Journal of Korea Multimedia Society Vol. 21, No. 7, July 2018(pp. 737-744)
https://doi.org/10.9717/kmms.2018.21.7.737

Speech/Music Classification Based on the Higher-Order Moments of Subband Energy

Jiin Soo Seo†

ABSTRACT

This paper presents a study on the performance of the higher-order moments for speech/music classification. For a successful speech/music classifier, extracting features that allow direct access to the relevant speech or music specific information is crucial. In addition to the conventional variance-based features, we utilize the higher-order moments of features, such as skewness and kurtosis. Moreover, we investigate the subband decomposition parameters in extracting features, which improves classification accuracy. Experiments on two speech/music datasets, which are publicly available, were performed and show that the higher-order moment features can improve classification accuracy when combined with the conventional variance-based features.

Key words: Speech/Music Classification, Audio Segmentation, Subband Energy, Skewness, Kurtosis

1. INTRODUCTION

Classification of audio under various criterions is indispensable to quickly and reliably respond to the users' search request on a large-size audio archive [1-2]. This paper focuses on one of the issues, speech/music classification [3-8] of an audio stream which is an essential front-end for applications exploiting only speech or music part of the audio signal. The speech part of an audio is necessary for speech or speaker recognition, while music part of it is needed for broadcast music monitoring. Recognizing sections of a signal that do not pertain to the task at hand reduces computation time and allows for more efficient resource allocation [4]. Therefore, speech/music classification accuracy, at front end of the applications, is utmost important, which has a direct influence on further processing stages.

The overview of the speech/music classification is shown in Fig. 1. A speech/music classifier is typically composed of three steps: frame-level spectral feature extraction (typically between 20 and 100 ms), segment-level feature aggregation on a longer duration (typically 1 sec), and statistical classification. As a frame-level spectral feature for speech/music classification, various short-time features have been studied, such as spectral centroid, rolloff and flux [3], zero-crossing rate [5], and the mel-frequency cepstral features [6, 7]. According to the comparative study [8] on the frame-level spectral features, the cepstral features showed best performance, followed by amplitude, pitch, and zero-crossing rate. Since a frame does not contain enough information for classification, the short-time low-level spectral features are integrated, on a longer duration, into a segment-level feature to incorporate temporal acoustic characteristics. Although various integration methods have been studied [3, 6, 7, 8], the variance of the short-time feature vectors in a segment has shown best performance [7]. Finally a statistical classifier is ap-

※ Corresponding Author : Jin Soo Seo, Address: (25457) 7 Jukhun-gil, Gangneung, Gangwon-do, Korea, TEL : +82-33-640-2428, FAX : +82-33-646-0740, E-mail : jsseo@gwnu.ac.kr

Receipt date : Mar. 31, 2018, Revision date : May 21, 2018
Approval date : Jun. 11, 2018

† Dept. of Electrical Eng., Gangneung-Wonju National University

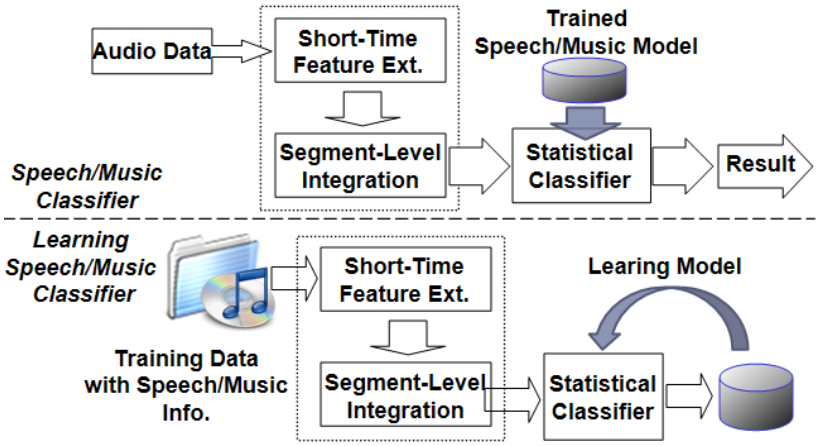

Fig. 1. Overview of the speech/music classification based on segment-level features.

plied to the segment-level feature. Any kind of state-of-the-art statistical classifiers, such as nearest neighbor, Gaussian mixture model, and support vector machine (SVM), can be used in training the speech/music model over the segment-level feature. Recent studies have been conducted to improve the performance of each processing step in Fig. 1. In the short-time feature extraction, speech-specific features representing the excitation source, vocal tract system and syllabic rate of speech have been studied [9]. In the segment-level feature integration, variance-based features have been investigated [7]. In the statistical classifier, different classifier structures, such as ensemble [10] and deep learning [11], have been applied. Among three research directions, this paper focuses on the segment-level feature aggregation.

This paper studies the performance of the higher-order moments, such as skewness and kurtosis, for musical genre classification. Higher-order moments contain supplementary statistical information of the signal over the conventional second-order moment (variance). For example, skewness is a measure of the asymmetry of the distribution, which may represent the relative disposition of the tonal and non-tonal components of the audio spectrum. We note that the higher-order moments are successfully applied to the musical genre classification [12] and the image defect detection [13]. By utilizing the higher-order moments, this paper extends the previous work [7], where variance is used in aggregating subband log energies. Experimental results show that the higher-order moments are conducive in improving the speech/music classification accuracy.

This paper is organized as follows. Section 2 describes the proposed speech/music classifier based on the temporal statistical moment features. Section 3 evaluates the performance of the proposed method. Finally, Section 4 concludes this paper.

2. SPEECH/MUSIC CLASSIFICATION BASED ON THE STATISTICAL MOMENT-BASED FEATURES

For a successful speech/music classifier, extracting features that allow direct access to the relevant speech- or music-specific information is crucial. The focus of this paper is segment-level aggregation of spectral subband log energies.

2.1 Extraction of Spectral Subband Energy

As a frame-level spectral feature, we consider the spectral subband log energy. The spectral subband energy extraction consists of five steps: 1)

an input audio clip, which may have come from a number of different formats, is converted to a single unified format (mono and sampling frequency 16000 Hz); 2) the converted audio signal is split into frames of a length 32 ms (512 samples) with 10 ms overlap (100 samples); 3) each frame is windowed by a Hamming window and transformed into the frequency domain; 4) the spectrum of each frame is processed by a triangular mel-scale filter bank, which is known to be relevant to human auditory perception); and 5) the spectral subband energy is calculated from the filter bank output.

As in the previous works [6, 7], a mel-scale filter bank is used for feature extraction. The mel scale, also called melodic scale, is a perceptual scale of pitches experimentally determined by assigning a perceptual pitch of 1000 mels to a 1000 Hz tone. The mel scale approximates the frequency resolution of the human ear being linear up to 1000 Hz and logarithmic after that. Fitting mel scale exactly in one closed-form equation is not possible. One popular approximation between frequency (f) and mel (m) by O'Shaugnessy is given as follows [14]:

$$m = 1127 \log_e (1 + f/700) \tag{1}$$

The widely-used mel-scale filters are shaped as triangular and equally spaced along the mel scale. Let $P[n, k]$ be the short-time power spectrum of an audio signal at frequency bin k of the n-th frame. As shown in Fig. 2, the log energy $E[n, b]$ of the b-th subband of the n-th frame are calculated by multiplying the magnitude spectrum by the corresponding triangular mel-filter coefficients F_b given by

$$E[n,b] = \log \sum_{k=0}^{L-1} P[n,k] F_b[k] \tag{2}$$

where L is the number of frequency bins (the size of the FFT). In Section 3, we perform a group of experiments to investigate the effect of the number B of subbands on classification performance.

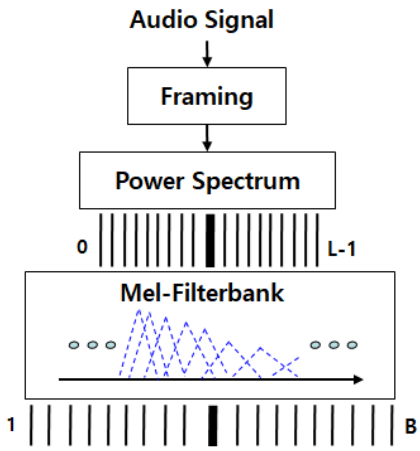

Fig. 2. Calculation of the subband energy using the triangular mel-scale filterbank.

2.2 Segment-Level Statistical Aggregation

As a B-dimensional subband energy vector from a frame does not contain enough information for classification, a segment, which is composed of W consecutive frames (typically 1 s), is used as an input to the classifier as shown in Fig. 1. Since a segment is composed of scores or hundreds of frames, we need to integrate the frame-level features into a segment-level feature. The segment-level features in the proposed method are based on the distributional characteristics of the subband energies.

The temporal characteristics of speech and music are quite different. Three differences between speech and music are noted in [7].

- Silent pause: Speech has more silent moments than music. While talking, people constantly make short pauses between words and syllables.
- Vowel duration: Vowels in music have longer durations than those in speech. The average duration of vowels in speech is 90~100 ms, while vowels in music can last for as long as a couple of seconds especially during choruses.
- Pitch variation: Pitch varies in music much less, and the rate of change is also slower. When a speaker is speaking the pitch can vary by as much as 160 Hz if for example, the speaker is sur-

prised, scared, or enraptured. The variations in music can also be rapid, typically with the tempo of the syllabic rate (4 Hz).

Due to the noted three differences, speech and music have quite different temporal distribution of spectrum. If we consider all three differences and use them in order to construct a new set of features, we can produce an efficient speech/music discriminator, which can also be an accurate discriminator for speech/non-speech classification. To count in the three differences, the previous work [6, 7] proposed a segment-level feature aggregation method based on the variance V of subband energies. The variance $V_s[b]$ of the b-th subband at the s-th segment is given by

$$V_s[b] = \frac{1}{W}\sum_{n=1}^{W}(E[n+N_s,b]-\mu_s[b])^2 \tag{3}$$

where N_s is the start frame position of the s-th segment, andμ_s[b] is the mean of the b-th subband energies of the s-th segment. However, in describing the energy distribution characteristics of a mel-scale subband, variance is not complete. In order to boost the classification accuracy further, we extract the skewness and the kurtosis of the energies in each subband of a segment. The skewness $S_s[b]$ of the b-th subband at the s-th segment is the third-order standardized moment defined as

$$S_s[b] = \frac{\frac{1}{W}\sum_{n=1}^{W}(E[n+N_s,b]-\mu_s[b])^3}{\left(\frac{1}{W}\sum_{n=1}^{W}(E[n+N_s,b]-\mu_s[b])^2\right)^{3/2}} \tag{4}$$

The kurtosis $K_s[b]$ of the b-th subband at the s-th segment is the fourth-order standardized moment defined as

$$K_s[b] = \frac{\frac{1}{W}\sum_{n=1}^{W}(E[n+N_s,b]-\mu_s[b])^4}{\left(\frac{1}{W}\sum_{n=1}^{W}(E[n+N_s,b]-\mu_s[b])^2\right)^{2}} \tag{5}$$

The skewness is a measure of the asymmetry of the distribution, which can depict the silent pauses of a subband spectrum. Since speech has more silent pauses, the distribution of its spectrum will be right-skewed (the mass of the distribution is on the left) [8]. The kurtosis is a measure of the peakedness of the distribution. Although the contribution of the kurtosis to the classification might not be clear, the kurtosis measure can depict the effective dynamic range of the spectrum in a subband.

To examine the validity of the temporal statistical moment for the classification, mean and standard deviation of the each moment were calculated using the GTZAN music/speech dataset dataset are shown in Table 1. By only looking at the mean and the standard deviation, we easily catch that the variance of the subband energies from speech is larger than that from music. Skewness and kurtosis showed tendency similar to variance. To quantitatively check the validity of the considered statistical moments, we calculate the symmetric KL divergence [15] of each moment between speech and music class in Table l by assuming that the moment features of each class follows

Table 1. Mean and standard deviation of the three considered temporal moment features of subband energy (B = 24, W = 100) for speech and music respectively for GTZAN music/speech dataset. The KL divergence between speech and music class assuming Normal distribution

Features	Class	Mean	Standard deviation	KL divergence
Variance	Speech	94.4	70.8	12.6
	Music	28.4	28.5	
Skewness	Speech	0.295	0.811	2.48
	Music	0.075	0.624	
Kurtosis	Speech	2.919	1.85	2.48
	Music	2.919	1.32	

Normal distribution. In terms of the KL divergence in Table 1, variance is the best in discriminating speech and music; skewness and kurtosis have almost the same discriminability.

3. EXPERIMENTAL RESULTS

The speech/music classification accuracy of the higher-order moments was evaluated on the two datasets, which are available online. The first dataset is GTZAN music/speech dataset [16]. The dataset consists of 120 tracks, each 30 s long. Each class (music/speech) has 64 examples. The second dataset is inhouse music/speech dataset, where the music data is from GTZAN genre dataset [17] (composed of 1000 songs over various genres, in total 8.3 hours long), and the speech data was gathered from three different datasets; LibriSpeech [18], Saivt-bnews [19], and speaker-change dataset [20]. The inhouse speech dataset, whose total length is 7.7 hours, comprises various sources of speech including news broadcast, audio book, and movie dialogues. In this experiment, all the audio recordings were resampled to 16000 Hz. Each recording in the dataset was divided into frames of 32 ms overlapped by 22 ms. We computed the mel-scale subband logarithmic energy. The number B of mel-scale subbands was adjusted from 4 to 36. The extracted frame-level energy features were temporally integrated over 100 consecutive frames (W = 100 corresponding to 1 s). Then the SVM classifier (using LIBSVM-library [21] default setting; RBF kernel and C-SVC) was trained and tested in classifying a segment-level feature. The classification accuracy in this paper was obtained using four-fold cross validation.

The classification accuracy as a function of the number of subbands is shown in Fig. 3 and 4 respectively for the GTZAN and the inhouse dataset. As the number of subbands increases, the classification accuracy for speech class increases, but that for music class decreases. With the number

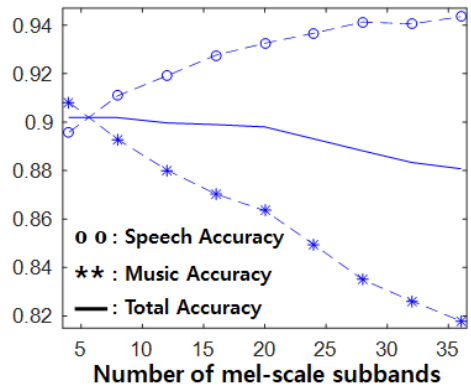

Fig. 3. Classification accuracy(%) versus number of mel-scale subbands for the GTZAN music/speech dataset.

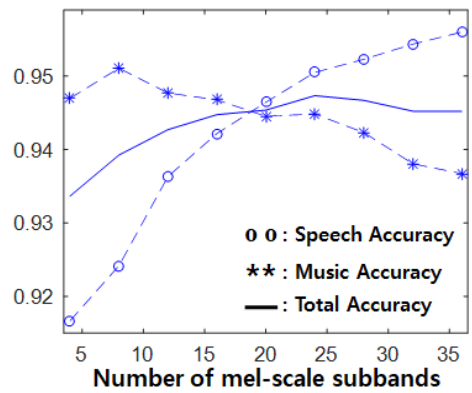

Fig. 4. Classification accuracy (%) versus number of mel-scale subbands for the inhouse dataset.

of subbands, there is a trade-off between accuracies of the speech and the music; modeling speech class needs finer subband decomposition. If an application, such as speaker identification or diarization, demands to parse speech segments accurately, the number of subbands should be more than 24. For the applications (such as music monitoring), where the accuracy on music class is more important, smaller number of subbands can be used. Interestingly, it is observed that the overall classification accuracy tends to be similar if we use more than 12 subbands. To verify the obtained experimental results, other than the mel-scale energy, the classification accuracy of the considered

two datasets were computed with the root-mean square energy and zero-crossing rate [22]. We choose the method in [22] since its source code by its authors is available online which is suitable for comparison. For the GTZAN dataset, the speech and music classification accuracy were 0.89 and 0.90 respectively. For the inhouse dataset, the speech and music classification accuracy were 0.90 and 0.74 respectively. By comparing them with the results in Fig. 3 and 4, the proposed method based on mel-scale energy performs speech and music discrimination reasonably well.

The overall classification accuracies of the GTZAN and the inhouse dataset are given in Table 2 and 3 respectively for the different temporal aggregation methods. As in the previous work [7], the number of subbands was 24 (B=24). When only one type of temporal moment feature was used in aggregating subband energy, variance showed the best performance followed by skewness and kurtosis. The difference of classification accuracy between variance and skewness was about 5% for both datasets. Concatenating the higher-order moments, skewness or kurtosis, and variance improved the classification accuracy up to 4.4% and 2.3% in Table 2 and 3 respectively for the GTZAN and the inhouse dataset. Skewness was more effective in improving accuracy than kurtosis. The performance of concatenating all three considered moments was similar to that of concatenating skewness and variance. Considering the baseline temporal aggregation method [7] using variance already achieved 89.3% and 94.7% for the GTZAN and the inhouse dataset respectively, the classification-accuracy improvement obtained by concatenating skewness and variance is noteworthy.

4. CONCLUSION

For speech/music classification, the effect of the number of subbands and the performance of the higher-order moments were investigated. The higher-order moments, such as skewness and kurtosis, are utilized in temporally aggregating the mel-scale subband energies. Experimental results show that the higher-order distributional moments

Table 2. Classification accuracy (%) of subband energy with temporal aggregation methods for the GTZAN music/speech dataset

Method	Speech Acc.	Music Acc.	Total Acc.
Variance (V) [7]	93.7	85.0	89.3
Skewness (S)	83.8	84.9	84.4
Kurtosis (K)	81.7	79.1	80.4
V+S	95.7	91.7	93.7
V+K	95.4	87.0	91.2
V+S+K	**96.0**	90.4	93.2

Table 3. Classification accuracy (%) of subband energy with temporal aggregation methods for the inhouse dataset

Method	Speech Acc.	Music Acc.	Total Acc.
Variance (V) [7]	95.1	94.5	94.7
Skewness (S)	87.7	90.6	89.3
Kurtosis (K)	80.6	89.1	85.4
V+S	95.9	97.4	96.7
V+K	95.5	96.4	96.0
V+S+K	**95.9**	**97.9**	**97.0**

are effective in improving classification accuracy when combined with the conventional second-order moment. Among the considered higher-order moments, the combination with skewness was more effective in improving the classification accuracy than that with kurtosis. Future work includes employing the higher-order distributional moments for other types of the frame-level features, such as zero-crossing rate and spectral flux, and reducing the dimensionality of the proposed method.

REFERENCE

[1] Z. Fu, G. Lu, K.M. Ting, and D. Zhang, "A Survey of Audio-based Music Classification and Annotation," *IEEE Transactions on Multimedia*, Vol. 13, No. 2, pp. 303-319, 2011.

[2] G. Park, S.Y. Park, and S.J. Kang, "Effective Mood Classification Method Based on Music Segments," *Journal of Korea Multimedia Society*, Vol. 10, No. 3, pp. 391-400, 2007.

[3] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1331-1334, 1997.

[4] G. Sell and P. Clark, "Music Tonality Features for Speech/Music Discrimination," *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2489-2493, 2014.

[5] J. Saunders, "Real Time Discrimination of Broadcast Speech/Music," *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 993-996, 1996.

[6] M. Kos, M. Grasic, and Z. Kacic, "Online Speech/Music Segmentation Based on the Variance Mean of Filter Bank Energy," *EURASIP Journal on Advances in Signal Processing*, pp. 1-13, 2009.

[7] M. Kos, Z. Kacic, and D. Vlaj, "Acoustic Classification and Segmentation Using Modified Spectral Roll-off and Variance-Based Features," *Digital Signal Processing*, Vol. 23, No. 2, pp. 659-674, 2013.

[8] M. Carey, E. Parris, and H. Thomas, "A Comparison of Features for Speech, Music Discrimination," *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 149-152, 1999.

[9] B.K. Khonglah and S.M. Prasanna, "Speech/Music Classification Using Speech-specific Features," *Digital Signal Processing*, Vol. 48, No. 1, pp. 71-83, 2016.

[10] K. Kim, A. Baijal, B. Ko, S. Lee, I. Hwang, and Y. Kim, et al., "Speech Music Discrimination Using an Ensemble of Biased Classifiers," *Proceeding of Audio Engineering Society Convention 139*, pp. 9457, 2015.

[11] A. Pikrakis and S. Theodoridis, "Speech-music Discrimination: A Deep Learning Perspective," *Proceeding of European Signal Processing Conference*, pp. 616-620, 2014.

[12] J. Seo and S. Lee, "Higher-order Moments for Musical Genre Classification," *Signal Processing*, Vol. 91, No. 8, pp. 2154-2157, 2011.

[13] E. Gu and K.H. Park, "Defect Detection Algorithm of TFT-LCD Polarizing Film Using the Probability Density Function Based on Cluster Characteristic," *Journal of Korea Multimedia Society*, Vol. 19, No. 3, pp. 633-641, 2016.

[14] D. O'Shaughnessy, *Speech Communications: Human and Machine*, Wiley-IEEE Press, Piscataway, 1999.

[15] J. Seo, "A Music Similarity Function Based on the Centroid Model," *IEICE Transactions on Information and Systems*, Vol. 96, No. 7, pp. 1573-1576, 2013.

[16] G. Tzanetakis, GTZAN music/speech collection, http://marsyasweb.appspot.com/download/data_sets/ (accessed July, 24, 2018).

[17] G. Tzanetakis and P. Cook, "Musical Genre

Classification of Audio Signals," *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 5, pp. 293-302, 2002.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR Corpus Based on Public Domain Audio Books," *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206-5210, 2015.

[19] H. Ghaemmaghami, D. Dean, and S. Sridharan, "Speaker Attribution of Australian Broadcast News Data," *Proceeding of Workshop on Speech, Language and Audio in Multimedia*, pp. 72-77, 2013.

[20] J. Seo, "Speaker Change Detection Based on a Weighted Distance Measure over the Centroid Model," *IEICE Transactions on Information and Systems*, Vol. 95, No. 5, pp. 1543-1546, 2012.

[21] C. Chang and C. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, No. 3, pp. 155-166, 2011.

[22] C. Panagiotakis and G. Tziritas, "A Speech/Music Discriminator Based on RMS and Zero-crossings," *IEEE Transactions on Multimedia*, Vol. 7, No. 1, pp. 155-166, 2005.

Jiin Soo Seo

He received the B.S., M.S., and Ph.D. degrees from Korea Advanced Institute of Science and Technology in 1998, 2000, and 2005 respectively, all in electrical engineering. He was a senior researcher at ETRI from 2006 to 2008. He joined the Department of Electrical Engineering at Gangneung-Wonju National University in 2008. His research interests are audio and image processing, multimedia retrieval, and pattern recognition.