스펙트럴 피크 트랙 분석을 이용한 음성/음악 분류

*금지수, 이현수
경희대학교 전자계산공학과
e-mail : *jskeum@khu.ac.kr, leehs@khu.ac.kr*

Speech/Music Discrimination Using Spectral Peak Track Analysis

*Ji-Soo Keum, Hyon-Soo Lee
Dept. of Computer Engineering, Kyung Hee University

Abstract

In this study, we propose a speech/music discrimination method using spectral peak track analysis. The proposed method uses the spectral peak track's duration at the same frequency channel for feature parameter. And use the duration threshold to discriminate the speech/music. Experiment result, correct discrimination ratio varies according to threshold, but achieved a performance comparable to another method and has a computational efficient for discrimination.

I. 서론

화자 인덱싱(Speaker Indexing)은 오디오 데이터에서 음성 구간을 대상으로 화자의 변화 위치를 검출해 내고, 동일한 화자의 발성 구간을 찾아내는 기술로 화자별 발성 내용을 요약하고 이해하기 위해 필요한 기술이다. 이러한 화자 인덱싱 기술은 음성 구간을 대상으로 수행되기 때문에 음성과 비음성(음악, 노래, 배경음) 그리고 묵음 구간의 분류가 선행되어야 한다.

기존의 음성, 음악 분류 연구에서는 단구간 에너지(Short Term Energy)와 영교차율(Zero Crossing Ratio)이 주로 사용되었고 주파수(Frequency) 분석을 적용한 방법들도 제안되었다[1][2]. 그리고 음성인식과 화자인식에서 사용하는 MFCC(Mel Frequency Cepstral Coefficient)와 음성과 음악의 주파수 특성을 분석한 스펙트럴 피크 트랙(Spectral Peak Track) 분석 방법도 오디오 내용 분석에 사용되었다[3][4].

본 논문에서는 한국어 방송 데이터의 화자 인덱싱을 위한 전처리 과정으로 스펙트럴 피크 트랙 분석을 이용한 음성과 음악(가요)의 분류 방법을 제안한다. 제안하는 방법은 음성과 음악의 대표적인 차이인 특정 주파수 대역의 스펙트럴 피크 분포와 그 분포 대역의 지속성을 특징 파라미터로 사용한다. 그리고 분류 알고리즘으로는 피크 트랙의 지속성에 대해 임계값을 적용하여 음성과 음악으로 분류한다.

II. 제안하는 음성/음악 분류

2.1 스펙트럴 피크 트랙 분석

오디오 내용 분석을 위한 다양한 특징 파라미터 중에서 스펙트럴 피크 트랙은 음성에 대해서는 자음과 모음의 교차에 의해 짧게 나타나고, 음악에 대해서는 특정 주파수 대역에서 지속성을 가지고 길게 나타나며 리듬에 의해 반복적으로도 나타난다[4]. 그림 1은 5초 길이의 음성과 음악에 대해 추출한 스펙트럴 피크 트랙이다.

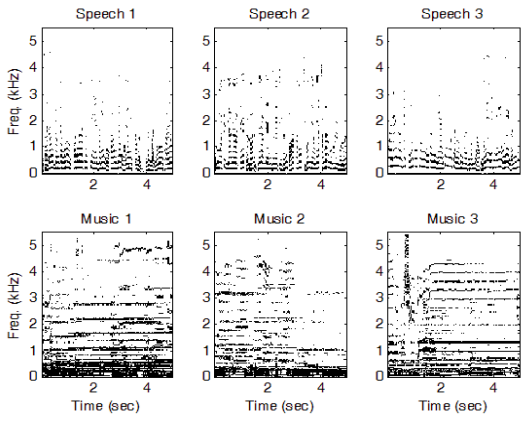

그림 1. 음성/음악의 스펙트럴 피크 트랙

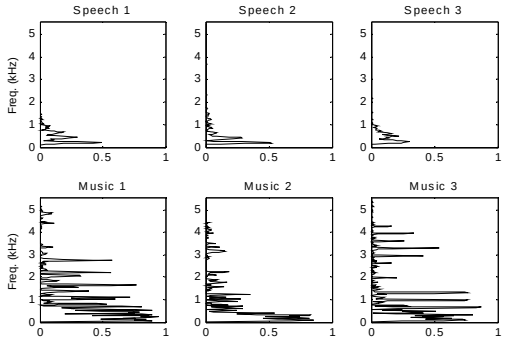

그림 2. 음성/음악에 대한 주파수 채널별 스펙트럴 지속성

그림 2는 [4]의 연구에서 제안한 방법을 이용하여 각 주파수 채널에서 92ms의 지속성을 갖는 스펙트럴 피크 트랙들의 누적을 세그먼트의 프레임수로 나눈 것으로 동일 주파수 채널에서의 지속성을 나타낸다.

2.2 음성/음악 분류 알고리즘

음성과 음악의 분류는 지속성의 임계값(Threshold)을 결정하여, 결정된 임계값보다 큰 채널의 수가 N개보다 같거나 적을 때는 음성, 채널의 수가 N개보다 많을 때는 음악으로 결정한다. 그림 3은 지속성 임계값은 0.5이고, 임계값보다 큰 채널의 수를 5로 정했을 때 음성과 음악으로 분류되는 것을 나타낸다.

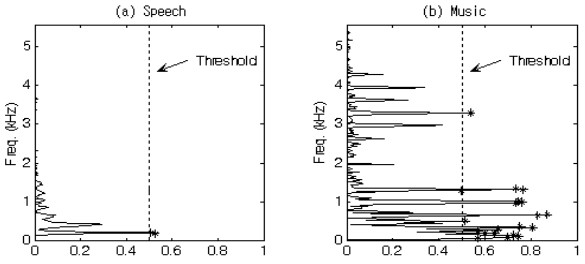

그림 3. 임계값과 채널 수에 의한 음성/음악 분류

III. 음성/음악 분류 실험 및 결과

3.1 실험 방법

음성과 음악 분류 실험을 위해 한국어 방송 뉴스에서 남성 10명과 여성 5명의 음성 데이터를 수집하였고, 음악 데이터로는 가요 20곡을 준비하였다. 수집된 음성 데이터는 1초, 2초, 5초 길이의 세그먼트로 화자별 100개씩 총 1,500개, 음악 데이터는 세그먼트 길이별로 곡당 400개씩 총 8,000개를 준비하였다. 수집된 데이터는 11kHz로 샘플링 되었고 16bit로 양자화 되었으며, 세그먼트의 길이에 의한 성능과 임계값에 의한 성능이 평가되었다.

표 1. 음성/음악 분류 결과

세그먼트 길이	분류	임계값(지속성 및 채널의 수)								
		0.4			0.5			0.6		
		3	4	5	3	4	5	3	4	5
1초	음성	93.8	98.3	99.5	98.9	99.8	100	99.9	100	100
	음악	93.4	90.9	88.6	89.5	86.6	83.4	82.8	78.4	73.8
2초	음성	97.9	99.5	99.8	99.8	100	100	100	100	100
	음악	92.6	90.5	88.3	88.2	84.9	81.7	82.3	78.2	73.3
5초	음성	99.2	99.9	100	100	100	100	100	100	100
	음악	92.3	90.3	87.9	87.2	84.0	80.1	78.2	72.9	67.8

3.2 실험 결과

음성의 주파수 채널별 지속성은 0.4 이상으로 결정할 때 분류의 경계로 적절함을 실험 결과에서 확인할 수 있다. 음악에 대한 실험은 악기로만 연주를 하는 부분과 연주와 노래를 같이하는 다양한 종류의 데이터들이 실험되었기 때문에 음성 결과에 비교하여 낮은 성능을 보였다. 그러나 화자 인덱싱에서는 음성이 음악으로 오분류 되지 않는 것이 더 중요하고, 음악이 음성으로 분류된 것은 후처리 과정을 추가하여 분류해 낼 수 있기 때문에 높은 분류 성능이라 할 수 있다.

IV. 결론 및 향후 연구 방향

본 연구에서는 스펙트럴 피크 트랙을 분석하여 음성과 음악을 분류하는 방법을 제안하였다. 실험결과 기존의 방법에 비해 간단한 알고리즘으로 높은 성능을 얻을 수 있었으며, 향후에는 성능을 보다 높이기 위해 분류된 결과의 후처리 과정의 연구가 진행될 것이다.

참고문헌

[1] Michael J. Carey, Eluned S. Parris and Harvet Lloyd-Thomas, “A Comparison of Feature Speech, Music Discrimination”, Proc. ICASSP, Vol. 1, pp. 149-152, 1999

[2] 이경록, 서봉수, 김진영, “오디오 인덱싱을 위한 음성/음악 분류 특징 비교”, 한국음향학회지 제20권 제2호, pp. 10-15, 2001

[3] Tong Zhang, Jay Kuo, “Audio content analysis for online audiovisual data segmentation and classification”, IEEE Transactions on Speech and Audio Processing, Vol. 9 no. 4, pp. 441-457 2001

[4] Ji-Soo Keum, Chan-Ho Park, Hyon-Soo Lee, “A New Text-independent Speaker Identification using Vector Quantization and Multi-Layer Perceptron”, Advances in Neural Networks - ISNN 2006, LNCS 3972, pp. 165-171, 2006