A novel audio stream segmentation method for audio signal discrimination

Mingu Lee\textsuperscript{a)}, Sang Bae Chon\textsuperscript{b)}, Jee Ho Park\textsuperscript{c)}, and Koeng-Mo Sung\textsuperscript{d)}

Applied Acoustics Laboratory, Institute of New Media and Communications, School of Electrical Engineering and Computer Science, Seoul National University
San 56–1, Shillim-dong, Gwanak-gu, Seoul, 151–744, Republic of Korea

\textsuperscript{a}) klugh34@acoustics.snu.ac.kr
\textsuperscript{b}) strlen@acoustics.snu.ac.kr
\textsuperscript{c}) idanced@acoustics.snu.ac.kr
\textsuperscript{d}) kmsung@acoustics.snu.ac.kr

Abstract: In this letter, an online segmentation algorithm for audio signal discrimination is presented. By detecting abrupt changes in audio signal features and decimating them followed by strength thresholding, segment boundaries of the audio stream are obtained. The resulting segment boundaries provide efficiency and accuracy for the classification stage of audio signal discrimination system.

Keywords: audio stream segmentation, audio signal discrimination, multi-stage grouping, strength thresholding

Classification: Science and engineering for electronics

References

1 Introduction

Audio segmentation is a technique that detects the boundaries of an audio stream that has different characteristics in signal. It is generally used as a very front part of many applications in audio signal processing. In the case-dependent coding schemes such as MPEG Unified Speech and Audio Coding (USAC) [1] or ITU-T Extended Adaptive Multi-Rate – Wideband Codec (AMR-WB+) [2] as an example, the speech-music discriminator can be used in order to decide which codec is more appropriate in a duration divided by an audio segmentation algorithm. In other applications such as content search, speech recognition, or genre classification, it is necessary to implement an audio segmentation in order to obtain proper information from a proper segment.

There have been many researches in audio signal segmentation. In [3], Zhang et al. used all the detected abrupt changes in several specific signal features as segment boundaries. However, too many boundaries cause over-segmentation which decreases the performance of classification by having the segments with very short length. Siegler et al. modeled the probability density function of each class by a Gaussian mixture, and set the segment boundaries at the local maxima of symmetric Kullback-Leibler distances of adjacent frames, but the failure rate of detecting the segment boundaries is over 35% [4]. Meinedo et al. improved Siegler’s method by using Kullback-Leibler distances in cepstral domain with various frame lengths, and lowered the failure rate by 14% [5]. Failing to detect the segment boundaries, however, is more undesirable than over-segmentation in the sense that a segment which contains more than one class of signal cannot be exactly classified by any means. Panagiotakis proposed a speech/music discriminator which models the probability densities of frame energies by a $\chi^2$ distribution and detect the changes of frame characteristics based on Matusita distances [6]. In only speech/music discrimination, Panagiotakis’ method has 3% of detect failure rate.

In this letter, an audio segmentation algorithm for audio signal discrimination system is proposed. The proposed algorithm has no restriction in the choice of signal features and in the categorization of signal. This is advantageous because, given the signal classes, users can choose appropriate signal features for the classification and those features can be used in the segmentation stage without increase of computation. Furthermore, since the segment length is algorithmically lower-limited, each segment has enough length to have characteristics of the class which it belongs, and, in consequence, the classification success rate increases in the classification stage. The proposed segmentation algorithm operates as follows. At first, abrupt changes in all of the features are detected to produce segment boundary candidates. Final segment boundaries are determined by combining the candidates using multi-stage grouping and then strength thresholding to the groups. The performance of a signal discrimination system with the proposed algorithm is verified by comparing with those with conventional segmentation algorithms.
2 Proposed segmentation algorithm for audio discrimination system

Audio discrimination system generally consists of a segmentation block and a classification block. After the segmentation block divides audio signal into a set of chunks from the changes of signal characteristics, the classification block determines each chunk to be a certain class from the signal characteristics. From the view point of complexity, it would be beneficial if the segmentation block uses same features that the classification block uses.

2.1 Detection of abrupt changes in audio signal features

Let \( x_1(l), x_2(l), \ldots, x_P(l) \) denote \( P \) short time features calculated at \( l^{th} \) frame of an audio stream \( s(n) \). The changes of the \( p^{th} \) features at \( l^{th} \) frame, \( d_p(l) \), are calculated as the difference in the mean values of the feature in previous \( N \) frames and next \( N \) frames as shown in Eq. (1).

\[
d_p(l) = \frac{1}{N} \left( \sum_{k=l+1}^{l+N} x_p(k) - \sum_{k=l-N}^{l-1} x_p(k) \right).
\]

Then, the segment boundary candidate \( c(l) \) is defined by

\[
c(l) = \sum_{p=1}^{M} PD(d_p(l))
\]

where \( PD(d(l)) \) is a peak detector which returns 1 if a peak is detected at \( l^{th} \) frame in \( d \). In the experiments in this paper, a simple detector, which pick positions of maximum amplitudes within predefined interval and thresholds them by their value, is used. As a result, segmentation boundary candidate \( c(l) \) represents how many features have abrupt changes at \( l^{th} \) frame.

2.2 Determination of segment boundaries

Usually the peaks detected in this way appear much more frequently than the true segment boundaries. Although all the detected peaks in changes of features may be used as segment boundaries for classification, it would lead misclassification because the over-segmented boundaries cause the short segment lengths. In the proposed method, the candidates are decimated before the classification stage by multi-stage grouping and strength thresholding.

Decimation of the candidates is designed to make each segment longer without missing the true boundaries. It decreases the chances of misclassification, unless true boundaries are missed, because the characteristics of the true class of audio signal are more accurately estimated in the longer segment. Moreover, it also saves computations in the classification stage with decreased number of segments to classify.

In the propose segmentation algorithm, decimations are performed in every \( N \) frames. Let \( G^0(s) = \{(t_0^i, S_0^i) : t_0^i < sN\} \) denotes the set of level zero groups where \( t_0^i \) and \( S_0^i \) are the temporal location in frame index domain of a level zero group and its strength, respectively, \( s \) is the integer hop factor.
and \(t_{i-1}^0 < t_i^0\) for all element index \(i\). The level zero groups are obtained simply from the nonzero elements of \(c(n)\). That is, \(t_i^0\) and \(S_i^0\) are the frame index and amplitude of the \(i^{th}\) nonzero element of \(c(n)\) where all \(t_i^0 < sN\).

Level 1 groups are obtained by combining adjacent level zero groups of which the frame index difference is less than the preset interval limit \(T_1\). That is, if \(t_i^0 - t_{i-1}^0 \leq T_1\) and \(t_{i+1}^0 - t_i^0 > T_1\), then \((t_i^0, S_i^0) \in (t_1^1, S_1^1)\) and \((t_{i-1}^0, S_{i-1}^0) \in (t_1^1, S_1^1)\) but \((t_i^0, S_i^0) \in (t_1^j, S_1^j)\) and \((t_{i+1}^0, S_{i+1}^0) \in (t_1^j, S_1^j)\). Likewise obtaining the set of level \(m\) groups \(G_m^m\) uses \(T_m^m\) as the limit of time differences for adjacent level \(m-1\) groups to belong to them. And the time location and the strength of \(k^{th}\) level \(m\) group can be calculated as

\[
t_k^m = \frac{\sum_{l=q}^{q+Q-1} S_{l-1}^{m-1} t_l^{m-1}}{\sum_{l=q}^{q+Q-1} S_{l-1}^{m-1}},
\]

(3)

\[
S_k^m = \sum_{l=q}^{q+Q-1} S_l^{m-1},
\]

(4)

when only \(\{(t_l^{m-1}, S_l^{m-1}) : l = q, \ldots, q + Q - 1\} \in (t_k^m, S_k^m)\). In other words, they are the centroid and the number of the candidates in the group, respectively. For example, for a level 1 group \((t_1^j, S_1^j)\),

Table I. Pseudo code of the decimation algorithm

In every \(N\) frames, \((n = sN\) and \(s\) is an increasing integer)\n
1. Calculate \(c(r) = \sum_{j=1}^{M} PD(d_j, r)\) where \(r = n - N, n - N + 1, \ldots, n\).

2. Obtain \(G^s(s) = \{(t_i^k, S_i^k) : t_i^k < sN\) and \(t_i^{k+1} > t_i^k\) for all \(i\) from \(c(r)\) as describe above.

3. Given \(M\), obtain \(G^m(s)\) as follows,

\[
G^{\text{temp}} = G^s(s-\eta) \cup G(s) = \{t_i^{\text{temp}}, S_i^{\text{temp}}\}
\]

for \(m = 1, \ldots, M\)

\[
\text{flag} = |G^m(s - 2)\|
\]

\[
\text{count} = |G^m(s - 2)|
\]

for \(i = \text{count} + 1, \ldots, |G^{\text{temp}}|\)

\[
\text{if } t_i^{\text{temp}} > t_{i-1}^{\text{temp}} \geq T_m
\]

\[
S_{\text{count}} = \frac{\sum_{l=1}^{i-1} S_l^{\text{temp}}}{i - \text{flag}}
\]

\[
t_{\text{count}} = \frac{\sum_{l=1}^{i-1} S_l^{\text{temp}} t_l^{\text{temp}}}{S_{\text{count}}}
\]

\[
\text{flag} = i
\]

\[
\text{count} = \text{count} + 1
\]

end

end

\[
G^{m+1}(s) = G^s(s-\eta) \cup \{(t_i^m, S_i^m) : i = 1, \ldots, \text{count} - 1\}
\]

\[
\cup \{(t_k^{m+1}, S_k^{m+1}) : k = \text{flag}, \ldots, i\}
\]

end

\[
G^s = G^{M+1}(s).
\]
\[ t_j^1 = \sum_{i=p}^{p+P-1} t_i^0 / P, \]  
\[ S_j^1 = P, \]  
(5)  
(6)

when only \( \{(t_i^0, S_i^0) | i = p, \ldots, p + P - 1 \} \in (t_j^1, S_j^1) \). The pseudo code of the decimation algorithm is shown in Table I where |G| is the cardinality of the set G. Grouping the candidates in multi-stage is essential because it prevents irrelevantly detected peaks which spread very widely but rather uniformly clustered together.

Then segment boundaries are determined as the final level groups which have strengths larger than a preset threshold \( H \). The threshold can be set to any value considering the number of used features and the accuracy of peak detection. In the experiment in section 3, it is set to 1/2 of the number of features. That roughly means that majority of the features determine whether the instant is segment boundary or not.

3 Experimental results

We conducted an experiment to assess the performance of the proposed algorithm from two viewpoints, segmentation accuracy and overall classification performance. In the experiments, a test stream is made as a mixture of 105 sound sources composed of music, speech, and speech on music with various lengths around four seconds, and the sampling rate is 12 kHz.

In the discrimination system, non-overlapped signal frame of 10 ms is used. Two stage grouping scheme is used for the candidate decimation with the parameters \( N, T_1 \) and \( T_2 \) of 100, 0.2 s and 0.5 s, respectively. For both segmentation and classification, 10 audio signal features, short-time energy, zero crossing rate, spectral centroid, spectral roll-off, spectral flux and first five mel-frequency cepstral coefficients, are used [7]. The strength threshold \( H \) is set to 5, half of the number of the features.

3.1 Segmentation results

With a test stream where adjacent clips are different kind, the proposed segmentation method detects 100% and 82.6% of the real segment boundaries within 0.5 s and 0.3 s accuracy, respectively, while the detection rates are 95.6% and 82.6% for Zhang’s method [3]. Although the stream is over-segmented, each segment has its length at least 0.5 s with which decreased misclassification rate can be expected compared to Zhang’s method which has 41.9% of segments with length less than 0.5 s. An example of the decimation results is depicted in Fig. 1, where the vertical lines and the red stars represent the true segment boundaries and the determined segment boundaries in each step.

3.2 Classification results

The classification performance of the audio stream using the proposed method in the segmentation stage is compared to 3 other segmentation schemes –
Fig. 1. An example of decimation of detected peaks

Fig. 2. Examples of segmentation and classification results using the true segment boundaries, using Zhang’s method and using all detected peaks in 10 features as segment boundaries. Since the optimal choice of classifier is not within the scope of this letter, kNN (k-nearest neighbor) classifier which is very simple and popular classification technique is used in the classification stage only for comparison. 2/3 of the test clips are used to train the classifier and the remainder are used to test. Fig. 2 illustrates the advantages of the proposed method. Using all the detected peaks and Zhang’s
method for segment boundaries causes short false transitions of class in the middle of one kind of signal. They are very undesirable in the cases that there exist some audible artifacts from changes of codec or post-processing according to the class of audio signals. The overall success rates, calculated as the number of correctly classified frames divided by the number of all frames in the test stream, are shown in Table II. Note that only the relative performance of classification is important – the absolute performance can be enhanced by choosing better features and classifier.

### 4 Conclusion

In this letter, an online segmentation method for audio signal discrimination systems is proposed. For classification task which use several signal features, abrupt changes of the features are detected to produce candidates of signal boundaries. These candidates are then decimated by multi-stage grouping and strength thresholding. The proposed method has improved the detection rate in the experiment and let each segment secure enough length for enhanced classification performance as well as decreased computational load.

When the proposed method is used in real-time discrimination systems, it has an algorithmic delay with length of $2^N$. The segmentation performances of signal features are related to the classification performances of them. That is, better features for classification tend to have far distant values between classes which are also advantageous for segmentation.