature, such as [2]. Many electronic and mathematic models have been defined to mimic the traveling wave, auditory filters, and frequency responses of the basilar membrane. Furthermore, based on the concept of the cochlea, many feature extraction algorithms have been developed for speech recognition, such as MFCC and others. Since all the above approaches are focused on studying and modeling the auditory periphery system. They provide the analysis, but no synthesis. However, an auditory-based transform with both forward and inverse transforms for digital computers is needed for many audio applications, such as noise reduction, hearing aid, coding, speech and music synthesis, speaker and speech recognition, etc.

Gammatone filter banks [3] has been proposed to model the impulse responses of the basilar membrane and has been used to decompose time-domain signals into different frequency bands; however, there is no mathematical proof of how to synthesize the decomposed multichannel signals back to a time-domain signal. Although some suggestions on resynthesis have been given in plain language [4] or simply at the conceptual level, there remain no details or mathematical proof to validate the accuracy and computation efficiency. Actually, from our analysis, the suggested computation in [4] is at least redundant. In [5], a Gammatone based transform with analysis and synthesis was presented, but the filter bank derives a complex valued output which is not only different from the real cochlea but further complicates its implementation.

To employ the concept of the auditory system to audio signal processing, Li proposed an auditory-based transform in [6]. The detailed results are presented here. We note that the purpose of this research is not to model the cochlea precisely and completely; instead, it is to provide a simple and fast transform for real application, as an alternative selection to the FT and WT.

2. DEFINITION OF THE PROPOSED TRANSFORM

When sound enters the human ear, acoustic energy from the outer ear is converted to mechanical energy via the middle ear which consists of three small bones. When the last bone in the middle ear, the stapes, moves, it sets the fluid inside the cochlea in motion creating traveling waves on the basilar membrane. The impulse response of the basilar membrane (BM) in the cochlea can be represented by function $\psi(t) \in \mathbf{L}^2(\mathcal{R})$. The function satisfies the following conditions:

1. It integrates to zero:

$$\int_{-\infty}^{\infty} \psi(t) dt = 0. \quad (1)$$

This work was supported in part by the AFRL under grant FA8750-08-C-0028.

2. It is square integrable or has finite energy:

$$\int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty. \quad (2)$$

3. It satisfies:

$$\int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{\omega} d\omega = C. \quad (3)$$

where $0 < C < \infty$ and

$$\Psi(\omega) = \int_{-\infty}^{\infty} \psi(t) e^{-j\omega t} d\omega. \quad (4)$$

4. It tapers off to zero on both ends just as it is observed in psychoacoustic experiments with the BM [1].
5. It has one major modulation frequency and its frequency response is a triangle-like, band-pass filter.

The first three conditions are required by mathematics for further derivation. The last two are required in order to match previous psychoacoustic and physiological experimental results and to approximate numerical computations presented later.

Let $f(t)$ be any square integrable function. The forward transform of $f(t)$ with respect to a function representing the BM impulse response $\psi(t)$ is defined as:

$$T(a, b) = \int_{-\infty}^{\infty} f(t) \frac{1}{\sqrt{|a|}} \psi\left(\frac{t-b}{a}\right) dt \quad (5)$$

where a and b are real, and both $f(t)$ and $\psi(t)$ belong to $\mathbf{L}^2(\mathbf{R})$, and $T(a, b)$ represents the traveling waves in the BM. The above equation can also be written as:

$$T(a, b) = \int_{-\infty}^{\infty} f(t) \psi_{a,b}(t) dt \quad (6)$$

where

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}} \psi\left(\frac{t-b}{a}\right). \quad (7)$$

Note that $1/\sqrt{|a|}$ is an energy normalizing factor. It ensures that the energy stays the same for all a and b ; therefore, we have:

$$\int_{-\infty}^{\infty} |\psi_{a,b}(t)|^2 dt = \int_{-\infty}^{\infty} |\psi(t)|^2 dt \quad (8)$$

The factor a is a scale or dilation variable. By changing a , we can shift the central frequency of an impulse response function. The factor b is a time shift or translation variable. For a given value of a , the factor b shifts the function $\psi_{a,0}(t)$ by an amount b along the time axis.

A typical cochlear impulse response function or cochlear filter can be defined as:

$$\begin{aligned} \psi_{a,b}(t) &= \frac{1}{\sqrt{|a|}} \left(\frac{t-b}{a}\right)^\alpha \exp\left[-2\pi f_L \beta \left(\frac{t-b}{a}\right)\right] \\ &\quad \cos\left[2\pi f_L \left(\frac{t-b}{a}\right) + \theta\right] u(t) \end{aligned} \quad (9)$$

where $\alpha > 0$ and $\beta > 0$, $u(t)$ is the unit step function, $u(t) = 1$ for $t \geq 0$ and 0 otherwise. The value of θ should be selected such that (1) is satisfied. b is the time shift variable, and a is the scale variable. The value of a can be determined by the current filter

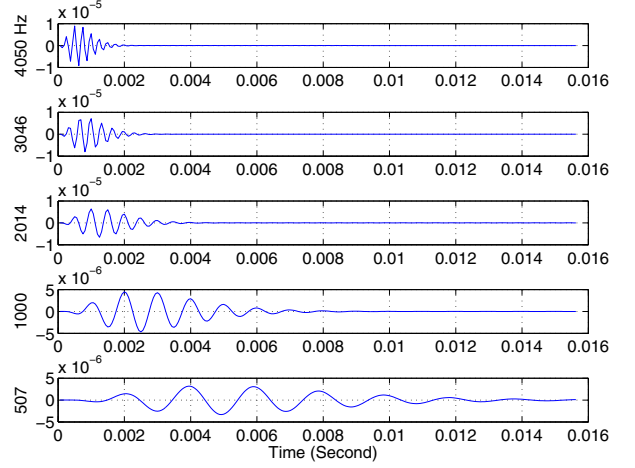


Figure 1: Impulse responses of the BM in the proposed transform when $\alpha = 3$ and $\beta = 0.2$. They are very similar to psychological measurements, such as the figures in [1].

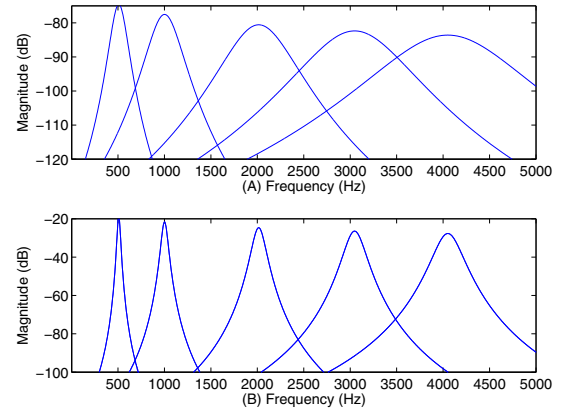


Figure 2: The frequency responses of the cochlear filters when $\alpha = 3$: (A) $\beta = 0.2$; and (B) $\beta = 0.035$.

central frequency f_c and the lowest central frequency f_L in the cochlear filter bank:

$$a = f_L / f_c. \quad (10)$$

Since we contract $\psi_{a,b}(t)$ with the lowest frequency along the time axis, the value of a is in $0 < a \leq 1$. If we stretch ψ , the value of $a > 1$. The frequency distribution of the cochlear filter can be in the form of linear or nonlinear scales such as ERB (equivalent rectangular bandwidth) Bark, Mel, log, etc. Note that the values of the a_i needs to be pre-calculated for the required central frequency of the cochlear filter. Fig. 1 shows the impulse responses for 5 cochlear filters and Fig. 2, their corresponding frequency responses. Normally, we use $\alpha = 3$. The value of β can be selected by applications. We used $\beta = 0.2$ for noise reduction and smaller values for feature extraction.

3. THE INVERSE TRANSFORM

Just as the Fourier transform requires an inverse transform a similar inverse transform is also needed for the proposed transform.

The need arises when the processed frequency decomposed signals need to be converted back real signals, such as speech and music synthesis and noise reduction; and second, to prove that no information is lost through the proposed forward transforms. This is necessary to a transform.

If (3) is satisfied, the inverse transform exists:

$$f(t) = \frac{1}{C} \int_{a=0}^{\infty} \int_{b=0}^{\infty} \frac{1}{|a|^2} T(a, b) \psi_{a,b}(t) da db \quad (11)$$

The derivation of the above transform is similar to inverse continuous wavelet transform [7]. Equation (6) can be written in the form of convolution with $f(b)$:

$$T(a, b) = f(b) * \psi_{a,0}(-b) \quad (12)$$

Take the Fourier transform of both sides, the convolution becomes multiplication:

$$\int_{b=-\infty}^{\infty} T(a, b) e^{-j\omega b} db = \sqrt{|a|} F(\omega) \Psi^*(a\omega) \quad (13)$$

where $F(\omega)$ and $\Psi(\omega)$ represents the Fourier transforms of $f(t)$ and $\psi(t)$, respectively. We now multiply both sides of the above equation by $\Psi(a\omega)/|a|^{3/2}$ and integrate with a :

$$\int_{a=-\infty}^{\infty} \int_{b=-\infty}^{\infty} \frac{1}{|a|^{3/2}} T(a, b) \Psi(a\omega) e^{-j\omega b} da db = F(\omega) \int_{a=-\infty}^{\infty} \frac{|\Psi(a\omega)|^2}{|a|} da. \quad (14)$$

The integration on the right hand side can be further written as:

$$\int_{a=-\infty}^{\infty} \frac{|\Psi(a\omega)|^2}{|a|} da = \int_{\omega=-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{|\omega|} d\omega = C \quad (15)$$

to meet the admissibility condition in (3) where C is a constant. Rearrange (14), we can then have

$$F(\omega) = \frac{1}{C} \int_{a=-\infty}^{\infty} \int_{b=-\infty}^{\infty} \frac{1}{|a|^{3/2}} T(a, b) \Psi(a\omega) e^{-j\omega b} da db. \quad (16)$$

We now can take inverse Fourier transform on both sides of the above equation to achieve (11).

4. THE DISCRETE-TIME AND FAST TRANSFORM

In practical applications, the discrete-time cochlear transform is necessary. The forward discrete transform can be written as:

$$T[a_i, b] = \sum_{n=0}^N f[n] \frac{1}{\sqrt{|a_i|}} \psi \left[\frac{n-b}{a_i} \right] \quad (17)$$

where $a_i = f_L/f_{c_i}$ is the scaling factor for the i th frequency band f_{c_i} and N is the length of digital signal $f[n]$. The scaling factor a_i can be a linear or nonlinear scale. For the discrete transform, a_i can also be in ERB-scale, Bark, Log, or other nonlinear scales.

The corresponding discrete-time inverse transform is:

$$\tilde{f}[n] = \frac{1}{C} \sum_{a_i=a_1}^{a_k} \sum_{b=1}^N \frac{1}{|a_i|} T[a_i, b] \psi \left[\frac{n-b}{a_i} \right] \quad (18)$$

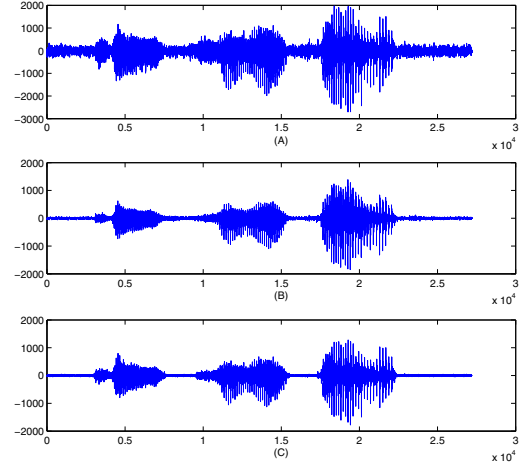


Figure 3: (A) and (B) are speech waveforms simultaneously recorded in a moving car. The microphones are located at the visor and drivers lapel, respectively. (C) is after noise reduction using the proposed transform from the waveform in (A).

where $0 \leq t \leq N$, $a_1 \leq a_i \leq a_k$, and $1 \leq b \leq N$. We note that $\tilde{f}[n]$ approximates $f[n]$ given the limited number of decomposed frequency bands. The above formulas have been verified by the following experiments. Also note that (18) can also be applied to compute the inverse continuous WT, where ψ needs to be replaced.

Just as the Fourier Transform has a fast algorithm, the FFT, the fast algorithms for the proposed transform also exists. The most computational intensive components in (17) and (18), the convolutions, of both the forward and the inverse transforms can be implemented by the FFT. Also, depending on the application, the resolution of the lower frequency bands can be reduced to save the computation.

5. EXPERIMENTS AND DISCUSSIONS

Transform Validation: In addition to the theoretical proof, we also validated the proposed transform via real audio data. One of our experiments is to use the speech waveform of a males voice saying the words: two, zero, five, as shown in Fig. 3 (A) as the original data. In the forward transform using (17), we used frequencies between 80 Hz to 5KHz. $\alpha = 3$ and $\beta = 0.2$ to decompose the original data into multiple frequency bands in the Bark scale. In the inverse transform using (18), we synthesized the multiple band output back to speech. When plotted both waveforms before and after the transform, we cannot visualize any difference. We then use the correlation coefficients σ_{12}^2 as the measurement to compute

Table 1: Correlation Coefficients σ_{12}^2 between the original and synthesized signals using the proposed inverse transform.

No. of Filters	8	16	32	64
σ_{12}^2	0.74	0.96	0.99	0.99

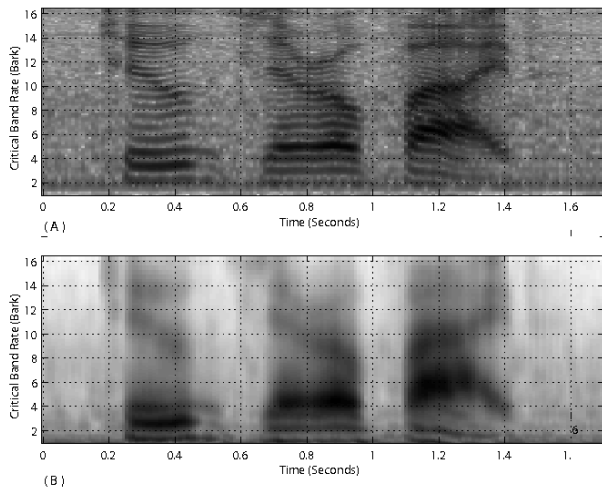


Figure 4: (A) and (B) are from the FFT and the proposed spectrogram, respectively. They were displayed in the Bark scale from 0 to 6.4 Barks (0 to 3500 KHz). The proposed transform is harmonic free and has less computational noise. The original speech waveform was from Fig. 3 (A).

the difference. The results are shown in Table 1. It validated the proposed forward and inverse transforms derived. It also indicates that 32 filters are good for most applications.

Noise Robustness: Using the data in Fig. 3 (A), we calculated the FFT spectrograms as shown in Fig. 4 (A), with 30 ms Hamming window shifting every 10 ms. To facilitate the comparison, we swapped the frequency distribution from linear scale to the Bark scale using the method in [8]. The spectrograms from the proposed auditory-based transform is shown in Fig. 4 (B). They were generated from the output of the proposed transform using the same window size to compute the average densities for each band. Comparing the two spectrograms in Fig. 4, we found that there are no pitch harmonics and less computational noise in the spectrums generated from the proposed transform. This is due to the variable length of filters. Furthermore, when comparing the spectrums, we found that the proposed transform can generate clear pitch signal for detection and has much less distortion to background noises.

Applications: Fig. 3 is also an example of applying the proposed transform to noise reduction. The original speech, as shown in Fig. 3 (A), was first decomposed into 32 frequency bands using (17). A voice detector was used to recognize noise and speech using a moving window. A denoising function was then applied to each of the decomposed frequency bands. The inverse transform in (18) is then applied to convert the processed signals back to clean speech signal as shown in Fig. 3 (C), which is similar to the original close talking data in Fig. 3 (B).

Recently, based on the proposed transform, we have developed a new feature extraction algorithm which has shown significant noise robustness over the MFCC feature in speaker recognition. The result will be published later.

Discussions: In comparison, Eq. (9) is different than the Gammatone filters:

$$G_{f_c}(t) = t^{N-1} \exp[-2\pi b(f_c)t] \cos(2\pi f_c t + \psi) u(t) \quad (19)$$

where $b(f_c)$ is a function where β in (9) is a constant. The derivation of (9) comes from the impulse response experimental results in [1, 9]. The suggested resynthesis approach for Gammatone [4] is different from the results of the proposed inverse transform.

Compared to the FT, the proposed transform only uses real number computation and is more robust to noises. Its frequency distribution can be in any linear or nonlinear scale. The proposed transform is similar to the continuous WT (CWT); however, the filter in (9) is different than existing wavelets. Also, there is no formula to compute the inverse CWT numerically. We note that (18) can also be applied to compute the ICWT. Compared to discrete-time WT, the frequency response of proposed transform is not limited to the dyadic scale. It can be in any linear or nonlinear scale.

6. CONCLUSIONS

The concept of the proposed transform is to mimic the impulse responses of the basilar membrane and its nonlinear frequency distribution characteristics. As the results show, the proposed transform has significant advantages in its noise robustness and its freedom from harmonic distortion and computational noise. These advantages can lead to many new applications, such as robust features for speech and speaker recognition, new algorithms for noise reduction and denoising, speech and music synthesis, audio coding, hearing aid, audio signal processing, etc. In summary, the proposed transform is a new time-frequency transform for audio and speech processing and many other applications.

7. ACKNOWLEDGMENT

The author would like to thank Manli Zhu, Yan Huang, and Joshua J. Hajicek for useful discussions.

8. REFERENCES

- [1] G. von Békésy, *Experiments in hearing*. New York: McGRAW-HILL, 1960.
- [2] J. Allen, "Cochlear modeling," *IEEE ASSP Magazine*, pp. 3–29, Jan. 1985.
- [3] P. I. M. Johannesma, "The pre-response stimulus ensemble of neurons in the cochlear nucleus," *The proceeding of the symposium on hearing Theory*, vol. IPO, pp. 58–69, June 1972.
- [4] M. Weintraub, *A theory and computational model of auditory monaural sound separation*. PhD thesis, Stanford University, CA, August 1985.
- [5] V. Hohmann, "Frequency analysis and synthesis using a Gammatone filterbank," *Acta Acoustica United with Acustica*, vol. 88, pp. 433–442, 2002.
- [6] Q. Li, "Solution for pervasive speaker recognition," SBIR Phase I Proposal, Submitted to NSF IT.F4, Li Creative Technologies, Inc., NJ, June 2003.
- [7] R. Rao and A. Bopardikar, *Wavelet Transforms*. MA: Adison-Wesley, 1998.
- [8] Q. Li, F. K. Soong, and S. Olivier, "An auditory system-based feature for robust speech recognition," in *Proc. 7th European Conf. on Speech Communication and Technology*, (Denmark), pp. 619–622, Sept. 2001.
- [9] J. P. Wilson and J. Johnstone, "Capacitive probe measures of basilar membrane vibrations in," *Hearing Theory*, 1972.