PITCHED INSTRUMENT ONSET DETECTION BASED ON AUDITORY SPECTRA

Emmanouil Benetos, André Holzapfel, and Yannis Stylianou
Institute of Computer Science, FORTH, Greece,
and Multimedia Informatics Lab, Computer Science Department, University of Crete, Greece
{benetos,hannover,yannis}@csd.uoc.gr

ABSTRACT

In this paper, a novel method for onset detection of music signals using auditory spectra is proposed. The auditory spectrogram provides a time-frequency representation that employs a sound processing model resembling the human auditory system. Recent work on onset detection employs DFT-based features, such as the spectral flux and group delay function. The spectral flux and group delay are introduced in the auditory framework and an onset detection algorithm is proposed. Experiments are conducted on a dataset covering 11 pitched instrument types, consisting of 1829 onsets in total. Results indicate the superiority of the auditory representations over the DFT-based ones, with the auditory spectral flux exhibiting an onset detection improvement by 2% in terms of F-measure when compared to the DFT-based feature.

1. INTRODUCTION

The detection of the starting time of each musical note plays an important role in the analysis of music signals. This process is referred to as musical instrument onset detection and it is an essential step for music transcription applications, as well as for music signal compression, beat tracking, and music information retrieval. The goal of an onset detection system is the accurate estimation of note onset times, regardless of the instrument type or performance style. Several approaches for pitched instrument onset detection have been proposed in the literature, however they are mostly limited to a small number of instrument classes.

In [1], an onset detection system combining both energy and phase information was proposed. The employed dataset contained pitched nonpercussive, pitched percussive, nonpitched percussive, and complex sounds. Reported results indicated an improvement over energy and phase-based approaches. An improved version of the system in [1] was proposed in [4], tested on the same dataset. In [3], a system for onset detection employing a constant-Q pitch detector was proposed, tested on the pitched nonpercussive sounds also employed in [1]. It is also suggested in [3] that a detector based on a computational auditory model might improve onset detection performance. Gainza et al. employed FIR comb filters on a frame by frame basis combining the inharmonicity properties with the energy increases of the signal onset [5]. Results report an improvement offer energy-based and phase-based approaches. Finally in [6], the group delay function was proposed for onset detection in a beat tracking application. Multiband analysis was performed on two datasets, the first from the MIREX 2006 beat tracking task and the second containing samples of traditional Cretan music.

In this paper, a novel approach for onset detection is proposed by employing auditory spectrograms instead of DFT-derived spectrograms for the computation of onsets detection features. The auditory spectra, based on the model presented in [11], are designed to mimic the functions of the human auditory system. In the auditory domain, the group delay and spectral flux features are introduced, and an onset detection system is proposed. Comparative experiments on onset detection were performed using the same features in the DFT domain. The dataset used for experimentation contains a wide variety of pitched instrument types, not limited to western instruments, containing 1829 onsets in total. Results indicate that the auditory features outperform DFT-based features for onset detection, with the auditory spectral flux reaching an F-measure of 75.9%.

The outline of the paper is as follows. Section 2 is devoted to the DFT-based features and system for onset detection. In Section 3, the auditory model and features are presented, along with the proposed onset detection system. The employed dataset, the methods used for evaluation and the experimental results are discussed in Section 4. Conclusions are drawn and future directions are indicated in Section 5.

2. DFT-BASED ONSET DETECTION

2.1 Group Delay

As described in [6], phase information can be used for onset detection by considering the group delay $\tau(\omega)$, which for a given signal $x[n]$ with a phase spectrum $\Phi(\omega)$ is defined as the derivative of phase over frequency:

$$\tau(\omega) = -\frac{d\phi(\omega)}{d\omega} \quad (1)$$

The average of the group delay is determined by the distance between the center of the analysis window and the position of an impulse within the window, even when the impulse has been filtered by a causal and stable filter. As the onset of a musical instrument might be modelled by an impulse sent into a causal and stable system, in [6] the average group delay is used as an onset detection function: using a large overlap, an analysis window is shifted over the signal and for each window position the average group delay is computed. The obtained sequence of average group delays is referred to as phase slope function. In Figure 1, an example of a phase slope function is depicted by the dashed line which has positive zero crossings at the position of impulses in the signal. In order to avoid error problems when unwrapping phase in the group delay computation, the slope of the phase function can be computed as [10]:

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \qquad (2)$$

where

$$\begin{aligned} X(\omega) &= X_R(\omega) + jX_I(\omega) \\ Y(\omega) &= Y_R(\omega) + jY_I(\omega) \end{aligned}$$

are the Fourier Transforms of $x[n]$ and $nx[n]$, respectively. The phase slope is then computed as the negative of the average of the group delay function. In this paper, the implementation of the phase slope onset detector as presented in [7] has been used, which includes a multiband processing of the complex DFT spectra and band-wise zero-crossing selection for increased accuracy. The resulting group delay onset detection signal, computed from the band-wise zero-crossing selection, contains peaks located at the time instants of the detected onsets.

2.2 Spectral Flux

Spectral flux (SF) is based on the detection of sudden positive energy changes in the signal which indicate attack parts of new notes. The accuracy of onset detection using SF and its computational simplicity were presented in [2, 4]. SF is computed as:

$$SF(k) = \sum_{\omega} HW(|X(\omega, k)| - |X(\omega, k-1)|) \qquad (3)$$

where $HW(x) = \frac{x+|x|}{2}$ is the half wave rectifier function, and $X(\omega, k)$ is the STFT of the signal with $5.6ms$ hop size and a window length h of $46ms$. For the experiments in this paper, the L1-norm SF is used as shown in (3), since it was shown in [4] that it outperforms the L2-norm.

2.3 DFT-based Onset Detection System

Onsets are detected by selecting the zero crossings of the phase slope and the local maxima of the spectral ux detection signals. The onset detection method has been motivated by the processing steps proposed in [1]: first, the detection signals are smoothed using a Hanning window of length 51ms, which was found to be crucial for improving onset detection results. Afterwards, the signals are normalized using z-score. In [7], the application of an adaptive threshold has been shown to improve accuracy for SF, while it was found be impaired in case of PS. For that reason, an adaptive threshold is applied to SF only. It is computed by applying a moving median lter of length 97ms which is subtracted from the SF detection signals. Finally, a peak selection algorithm is performed in order to produce the detected onsets, by selecting peaks that are separated by a minimum peak distance of 40ms.

3. AUDITORY SPECTRUM-BASED ONSET DETECTION

In this Section the auditory model is presented, followed by the definition of the group delay function and spectral flux in the auditory spectrum domain. Finally, an onset detection system using auditory spectra is proposed.

3.1 Auditory Model

The auditory model was first introduced in [13] and formalized in [11]. It is inspired by physiological, psychoacoustical and computational studies in the human primary auditory cortex. The model consists of two stages, a spectral estimation model (designed to mimic the cochlea in the auditory system) and spectral analysis model (which mimics the primary auditory cortex). The spectral estimation model produces the so-called auditory spectrogram.

The auditory spectrum produces a time-frequency representation of the signal on a logarithmically scaled frequency axis, referred as the tonotopic axis. The auditory spectrogram consists of 128 log-frequency bins and can be approximated as:

$$X_A[n, l] = \max(\partial_l g(\partial_n x[n] *_n h[n, l]), 0), \qquad (4)$$

where $x[n]$ is the original signal and $h[n, l]$ is a minimum-phase seed bandpass filter where $h[n, l] = \alpha h[\alpha n, l_0]$, with scaling factor $\alpha = 2^{l-l_0}$ and $l = 1, \ldots, 129$. The convolution of $x[n]$ with $h[n, l]$ is an application of a constant-Q filter-bank wavelet transform. ∂_i stands for differentiation over i, and $g(m) = \frac{1}{1+e^{-m}} - \frac{1}{2}$ is a sigmoid-like function, which is used to model the hair cell response in the human auditory system. It should be noted that in (4) two operations are not mentioned for simplicity purposes, they are however employed for the auditory spectra computation. The first consists of a temporal smoothing operation which filters out responses beyond 4 kHz and the second consists of a temporal integration of $X_A[n, l]$, which is followed by subsampling.

3.2 Auditory Group Delay

According to (2), and by noting that $X_A[n, l]$ has no imaginary values like the DFT-based group delay, the proposed function for computing the group delay in the auditory

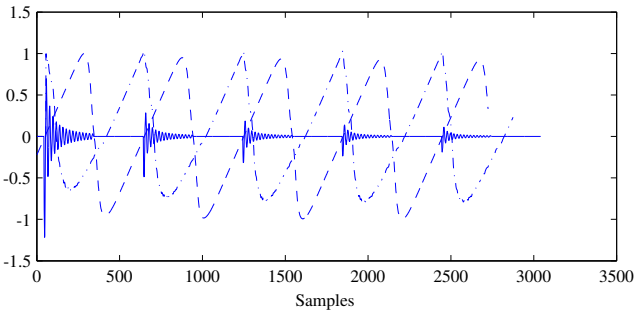

Figure 1. A sequence of impulses with linearly time varying amplitudes, the associated DFT-based group delay function (dashed line), and the associated auditory spectrum-based group delay function (dashed-dotted line).

spectrum is defined as:

$$AUD\_GRD[n,l] = \frac{Y_A[n,l]}{X_A[n,l]}, \quad (5)$$

where $Y_A[n,l]$ is the auditory spectrum of $nx[n]$. Due to the differentiation factor ∂_n in (4), onsets are detected by determining the positions of positive peaks rather than positive zero-crossings. In Figure 1, the auditory spectrum-based group delay that is obtained when shifting an analysis window over a sample signal is depicted as a dashed-dotted line. Note that the term group delay was preferred for the detection function instead of auditory phase slope, because no average value has been computed for neighboring bands as is the case for the DFT-based phase slope.

The processing steps for the computation of the onset detection signal, based on the auditory spectrum group delay function, can be seen in Figure 2. The auditory spectrum was computed using the NSL toolbox [9]. For the computation of the auditory spectrum the window length is set to 0.1s, with 4.5ms hop size and the resulted spectrogram is computed for a bandwidth of 76-3242 Hz. In processing block 2, the auditory group delay function is computed from auditory spectrograms $X_A[n,l]$ and $Y_A[n,l]$ using (5). For our analysis, tonotopic bands $b = 10,\ldots,39$ of the auditory spectrogram were utilized, thus ignoring bands containing high-frequency noise, as well as bands ranging from 76-104 Hz which are not crucial for onset detection purposes, because these frequencies are below the F0 range of the investigated instruments. In processing block 3 of Figure 2, each band is smoothed in time using a 3rd degree Savitzky-Golay filter with window size equal to 12 samples [12]. The Savitzky-Golay filter uses local polynomial regression and is considered superior compared to FIR filters or moving average filters, preserving the local maxima of the signal while rejecting noise. In processing block 4, for each group delay band, peak picking is performed in order to select candidate onsets. For each band, an onset detection signal is constructed containing either the value zero when no peak has been detected, or the amplitude of the detected peak. In each band b, a threshold for peak detection is determined separately by the mean value of the half-wave rectified group delay function for the particular band. Finally, all band-wise detection signals are summed, creating a single onset detection signal based on the auditory group delay.

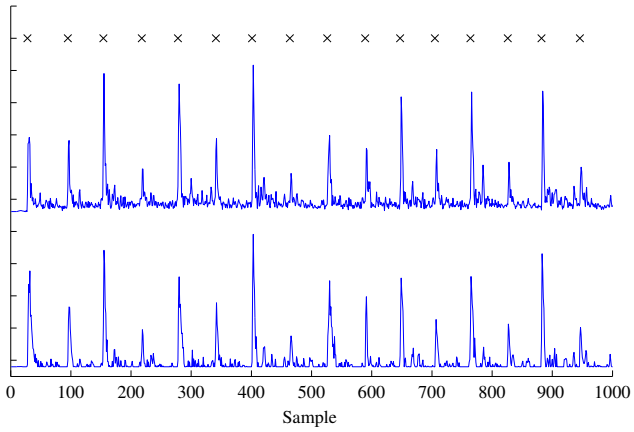

Figure 3. The spectral flux onset strength signals of a tanbur recording. The lower-placed signal depicts the auditory spectrum-derived spectral flux, while the higher-placed signal depicts the DFT-based spectral flux. The 'x' marker corresponds to the annotated onset time.

3.3 Auditory Spectral Flux

The spectral difference in the auditory domain is defined in a similar manner to the group delay. The spectral flux in the auditory spectrum is defined using the L1 norm:

$$AUD\_SF[n] = \sum_l HW(X_A[n,l] - X_A[n-1,l]). \quad (6)$$

For the auditory spectral flux, the original signal is resampled to 8kHz and the spectral flux is computed with a step size of 8*ms*. It should be noted that no band-wise smoothing or band selection was performed on the auditory spectral flux, since it was found to degrade onset detection performance. In Figure 3, the auditory spectrum-based and DFT-based spectral flux onset strength signals of a tanbur (plucked string instrument) recording are depicted. The annotated onset times can also be seen, as well as a false detection for the DFT-based spectral flux at sample 790.

3.4 Auditory Spectrum-based Onset Detection System

Onsets from the auditory group delay and spectral flux detection signals are detected using roughly the same approach as for the DFT representations, by selecting the local maxima of the signals. First, each detection function is normalized using z-score standardization. Afterwards, a moving median filter of length 0.2s is computed as an adaptive threshold, which is a robust method for detecting impulses in audio signals [8]. The adaptive threshold is then subtracted from the detection signals. Finally, peak picking is performed, by selecting peaks that are higher than threshold δ and are separated by a minimum peak distance of 40ms.

Figure 2. Block diagram of the computation of the auditory spectrum-based group delay.

Instrument	No. of onsets	No. of files
Cello	150	5
Clarinet	149	5
Guitar	174	5
Kemençe	186	5
Ney	147	7
Ud	211	5
Piano	195	5
Saxophone	148	5
Tanbur	156	5
Trumpet	140	5
Violin	173	5
Total	**1829**	**57**

Table 1. Onset dataset details.

4. EXPERIMENTS

4.1 Dataset

In our experiments, the dataset introduced in [7] was employed. It consists of 57 recordings of pitched instruments, including 11 instrument types, as seen in Table 1. The various instrument types can be organized into three classes: pitched-percussive instruments (guitar, ud, piano, and tanbur), wind instruments (clarinet, ney, saxophone, and trumpet), and bowed string instruments (cello, kemençe, and violin). It should be noted that the set is not limited to western instruments, but also contains middle-eastern instrument samples. In total, the recordings contain 1829 annotated onsets, while each instrument type contains roughly the same number of onsets. All recordings are monophonic, sampled at 44.1kHz.

4.2 Evaluation Methods

For evaluating the results of the proposed onset detection systems, the recall (R), precision (P), and F-measure (F) are employed as figures of merit. Let N_{tp} stand for the number of correctly detected onsets, N_{fp} the number of false positives, and N_{fn} the number of missed onsets. P and R are defined as:

$$P = \frac{N_{tp}}{N_{tp} + N_{fp}}, \quad R = \frac{N_{tp}}{N_{tp} + N_{FN}} \tag{7}$$

while the F-measure is computed from P and R:

$$F = \frac{2PR}{P + R} \tag{8}$$

It should be noted that P, R, and F are utilized for evaluation in the MIREX onset detection contests. An onset is correctly matched if it is detected within 50ms of the ground truth onset time. By varying parameter δ in small steps, P/R-curves can be created by placing R values on the horizontal axis and P values on the vertical one. The P/R-curve which is closer to the upper right corner of the diagram is considered to be the best detector with regards to F.

Feature	GRD	SF	AUD_GRD	AUD_SF
F-measure	73.7%	73.9%	73.8%	75.9%

Table 2. F-measures for the various onset detection features.

4.3 Results

The performance of the various onset detection features is shown in P/R-curves in Figure 4. In Figure 3(a) the performance of the complete dataset as described in Table 1 is shown. Regarding the optimum F-measure, the DFT-based group delay and spectral flux along with the auditory group delay seem to perform almost equally good, but they are surpassed by the auditory spectral flux. The best F-measures on the complete dataset can be seen in Table 2, where it can be seen that the auditory spectral flux outperforms the other three features by about 2% in terms of F-measure. The auditory group delay performs marginally better than its DFT-based counterpart, achieving high precision rates. In general, the auditory-based features outperform their respective DFT-based features.

As far as the individual instrument types are concerned, the auditory group delay exhibits very high precision rates for the set of string instruments in Figure 3(b), making it useful for beat tracking tasks. However, the auditory group delay is vastly outperformed by the remaining three features when pitched percussive instruments are employed in Figure 3(c), with the DFT-based spectral flux achieving very high precision and recall rates. The DFT-based spectral flux slightly outperforms the auditory spectrum-based spectral flux for pitched percussive instruments, which can be attributed to the limited frequency range of the auditory spectrum, since percussive onsets are detected in high frequency bands [2]. It should be noted that all features report high rates for pitched percussive instruments compared to string and wind instruments. Finally, the set of wind instruments in Figure 3(d) shows lower precision rates compared to the other sets. The auditory features achieve roughly the same best F-measure, outperforming the DFT-based features.

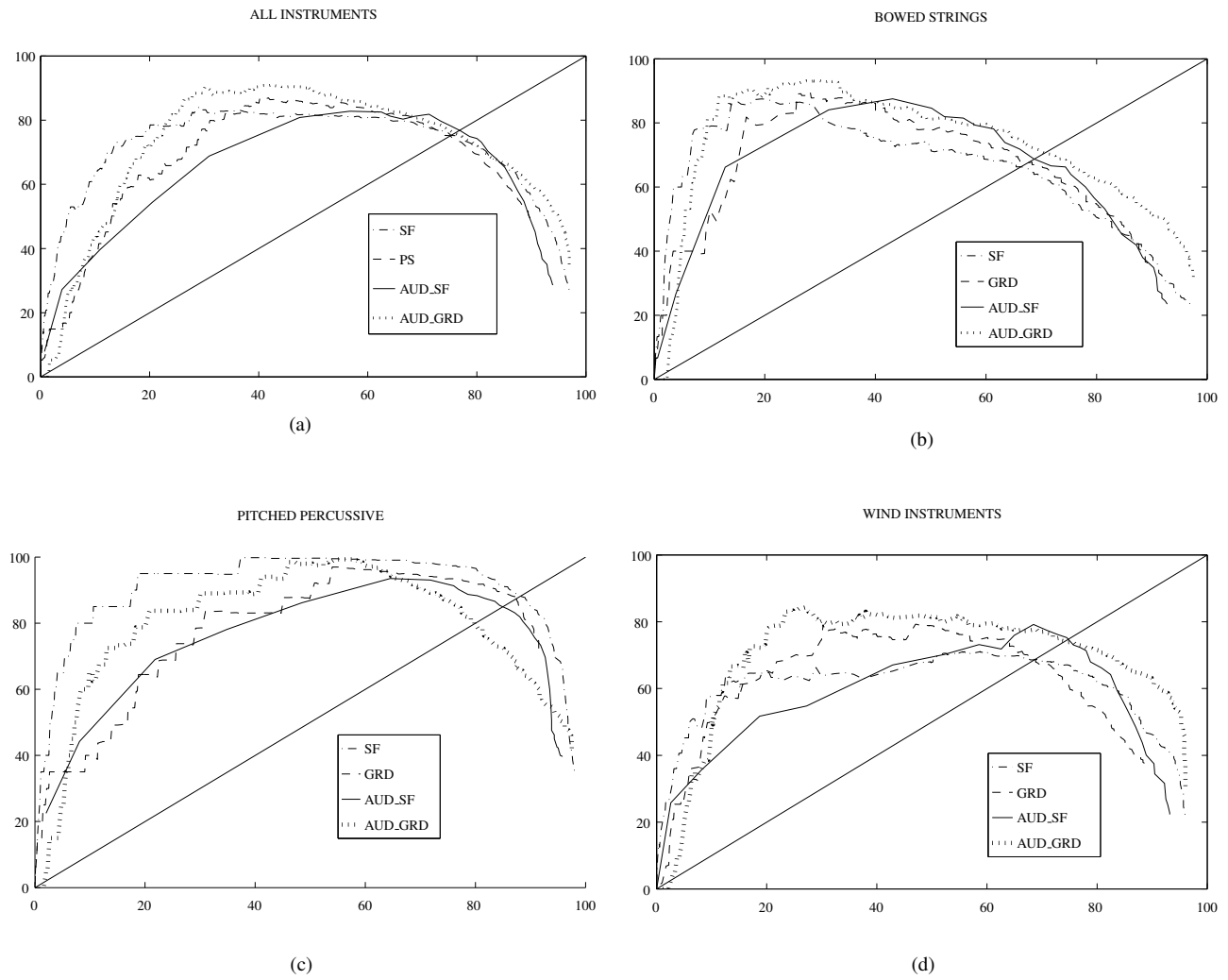

Figure 4. Performance curves of the various onset detection features. Recall and Precision values are plotted on the horizontal and vertical axis, respectively.

5. CONCLUSIONS

In this paper a new approach for onset detection using auditory spectra was proposed. The group delay function and spectral flux in the auditory domain were introduced as features for onset detection, and a system was proposed. The onset detection performance of the auditory spectral flux was found to be superior compared to the DFT-based feature, reaching an F-measure of 75.9% compared to 73.9% of the DFT-based spectral flux. While the performance of the auditory spectral flux for pitched percussive instruments was inferior compared to DFT-based features, it is relatively superior when string and wind instruments are tested.

In the future, a fusion of the onset detection features in the auditory domain will be performed, in an attempt to maximize onset detection performance. The system could also consider onsets produced by non-pitched percussive instruments, which can be easily detected using energy descriptors. In addition, the creation of an onset detection system which is dependent of the instrument family can lead to improved results. Finally, the aforementioned techniques can be developed for usage in polyphonic recordings.

6. REFERENCES

[1] J. P. Bello, C. Duxbury, M. Davies, and M. Sandler, "On the use of phase and energy for musical onset detection in the complex domain," *IEEE Signal Proc. Letters*, Vol. 11, No. 6, pp. 553-556, June 2004.

[2] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection of music signals," *IEEE Trans. Speech and Audio Proc.*, Vol. 13, No. 5, pp. 1035-1047, Sep. 2005.

[3] N. Collins, "Using a pitch detector for onset detection," in Proc. *6th Int. Conf. Music Information Retrieval*, pp. 100-106, September 2005.

[4] S. Dixon, "Onset detection revisited," in Proc. *9th Int. Conf. Digital Audio Effects*, pp. 133-137, 2006.

[5] M. Gainza, E. Coyle, and B. Lawlor, "Onset detection

using comb filters," in Proc. *IEEE Workshop Applications of Signal Processing to Audio and Acoustics*, pp. 263-266, 2005.

[6] A. Holzapfel and Y. Stylianou, "Beat tracking using group delay based onset detection," in Proc. *9th Int. Conf. Music Information Retrieval*, Sep. 2008.

[7] A. Holzapfel, Y. Stylianou, A. C. Gedik, and B. Bozkurt "Three dimensions of pitched instrument onset detection," *IEEE Trans. Audio, Language, and Speech Processing*, accepted for publication.

[8] I. Kauppinen, "Methods for detecting impulsive noise in speech and audio signals," in Proc. *14th Int. Conf. Digital Signal Proc.*, Vol. 2, pp. 967-970, July 2002.

[9] T. Chi and S. A. Shamma, "NSL Matlab Toolbox," `http://www.isr.umd.edu/Labs/NSL/Software.htm`, Neural Systems Lab., Univ. Maryland.

[10] A. V. Oppenheim, R. W Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, Prentice Hall, 1998.

[11] P. Ru, "Multiscale multirate spectro-temporal auditory model," *PhD Thesis, Univ. Maryland College Park*, 2001.

[12] A. Savitzky, and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, Vol. 36, No. 8, pp. 1627-1639, July 1964.

[13] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Information Theory*, Vol. 38, No. 2, pp. 824-839, March 1992.