A NOISE-ROBUST FFT-BASED SPECTRUM FOR AUDIO CLASSIFICATION

Wei Chu and Benoît Champagne

Department of Electrical and Computer Engineering
McGill University, Montreal, Quebec, Canada, H3A 2A7
wchu@tsp.ece.mcgill.ca, champagne@ece.mcgill.ca

ABSTRACT

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, inspired by the inherent self-normalization property of the early auditory model, we propose a simplified FFT-based spectrum which is noise-robust in audio classification. To evaluate the comparative performance of the proposed FFT-based spectrum, a three-class (i.e., speech, music and noise) audio classification task is carried out wherein a support vector machine (SVM) is employed as the classifier. Compared to a conventional FFT-based spectrum, both the original auditory spectrum and the proposed self-normalized FFT-based spectrum show more robust performance in noisy test cases. Test results also indicate that the performance of the self-normalized FFT-based spectrum is close to that of the original auditory spectrum, while its computational complexity is significantly lower.

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 algorithms. Saunders [2] proposed a technique, which is based on a measure of energy contour and the distribution of zero-crossing rate (ZCR), to discriminate speech from music on broadcast FM radio. 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 a fast and effective 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 in a classification task, the effect of background noise on the performance of classification has not been investigated widely. A classification algorithm trained using clean 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 auditory spectrum, which is calculated from an early auditory model [1], was proved to be robust in noisy environment due to an inherent self-normalization property which causes spectral enhancement. Recently, this early auditory model has been employed in audio classification and excellent performance has been reported [6] [7]. However, this model is not well-suited for some practical applications due to its high computational requirements and the use of nonlinear processing. Therefore, it would be desirable that this early auditory model be simplified, or even approximated in the frequency domain wherein efficient FFT algorithms are available.

In this paper, inspired by the inherent self-normalization property of the early auditory model [1], we propose a simplified model to calculate a novel self-normalized FFT-based spectrum. To evaluate the comparative performance of the proposed self-normalized FFT-based spectrum, a speech/music/noise 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 auditory spectrum and the self-normalized FFT-based spectrum show more robust performance in noisy test cases. Experimental results also show that the performance of the self-normalized FFT-based spectrum is close to that of the original auditory spectrum, while its computational complexity is reduced by an order of magnitude.

The paper is organized as follows. Section 2 briefly introduces the self-normalization scheme inherent in the early auditory model [1]. Inspired by this self-normalization property, a new model is proposed in Section 3 to calculate a noise-robust FFT-based spectrum. Section 4 explains the extraction of audio features and the setup of the classification tests. The test results are presented in Section 5.

2. SELF-NORMALIZATION INHERENT IN AN EARLY AUDITORY MODEL

2.1. Background on the Early Auditory Model

The auditory spectrum used in this work is calculated from an early auditory model introduced in [1] and [8]. This model, which can be simplified as three-stage process shown in Fig. 1, describes the transformation of an acoustic signal into an internal neural representation referred to as auditory spectrogram. A signal entering the ear 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 nerve and the biophysical process is modeled by the following three steps: a temporal derivative, a nonlinear sigmoid-like compressive function $g(\cdot)$, and a low-pass filter $w(t)$. At the last stage, a lateral inhibitory network (LIN) detects discontinuities in the responses across the tonotopic axis of the auditory nerve array. The operations can be effectively divided into the following steps: a derivative with respect to the tonotopic axis s, a local smoothing $v(s)$, a half-wave rectifying (HWR), and a

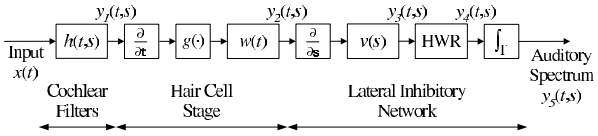


Fig. 1. Schematic description of the early auditory model [1].

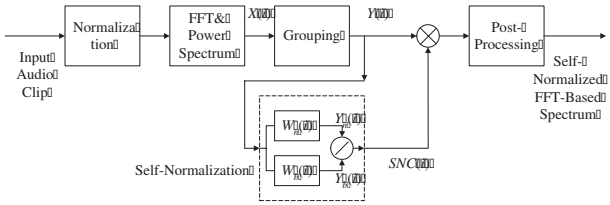


Fig. 2. Schematic description of the proposed FFT-based model.

temporal integration. 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].

2.2. The Inherent Self-Normalization Scheme

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

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

where E denotes statistical expectation, $E[y_5(t,s)]$ is the output 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 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 [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 the fact that the spectral components of the sound signal receive unproportional scaling. 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.

3. A NEW SELF-NORMALIZED FFT-BASED 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 higher than that of a conventional FFT-based spectrum. Therefore, it is desirable to approximate this model in the frequency domain wherein efficient FFT algorithms are available. In this work, by integrating the self-normalization property of the above early auditory model, we propose a new frequency-domain model to calculate a self-normalized FFT-based spectrum. The details of this model, illustrated in Fig. 2, are presented below.

3.1. Normalization of the Input Signal

To make the algorithm adaptable to input signals with different energy levels, each input audio clip is normalized with respect to the square-root value of its average energy.

3.2. Power Spectrum Grouping

Using the normalized audio signal, narrow-band (30 ms) spectra are calculated using a 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

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

where i is the frequency index, and $X(i)$ and $Y(i)$ represent the power spectrum before and after grouping, respectively. From formula (2), 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.

3.3. Spectral Self-Normalization

To apply self-normalization of the aforementioned early auditory model on the above 132-dimensional power spectrum vector, we first define a narrow filter $W_n(i)$ and a broad filter $W_b(i)$ as

$$\begin{aligned} W_n(i) &= \sum_{k=-1}^{1} a_k \delta(i-k) \\ W_b(i) &= \sum_{k=-2}^{2} b_k \delta(i-k) \end{aligned} \tag{3}$$

where a_k's and b_k's are coefficients, and i is the frequency index. Let $Y_n(i)$ and $Y_b(i)$ represent the outputs from filters $W_n(i)$ and $W_b(i)$ respectively, i.e.,

$$\begin{aligned} Y_n(i) &= Y(i) * W_n(i) \\ Y_b(i) &= Y(i) * W_b(i) \end{aligned} \tag{4}$$

where $*$ denotes convolution. Based on $Y_n(i)$ and $Y_b(i)$, a self-normalization coefficient at frequency index i, $SNC(i)$, is defined as

$$SNC(i) = \frac{Y_n(i)}{Y_b(i)} \qquad i = 1, 2, \ \cdots, 132 \tag{5}$$

Finally, the self-normalized spectrum at frequency index i is obtained by multiplying the power spectrum at that frequency index, i.e., $Y(i)$, with the corresponding self-normalization coefficient

$SNC(i)$, and applying a square-root operation. After discarding the first and the last two components, we obtain a 128-dimensional self-normalized spectrum vector.

4. FEATURE EXTRACTION AND CLASSIFICATION TEST

4.1. Audio Sample Database

To carry out performance tests, a generic audio database is built which includes 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. The training set and testing set each contain 3600 one-second audio clips including 1200 speech, 1200 music and 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, clean music and noise. A test with a specific SNR value refers to a test wherein the training set contains clean speech, clean music and noise while the testing set contains noisy speech (with that specific SNR value), noisy music (with that specific SNR value) and noise.

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.

Besides the proposed self-normalized FFT-based spectrum, the conventional FFT-based spectrum is also calculated. It is actually the logarithmic value of $Y(i)$ [1] without the input normalization (see Fig. 2). Based on the conventional and the proposed self-normalized FFT-based spectra, 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 in order to meet the needs from our simulation tests.

The support vector machine (SVM) was recently employed in audio classification task [6] [10]. In this work, we use the SVM^{struct} algorithm [11]– [13] to carry out the classification task.

5. PERFORMANCE ANALYSIS

5.1. Classification Test Results

The test results (i.e., the error classification rate) are listed in Table 1, wherein "AUD", "FFT", and "FFT_SN" represent the original auditory spectrum [1], the conventional FFT-based spectrum, and the proposed self-normalized FFT-based spectrum respectively.

Although the conventional FFT-based spectrum provides an excellent performance in the clean case, its performance degrades rapidly 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 self-normalized FFT-based spectrum are more robust in noisy test cases. Results in Table 1 also indicate that the performance of the proposed self-normalized FFT-based spectrum is close to that of the original auditory spectrum.

Table 1. Error rate (%): the auditory spectrum, the conventional FFT-based spectrum, and the self-normalized FFT-based spectrum.

SNR(dB)	∞	20	15	10	5	Average
AUD	3.06	3.42	3.78	5.92	12.19	5.67
FFT	2.42	22.97	37.39	47.64	55.50	33.18
FFT_SN	2.94	3.22	4.14	6.56	13.78	6.13

Two examples of audio features are shown in Figs. 3 and 4. Fig. 3 shows the FFT-based spectrum features (mean and variance) for a one-second speech clip in clean test case and in noisy test case with 10 dB SNR. At SNR=10 dB, the proposed self-normalized FFT-based spectrum features are close to those in the clean test case. However, this is not the case for the conventional FFT-based spectrum features wherein the change is large. A similar situation can be found in Fig. 4, which shows test results for a one-second music clip. The small difference shown in Figs. 3 and 4 indicates that a property of spectral enhancement or noise suppression, which is inherent in the original early auditory model, is now included in our proposed self-normalized FFT-based model.

5.2. Computational Complexity

The potential of the proposed self-normalized FFT-based model lies in its low computational complexity. A rough estimation of the computational load is carried out by counting the number of the required multiplications per unit time.

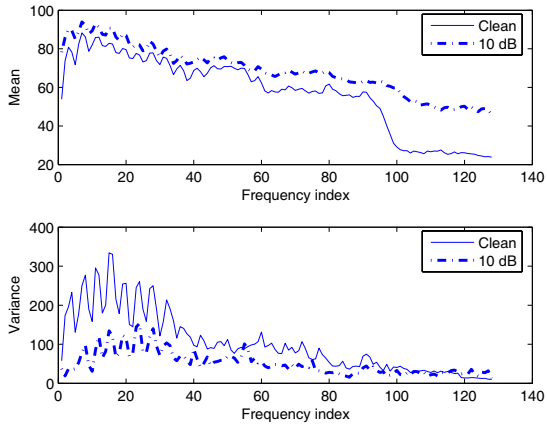
For a one-second audio clip with 16 kHz sampling frequency, based on the implementation [9], the original early auditory model requires more than 3×10^7 multiplications in bandpass filtering. We ignore all other calculations for early auditory model.

For the proposed self-normalized FFT-based model, with a safety margin, we first estimate the ratio of the average machine cycle used for a division to that of a multiplication as 10:1. The corresponding ratio for square-root operation is estimated as 20:1. With these results, for a one-second audio clip, the number of multiplications consumed for the proposed self-normalized FFT-based spectrum is about 1.5×10^6, which is less than 1/20 of the multiplications used for the calculation of the original auditory spectrum.

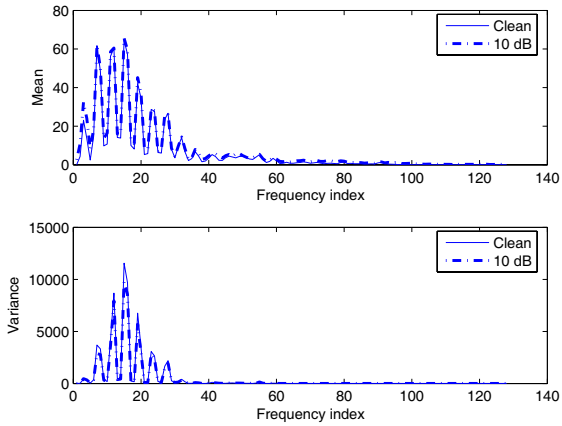
6. CONCLUSIONS

In this paper, inspired by the inherent self-normalization property of an early auditory model introduced in [1], we have proposed a new model to calculate a noise-robust FFT-based spectrum. To evaluate the performance of the original auditory spectrum and the proposed FFT-based spectrum, a three-class (i.e., speech, music and noise) audio classification task has been carried out wherein a support vector machine is used as the classifier. Compared to a conventional FFT-based spectrum, the original auditory spectrum and the proposed self-normalized FFT-based spectrum show more robust performance in noisy test cases. Test results also indicate that the performance of the proposed self-normalized FFT-based spectrum is close to that of the original auditory spectrum, while its computational complexity is significantly lower.

[1] The first and the last two components are discarded in order to keep the dimension as 128.

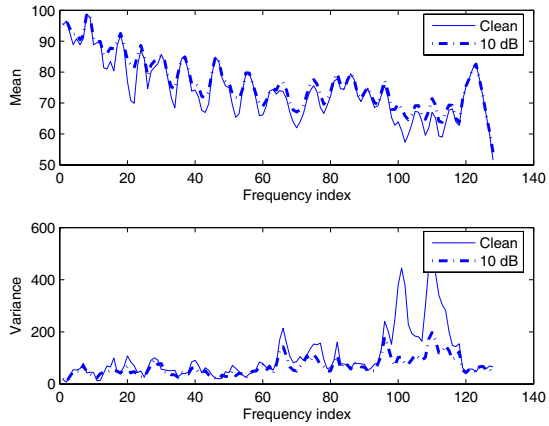


(a) Conventional FFT spectrum features

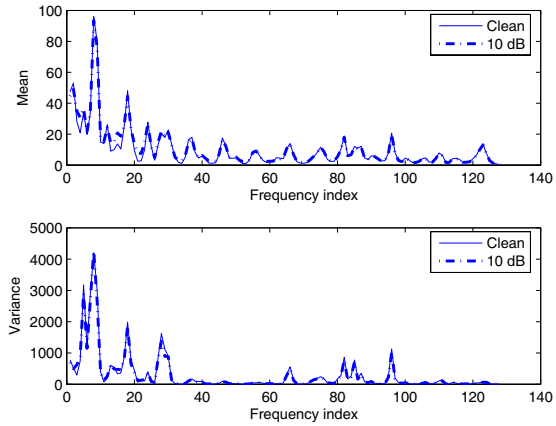


(b) Self-normalized FFT spectrum features

Fig. 3. FFT-based spectrum features for one-second speech clip.



(a) Conventional FFT spectrum features



(b) Self-normalized FFT spectrum features

Fig. 4. FFT-based spectrum features for one-second music clip.

7. REFERENCES

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

[2] J. Saunders, "Real-time discrimination of broadcast speech /music," in *Proc. ICASSP*, 1996, pp. 993–996.

[3] T. Zhang and C.-C. Jay Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. on Speech and Audio Proc.*, 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. on Speech and Audio Proc.*, vol. 10, no. 7, pp. 504–516, October 2002.

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

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

[7] S. Ravindran and D. Anderson, "Low-power audio classification for ubiquitous sensor networks," in *Proc. ICASSP*, 2004, vol. IV, 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/nsl.html.

[10] Y. Li and C. Dorai, "SVM-based audio classification for instructional video analysis," in *Proc. ICASSP*, 2004, vol. V, pp. 897–900.

[11] T. Joachims, "SVM-struct," http://www.cs.cornell.edu/People/tj/.

[12] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector learning for interdependent and structured output spaces," in *Proc. the 21 International Conference on Machine Learning*, 2004.

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