논문 2013-50-12-27

발생/소멸 패턴을 이용한 비정형 혼합 오디오의 주성분 검출

(Detecting Prominent Content in Unstructured Audio using
Intensity-based Attack/release Patterns)

김 사무엘*

(Samuel Kim[©])

요 약

이 논문에서는 비정형 혼합 오디오 신호에서 청취자에게 전달 되도록 의도된 주된 신호의 종류를 검출해 낼 수 있는 방법을 제안한다. 주된 신호의 종류는 음성, 음악, 음향효과로 정하였으며, 인텐시티 기반의 발생/소멸 패턴에서 추출할 수 있는 특징을 사용하여 그들을 구별할 수 있는 방법을 소개한다. 청취자가 주어진 오디오 신호에서 주된 신호를 받아들이는 주관적인 평가를 반영하기 위해서, 웹기반의 평가시스템을 도입하여 18시간의 다양한 종류의 장르 비디오의 오디오를 평가하였다. 실험을 통하여 비디오의 장르별로 각기 다른 성능을 보이지만 가능성 있는 (음성위주의 토크쇼의 경우 86.7%, 액션 영화 49.3%) 정확도를 보였다.

Abstract

Defining the concept of prominent audio content as the most informative audio content from the users' perspective within a given unstructured audio segment, we propose a simple but robust intensity-based attack/release pattern features to detect the prominent audio content. We also propose a web-based annotation procedure to retrieve users' subjective perception and annotated 18 hours of video clips across various genres, such as cartoon, movie, news, etc. The experiments with a linear classification method whose models are trained for speech, music, and sound effect demonstrate promising - but varying across the genres of programs - results (e.g., 86.7% weighted accuracy for speech-oriented talk shows and 49.3% weighted accuracy for {action movies}).

Keywords : audio classification, audio information retrieval, unstructured audio, genre classification

I. INTRODUCTION

In this work, we focus on detecting *prominent audio content* with applications to delivering optimized audio experience to the users: particularly automatic configuration of sound output settings (e.g. equalizers, reverberators, etc.)^[1] which needs to be managed manually by users otherwise.

Similar applications can be also found in the

literature on developing hearing aids; high-end yet low-profile hearing aids try to classify environment sound^[2] where the devices can automatically config their speech enhancement algorithms according to the users' environments. For example, Kates tried to classify eleven different types of background noise using four signal features, such as the fluctuations of the signal envelope, the mean frequency, low- and high-frequency slopes of the average spectrum^[3]. Nordqvist *et al.* used an HMM-based framework along with quantized cepstral coefficients to classify *speech in traffic noise*, *speech in babble noise* and

* 정회원, Given Zone LLC

© Corresponding Author(E-mail: samuel.kim@allthatsignal.com)
접수일자: 2013년10월24일, 수정완료일: 2013년11월27일

clean speech^[4].

To detect prominent audio signals in heterogeneous mixture of various sound sources is very challenging for several reasons. Firstly, the audio and video content is not well-segmented; an audio stream is continuous not having starting and ending point information while it is required to generate classification results in real-time so that consequent process can be performed appropriately. Secondly, a generic audio is a heterogeneous mixture of several sound sources which often leads to ambiguity issues^[5, 6]. For example, in music shows, suppose a crowd gives a musician a round of applause during musical performances. It is difficult to decide whether applause or music is the prominent audio content in the particular scene. It may vary according to the context of the program or users' subjective standard.

The proposed concept of prominent audio content is somewhat related to other approaches, such as detecting foreground sound from background sound^[7] and primary component retrieval^[8]. Chu *et al.* used a semi-supervised learning method along with matching-pursuit (MP) based features and cepstral features to detect unanticipated sound content from recurring and noise-like yet changing over time sounds, e.g. detecting passing ambulance with siren on a street.

In [8], Piquier *et al.* used 4 Hz modulation energy^[9], modulation entropy, number of segments and durations of segments toward classifying primary component (either speech or music) and detecting jingles.

Here, in this work, we define the prominent audio content as the most informative audio content from the users' perspective within a given audio segment. The rationale behind is that the most informative audio content should be emphasized through the automatic sound configuration. Assuming that the prominency is often revealed in terms of intensity of audio signals, we propose to use intensity-based features, particularly *attack/release* patterns. We also

그림 1. 제안된 시스템의 다이어그램

Fig. 1. A diagram of the proposed system.

prepared a web-based annotation procedure for a scalable data collection of users' subjective perception of audio content. For this particular application, we use three different classes of audio content: *speech*, *music* and *sound effect*. The *speech* category is for the audio content that users pay attention to the linguistic information, e.g., monologues by an anchor or a reporter, multiparty conversations between participants with or without background sounds. The *music* category is, on the other hand, for the audio content that users may appreciate the musical events such as popular music performances (i.e., rock, folk, acapella, etc.) and classical music concerts. Note that background music without any other prominent sound content can be categorized as *music*. There are many other types of sound content that cannot be categorized as neither *speech* nor *music*, and we categorize them as *sound effect*. They may include vocal/non-vocal human-generated sounds (e.g., cheer and applause) and musical/non-musical machine-generated sounds (e.g., jingles and machine guns).

The rest of this paper is organized as follows. In the following section, the description of methodology including feature extraction and machine learning algorithm will be provided. The data collection procedure along with the experimental results are given in Section III followed by conclusions and future plans in Section IV.

II. METHODOLOGY

Fig. 1 shows a brief diagram of the system; when a signal is fed into a system, it is segmented with a desired length to detect the prominent audio content. The system extracts frame-level (usually around 20 milliseconds) features then compute segment-level

features based on frame-level features to represent the corresponding segment. The segment-level features then are fed into supervised classifier to train (or, to predict) the prominent sound content of the segment. The details are following in the subsections.

1. Feature Extraction

In this work, we use a simple intensity-based features; in particular, we model the *attack/release* patterns in the intensity level of audio signals which can be computed directly from the audio signals in frame-levels. In this work, we use 10 millisecond-long sliding windows with 50% overlap to compute energy of the audio signals to represent the intensity.

Fig. 2 shows examples of normalized energy contour extracted from different types of audio content. The energy values are normalized so that they are zero mean and unit standard deviation. This procedure is required since the energy levels may vary across programs. As reported in the literature, we can observe that the energy contour of a speech signal varies slower than the ones of a music signal and a sound effect. On the other hand, the energy contour of a music signal seems to have some periodicity while varying faster than the one of a speech signal. In particular, there are regular abrupt increases/decreases (it may represent beats of the music signal) and the contour fluctuates between the abrupt changes. The energy contour of a sound effect varies faster and less periodic than the others.

In this work, we focus on modeling *attacks* and *releases* of energy contour. Attack is often used to describe an energy level increase when a sound event starts, while release is used to describe an energy level decrease when a sound event ends^[10]. Here we use the terms to represent energy level increase and decrease without concerning sound events (detecting sound event is beyond the scope of this paper). To parameterize the concept of attack and release, we

그림 2. 다른 종류의 오디오 콘텐츠의 에너지 컨투어 패턴 (a) 음성, (b) 음악, 그리고 (c) 음향효과

Fig. 2. Examples of normalized energy contour extracted from different types of audio content: (a) speech, (b) music and (c) sound effect.

first detect positive peaks and negative peaks (local maxima and local minima, respectively) in the energy contour of a given audio signal. Then, we consider the periods from negative peaks to positive peaks as attacks and the periods from positive peaks to negative peaks as releases. Here we use four parameters: durations and amplitudes of attacks and releases (see Fig. 3 for an example). Once we obtain all the parameters considering attacks and releases within a given segment (2 seconds in this paper), we calculate a set of statistics of those four parameters, such as mean, standard deviation, maximum, and

그림 3. 에너지를 이용한 발생/소멸 패턴의 모델링
Fig. 3. Modeling energy contour using the concept of *attack* and *release*.

minimum. In addition to the statistics, we also calculate the total number of peaks and the statistics of normalized energy within a given segment. Therefore, the final dimension of the feature vector is $18(4 \times 4 + 2)$. This simple feature vector is anticipated to represent dynamic characteristics of sound with low computational power. Although it is not within the scope of this paper, deploying other conventional audio features including spectrum-related features and tonal features is worth studying to improve the performance in future.

2. Classification

For the supervised classifier, we utilize a simple linear classification algorithm. In particular, we use the LIBLINEAR library implemented by Fan *et al.*^[11]. It is designed to deal with large-scale datasets and has been used widely in various domains.

Basically, it trains class models to minimize the following given that there are l pairs of instance-label, i.e., (x_i, y_i) ;

$$\min_{\omega} \frac{1}{2} \omega^T \omega + C \sum_{i=1}^l \xi(\omega; x_i, y_i) \quad (1)$$

where C is a penalty function, ξ is a loss function, and w is a weight vector. In this work, we use L2-logistic regression, i.e., $\log(1 + e^{-y_i w^T x_i})$ and set $C=1$ empirically. Since this approach is originally devised to deal with binary classes and the number of classes in this work is three, we apply an

one-to-other strategy for multi-class classification tasks.

III. EXPERIMENTS

1. Data Collection

We have collected approximately 18 hours of video clips across various genres including movies, cartoons, music shows, talk shows, etc., from TV channels and DVDs. Table 1 shows the statistics of collected database accruing to their genres. During the data collection, the video content is encoded with mpeg4 while the audio content is recorded in 48kHz stereo and archived in the mp3 format (192 kbps).

To annotated the collected database, we have developed a web-based annotation tool so that annotators can watch a video clip (see Fig. 4 for the graphical user interface). As the instructions given in Table 2, the annotators are given a video clip and the

표 1. 수집된 비디오 데이터베이스의 장르별 통계
Table 1. Collected data statistics according to genres of video clips.

Genre	Number of programs	Total length (in minutes)
Cartoon	7	106.0
Movie (action)	1	118.0
Movie (drama)	2	120.8
Music show	3	242.7
News	2	96.1
Talk show	1	107.4
Variety show	4	270.9
Total	20	1061.9

표 2. 어노테이션을 위한 인스트럭션
Table 2. Instructions for the annotation process.

1	Click 'play' button to start.
2	The video will be played for 2 seconds and paused.
3	Select the most prominent audio content of the video segments among the given options every 2 seconds.
4	Select other audio content, if any, among the given options (multiple choices are possible).
5	Click 'next' button to continue.
6	Go back Line #2 until the whole clip is annotated.

그림 4. 웹 기반의 어노테이션 유저 인터페이스

Fig. 4. Web-based annotation user interface.

표 3. 수집된 비디오 데이터베이스의 사운드 콘텐츠별 길이

Table 3. Length of collected training data with respect to sound content (in minutes).

	Separated DB	Segmented DB	Total
speech	465.0	42.7	507.7
music	320.4	44.2	364.6
sound effect	259.3	10.0	269.3
Total	1044.7	96.8	1141.5

media player pauses its playback every two seconds. During the pauses, the annotators are given a set of audio content, i.e., *sound effect*, *music*, *speech*, and *silence*, and asked to select one of them as the most prominent audio content. We also have *don't know* option for the cases that are too ambiguous. For background sounds, if any, the annotators can select multiple choices from the sound content list.

The rationale behind of this web-based annotation process is that it can be easily extended to the large database using a crowd to collect subjective labels. Furthermore, it is a lot simpler than the traditional annotation method which requires the annotators should decide the starting and ending point as well. It also matches with our goal that which is the most prominent sound content within a given audio signal rather than figuring out when the event exactly

그림 5. 수집된 데이터베이스의 비디오 장르별 오디오 콘텐츠의 분포도

Fig. 5. Sound content distributions of collected database with respect to different genres of video clips.

happened. We empirically set the length of segments as 2 seconds.

Fig. 5 illustrate the total length of prominent audio content according to genres of video clips. As one may expect, clips like news and talk show are dominated by speech sounds while music show includes considerable portion of music and action movies includes considerable portion of sound effect.

For training, we have collected approximately 19 hours' database (see Table 3 for the statistics of the collected training data). It consists of two different sets of database: separated DB and segmented DB. The separated DB was collected from sound clips that are designated to represent the corresponding sound content, i.e., personal music collections for music, BBC sound effect library^[12] for sound effect, and personal collections of podcast services for speech. The segmented DB, on the other hand, was collected through manual segmentation of video clips. One of the reasons to include segmented clips in the training data is to model mixture of sound different contents, e.g, speech with background music, which is rarely seen in designated database.

2. Results

To evaluate performance of the tasks, we use the unweighted accuracy (UA) and weighted accuracy (WA); UA represents an unweighted mean of per-class recall values while WA represent an weighted mean of per-class recall values. These

(a) 2개 분류
(a) Ternary

(b) 3개 분류
(b) Binary

그림 6. 비디오 장르별 성능 평가
(a) 2개 분류 (b) 3개 분류

Fig. 6. Unweighted accuracy and weighted accuracy according to genres of video clips for (a) ternary and (b) binary cases.

measurements are often used when a system deals with highly unbalanced data such as paralinguistic classification tasks, e.g.,^[13].

Fig. 6 (a) and (b) illustrate UA and WA according to genres of video clips for ternary and binary cases, respectively. The ternary case consider three different types of sound content as individual classes, while the binary case consider music and sound effect as one class so that it becomes a binary classification of speech/non-speech. Note that this is not a binary classification task that individual binary models are trained. Instead, the results are merged to be considered as the same results. As seen in the figures, the performance of the system varies significantly according to the genres of video clips. Speech-oriented programs like *talk show* and *news* show relatively higher performance, while programs like *action movie* which include considerable amount

표 4. 다른 종류의 비디오 장르별 시간으로 나타난 성능 평가 가로는 주관적 평가로 이루어진 레이블, 세로는 예측된 레이블

Table 4. Confusion matrices of classification tasks according to the genre of video clips in terms of minutes. The rows represent the ground-truth (human evaluated) classes while the columns represent the predicted classes.

(a) 만화 (a) cartoon				(b) 영화 (액션) (b) movie (action)			
	SE	MU	SP		SE	MU	SP
SE	2.7	10.5	2.7	SE	20.3	27.9	3.6
MU	0.9	13.0	1.6	MU	13.3	23.4	0.8
SP	4.7	20.2	49.7	SP	5.3	14.6	8.8
(c) 영화 (드라마) (c) movie (drama)				(d) 음악쇼 (d) music show			
	SE	MU	SP		SE	MU	SP
SE	2.0	7.3	0.4	SE	8.0	8.5	0.4
MU	0.4	25.2	0.9	MU	10.5	96.6	8.1
SP	2.2	45.7	36.7	SP	5.2	41.9	63.5
(e) 뉴스 (e) news				(f) 토크쇼 (f) talk show			
	SE	MU	SP		SE	MU	SP
SE	0.4	1.7	0.1	SE	0.0	0.0	0.0
MU	0.4	1.8	0.1	MU	0.3	0.4	0.1
SP	2.6	16.1	72.9	SP	5.3	20.2	81.1
(g) 버라이어티 쇼 (g) variety show				(h) 전체 (h) ALL			
	SE	MU	SP		SE	MU	SP
SE	5.1	19.0	3.1	SE	38.5	74.8	10.4
MU	2.5	20.4	6.0	MU	28.4	180.7	17.6
SP	8.9	81.5	124.4	SP	34.2	240.2	437.1

of sound effect show relatively lower performance.

For further investigation, we examine mis-classification patterns using confusion matrices. Table 4 and 5 represent confusion matrices classification tasks according to the genre of video clips in terms of minutes and per-class accuracy, respectively. The rows represent the ground-truth (human evaluated) classes while the columns represent the predicted classes. The confusion matrices confirm that the performance of the system varies according to the genres of video clips. It can be also observed that the confusion between *sound*

표 5. 다른 종류의 비디오 장르별 정확도로 나타난 성능 평가 가로는 주관적 평가로 이루어진 레이블, 세로는 예측된 레이블

Table 5. Confusion matrices of classification tasks according to the genre of video clips in terms of per-class accuracy (%). The rows represent the ground-truth (human evaluated) classes while the columns represent the predicted classes.

(a) 만화 (a) cartoon				(b) 영화 (액션) (b) movie (action)			
	SE	MU	SP		SE	MU	SP
SE	17.2	66.1	16.7	SE	39.2	53.8	7.0
MU	5.8	84.0	10.2	MU	35.4	62.4	2.2
SP	6.3	27.1	66.6	SP	18.6	50.8	30.6
(c) 영화 (드라마) (c) movie (drama)				(d) 음악쇼 (d) music show			
	SE	MU	SP		SE	MU	SP
SE	20.4	75.2	4.5	SE	47.3	50.1	2.6
MU	1.6	95.1	3.4	MU	9.1	83.8	7.0
SP	2.6	54.0	43.4	SP	4.7	37.9	57.4
(e) 뉴스 (e) news				(f) 토크쇼 (f) talk show			
	SE	MU	SP		SE	MU	SP
SE	17.4	76.4	6.2	SE	0.0	0.0	0.0
MU	19.0	77.1	3.9	MU	42.8	49.4	7.8
SP	2.9	17.6	79.6	SP	5.0	18.9	76.0
(g) 버라이어티 쇼 (g) variety show				(h) 전체 (h) ALL			
	SE	MU	SP		SE	MU	SP
SE	18.7	70.0	11.3	SE	31.3	60.5	8.4
MU	8.6	70.7	20.7	MU	12.5	79.7	7.8
SP	4.2	37.9	57.9	SP	4.8	33.8	61.4

effect and *music* is higher than the confusion between *speech* and others, which is desirable in many applications.

IV. CONCLUSIONS AND FUTURE WORK

Defining the concept of prominent audio content as the most informative audio content from the users' perspective within a given audio segment, we proposed a simple but robust method to detect the prominent audio content, particularly *speech*, *music*, and *sound effect*, in highly heterogeneous sound mixtures, such as the ones in video clips.

Assuming that the prominence is revealed in intensity levels of a signal, we proposed to use intensity-based *attack/release* patterns. The experiments with the collected database whose labels were annotated through a web-based annotation procedure to retrieve users' subjective experiences demonstrated promising results (67.5% overall WA and 61.7% overall UA in ternary cases and 76.2% overall WA and 78.5% overall UA in binary cases). The results also suggested the performance significantly varies according to the genre of the corresponding video clips (86.7% WA for *talk show* and 49.3% WA for *action movie* in ternary cases). This naturally leads us to our future work to integrate with an on-line genre classification algorithm^[14]. We will also investigate on-line audio scene change detection methods to provide starting and ending points^[15] along with integrating other promising approaches, e.g., singing voice detection using vibrato in [16].

REFERENCES

- [1] K.-K. Lee, Y.-H. Cho, and K.-S. Park, "Implementation of an intelligent audio graphic equalizer system," *Journal of The Institute of Electronics Engineers of Korea*, vol. 43, no. 3, pp. 76-83, 2006.
- [2] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: state of the art, challenges, and future trends," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 2915 - 2929, Jan. 2005. [Online]. Available: <http://dx.doi.org/10.1155/ASP.2005.2915>
- [3] J. M. Kates, "Classification of background noises for hearing-aid applications," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 461 - 470, 1995. [Online]. Available: <http://link.aip.org/link/?JAS/97/461/1>
- [4] P. Nordqvist and A. Leijon, "An efficient robust sound classification algorithm for hearing aids," *The Journal of the Acoustical Society of America*, vol. 115, no. 6, pp. 3033 - 3041, 2004. [Online]. Available: <http://link.aip.org/link/>

- JAS/115/3033/1
- [5] S. Kim, P. Georgiou, and S. Narayanan, "Latent acoustic topic models for unstructured audio classification," *APSIPA Transactions of Signal and Information Processing*, vol. 1, Nov. 2012.
 - [6] J.-E. Kim and I.-S. Lee, "Speech/mixed content signal classification based on GMM using MFCC," *Journal of The Institute of Electronics Engineers of Korea*, vol. 50, no. 2, pp. 185-192, 2013.
 - [7] S. Chu, S. Narayanan, and C.-C. J. Kuo, "A semi-supervised learning approach to online audio background detection," in *IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, 2009.
 - [8] J. Pinquier and R. André-Obrecht, "Audio indexing: primary components retrieval," *Multimedia tools and applications*, vol. 30, no. 3, pp. 313 - 330, 2006.
 - [9] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *The Journal of the Acoustical Society of America*, vol. 77, no. 3, pp. 1069 - 1077, 1985. [Online]. Available: <http://link.aip.org/link/?JAS/77/1069/1>
 - [10] E. D. Burnett and H. C. Schweitzer, "Attack and release times of automatic-gain-control gearing aids," *The Journal of the Acoustical Society of America*, vol. 62, no. 3, pp. 784-786, 1977. [Online]. Available: <http://link.aip.org/link/?JAS/62/784/1>
 - [11] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871 - 1874, 2008.
 - [12] The BBC sound effects library - original series. [Online]. Available: <http://www.sound-ideas.com>
 - [13] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The interspeech 2011 speaker state challenge," in *Proceedings of Interspeech*, 2011.
 - [14] S. Kim, P. Georgiou, and S. Narayanan, "On-line genre classification of TV programs using audio content," in *IEEE International Conference of Acoustics, Speech, and Signal Processing*, 2013.
 - [15] K. J. Han, S. Kim, and S. Narayanan, "Strategies to improve the robustness of

agglomerative hierarchical clustering under data source variation for speaker diarization," *IEEE transaction of speech, audio, and language processing*, pp. 1590 - 1601, Nov 2008.

- [16] H. Lachambre, R. André-Obrecht, and J. Pinquier, "Singing voice detection in monophonic and polyphonic contexts," *15th European Signal Processing Conference (EUSIPCO)*, pp. 1344 - 1348, 2009.

— 저 자 소 개 —

김사무엘(정회원)

2003년 연세대학교 전기전자 공학과 학사 졸업.

2005년 연세대학교 전기전자 공학과 석사 졸업.

2010년 미국 남가주대학교 전기 전자공학과 박사 졸업.

2011년 1월 ~ 2011년 6월 미국 남가주대학교 박사 후 과정.

2011년 6월 ~ 2013년 1월 스위스 IDIAP 연구소 박사후 과정.

2013년 1월 ~ 현재 Given Zone, LLC 대표.

<주관심분야 : Information processing and machine learning for big data; specifically human-centered multimedia signals including video, audio, and text with applications to data mining such as modeling, classification, detection, information retrieval >