A simplified early auditory model with application in audio classification

Un modèle auditif simplifié avec application à la classification audio

Wei Chu and Benoît Champagne*

The past decade has seen extensive research on audio classification and segmentation algorithms. However, the effect of background noise on classification performance has not been widely investigated. Recently, an early auditory model that calculates a so-called auditory spectrum has achieved excellent performance in audio classification along with robustness in a noisy environment. Unfortunately, this early auditory model is characterized by high computational requirements and the use of nonlinear processing. In this paper, certain modifications are introduced to develop a simplified version of this model which is linear except for the calculation of the square-root value of the energy. Speech/music and speech/non-speech classification tasks are carried out to evaluate the classification performance, with a support vector machine (SVM) as the classifier. Compared to a conventional fast Fourier transform–based spectrum, both the original auditory spectrum and the proposed simplified auditory spectrum show more robust performance in noisy test cases. Test results also indicate that despite a reduced computational complexity, the performance of the proposed simplified auditory spectrum is close to that of the original auditory spectrum.

La dernière décennie a connu une expansion de la recherche sur les algorithmes de classification audio et de segmentation. Cependant, l'effet du bruit de fond sur les performances de la classification n'a pas été largement étudié. Récemment, un modèle auditif qui calcule un spectre auditif a atteint une performance excellente en classification audio ainsi qu'une robustesse dans un environnement bruité. Malheureusement, ce modèle auditif est caractérisé par des besoins élevés en calcul et par un traitement non-linéaire. Dans ce papier, quelques modifications sont introduites afin de développer une version simplifiée de ce modèle qui est linéaire à l'exception du calcul de la valeur de la racine carrée de l'énergie. Des tâches de classification de la parole/musique de même que de la parole/non-parole sont effectuées pour évaluer la performance de la classification, en utilisant un classifieur à automate à support vectoriel. Comparé à une transformation rapide de Fourier conventionnelle, les deux spectres auditifs – celui d'origine et celui simplifié proposé – montrent des performances plus robustes dans les tests avec bruit. Les résultats des tests montrent également qu'en dépit d'une complexité de calcul réduite, la performance du spectre auditif simplifié qui a été proposé est proche du spectre auditif d'origine.

Keywords: audio classification; auditory spectrum; early auditory model; noise robustness

I. Introduction

Audio classification and segmentation can provide useful information for understanding both audio and video content. In recent years many studies have been carried out on audio classification. In work by Scheirer and Slaney [1] to classify speech and music, as many as 13 features are employed, including 4 Hz modulation energy, spectral rolloff point, spectral centroid, spectral flux (delta spectrum magnitude), and zero-crossing rate (ZCR). Using audio features such as energy function, ZCR, fundamental frequency, and spectral peak tracks, Zhang and Kuo [2] proposed an approach to automatic segmentation and classification of audiovisual data. Lu et al. [3] proposed a two-stage robust approach that is capable of classifying and segmenting an audio stream into speech, music, environment sound, and silence. In a recent work, Panagiotakis and Tziritas [4] proposed an algorithm for audio segmentation and classification using mean signal amplitude distribution and ZCR.

Although in some previous research the background noise has been considered as one of the audio types or as a component of some hybrid sounds, the effect of background noise on the performance of classification has not been widely investigated. A classification algorithm trained using clean test sequences may fail to work properly when the actual testing sequences contain background noise with certain SNR levels (see test results in [5] and [6]). The so-called early auditory model proposed by Wang and Shamma [7] has proved to be robust in noisy environments because of an inherent self-normalization property which causes noise suppression. Recently, this early auditory model has been employed in audio classification, and excellent performance has been reported in [6]. However, this model is characterized by high computational requirements and the use of nonlinear processing. It would be desirable to have a simplified version of this early auditory model, or even to have an approximated model in the frequency domain, where efficient fast Fourier transform (FFT) algorithms are available.

In this paper we propose, based on certain modifications, a simplified version of this early auditory model which is linear except for the calculation of the square-root value of the energy. To evaluate the classification performance, speech/music and speech/non-speech classification tasks are carried out, in which a support vector machine (SVM) is used as the classifier. Compared to a conventional FFT-based spectrum, both the original auditory spectrum and the proposed simplified auditory spectrum show more robust performance in noisy test cases. Experimental results also show that despite its reduced computational complexity, the performance of the proposed simplified auditory spectrum is close to that of the original auditory spectrum.

The paper is organized as follows. Section II briefly introduces the early auditory model [7] considered in this work. A simplified version of this model is proposed in Section III. Section IV explains the extraction of audio features and the setup of the classification tests. Test results are presented in Section V, and conclusions appear in Section VI.

*Wei Chu and Benoît Champagne are with the Department of Electrical and Computer Engineering, McGill University, Montreal, Quebec H3A 2A7. E-mail: wchu@tsp.ece.mcgill.ca, champagne@ece.mcgill.ca. This paper was awarded a prize in the Student Paper Competition at the 2006 Canadian Conference on Electrical and Computer Engineering. It is presented here in a revised format.

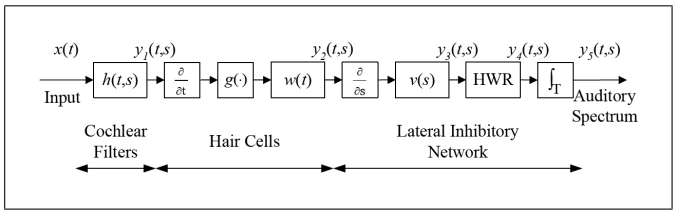


Figure 1: *Schematic description of the early auditory model [7].*

II. Early auditory model

The auditory spectrum used in this work is calculated from a so-called early auditory model introduced in [7] and [8]. This model, which can be simplified as a three-stage processing sequence (see Fig. 1), describes the transformation of an acoustic signal into an internal neural representation referred to as an auditory spectrogram. A signal entering the ear first produces a complex spatio-temporal pattern of vibrations along the basilar membrane (BM). A simple way to describe the response characteristics of the BM is to model it as a bank of constant-Q highly asymmetric bandpass filters $h(t, s)$, where t is the time index and s denotes a specific location on the BM (or equivalently, s is the frequency index).

In the next stage, the motion on the BM is transformed into neural spikes in the auditory nerves, and the biophysical process is modelled by the following three steps: a temporal derivative, which is employed to convert instantaneous membrane displacement into velocity; a nonlinear sigmoid-like function $g(\cdot)$, which models the nonlinear channel through the hair cell; and a low-pass filter $w(t)$, which accounts for the leakage of the cell membranes.

In the last stage, a lateral inhibitory network (LIN) detects discontinuities along the cochlear axis, s. The operations can be effectively divided into the following stages: a derivative with respect to the tonotopic axis s which mimics the lateral interaction among LIN neurons; a local smoothing filter, $v(s)$, due to the finite spatial extent of the lateral interactions; a half-wave rectification (HWR) modelling the nonlinearity of the LIN neurons; and a temporal integration which reflects the fact that the central auditory neurons are unable to follow rapid temporal modulations.

These operations effectively compute a spectrogram of an acoustic signal. At a specific time index t, the output $y_5(t, s)$ is referred to as an auditory spectrum. For simplicity, the spatial smoothing, $v(s)$, is ignored in the implementation [7].

III. Simplified early auditory model

Because of a complex computation procedure and the use of nonlinear processing in the above early auditory model, the computational complexity of the auditory spectrum is expected to be much higher than that of a conventional FFT-based spectrum. It is thus desirable that the model be simplified.

A. Pre-emphasis and nonlinear compression

This early auditory model has proved to be noise-robust because of an inherent self-normalization property. According to the stochastic analysis carried out in [7], the following relationships hold:

$$\begin{aligned} E[y_5(t,s)] &= E[y_4(t,s)] *_t \Pi(t), \\ E[y_4(t,s)] &= E[g'(U)E[\max(V,0) \mid U]], \\ V &= (\partial_t x(t)) *_t \partial_s h(t,s), \\ U &= (\partial_t x(t)) *_t h(t,s), \end{aligned} \tag{1}$$

where E denotes statistical expectation, $E[y_5(t, s)]$ is the output average auditory spectrum, $\Pi(t)$ is a temporal integration function, and $*_t$ denotes time-domain convolution. According to [7], $E[y_4(t, s)]$ is a quantity that is proportional to the energy[1] of V and inversely proportional to the energy of U. The definitions of U and V given in (1) further suggest that the auditory spectrum is an averaged ratio of the signal energy passing through the differential filters $\partial_s h(t, s)$ and the cochlear filters $h(t, s)$, or equivalently, the auditory spectrum is a self-normalized spectral profile. Considering that the cochlear filters are broad while the differential filters are narrow and centred around the same frequencies, this self-normalization property leads to unproportional scaling for spectral components of the sound signal. Specifically, a spectral peak receives a relatively small normalization factor, whereas a spectral valley receives a relatively large normalization factor. The difference in the normalization is known as spectral enhancement or noise suppression.

When the hair-cell nonlinearity is replaced by a linear function, e.g., $g'(x) = 1$ (see Fig. 1), we have $E[y_4(t, s)] = E[\max(V, 0)]$, where $E[y_4(t, s)]$ represents the spectral energy profile of the sound signal $x(t)$ across the channels indexed by s. With a linear function $g(x)$, it is found in our test that if the input signal is not pre-emphasized, the classification performance of the modified auditory spectrum is close to that of the original auditory spectrum. A close performance may suggest that a scheme for noise suppression is implicitly part of this modified auditory model. However, according to [7], with a linear function $g(x)$, the whole processing scheme is viewed as estimating the energy resolved by the differential filters alone without self-normalization. It seems that the self-normalization alone cannot be used to explain the noise suppression for this modified model. The actual cause of the noise suppression in this case is under investigation.

B. HWR and temporal integration

Referring to Fig. 1, we note that the LIN stage consists of a derivative with respect to the tonotopic axis s, a local smoothing, $v(s)$, a half-wave rectification, and a temporal integration (implemented via low-pass filtering and downsampling at a frame rate [9]). The HWR and temporal integration serve to extract a positive quantity corresponding to a specific frame and a specific channel (i.e., a component of the auditory spectrogram). A simple way to interpret this positive quantity is as the square-root value of the frame energy in a specific channel. Based on these considerations, an approximation to the HWR and temporal integration is proposed, where the original processing is replaced by the calculation of the square-root value of the frame energy. Fig. 2 shows the auditory spectrograms of a one-second speech clip calculated using the original early auditory model and the modified model (i.e., the original model with proposed modifications on HWR and temporal integration). The two spectral-temporal patterns are very close.

C. Simplified model

By introducing modifications to the original processing steps of pre-emphasis, nonlinear compression, half-wave rectification, and temporal integration, we propose a simplified version of this model. Except for the calculation of the square-root value of the energy, this simplified model is linear. Considering the relationship between time-domain energy and frequency-domain energy as per Parseval's theorem [10], it is possible to further implement this simplified model in the frequency domain so that significant reductions in computational complexity can be achieved. Such a self-normalized FFT-based model has been further proposed and applied in a speech/music/noise classification task in [11].

IV. Audio classification test

A. Audio sample database

To carry out performance tests, a generic audio database is built which

[1] $E[y_4(t, s)]$ is related to $E[\max(V, 0)]$, a quantity proportional (though not necessarily linearly) to the standard deviation, σ, of V when V is zero mean. In [7], the quantity $E[\max(V, 0)]$ is referred to as energy, considering the one-to-one correspondence between σ and σ^2.

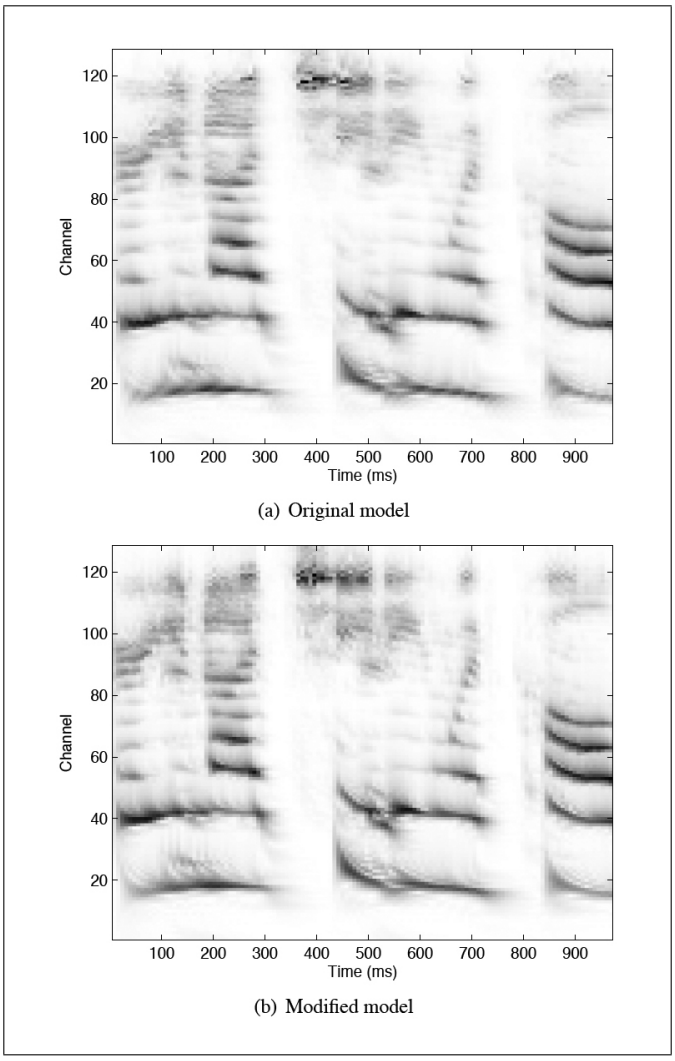


Figure 2: *Auditory spectrograms of a one-second speech clip.*

includes speech, music, and noise clips, sampled at the rate of 16 kHz. The music clips consist of different types, including blues, classical, country, jazz, and rock. The music clips also contain segments that are played by certain Chinese traditional instruments. Noise samples are selected from the NOISEX database, which contains recordings of various noises. The total length of all the audio samples is 200 minutes. These samples are divided equally into two parts for training and testing. The audio classification decisions are made on a one-second basis.

In the following, for the speech/music classification task, a clean test is a test in which both the training and the testing sets contain clean speech and clean music. A specific SNR value indicates a test in which the training set contains clean speech and clean music while the testing set contains noisy speech and noisy music (both with the stated SNR value). As for the speech/non-speech classification task, music and noise clips are grouped together as the non-speech set. The clean and noisy tests are carried out in a way similar to that for speech/music classification, except that noise clips are added in the training and testing.

B. Audio features

In this work, audio features are extracted based on the aforementioned auditory spectrum and the FFT-based spectrum. Using auditory spectrum data, we further calculate mean and variance in each channel over a one-second time window. Corresponding to each one-second audio clip, the auditory feature set is a 256-dimensional mean-plus-variance vector.

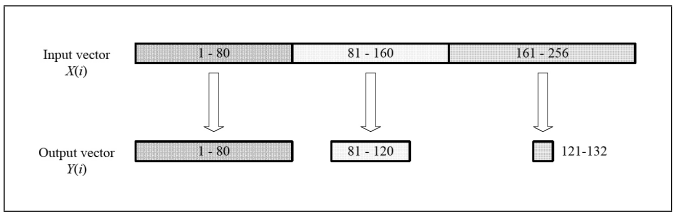


Figure 3: *The power spectrum grouping scheme.*

For the FFT-based spectrum, a narrowband (30 ms) spectrum is calculated using 512-point FFT with an overlap of 20 ms. To reduce the dimension of the obtained power spectrum vector, we may use methods such as principal component analysis. In this work, to simplify the processing, we propose a simple grouping scheme to reduce the dimension. The grouping is carried out according to the following formula:

$$Y(i) = \begin{cases} X(i) & 1 \le i \le 80, \\ \frac{1}{2}\sum_{k=0}^{1} X(2i-80-k) & 81 \le i \le 120, \\ \frac{1}{8}\sum_{k=0}^{7} X(8i-800-k) & 121 \le i \le 132, \end{cases} \tag{2}$$

where i is the frequency index and $X(i)$ and $Y(i)$ represent the power spectrum before and after grouping, respectively. This grouping scheme places the emphasis on low-frequency components. As shown in Fig. 3, based on this grouping scheme, a set of 256 power spectrum components is transformed into a 132-dimensional vector. After discarding the first and the last two components and applying logarithmic operation, we obtain a 128-dimensional power spectrum vector. Furthermore, mean and variance are calculated similarly on different frequency indices over a one-second time window.

C. Implementation

In this work, we use a MATLAB toolbox developed by Neural Systems Laboratory, University of Maryland [9], to calculate the auditory spectrum. Relevant modifications are introduced to this toolbox to meet the needs of our study.

The support vector machine, which is a statistical machine learning technique applied in pattern recognition, has been recently employed in the audio classification task [5], [12]. An SVM first transforms input vectors into a high-dimensional feature space using a linear or nonlinear transformation and then conducts a linear separation in feature space. In this work, we use the SVM$^{\mathrm{struct}}$ algorithm [13]–[15] to carry out the classification task.

V. Performance analysis

The FFT-based spectrum features are used as a reference to compare the performance of the auditory spectrum features. The speech/music classification test results are listed in Table 1, where AUD, AUD_S, and FFT represent the original auditory spectrum, the simplified auditory spectrum, and the FFT-based spectrum respectively. The speech/non-speech classification test results are listed in Table 2.

Although the conventional FFT-based spectrum provides excellent performance in the clean test case, its performance degrades rapidly and significantly as the SNR decreases, leading to a very poor overall performance. Compared to the conventional FFT-based spectrum, the original auditory spectrum and the proposed simplified auditory spectrum are more robust in noisy test cases. Results in Tables 1 and 2 also indicate that despite a reduced computational complexity, the performance of the proposed simplified auditory spectrum is close to that of the original auditory spectrum, especially when SNR is greater than 10 dB.

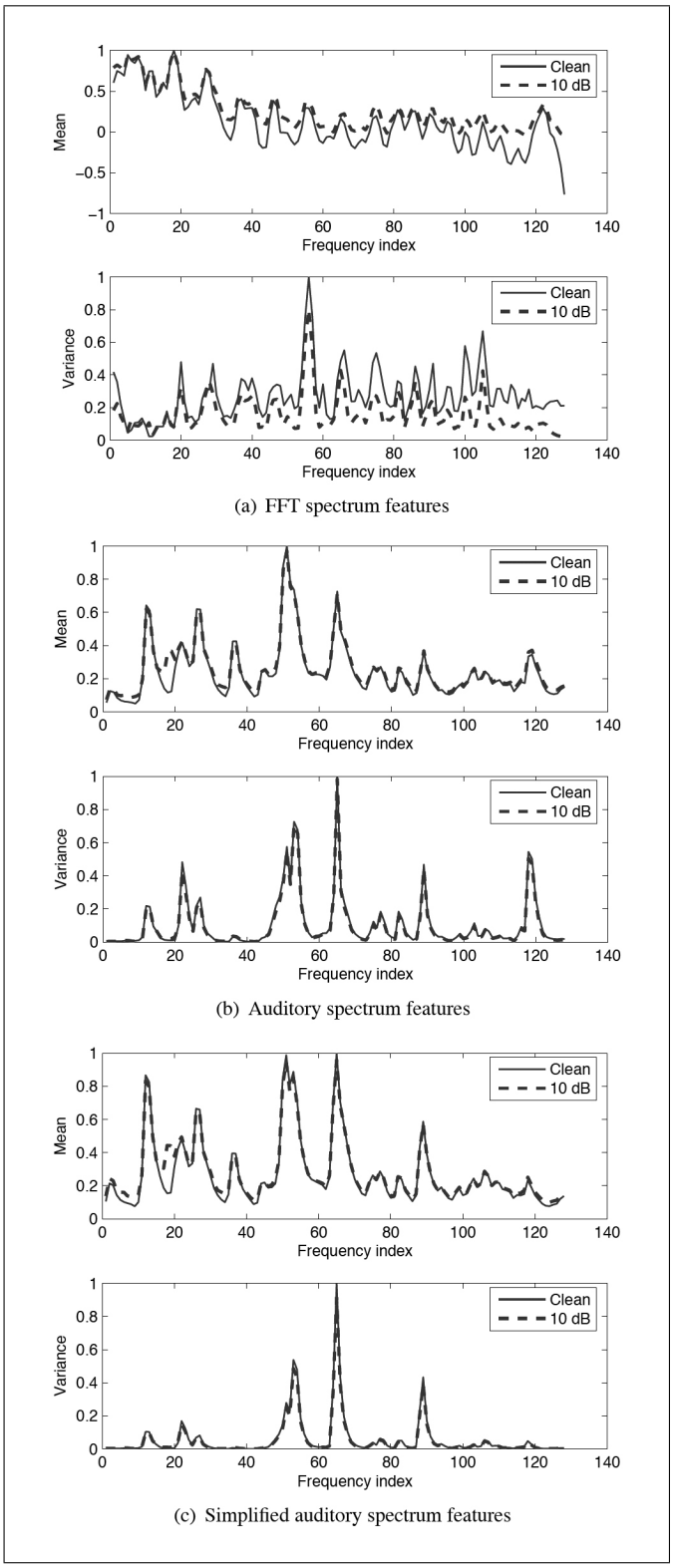
Table 1
Speech/music classification error rate for auditory spectrum (AUD), simplified auditory spectrum (AUD_S), and FFT-based spectrum (FFT)

SNR (dB)	AUD (%)	AUD_S (%)	FFT (%)
∞	2.2	2.7	1.0
20	2.5	3.1	20.6
15	3.3	3.9	37.3
10	5.9	7.4	42.9
5	14.3	19.3	44.2
Average	5.6	7.3	29.2

Table 2
Speech/non-speech classification error rate for auditory spectrum (AUD), simplified auditory spectrum (AUD_S), and FFT-based spectrum (FFT)

SNR (dB)	AUD (%)	AUD_S (%)	FFT (%)
∞	1.4	1.7	0.8
20	1.7	2.0	15.3
15	2.3	2.5	27.4
10	4.0	4.8	31.3
5	10.8	13.6	32.3
Average	4.0	4.9	21.4

An example of audio features (mean and variance values in relative scales) is given in Fig. 4, which shows the FFT-based spectrum, the original auditory spectrum, and the proposed simplified auditory spectrum features for a one-second music clip in a clean test case and in a noisy test case with 10 dB SNR. For the original auditory spectrum features and the proposed simplified auditory spectrum features, the results when SNR equals 10 dB are close to those for the clean test case. However, this is not the case for conventional FFT-based spectrum features, which show a relatively large change. The results presented in Fig. 4 demonstrate the noise-robustness of the original auditory spectrum features and the proposed simplified auditory spectrum features.



(a) FFT spectrum features

(b) Auditory spectrum features

(c) Simplified auditory spectrum features

Figure 4: *Audio features (mean and variance values) for a one-second music clip.*

VI. Conclusions

In this paper, we proposed a simplified version of an early auditory model [7] by introducing modifications to the original processing steps of pre-emphasis, nonlinear compression, half-wave rectification, and temporal integration. Except for the calculation of the square-root value of the energy, the proposed simplified early auditory model is linear. To evaluate the classification performance, speech/music and speech/non-speech classification tasks were carried out, with a support vector machine as the classifier. Compared to the conventional FFT-based spectrum, the original auditory spectrum and the proposed simplified auditory spectrum are more robust in noisy test cases. Experimental results also indicate that despite a reduced computational complexity, the performance of the proposed simplified auditory spectrum is close to that of the original auditory spectrum.

References

[1] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, Apr. 1997, pp. 1331–1334.

[2] T. Zhang and C.-C. Jay Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 4, May 2001, pp. 441–457.

[3] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 7, Oct. 2002, pp. 504–516.

[4] C. Panagiotakis and G. Tziritas, "A speech/music discriminator based on RMS and zero-crossings," *IEEE Trans. Multimedia*, vol. 7, Feb. 2005, pp. 155–166.

[5] N. Mesgarani, S. Shamma, and M. Slaney, "Speech discrimination based on multiscale spectro-temporal modulations," in *Proc. IEEE Int. Conf. Acoust., Speech,*

Signal Processing, vol. 1, May 2004, pp. 601–604.

[6] S. Ravindran and D. Anderson, "Low-power audio classification for ubiquitous sensor networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, May 2004, pp. 337–340.

[7] K. Wang and S. Shamma, "Self-normalization and noise-robustness in early auditory representations," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 3, July 1994, pp. 421–435.

[8] M. Elhilali, T. Chi, and S.A. Shamma, "A spectrotemporal modulation index (STMI) for assessment of speech intelligibility," *Speech Communication*, vol. 41, Oct. 2003, pp. 331–348.

[9] "NSL Matlab Toolbox" [online], College Park, Md.: Neural Systems Laboratory, University of Maryland, [cited Oct. 2006], available from World Wide Web: <http://www.isr.umd.edu/Labs/NSL/nsl.html>.

[10] A.V. Oppenheim, R.W. Schafer, and J.R. Buck, *Discrete-Time Signal Processing*, 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1999.

[11] W. Chu and B. Champagne, "A noise-robust FFT-based spectrum for audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2006, pp. 213–216.

[12] Y. Li and C. Dorai, "SVM-based audio classification for intructional video analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 5, May 2004, pp. 897–900.

[13] T. Joachims, "SVM$^{\texttt{struct}}$" [online], Ithaca, N.Y.: Dept. of Computer Science, Cornell University, July 2004 [cited Sept. 2006], available from World Wide Web: <http://www.cs.cornell.edu/People/tj/>.

[14] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector learning for interdependent and structured output spaces," in *Proc. 21st Int. Conf. Machine Learning*, July 2004.

[15] K. Crammer and Y. Singer, "On the algorithmic implementation of multi-class kernel-based vector machines," *J. Machine Learning Research*, vol. 2, Dec. 2001, pp. 265–292.