COMM. MATH. SCI.
Vol. 3, No. 1, pp. 745–754

AN INVERTIBLE DISCRETE AUDITORY TRANSFORM *

JACK XIN † AND YINGYONG QI ‡

Abstract. A discrete auditory transform (DAT) from sound signal to spectrum is presented and shown to be invertible in closed form. The transform preserves energy, and its spectrum is smoother than that of the discrete Fourier transform (DFT) consistent with human audition. DAT and DFT are compared in signal denoising tests with spectral thresholding method. The signals are noisy speech segments. It is found that DAT can gain 3 to 5 decibel (dB) in signal to noise ratio (SNR) over DFT except when the noise level is relatively low.

Key words. Invertible Discrete Auditory Transform.

AMS subject classifications: 94A12, 94A14, 65T99.

1. Introduction

Audible acoustic signal processing often consists of frame by frame discrete Fourier transform (DFT) of input signal followed by spreading in Fourier spectrum using the critical band filter, see M. Schroeder et al [4]. These two steps mimic the responses of human audition to the input signal, facilitating the computation of excitation pattern over critical bands and psychoacoustics-based processing [4]. However, the spreading operation in step two is not invertible.

In this work, an invertible discrete auditory transform (DAT) is formulated to combine the two steps into one. DAT incorporates the spectral spreading functions of Schroeder et al [4], which leads to smoother spectrum than that of the discrete Fourier transform (DFT). As a result, DAT has better localization properties in the time domain. DAT bears some resemblance to the wavelet transform [2] in that a function of one variable (time) is transformed into a function of two variables (time and frequency). It is this redundancy that makes the inversion possible and explicit.

The paper is organized as follows. In section 2, a general form of DAT is introduced and its inversion established. In section 3, a specific DAT is given based on the known auditory spectral energy spreading functions of Schroeder et al [4]. DAT spectrum is defined and compared with DFT spectrum. The time localization property of DAT basis functions is illustrated as well. In section 4, noisy signal reconstruction from thresholded spectrum is carried out and its signal to noise ratio is computed by using the reconstructed signal and the clean (noise free) signal. The signal is a segment of male or female speech, and can be either voiced (e.g. vowels) or unvoiced (consonants such as s and f). DAT is found to gain by 3 to 5 decibel (dB) in signal to noise ratio (SNR) over DFT in such a denoising task. Concluding remarks are made in section 5.

2. Discrete Auditory Transform

Let $s=(s_0,\cdots,,s_{N-1})$ be a discrete signal, and $\hat{s}$ its discrete Fourier transform (DFT) [1]:

$$\hat{s}_k=\sum_{n=0}^{N-1} s_n\, e^{-i(2\pi nk/N)}. \tag{2.1}$$

*Received: December 1, 2004. Accepted (in revised version): December 24, 2004. Communicated by Shi Jin.

†Department of Mathematics and ICES, University of Texas at Austin, Austin, TX 78712, USA (jxin@math.utexas.edu).

‡Qualcomm Inc, 5775 Morehouse Drive, San Diego, CA 92121, USA.

The DFT inversion formula is:

$$s_n = \frac{1}{N} \sum_{k=0}^{N-1} \hat{s}_k \, e^{i(2\pi nk/N)}. \tag{2.2}$$

For two signals s and t of length N, the Plancherel-Parseval equality is:

$$\sum_{n=0}^{N-1} s_n t_n^* = \frac{1}{N} \sum_{k=0}^{N-1} \hat{s}_k \hat{t}_k^*, \tag{2.3}$$

implying the energy identity:

$$\sum_{n=0}^{N-1} |s_n|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |\hat{s}_k|^2. \tag{2.4}$$

Define the discrete auditory transform (DAT):

$$S_{j,m} \equiv \sum_{l=0}^{N-1} s_l \, K_{j-l,m}, \tag{2.5}$$

where the double indexed discrete kernel function is given by:

$$K_{l,m} = \sum_{n=0}^{N-1} X_{m,n} \, e^{i(2\pi ln/N)}; \tag{2.6}$$

where the matrix $X_{m,n}$ has square sum equal to one in m:

$$\sum_{m=0}^{M-1} |X_{m,n}|^2 = 1, \ \forall n. \tag{2.7}$$

We remark that DFT is recovered from DAT if $M = N$, and $X_{m,n}$ is the $N \times N$ identity matrix. The dependence on j is factored out explicitly to become a phase factor, and $S_{j,m} = e^{2\pi ijm/N} \hat{s}_m$. We are interested in the case when $X_{m,n}$ is not a diagonal matrix, yet still maintains the largest entry in absolute value on its diagonal along each row. Such a structure is shown in auditory filter responses (see details in section 3). Mathematically, the matrix $X_{m,n}$ introduces a weighted (selective) averaging of Fourier spectrum. The spectral averaging feature resembles the classical Cesàro sum of Fourier series, or in convolution form the use of Fejér kernel in lieu of Dirichlet kernel, for improving the convergence of Fourier series [5]. The classical averaging is however uniform (with equal weights).

2.1. Energy Identity. Let us show the energy conservation property of the transform. Upon substituting (2.6) into (2.5), the transform can be written as:

$$S_{j,m} = \sum_{n=0}^{N-1} \hat{s}_n \, X_{m,n} \, e^{i(2\pi nj/N)}, \tag{2.8}$$

which is similar to the representation of time domain solutions of cochlear models as a sum of time harmonic solutions [6, 7].

It follows from (2.8), (2.4), and (2.2) that:

$$\frac{1}{N}\sum_{j=0}^{N-1}|S_{j,m}|^2=\sum_{n=0}^{N-1}|\hat{s}_n|^2\,|X_{m,n}|^2,$$

implying:

$$\frac{1}{N^2}\sum_{m=0}^{M-1}\sum_{j=0}^{N-1}|S_{j,m}|^2=\frac{1}{N}\sum_{n=0}^{N-1}|\hat{s}_n|^2=\sum_{j=0}^{N-1}|s_j|^2. \tag{2.9}$$

Polarizing with (2.9), one finds the analogous Plancherel-Parseval identity of DAT for two signals s and t:

$$\frac{1}{N^2}\sum_{m=0}^{M-1}\sum_{j=0}^{N-1}S_{j,m}T^*_{j,m}=\sum_{j=0}^{N-1}s_jt^*_j. \tag{2.10}$$

2.2. Inversion. The explicit inversion formula is:

$$s_j=\frac{1}{N^2}\sum_{m=0}^{M-1}\sum_{l=0}^{N-1}S_{l,m}\sum_{n=0}^{N-1}X^*_{m,n}\,e^{i(2\pi(j-l)n/N)}. \tag{2.11}$$

Proof: Consider the sum in l. In view of (2.8), we see that

$$\frac{1}{N}\sum_{l=0}^{N-1}S_{l,m}\,e^{i(2\pi(j-l)n/N)}=e^{2\pi ijn/N}\,\hat{s}_n\,X_{m,n}.$$

So the right hand side of (2.11) is equal to:

$$\frac{1}{N}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}|X_{m,n}|^2\,e^{2\pi ijn/N}\,\hat{s}_n,$$

which equals upon summing over m and using (2.7):

$$\frac{1}{N}\sum_{n=0}^{N-1}e^{2\pi ijn/N}\,\hat{s}_n=s_j.$$

3. Transform Kernel and Spectrum

The role of the transform kernel X_{mn} is to spread the DFT vector $\hat{s}_n$. Here our knowledge of human audition will be utilized. Let us adopt the real nonnegative energy spreading function of Schroeder et al [4], denoted by $S(b(f_m),b(f_n))$, where f_m is the frequency to spread from, f_n is the frequency to spread to, and b is the standard mapping from Hertz (Hz) to Bark scale [3]. The functional form of $S(\cdot,\cdot)$ is given in Schroeder et al [4]. The DFT of a real vector s satisfies the symmetry property $\hat{s}_k=\hat{s}^*_{N-k}$, $k=1,2\cdots,N-1$. It is natural for the spreading kernel $X_{m,n}$ to respect this symmetry. Suppose the discrete signal s has sampling frequency F_s (Hz). The DFT component $\hat{s}_n$ ($0\le n\le N/2$, N a power of 2) corresponds to frequency:

$$f_n=F_s\cdot n/N,\ \ n\le N/2. \tag{3.1}$$

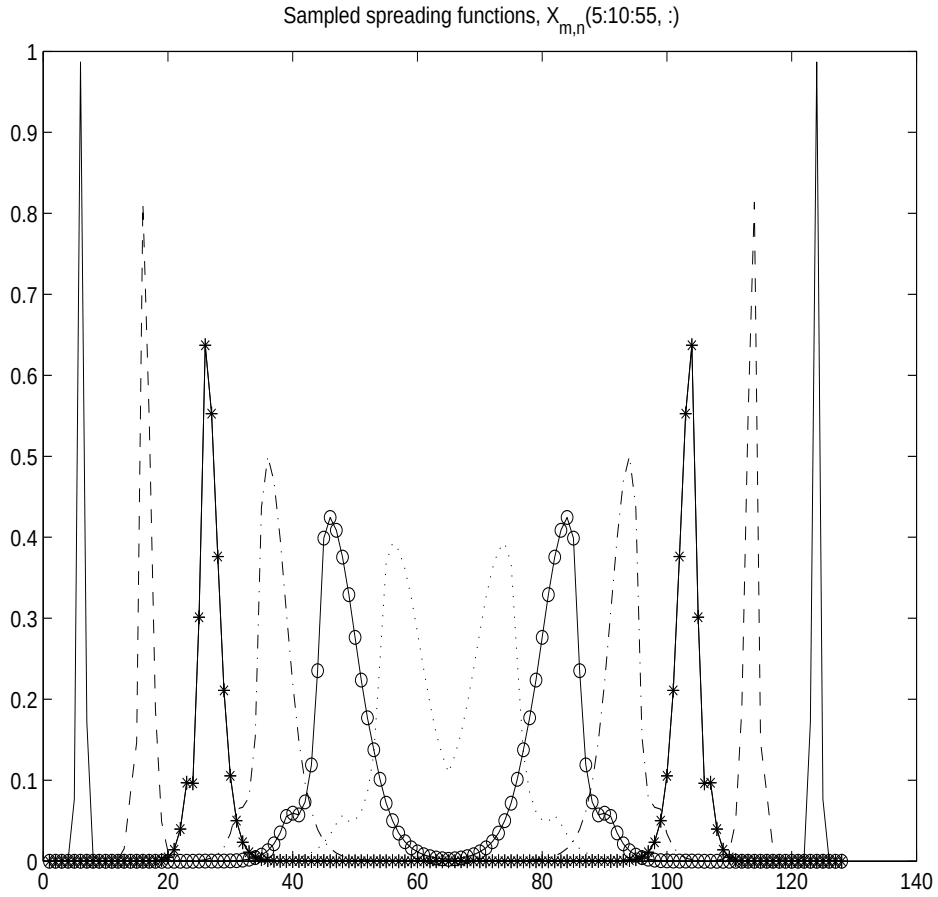

FIG. 3.1. *The transform kernel $X_{m,n}$ as a function of integer $n = 1$ to 128. The m variable equals 5 (solid), 15 (dash), 25 (line-star), 35 (dashdot), 45 (line-circle), 55 (dot).*

Let $U_{m,n} = U(f_m, f_n) = S^{1/2}(b(f_m), b(f_n))$, $0 \le m \le M-1$, $M = N/2$. The square root is to convert spreading from energy to amplitude scale. Then normalize $U_{m,n}$ to define $X_{m,n}$ as follows:

$$X_{m,0} = \frac{U(f_m, f_1)}{m_f(f_1)},$$

$$X_{m,n} = \frac{U(f_m, f_n)}{m_f(f_n)}, \quad 1 \le n \le N/2 - 1,$$

$$X_{m,n} = \frac{U(f_m, f_{N-n})}{m_f(f_{N-n})}, \quad N/2 \le n \le N-1,$$

where the m_f function is:

$$m_f(f) = \left(\sum_{m=0}^{M-1} |U(f_m, f)|^2 \right)^{1/2}. \tag{3.2}$$

We see that $X_{m,n}$ is symmetric in n with respect to $N/2$, and periodic in n ($X_{m,0} = X_{m,1} = X_{m,N-1}$ by construction). The normalization property (square sum equal to one) with respect to m holds. See Figure 1 for a plot of $X_{m,n}$ in n, at $m = 5:10:55$, where $N = 128$, $Fs = 16000$ Hz.

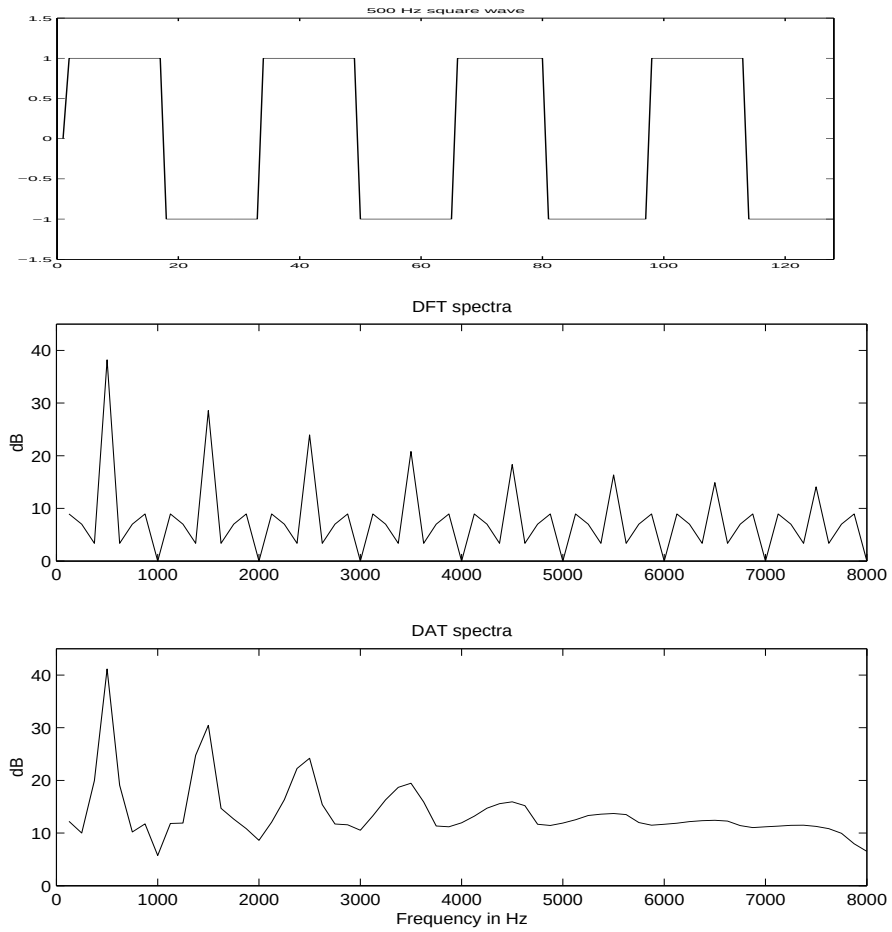

FIG. 3.2. *A 500 Hz square wave (top), its DFT spectrum (middle) and DAT spectrum (bottom).*

3.1. DAT Spectrum. In view of formula (2.8), we extract the amplitude (intensity) of the $S_{j,m}$ ($j \in [0, N-1]$) for each m and define the DAT spectrum:

$$\text{spec}(m) \equiv \left(\sum_{n=0}^{N-1} |\hat{s}_n X_{m,n}|^2 \right)^{1/2}. \tag{3.3}$$

As the input sound signal is divided into frames of length N, DAT spectrum can vary from frame to frame in time.

3.2. Transform Properties. Let us illustrate the DAT properties by considering a 500 Hz square wave (top panel of Figure 2). Middle and bottom panels of Figure 2 show DFT and DAT spectra of the 500 Hz square wave in the first frame, for $N=128$, and $M=N/2=64$. The later frames are similar. Compared with DFT spectrum, DAT spectrum is smoother, especially towards higher frequencies.

In the time domain representation (2.11), N times the inverse DFT of $X^*_{m,n}$ in n for each m plays the role of basis functions. Figure 3 shows such functions for $m=20$ and $m=40$, their time domain localization property reflects the smoother DAT spectrum.

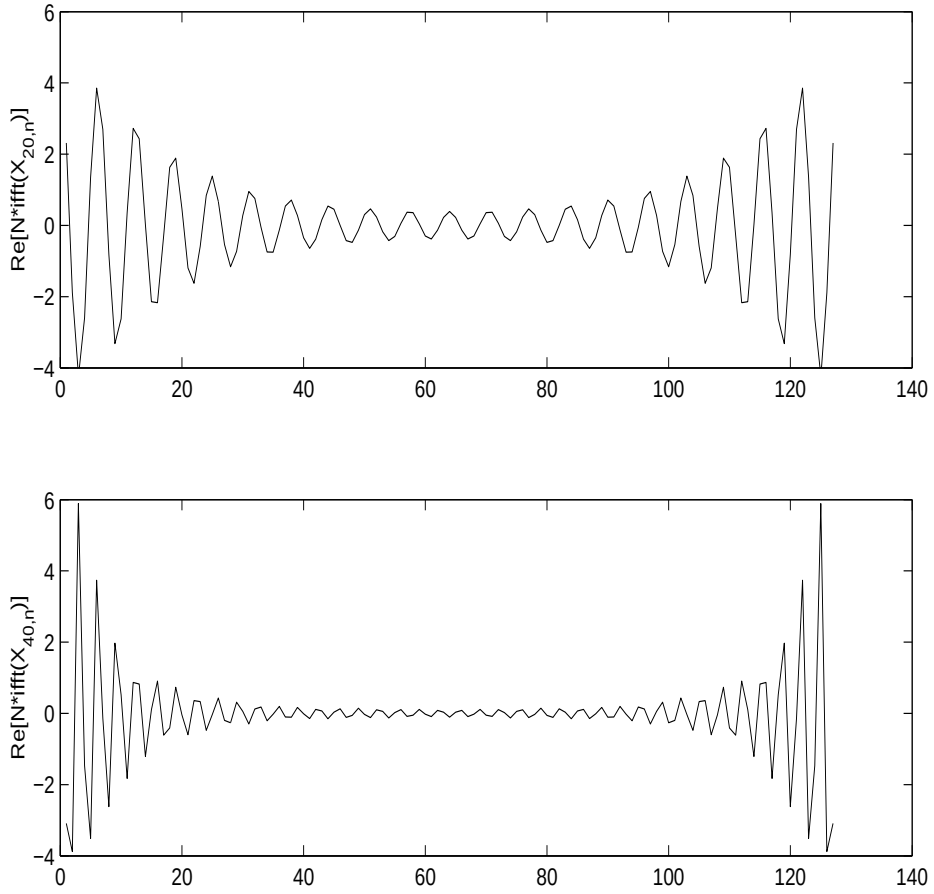

FIG. 3.3. *An illustration of localization property of DAT basis functions in the time domain.*

4. DAT and DFT in Signal Denoising

DAT and DFT were used to denoise speech signals. Both were numerically implemented with FFT. A simple thresholding method in the transformed domain was

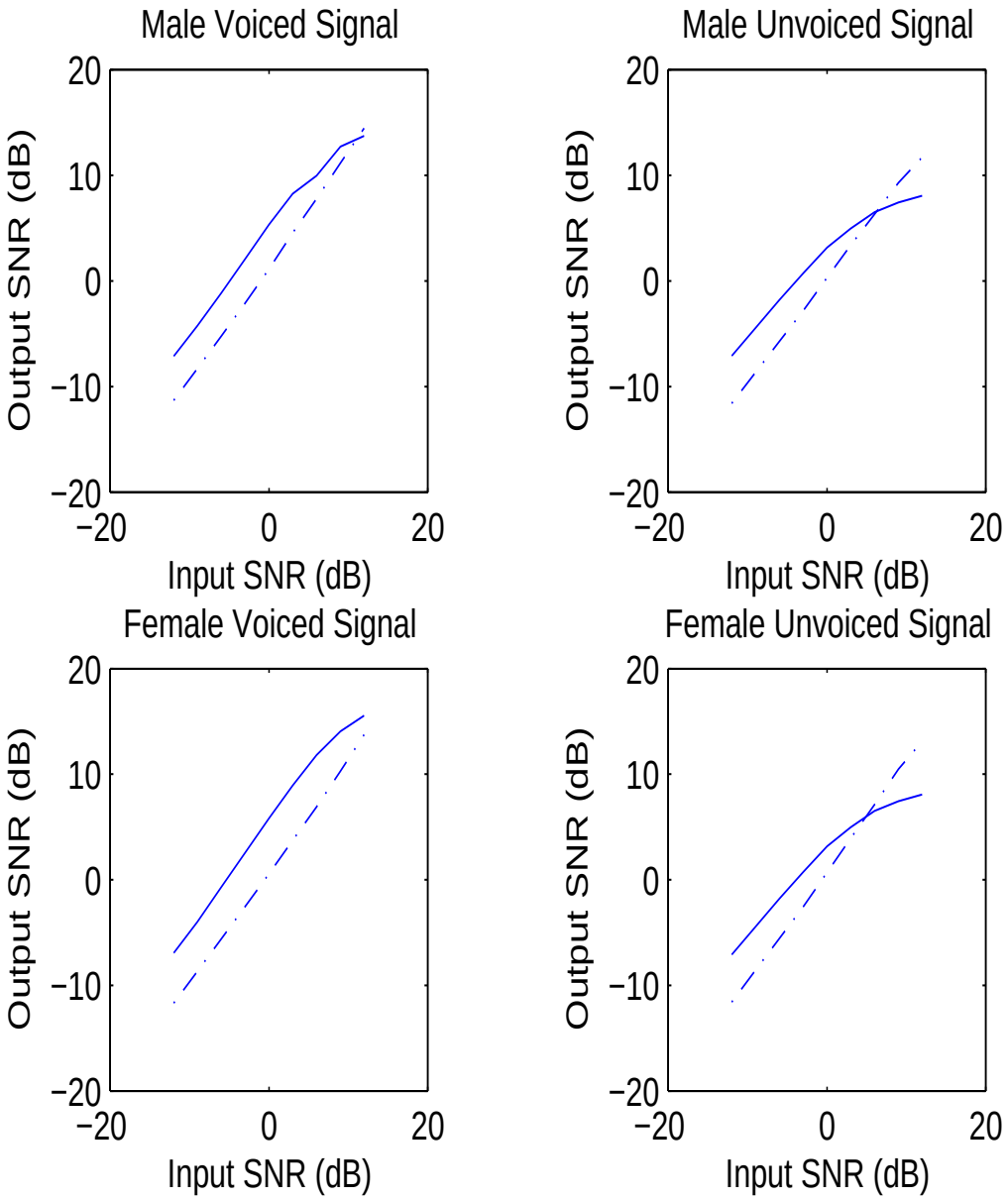

FIG. 3.4. *Comparison of DAT (solid) and DFT (dashdot) denoising by spectral thresholding for male/female, voiced/unvoiced speech segments.*

applied to improve the signal-to-noise ratio (SNR) of noisy speech. The underlying assumption of the method is that low level components in the transformed domain are more likely to be noise than signal plus noise. Thresholding, therefore, could improve

the overall SNR of the signal. It is a simple denoise method. More advanced methods exist for noise reduction and will be studied in the future. The simple thresholding method serves as a tool here to reveal the difference between DAT and DFT in signal processing.

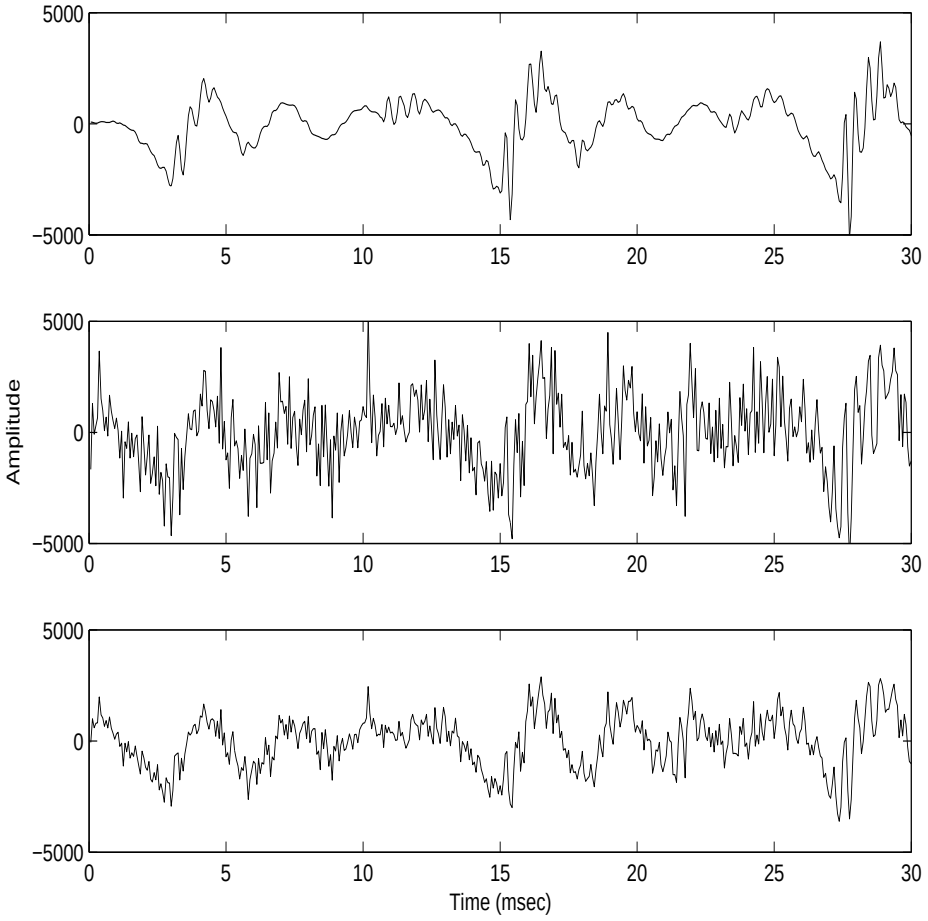

FIG. 4.1. *(Top down) voiced speech signal, noisy signal (SNR = 0 dB), denoised signal.*

Voiced and unvoiced speech segments were selected from a male and a female speaker respectively. Each segment has 512 data points. Noisy speech was created by adding Gaussian noise to the selected segments. The level of noise was set to produce the SNR ranging from -12 dB to +12 dB with a 3 dB step size. DAT and DFT were applied to the noisy speech signals. The magnitude of transformed components were then compared to a threshold. All components with magnitude smaller than the threshold were ignored for the reconstruction of the signal. Here, the threshold was computed as the average of the DFT magnitude spectrum. Signal was reconstructed directly by the inverse DAT and DFT, respectively. The SNRs of the reconstructed signal was computed and shown in Figure 4. Samples of original signal, its noise

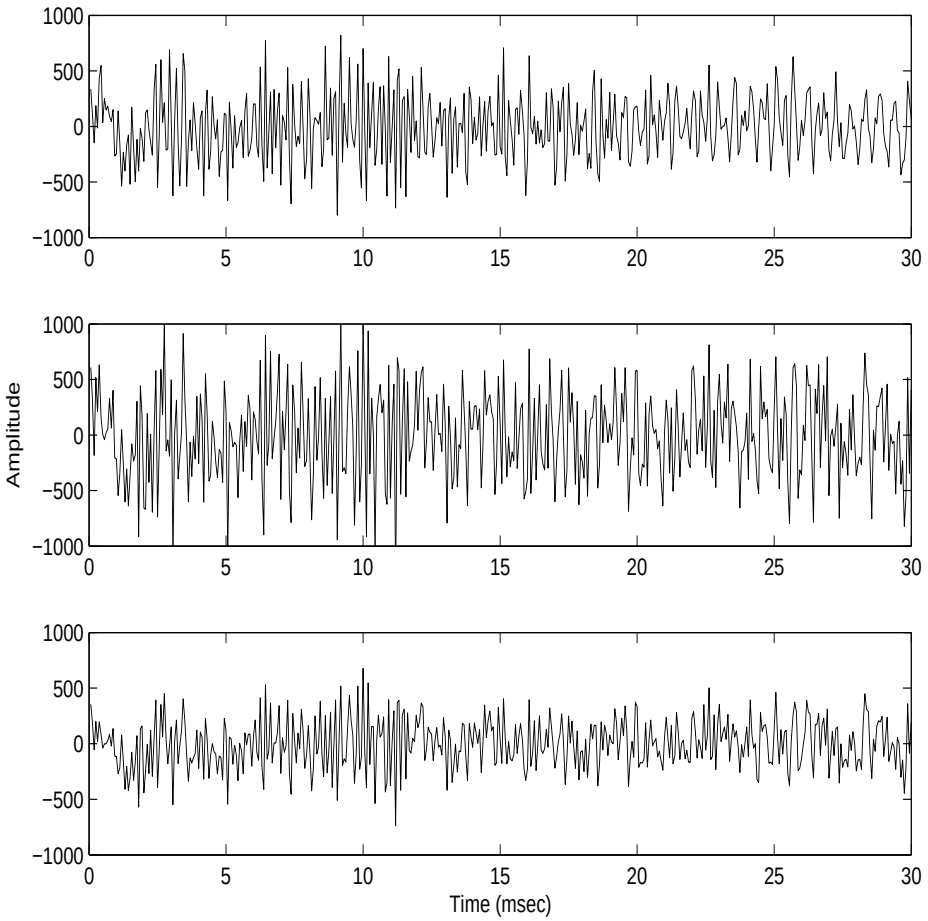

FIG. 4.2. *(Top down) unvoiced speech signal, noisy signal (SNR = 0 dB), denoised signal.*

added signal (SNR = 0 dB), and the signal denoised by thresholding were shown for a voiced (Figure 5) and unvoiced (Figure 6) speech segment.

These results indicate that DAT thresholding has about 3 to 5 dB SNR gain over DFT thresholding for voiced speech signals. The improvement is larger when the noise level is relatively high. For unvoiced (noise-like) speech signal, DAT thresholding also has a SNR gain when the noise level is relatively high. The DFT thresholding, however, has higher SNRs when noise level is relatively low. The DAT thresholding appears to have difficulty in discriminating between the original and added noise for unvoiced speech segments when the noise added is relatively low. This should not be as a handicap, in part because it may not be as necessary to denoise low-level noise when the speech itself is noise-like (see Figure 6). Denoise is mostly needed for voiced signals with high-level noise. The potential advantage of DAT thresholding is well demonstrated by the 3 to 5 dB improvement of SNR when noise level is relatively high. The noise-reduction advantage is likely a result of the spectral spreading (weighted

local spectral averaging) operation of DAT.

5. Concluding Remarks

Discrete auditory transforms (DAT) are introduced and shown to be invertible and energy preserving. DAT spectra are smoother than DFT's, and DAT basis functions are more localized than DFT's in the time domain. In signal denoising with the spectral thresholding method, it is observed that DAT increased SNR by 3 to 5 dB over DFT. Further study of DAT will be worthwhile in more complicated signal processing tasks.

Acknowledgements: This work was supported in part by NSF grant ITR-0219004, a Fellowship from the John Simon Guggenheim Memorial Foundation, and the Faculty Research Assignment Award at the University of Texas at Austin. We thank Prof. G. Papanicolaou and Prof. S-T Yau for their interest and helpful comments, and the anonymous referee for pointing out the connection of DAT with the Fejér kernel of Fourier analysis.

REFERENCES

[1] P. Brémaud, *Mathematical Principles of Signal Processing: Fourier and Wavelet Analysis*, Springer-Verlag, 2002.

[2] I. Debauchies, *Ten Lectures on Wavelets*, CMS-NSF Regional Conference in Applied Mathematics, SIAM, Philadelphia, 1992.

[3] W. M. Hartmann, *Signals, Sound, and Sensation*, Springer, 251-254, 2000.

[4] M. R. Schroeder, B. S. Atal and J. L. Hall, *Optimizing digital speech coders by exploiting properties of the human ear*, Journal Acoust. Soc. America, 66, 6, 1647-1652, 1979.

[5] A. Torchinsky, *Real-Variable Methods in Harmonic Analysis*, Chapters I and II, Academic Press, 1986.

[6] J. Xin and Y. Qi, *Global well-posedness and multi-tone solutions of a class of nonlinear nonlocal cochlear models in hearing*, Nonlinearity 17, 711-728, 2004.

[7] J. Xin, Y. Qi and L. Deng, *Time domain computation of a nonlinear nonlocal cochlear model with applications to multitone interaction in hearing*, Comm. Math. Sci., 1, 2, 211–227, 2003.