A Many to One Discrete Auditory Transform

Jack Xin * and Yingyong Qi*

Abstract

A many to one discrete auditory transform is presented to map a sound signal to a perceptually meaningful spectrum on the scale of human auditory filter band widths (critical bands). A generalized inverse is constructed in closed analytical form, preserving the band energy and band signal to noise ratio of the input sound signal. The forward and inverse transforms can be implemented in real time. Experiments on speech and music segments show that the inversion gives a perceptually equivalent though mathematically different sound from the input.

Keywords: Many to One Discrete Auditory Transform, Perceptual Inversion, Fast Computation.

AMS Subject Classification: 65T50, 94A12, 62P15.

*Department of Mathematics, UC Irvine, Irvine, CA 92697, USA. Email: jxin@math.uci.edu.

1 Introduction

Short term discrete Fourier transform (DFT) is a common tool to map sound signals from time domain to spectral domain for analysis and synthesis [1]. However, the spectral resolution of DFT over a standard short time window of 5 to 15 milleseconds (ms) in duration is much more refined than the resolution of human auditory filters that have band widths referred to as critical bands [4, 11]. Critical bands are nearly uniform in widths similar to DFT for frequencies under 500 Hz, yet the widths increase rapidly towards higher frequencies. The nonuniform frequency resolution of the ear resembles that of wavelets [2, 9], though critical band widths do not follow a simple power law, and auditory filter shapes may not obey the requirements of the wavelet basis functions. An orthogonal discrete transform with broader and smoother spectrum towards higher frequencies than that of DFT is recently constructed [10] to mimic the auditory filtering. Due to the limitation of orthogonality, the variation of the spectrum does not match the scale of critical bands. In addition, the spectrum of the transform does not carry enough perceptual meaning and so makes it inconvenient to perform psychoacoustically based spectral analysis and processing.

In this paper, we present a novel many-to-one discrete auditory transform (MDAT) that maps sound signals from the time domain to a perceptually meaningful spectral domain on the scale of critical bands. The many-to-one mapping is consistent with the fact that physically and mathematically different signals can sound the same to human ears [4, 11, 8]. The frequency resolution for the perception of sound in our brain is much lower than that is required to fully describe a signal mathematically [8]. The perception variables of MDAT are band energies and band signal to noise ratios (SNRs), motivated by perceptual coding in AAC (Advanced Audio Coding) and MP3 technology of digital music compression [5, 6]. The SNRs depend on two neighboring frames of a signal and so MDAT spectrum also encodes temporal information, different from DFT. As a test of the efficiency of these variables, and for the synthesis of sounds post spectral processing, we show how to construct an inverse which is perceptually equivalent to the input sound though mathematically not identical. Both the forward and inverse operations are in closed analytical form, and allow real time implementation of the resulting algorithms.

Compared with DFT which is implemented by FFT (Fast Fourier Trans-

form), MDAT has better temporal resolution due to its lower spectral resolution in higher frequencies. In terms of the 256 point FFT used in this paper, the number of frequency bands of MDAT is in the 40's (see Tables 1 and 2 for signals with different sampling frequencies) while DFT has 128 frequency components. Compared to time domain filter bank with a relatively small number of band pass filters (4 to 16 channels) as in body-worn hearing devices [3], MDAT has better frequency resolution yet does not have the delays encountered when a larger number of frequency separating band pass filters are needed. Hence MDAT is expected to be a useful tool in applications where a spectral processing strategy is necessary on the critical band scale, and a trade-off of spectral accuracy and temporal precision is to be optimized.

The paper is organized as follows. First, the MDAT is formulated and the associated perceptual variables are defined. A perceptually equivalent inverse is constructed in closed analytical form based on band energies and SNRs. Then MDAT is applied to speech (sampled at 16 kHz) and music (sampled at 44.1 kHz) signals, and properties of perceptual spectral variables are illustrated. The reconstructed signals are compared with the input signals both spectrally and in waveforms. The sound files of these signals are available on line at www.math.uci.edu/∼jxin/sounds.html.

2 MDAT and Perceptual Inversion

2.1 Transform from Sound to Perception

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

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

We shall refer to the $k = 0$ component of DFT as DC (direct current) and the other components as AC (alternating current) for short.

Let us further map the $\hat{s}_k$'s to a spectral domain of lower resolution where perception variables can be better defined. Such a spectral domain is obtained from binning the DFT components into bands of various widths, similar to the critical band width distribution of human auditory filters. The

detailed partition of DFT components, the band widths, and psychoacoustic bark values of the bands are listed in Table 1 and Table 2. Table 1 is at sampling frequency $Fs = 16$ kHz for speech sounds, and Table 2 is at $Fs = 44.1$ kHz for music sounds. Let $b = 1, 2, \cdots, J$, denote the number of bands, and let $B(b)$ denote the DFT wave numbers k in the b-th band. In case of Table 1, $N/J \approx 5.56$; and in Table 2, $N/J \approx 6.24$.

The signal energy in the b-th band is:

$$e(b) = \sum_{k \in B(b)} |\hat{s}_k|^2. \tag{2}$$

Let snr^b be the signal to noise ratio (SNR) in the b-th band, the perception domain consists of nonnegative $2J$-dimensional vectors whose components are band energies and band SNRs:

$$V_{perc} = \{(e(1), snr^1, e(2), snr^2, \cdots, e(J), snr^J)\}. \tag{3}$$

The snr^b are calculated following the AAC coding [5], an improvement of MP3 coding [6]. Let $r(k,t)$ and $f(k,t)$ be the amplitude and phase of $\hat{s}(k)$ at time frame t denoted by $\hat{s}(k,t)$. The predicted amplitude and phase at time frame t are:

$$\begin{aligned} r_{pred}(k,t) &= r(k,t-1) + \Delta r,\ \Delta r \equiv r(k,t-1) - r(k,t-2), \\ f_{pred}(k,t) &= f(k,t-1) + \Delta f,\ \Delta f \equiv f(k,t-1) - f(k,t-2). \end{aligned} \tag{4}$$

The unpredictability measure of the signal, a quantity for measuring the noisy (uncertain) part of signal, is:

$$c(k,t) = \frac{\text{abs}(\hat{s}(k,t) - \hat{s}_{pred}(k,t))}{\text{abs}(\hat{s}(k,t)) + \text{abs}(\hat{s}_{pred}(k,t))}, \tag{5}$$

where $\hat{s}_{pred}(k,t) = r_{pred}(t)\, e^{if_{pred}(t)}$. It is clear that $c(k,t) \in [0,1]$. Note that $c(k,t)$ encodes the time domain information of the signal s, which is not available in DFT. As a result, the perceptual variables (3) has both spectral and temporal information of the input signal. We shall omit the t dependence from now on, as all subsequent operations will not explicitly use t.

The weighted unpredictability measure is:

$$ec(b) = \sum_{k \in B(b)} r^2(k)\, c(k). \tag{6}$$

Next, convolve $e(b)$ and $ec(b)$ with spreading functions [8] on the bark scale [4] as:

$$ecb(b) = \sum_{b'=1}^{J} e(b') \,\text{spread}\,(\text{bark}\,(b'), \text{bark}\,(b)), \tag{7}$$

$$ct(b) = \sum_{b'=1}^{J} ec(b') \,\text{spread}\,(\text{bark}\,(b'), \text{bark}\,(b)), \tag{8}$$

where bark(b) is the bark value of the b-th partition (band). The bark scale [4] is nearly uniform on the logarithmic frequency scale. The spreading functions [8] carry the shape information of human auditory filters.

Normalizing ct by energy ecb gives:

$$cb(b) = ct(b)/ecb(b), \tag{9}$$

a noise to signal ratio, which in turn defines tonality index as:

$$tb(b) = -0.299 - 0.43\,\log(cb(b)), \tag{10}$$

if the value is in $(0, 1)$, otherwise equal to zero if the value is below zero, or one if the value is above 1. Finally, the signal to noise ratio in decibel (dB) is:

$$snr^b = tb(b)\,\text{TMN} + (1 - tb(b))\,\text{NMT}, \tag{11}$$

where TMN = 18 dB (tone masking noise), NMT = 6 dB (noise masking tone). The forward transform denoted by T from signal s to its image in the perception domain V_{perc} is a many-to-one mapping. Clearly, $T(-s) = Ts$.

We notice that each snr^b is a monotone function of $cb(b)$ which in turn depends on $ec(b)$ and $ct(b)$. So two other ways of characterizing the perception domain are:

$$V_{perc}^{(1)} = \{(e(1), cb(1), e(2), cb(2), \cdots, e(J), cb(J))\}, \tag{12}$$

$$V_{perc}^{(2)} = \{(e(1), ec(1), e(2), ec(2), \cdots, e(J), ec(J))\}. \tag{13}$$

In other words, $V_{perc}^{(1)}$ or $V_{perc}^{(2)}$ is sufficient to describe the perception variables, i.e. the band energies and band SNRs. Below we show how to reconstruct a sound signal from $V_{perc}^{(1)}$ or $V_{perc}^{(2)}$ and obtain a perceptually equivalent inverse.

2.2 Inversion

The inversion from a subset of $2J$ dimensional space to the signal space R^N ($N > 2J$) is non-unique. The inversion is through reconstructing the DFT vector $\hat{s}_k$. Let us write the reconstructed DFT vector as:

$$a_k^b = w_k^b \, e^{1/2}(b) \, e^{i\varphi_k^b}, \;\; k \in B(b), \tag{14}$$

where the real weighting factors w_k^b satisfy for all b:

$$\sum_{k \in B(b)} |w_k^b|^2 = 1, \tag{15}$$

to preserve the band energy $e(b)$. The real phase factors φ_k^b, and the DC component of DFT are assumed to be known for the reconstruction of the AC part of the DFT amplitude.

The second conserved quantity (constraint) is $ec(b)$ in (6):

$$ec(b) = \sum_{k \in B(b)} |w_k^b|^2 \, e(b) \, c(k) = e(b) \sum_{k \in B(b)} |w_k^b|^2 \, c(k). \tag{16}$$

Define:

$$< w^b >_c^2 = \sum_{k \in B(b)} |w_k^b|^2 \, c(k), \tag{17}$$

which equals

$$< w^b >_c^2 = ec(b)/e(b) \in (\min_{k \in B(b)} \; c(k), \max_{k \in B(b)} \; c(k)) \subset [0, 1]. \tag{18}$$

If the inversion is from $V_{perc}^{(2)}$, then the two spectral constraints (15) and (17) are available to be imposed in each band containing at least two DFT components. If the inversion is from $V_{perp}^{(1)}$, then $< w^b >_c^2$ has to be recovered from $e(b)$ and $cb(b)$. By (9), we have for each $b \in [1, J]$:

$$cb(b) = \frac{\sum_{b'=1}^{J} \; e(b') \; < w^{b'} >_c^2 \; \text{spread}(\text{bark}(b'), \text{bark}(b))}{\sum_{b'=1}^{J} \; e(b') \, \text{spread}(\text{bark}(b'), \text{bark}(b))}, \tag{19}$$

or

$$\begin{aligned} & \sum_{b'=1}^{J} e(b') \; < w^{b'} >_c^2 \; \text{spread}(\text{bark}(b'), \text{bark}(b)) \\ = \;\; & cb(b) \sum_{b'=1}^{J} e(b') \, \text{spread}(\text{bark}(b'), \text{bark}(b)). \end{aligned} \tag{20}$$

Equation (20) can be recast as a matrix equation $S\vec{x} = \vec{z}$, where $S = (\text{spread}(\text{bark}(b'), \text{bark}(b))$ is a square matrix, $\vec{x}$ is the column vector with entries $e(b) < w^b >_c^2$, $\vec{z}$ the right hand side column vector. The commonly used spreading matrix S (based on e.g. Schroeder's spreading functions [8]) does not have a nonnegative inverse. In order to find nonnegative solutions in general, one may solve a quadratic programming problem from (19). Define the matrix $Q = (q_{ij})$ with its entries:

$$q_{ij} = \frac{e(j)\,\text{spread}(\text{bark}(j), \text{bark}(i))}{\sum_{j=1}^{J} e(j)\,\text{spread}(\text{bark}(j), \text{bark}(i))}.$$

The matrix Q is invertible. A column vector $\vec{y} = (< w^b >_c^2)$ is sought to minimize the l^2 norm $\|\vec{cb} - Q\vec{y}\|_2$ subject to the constraint $yl(b) \leq y(b) \leq yu(b)$, $yl(b) = \min_{k \in B(b)}\, c(k)$, $yu(b) = \max_{k \in B(b)}\, c(k)$.

In signal processing tasks that keep the band SNRs invariant as in hearing aids gain prescriptions, the quadratic programming is not needed, directly inverting S will suffice to find $(< w^b >_c^2)$.

Next we solve for w_k^b from the two equations (15) and (17), using information of $c(k)$, $k \in B(b)$. Let N_b be the number of DFT components in $B(b)$, $\vec{\rho} = (|w_{k_1}^b|^2, |w_{k_2}^b|^2, \cdots, |w_{k_{N_b}}^b|^2)^T$, $\vec{\psi} = (c(k_1), c(k_2), \cdots, c(k_{N_b}))^T$, $k_j \in B(b)$, $\theta_b =< w^b >_c^2$, $\vec{e} = (1, 1, \cdots, 1)^T \in R^{N_b}$, T denoting transpose. Equations (15) and (17) now read (dot refers to inner product):

$$\vec{e} \cdot\, \vec{\rho} \,=\, 1, \tag{21}$$

$$\vec{\psi} \cdot\, \vec{\rho} \,=\, \theta_b. \tag{22}$$

If $\vec{\psi}$ is parallel to $\vec{e}$, equation (22) is redundant with $\theta_b = c(1)$ by definition and equation (21). This is true in particular if $N_b = 1$. The simplest smooth solution to (21) is $\vec{\rho} = \frac{1}{N_b}\vec{e}$.

If $N_b \geq 2$ and $\vec{\psi}$ is not parallel to $\vec{e}$, define vector:

$$\vec{v} = \vec{e} - \frac{\vec{e} \cdot \vec{e}}{\vec{e} \cdot \vec{\psi}}\,\vec{\psi} \neq 0, \tag{23}$$

clearly $\vec{v} \cdot \vec{e} = 0$, and $\vec{e} \cdot \vec{\psi} > 0$. Equations (21) and (22) imply that:

$$\vec{v} \cdot \vec{\rho} = 1 - \frac{\vec{e} \cdot \vec{e}}{\vec{e} \cdot \vec{\psi}}\,\theta_b. \tag{24}$$

Equation (21) and equation (24) say that in the orthonormal basis with $\vec{e}$ and $\vec{v}$ as two directions, the coordinates along $\vec{e}$ and $\vec{v}$ are constrained, the other coordinates are free. The simplest two dimensional solution is obtained by setting the free coordinates to zero ($\|\cdot\|_2$, l^2 norm or the Euclidean distance):

$$\vec{\rho} = \frac{1}{\sqrt{N_b}}\frac{\vec{e}}{\sqrt{N_b}} + \frac{1}{\|\vec{v}\|_2}(1 - \frac{\vec{e}\cdot\vec{e}}{\vec{e}\cdot\vec{\psi}}\theta_b)\frac{\vec{v}}{\|\vec{v}\|_2}, \tag{25}$$

which becomes upon substituting in (23):

$$\vec{\rho} = \left[\frac{1}{N_b} + \frac{1}{\|\vec{v}\|_2^2}\left(1 - \frac{\vec{e}\cdot\vec{e}}{\vec{e}\cdot\vec{\psi}}\theta_b\right)\right]\vec{e} - \frac{\vec{e}\cdot\vec{e}}{\vec{e}\cdot\vec{\psi}}\frac{1}{\|\vec{v}\|_2^2}\left(1 - \frac{\vec{e}\cdot\vec{e}}{\vec{e}\cdot\vec{\psi}}\theta_b\right)\vec{\psi}. \tag{26}$$

The regularity of solution (25) or (26) is no worse than that of $\vec{\psi}$ which is oscillatory in general. With the w_k^b's so determined, a time domain signal is reconstructed by inverse DFT using the reconstructed a_k^b, $k \in B(b)$, $b = 1, 2, \cdots, J$.

If $N_b = 2$, (26) is the unique solution. If $N_b \geq 3$ (true if b is above some critical number, see Table 1 and Table 2), there are infinitely many solutions to (21)-(22). It is desirable to seek a smoother solution because spectral smoothness improves temporal localization of the inverse transform. One way to obtain a smoother solution over the frequency bands ($N_b \geq 3$) starting with FFT wave number k_0 is to minimize the following quadratic function:

$$f = \frac{1}{2}\sum_{k=k_0}^{k_0+M-1}(\rho_{k+1} - \rho_k)^2, \tag{27}$$

where $M+1$ is the total number of DFT components in those bands $B(b)$ with $N_b \geq 3$, subject to the two constraints (21)-(22) in each such band $B(b)$. Let $\vec{u} = (\rho_{k_0}, \cdots, \rho_{k_0+M})^T$, and define:

$$g_b(\vec{u}) = -1 + \sum_{k\in B(b)}\rho_k, \tag{28}$$

$$h_b(\vec{u}) = -\theta_b + \sum_{k\in B(b)} c_k\,\rho_k, \tag{29}$$

then the constraints are of the form $g_b = 0$ and $h_b = 0$. The minimizer can be approached as a steady state in a constrained gradient descent method [7].

2.3 Solutions to Constrained Gradiant Descent

The constrained gradient descent method of optimization is analyzed below to yield a closed form solution. Let $\vec{u} = \vec{u}(\tau)$ solve the equation:

$$\vec{u}_\tau = -\nabla_{\vec{u}} f - \sum_{b, N_b \geq 3} \lambda_b \nabla_{\vec{u}} g_b - \sum_{b, N_b \geq 3} \eta_b \nabla_{\vec{u}} h_b, \tag{30}$$

where the Lagrange multipliers λ_b and η_b are chosen so that the constraint values in each band are preserved in τ:

$$\begin{aligned} \frac{d}{d\tau} g_b(\vec{u}) &= \nabla_{\vec{u}} g_b \cdot \vec{u}_t \\ &= -\nabla_{\vec{u}} g_b \cdot \nabla_{\vec{u}} f - \lambda_b |\nabla_{\vec{u}} g_b|^2 - \eta_b \nabla_{\vec{u}} g_b \cdot \nabla_{\vec{u}} h_b = 0, \end{aligned} \tag{31}$$

$$\begin{aligned} \frac{d}{d\tau} h_b(\vec{u}) &= \nabla_{\vec{u}} h_b \cdot \vec{u}_t \\ &= -\nabla_{\vec{u}} h_b \cdot \nabla_{\vec{u}} f - \lambda_b \nabla_{\vec{u}} h_b \cdot \nabla_{\vec{u}} g_b - \eta_b |\nabla_{\vec{u}} h_b|^2 = 0. \end{aligned} \tag{32}$$

We have used the fact that $\nabla_{\vec{u}} h_b$ or $\nabla_{\vec{u}} g_b$ only have nonzero components in the band $B(b)$. To solve (31)-(32) band by band, it is convenient to consider

$$\tilde{c}_k = 1 - \frac{c_k N_b}{\sum_{j \in B(b)} c_j}, \ k \in B(b). \tag{33}$$

If $\tilde{c}_k = 0$, for all $k \in B(b)$, then the second constraint $h_b = 0$ is redundant, $\eta_b = 0$, and

$$\lambda_b = -\frac{\nabla_{\vec{u}} g_b \cdot \nabla_{\vec{u}} f}{|\nabla_{\vec{u}} g_b|^2}. \tag{34}$$

If $\tilde{c}_k \neq 0$, for some $k \in B(b)$, replace the constraint $h_b = 0$ by:

$$\tilde{h}_b(\vec{u}) = \sum_{k \in B(b)} \tilde{c}_k \rho_k - 1 + \frac{N_b < w^b >_c^2}{\sum_{k \in B(b)} c_k} = 0. \tag{35}$$

Then the $\vec{u}$ equation is (30) with $\tilde{h}_b$ in place of h_b. Due to $\nabla_{\vec{u}} g_b \cdot \nabla_{\vec{u}} \tilde{h}_b = 0$, λ_b is as given in (34), and:

$$\eta_b = -\frac{\nabla_{\vec{u}} \tilde{h}_b \cdot \nabla_{\vec{u}} f}{|\nabla_{\vec{u}} \tilde{h}_b|^2}, \tag{36}$$

where $|\nabla_{\vec{u}}\tilde{h}_b|^2 = \sum_{k\in B(b)} \tilde{c}_k^2$, and $|\nabla_{\vec{u}} g_b|^2 = N_b$ in (34).

Finally, let us put the $\vec{u}$ equation in matrix form. Let A be the symmetric tridiagonal matrix with 1's on the off-diagonals, and $(-1, -2, \cdots, -2, -1)$ on the diagonal ($\cdots$ refer to -2's), then $\nabla_{\vec{u}} f = A\vec{u}$. Let R be the block diagonal matrix where each block is the symmetric $N_b \times N_b$ matrix with the (i, j)-th entry being $N_b^{-1} + \frac{\tilde{c}_i \tilde{c}_j}{\sum_{k\in B(b)} \tilde{c}_k^2}$. If $\sum_{k\in B(b)} \tilde{c}_k^2$ is zero, the second term in the sum is understood to be absent. The matrix form of $\vec{u}$ equation is (I the identity matrix): $\vec{u}_\tau = (I - R)A\vec{u}$, whose solution is in closed form $\vec{u}(\tau) = \exp\{(I - R)A\tau\}\vec{u}_0$. The initial data $\vec{u}_0$ is given by the values of ρ_{k_0}, $\cdots$, ρ_{k_0+M} in the explicit formula (26).

2.4 Computing and Experiments

The forward and inverse transforms are implemented with the 256 point FFT. For speech signals, Table 1 is used at sampling frequency 16 kHz. For music signals, Table 2 is used at sampling frequency 44.1 kHz. Top (bottom) panel of Fig. 1 shows the oscillatory unpredictability measure $c(k)$ of a speech (music) frame. Top (bottom) panel of Fig. 2 is the corresponding weighted unpredictability measure $ec(b)$ for the speech (music) frame, oscillation is slower over the coarser scale b. In Fig. 3 (Fig. 4), we compare the original and reconstructed FFT amplitude spectra ($k \in [20, 128]$) of a speech (music) frame. The difference is negligible for $k \in [0, 20]$. We see that the reconstructed FFT spectra captured well the upper envelope of the original FFT spectra of the speech frame. For the music frame, much more details of the FFT spectra are recovered. Except for a mismatched peak and a valley over $k \in [20, 40]$, the dashed and solid curves nearly agree. If one zooms in further, one may see differences over smaller scales yet the reconstructed (dashed) curve again keeps track of the envelope of the original spectral shape well. Fig. 5 compares the smoother spectral solution ($\tau = 2$, dashed) with the simple solution ($\tau = 0$, solid) in case of a speech frame over $k \in [30, 128]$ where constrained optimization (smoothing) takes place. The steady state is almost approached at $\tau = 2$. The smoothing is similar for music frames.

Fig. 6 (Fig. 7) compares the original and reconstructed speech (music) waveforms. The total relative l^2 error for the speech signal in Fig. 6 is 12 %, and is only 1.5% for music signal of Fig. 7. This is consistent with the

better spectral fit of Fig. 4 than that of Fig. 3. The improvement by the optimization (27)-(29) is however found to be minor both in terms of the relative l^2 error of reconstructed signals and perceptual difference in hearing the signals. The optimization step may be helpful however in other signal processing tasks to be evaluated in the future.

The original and reconstructed ($\tau = 0$) speech (music) signals in Fig. 6 and Fig. 7 can be heard from sound files (www.math.uci.edu/∼jxin/sounds.html). Inspite of the errors (loss) incurred in the reconstruction, there is little perceptual difference between the original and the reconstructed signals, thanks to the masking effects present in the human ears [8]. Hence we have achieved the perceptually equivalent inversion of the many-to-one transform.

3 Discussion and Conclusion

A many-to-one auditory transform is introduced so that the resulting spectrum, especially towards the higher frequency regime, is much less refined than the FFT spectrum, yet just enough to resolve the band widths of human auditory filters (critical bands). A reconstruction of perceptually equivalent inverse is given so that the inverted signal makes little perceptual difference from the input signal even though there is a loss mathematically. The inversion preserves the band energies and band signal to noise ratios, which prove to be essential in capturing the perception of sounds. Both the forward and inverse transforms are in closed analytical form and can be carried out in real time. Test examples on speech and music signals illustrated the properties of the transform and its inversion. The transform is a promising new tool for sound compensation or enhancement that requires spectral manipulations over the scale of critical bands.

A future study may concern with inversion when there is loss in phase information of FFT spectra. Phase is stochastic in nature, and challenging to reduce its dimensions. Phase is related to sound quality, though less is known about it than amplitude. Another is to further develop MDAT in specific applications such as hearing aids and hearing implants.

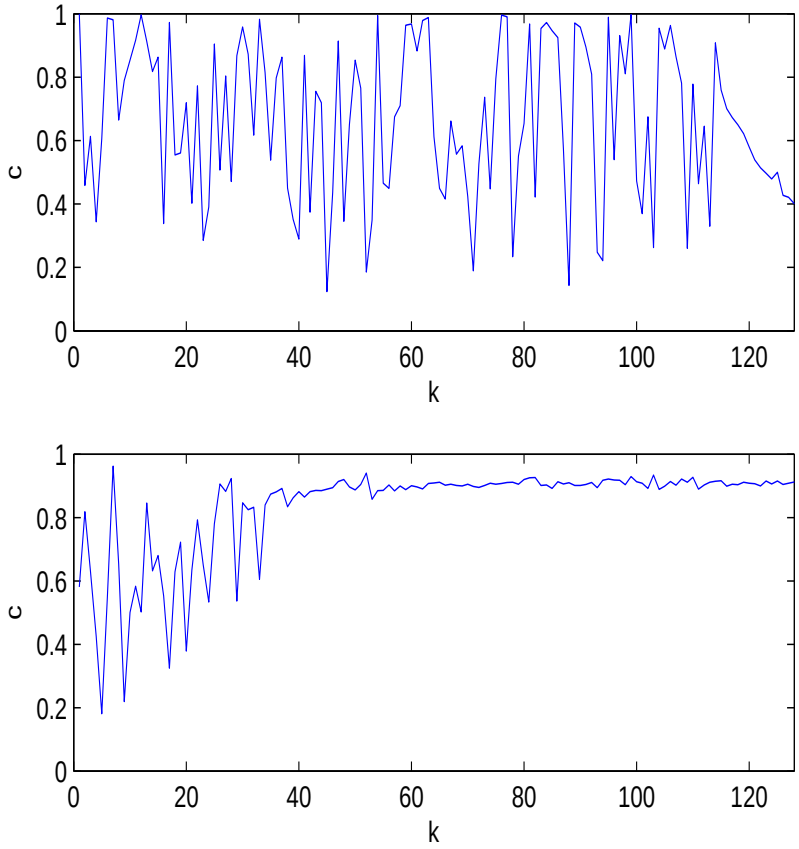

Figure 1: Top panel: unpredicatibility measure $c(k)$ of a speech frame, illustrating its oscillatory nature in FFT wave number k, $k \in [0, 128]$. Bottom panel: unpredicatibility measure $c(k)$ of a music frame, $k \in [0, 128]$.

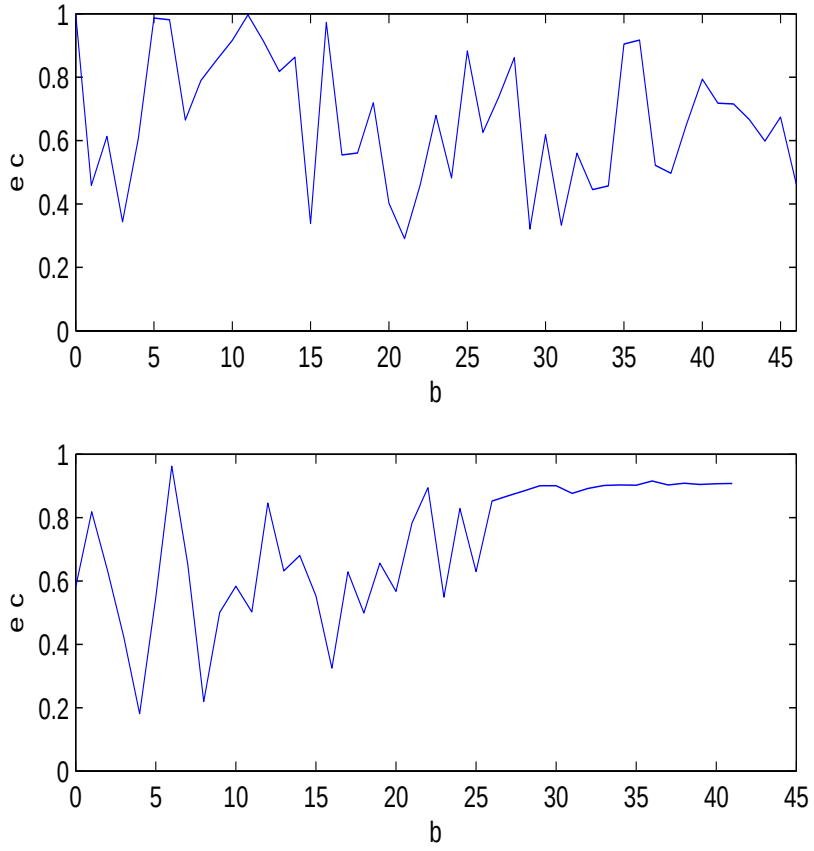

Figure 2: Top panel: weighted unpredicatibility measure $ec(b)$ of a speech frame, $b \in [0, 46]$. Bottom panel: weighted unpredicatibility measure $ec(b)$ of a music frame, $b \in [0, 41]$.

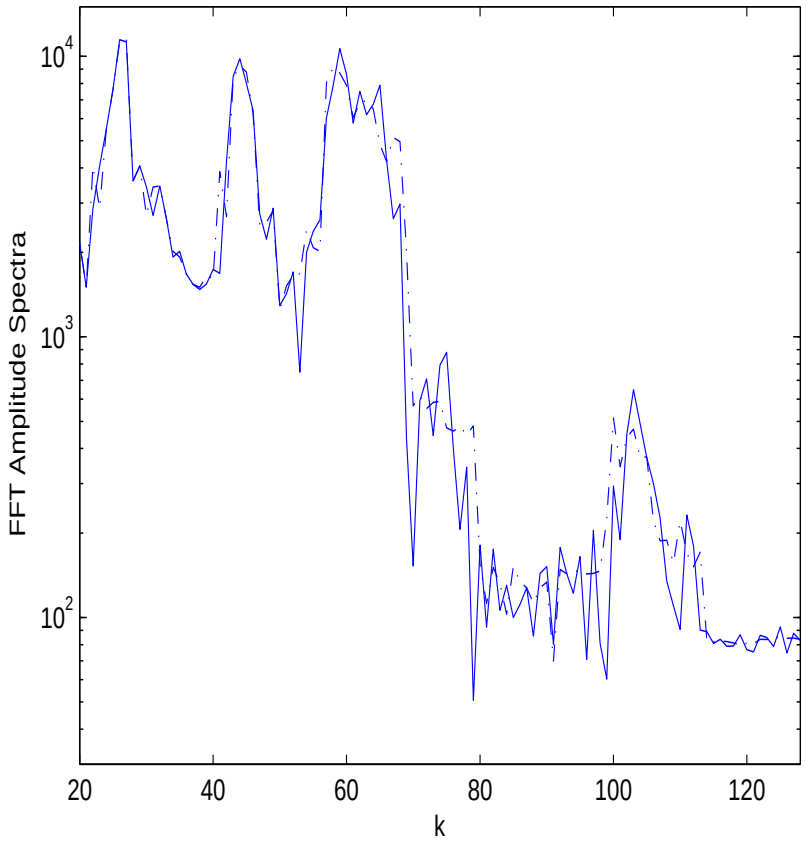

Figure 3: Original (solid) and reconstructed (dashed, $\tau = 0$) FFT amplitude spectra of a speech frame.

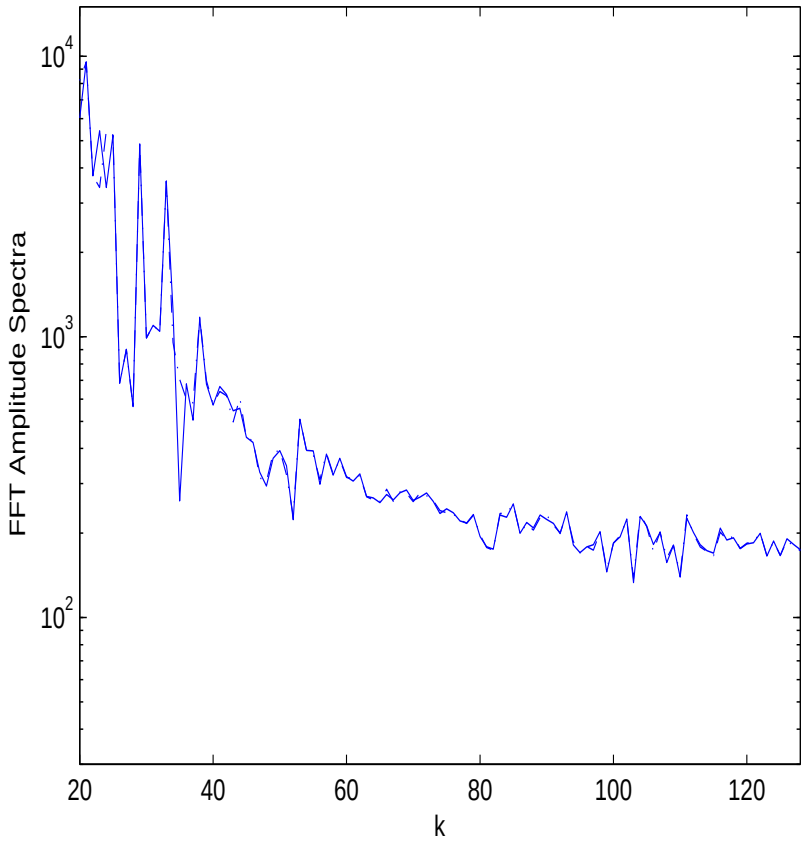

Figure 4: Original (solid) and reconstructed (dashed, $\tau = 0$) FFT amplitude spectra of a music frame.

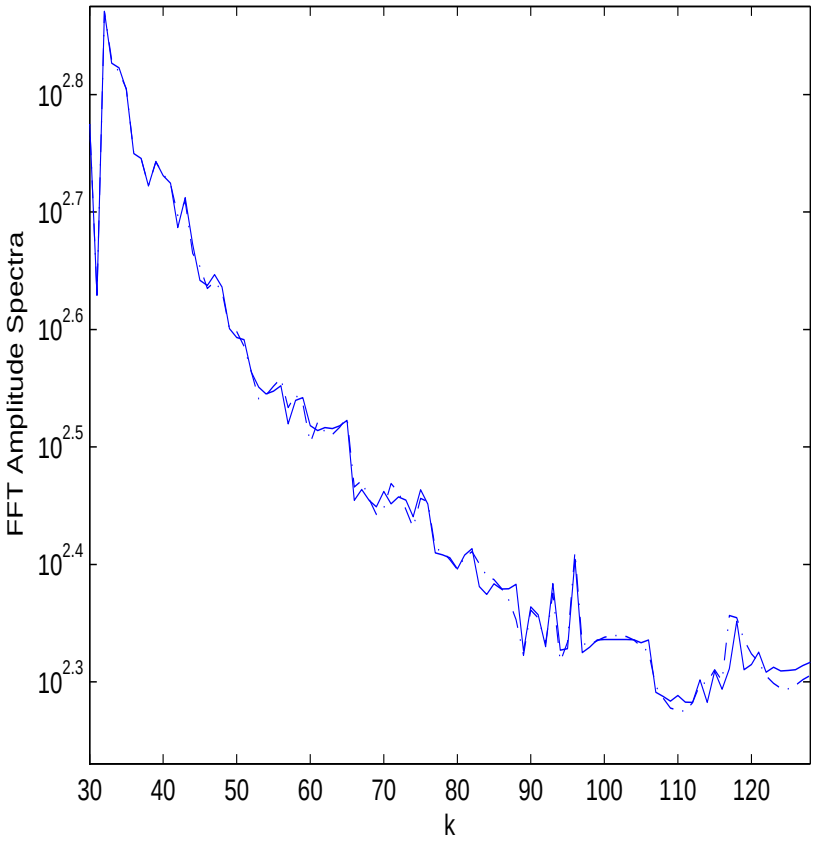

Figure 5: Comparison of reconstructed (solid, $\tau = 0$) and (dashed, $\tau = 2$) FFT amplitude spectra of a speech frame. The dashed curve is smoother while satisfying the same spectral constraints.

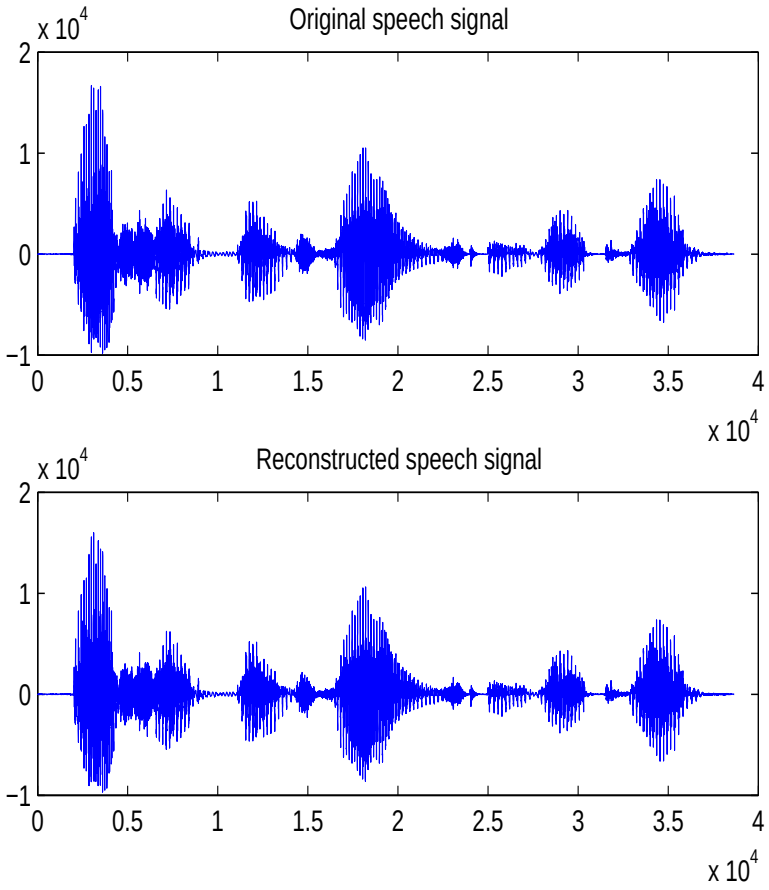

Figure 6: Comparison of the input (top) and reconstructed (bottom, $\tau = 0$) speech signals in waveforms.

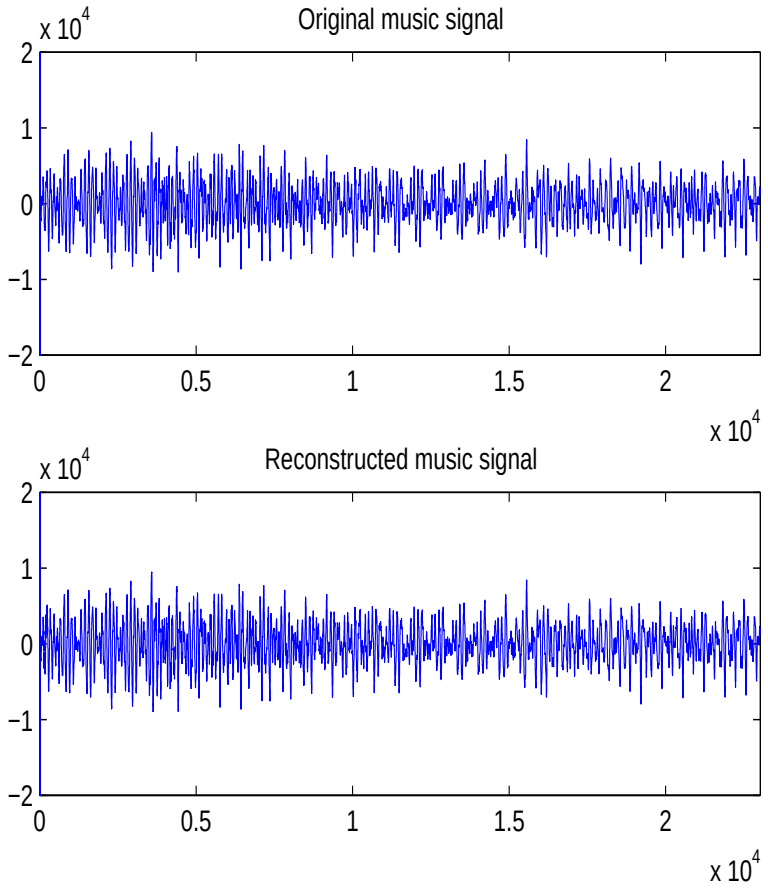

Figure 7: Comparison of the input (top) and reconstructed (bottom, $\tau = 0$) music signals in waveforms.

4 Acknowledgements

We thank G. Papanicolaou and H-K Zhao for helpful comments. This work was supported in part by National Science Foundation Information Technology Research Grant ITR-0219004, and National Institute of Health Grant 2R43DC005678-02A1 on psychoacoustic-based voice quality assessment.

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] S. Greenberg, W. Ainsworth, A. Popper, R. Fay, eds, "Speech Processing in the Auditory System", Springer Handbook of Auditory Research, Springer, 2004.

[4] W. Hartmann, "Signals, Sound, and Sensation", Springer, 2000, pp 251-254.

[5] International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), "Psychoacoustic model for AAC encoder", ISO/IEC 14496-3:2001(E), pp 269-292, 2001.

[6] K. Pohlmann, "Principles of Digital Audio", 4th edition, McGraw-Hill Video/Audio Professional, 2000.

[7] S. Osher, R. Fedkiw, "Level Set Methods and Dynamic Implicit Surfaces", Chapter III, Applied Mathematical Sciences 153, Springer, 2003.

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

[9] G. Strang, T. Nguyen, "Wavelets and Filter Banks", Wesley-Cambridge Press, 1997.

[10] J. Xin and Y. Qi, *An Orthogonal Discrete Auditory Transform*, Comm. Math. Sciences, Vol. 3, No. 2, pp 251-259, 2005.

[11] E. Zwicker, H. Fastl, "Psychoacoustics: Facts and Models", Springer Series in Information Sciences, 22, 2nd edition, 1999.

Table 1: Partition and psychoacoustic parameters for the 256 point FFT at 16 kHz sampling frequency. The columns are (from left to right) band index, low FFT index of the band, high FFT index of the band, number of FFT components in the band (width), the bark value of the band. The symmetric part of the AC components of FFT are not listed. Zero index refers to DC component of FFT.

Band Index	Low FFT Index	High FFT Index	Width	Bark Value
0	0	0	1	0
1	1	1	1	0.63
2	2	2	1	1.26
3	3	3	1	1.88
4	4	4	1	2.50
5	5	5	1	3.11
6	6	6	1	3.70
7	7	7	1	4.28
8	8	8	1	4.85
9	9	9	1	5.39
10	10	10	1	5.92
11	11	11	1	6.43
12	12	12	1	6.93
13	13	13	1	7.40
14	14	14	1	7.85
15	15	15	1	8.29
16	16	16	1	8.70
17	17	17	1	9.10
18	18	18	1	9.49
19	19	19	1	9.85
20	20	20	1	10.20
21	21	22	2	10.85
22	23	24	2	11.44
23	25	26	2	11.99
24	27	28	2	12.50
25	29	30	2	12.96
26	31	32	2	13.39
27	33	34	2	13.78

Table 1: Continued.

Band Index	Low FFT Index	High FFT Index	Width	Bark Value
28	35	36	2	14.15
29	37	39	3	14.57
30	40	42	3	15.03
31	43	45	3	15.45
32	46	48	3	15.84
33	49	51	3	16.19
34	52	55	4	16.57
35	56	59	4	16.97
36	60	63	4	17.33
37	64	68	5	17.71
38	69	73	5	18.09
39	74	78	5	18.44
40	79	84	6	18.80
41	85	90	6	19.17
42	91	97	7	19.53
43	98	104	7	19.89
44	105	112	8	20.25
45	113	120	8	20.61
46	121	127	7	20.92

Table 2: Partition and psychoacoustic parameters for the 256 point FFT at 44.1 kHz sampling frequency. The columns are (from left to right) band index, low FFT index of the band, high FFT index of the band, number of FFT components in the band (width), the bark value of the band. The symmetric part of the AC components of FFT are not listed. Zero index refers to DC component of FFT.

Band Index	Low FFT Index	High FFT Index	Width	Bark Value
0	0	0	1	0
1	1	1	1	1.73
2	2	2	1	3.41
3	3	3	1	4.99
4	4	4	1	6.45
5	5	5	1	7.75
6	6	6	1	8.92
7	7	7	1	9.96
8	8	8	1	10.87
9	9	9	1	11.68
10	10	10	1	12.39
11	11	11	1	13.03
12	12	12	1	13.61
13	13	13	1	14.12
14	14	14	1	14.59
15	15	15	1	15.01
16	16	16	1	15.40
17	17	17	1	15.76
18	18	19	2	16.39
19	20	21	2	16.95
20	22	23	2	17.45
21	24	25	2	17.89
22	26	27	2	18.30
23	28	29	2	18.67
24	30	31	2	19.02
25	32	34	3	19.41
26	35	37	3	19.85
27	38	40	3	20.25

Table 2: Continued.

Band Index	Low FFT Index	High FFT Index	Width	Bark Value
28	41	43	3	20.62
29	44	47	4	21.01
30	48	51	4	21.43
31	52	55	4	21.81
32	56	59	4	22.15
33	60	64	5	22.51
34	65	69	5	22.87
35	70	75	6	23.23
36	76	81	6	23.59
37	82	88	7	23.93
38	89	96	8	24.00
39	97	105	9	24.00
40	106	115	10	24.00
41	116	127	12	24.00