A Simplified Early Auditory Model with Application in Speech/Music Classification

Wei Chu and Benoît Champagne
Department of Electrical and Computer Engineering
McGill University, Montreal, Quebec, Canada, H3A 2A7
e-mail: wchu@tsp.ece.mcgill.ca, champagne@ece.mcgill.ca

Abstract

The past decade has seen extensive research on audio classification and segmentation algorithms. However, the effect of background noise on the performance of classification has not been investigated widely. Recently, an early auditory model [1] that calculates a so-called auditory spectrum, has been employed in audio classification where excellent performance is reported along with robustness in noisy environment. Unfortunately, this early auditory model is characterized by high computational requirements and the use of nonlinear processing. In this paper, by introducing certain modifications we propose a simplified version of this model which is linear except for the calculation of the square-root value of the energy. A speech/music classification task is carried out to evaluate the classification performance wherein 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. Test results also indicate that, with a reduced computational complexity, the performance of the proposed simplified auditory spectrum is close to that of the original auditory spectrum.

Keywords *— Audio classification; early auditory model; auditory spectrum; noise-robustness.*

1 Introduction

Audio classification and segmentation can provide useful information for both audio and video content understanding. In recent years many studies have been carried out on audio classification. In a work by Scheirer and Slaney [2], to classify speech and music, as many as 13 features have been employed which include 4Hz modulation energy, spectral rolloff point, spectral centroid, spectral flux (delta spectrum magnitude), ZCR, etc. By using audio features such as energy function, ZCR, fundamental frequency, and spectral peak tracks, Zhang and Kuo [3] proposed an approach to automatic segmentation and classification of audiovisual data. Lu *et al.* [4] 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 [5] 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 investigated widely. 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 [6] and [7]). The so-called early auditory model proposed by Wang and Shamma [1] is proved to be robust in noisy environment due to 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] and [7]. However, this model is characterized by high computational requirements and the use of nonlinear processing. It would be desirable that this early auditory model be simplified, or even approximated in frequency domain wherein efficient FFT algorithms are available.

In this paper, by introducing certain modifications, we propose 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, a speech/music classification task is carried out wherein 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, in spite of 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 2 briefly introduces the early auditory model [1] considered in this work. A simplified version of this model is proposed in Section 3. Section 4 explains the extraction of audio features and the setup of the classification tests. The test results are presented in Section 5.

2 Early Auditory Model

The auditory spectrum used in this work are calculated from a so-called early auditory model introduced in [1] 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 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).

At the next stage, the motion on the BM is transformed into neural spikes in the auditory nerves and the biophysical

1-4244-0038-4 2006
IEEE CCECE/CCGEI, Ottawa, May 2006

process is modeled 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 lowpass filter $w(t)$ which accounts for the leakage of the cell membranes [1].

At the last stage, a lateral inhibitory network (LIN) detects discontinuities along the cochlear axis s. The operations can be effectively divided into the following steps: a derivative with respect to the tonotopic axis s that mimics the lateral interaction among LIN neurons, a local smoothing $v(s)$ due to the finite spatial extent of the lateral interactions, a half-wave rectification (HWR) modeling 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 [1].

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 [1].

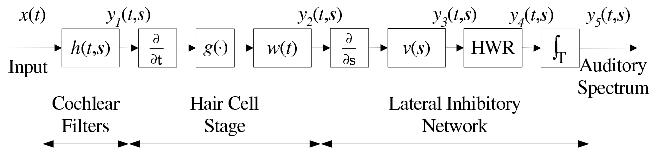


Figure 1. Schematic description of the early auditory model [1]

3 Simplified Early Auditory Model

Due to 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.

3.1 Pre-emphasis and Nonlinear Compression

This early auditory model is proved to be noise-robust due to an inherent self-normalization property [1]. According to the stochastic analysis carried out in [1], the following relationships hold

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

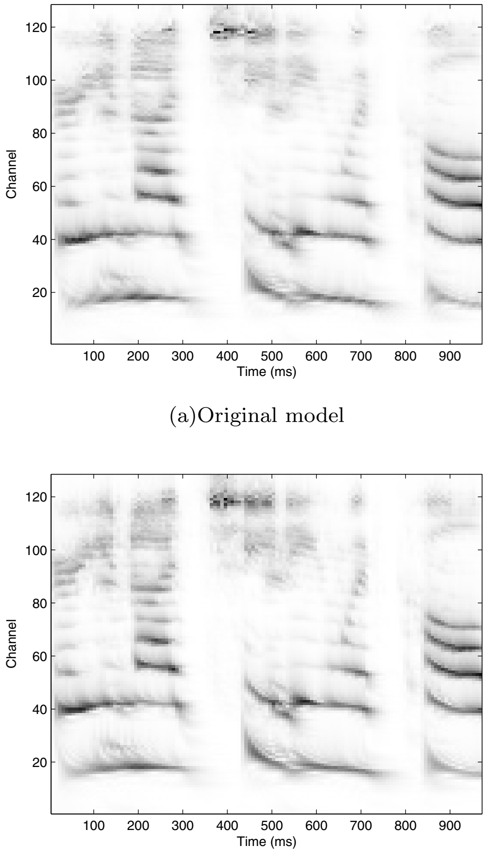
where E denotes statistical expectation, $E[y_5(t,s)]$ is the output average auditory spectrum, $\Pi(t)$ is a temporal integration function, and $*_t$ denotes time-domain convolution. According to [1], $E[y_4(t,s)]$ is a quantity that is proportional to the *energy*[1] of V and inversely proportional 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 differential filters $\partial_s h(t,s)$ and cochlear filters $h(t,s)$, or equivalently, the auditory spectrum is a self-normalized spectral profile [1]. Considering that the cochlear filters are broad while the differential filters are narrow and centered 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 [1].

In case when the hair cell nonlinearity is replaced by a linear function, e.g., $g'(x) = 1$ (see Fig. 1), we have $E[y_4(t,s)] = E[\max(V,0)]$ [1]. $E[y_4(t,s)]$ is the spectral energy profile of the sound signal $x(t)$ across the channels indexed by s [1]. 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 [1], 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 as mentioned in [1] cannot be employed to explain the noise suppression for this modified model. The actual cause of the noise suppression in this case is under investigation.

3.2 HWR and Temporal Integration

Referring to Fig. 1, 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 lowpass filtering and a 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 that it is 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.

[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 [1], quantity $E[\max(V,0)]$ is referred to as *energy* considering the one-to-one correspondence between σ and σ^2.



(a)Original model

(b)Modified model

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

3.3 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 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].

4 Audio Classification Test

4.1 Audio Sample Database

To carry out performance test, a generic audio database is built which include speech, music and noise clips. Music clips include five different types, i.e., blues, classical, country, jazz, and rock. Eleven types of noise, which include speech babble, car interior noise, copy center noise, etc., are employed to form the noise set. These noise data are used to generate noisy speech and noisy music clips with different SNR values. The training set and testing set each contain 1200 one-second speech clips and 1200 one-second music clips. Testing set also contains 1200 noise clips. The sampling rate is 16 kHz.

In the following, a clean test refers to a test wherein both the training set and testing set contain clean speech and clean music. A test with a specific SNR value refers to a test wherein the training set contains clean speech and clean music while the testing set contains noisy speech and noisy music (both with that specific SNR value).

4.2 Audio Features

In this work, audio features are extracted based on the aforementioned auditory spectrum and FFT-based spectrum. Using auditory spectrum data, mean and variance are further calculated 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+variance vector.

For FFT-based spectrum, narrow-band (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 like principal component analysis (PCA). 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

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

where i is the frequency index, and $X(i)$ and $Y(i)$ represent the power spectrum before and after grouping, respectively. This grouping scheme gives emphasis to low-frequency components. 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. Further, mean and variance are calculated similarly on different frequency indices over a one-second time window.

4.3 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 was recently employed in audio classification task [6] [12]. In this work, we use SVM^{struct} algorithm [13]– [15] to carry out the classification task.

5 Performance Analysis

The FFT-based spectrum features are used as a reference to compare the performance of the auditory spectrum features. The test results are listed in Table I, where "AUD", "AUD_S" and "FFT" represent the original auditory spectrum, the simplified auditory spectrum, and the FFT-based spectrum respectively.

TABLE I

Classification error rate (%) for the auditory spectrum (AUD), its simplified version (AUD_S), and the FFT-based spectrum (FFT).

SNR (dB)	AUD	AUD_S	FFT
∞	2.63	3.04	1.42
20	3.04	3.08	32.21
15	3.42	3.75	40.67
10	6.00	7.38	44.17
5	18.25	24.00	46.13
Average	6.67	8.25	32.92

Although the conventional FFT-based spectrum provides an excellent performance in 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 Table I also indicate that, with a reduced computational complexity, the performance of the proposed simplified auditory spectrum is close to that of the original auditory spectrum, especially when SNR$\geq$10 dB.

6 Conclusions

In this paper, we have proposed a simplified version of an early auditory model [1] by introducing modifications to the original processing steps of pre-emphasis, nonlinear compression, half-wave rectification (HWR), 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, a speech/music classification task has been carried out wherein a support vector machine is used 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, in spite of a reduced computational complexity, the performance of the proposed simplified auditory spectrum is close to that of the original auditory spectrum.

References

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

[2] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. ICASSP'97*, April 1997, vol. 2, pp. 1331–1334.

[3] 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, pp. 441–457, May 2001.

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

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

[6] N. Mesgarani, S. Shamma, and M. Slaney, "Speech discrimination based on multiscale spectro-temporal modulations," in *Proc. ICASSP'04*, May 2004, vol. 1, pp. 601–604.

[7] S. Ravindran and D. Anderson, "Low-power audio classification for ubiquitous sensor networks," in *Proc. ICASSP'04*, May 2004, vol. 4, pp. 337–340.

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

[9] Neural Systems Laboratory, University of Maryland, "NSL Matlab Toolbox," http://www.isr.umd.edu /Labs/NSL.

[10] A. V. Oppenheim, R. W. Schafer, with J. R. Buck, *Discrete-Time Signal Processing*, Prentice-Hall, second edition, 1999.

[11] W. Chu and B. Champagne, "A noise-robust FFT-based spectrum for audio classification," in *Proc. ICASSP'06*, May 2006.

[12] Y. Li and C. Dorai, "SVM-based audio classification for intructional video analysis," in *Proc. ICASSP'04*, May 2004, vol. 5, pp. 897–900.

[13] T. Joachims, "SVM-struct," 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. of the 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, pp. 265–292, 2001.