논문 2010-47SP-5-10

방송 오디오 신호로부터 음악 신호 검출에 관한 연구

(A Study of Automatic Detection of Music Signal from Broadcasting Audio Signal)

윤 원 중*, 박 규 식*

(Won-Jung Yoon and Kyu-Sik Park)

요 약

본 논문에서는 실제 방송 환경에 적용 가능한 방송용 음원 모니터링 시스템을 구축하기 위한 사전연구로 방송 오디오 신호로부터 음악신호 구간을 자동으로 검출할 수 있는 시스템을 제안하였다. 음악구간과 비음악구간의 구분을 위한 특징으로는 사람의 음성 발화 특성을 반영하여 에너지 표준편차와 log 에너지 표준편차 그리고 log 에너지 평균 등 3개의 간단한 시간영역 특징들을 사용하였으며 최종 음악신호 구간 판별은 각 에너지 한계값(threshold)을 이용한 Rule-base 분류를 기반으로 하였다. 실제 FM 라디오 방송 신호를 24시간 녹음하여 진행한 모의실험에서 음악구간 인식률은 96%, 비-음악구간 인식률은 87%를 나타내어 방송용 음원 모니터링 시스템의 전처리기로 손색이 없음을 확인할 수 있었다.

Abstract

In this paper, we proposed an automatic music/non-music signal discrimination system from broadcasting audio signal as a preliminary study of building a sound source monitoring system in real broadcasting environment. By reflecting human speech articulation characteristics, we used three simple time-domain features such as energy standard deviation, log energy standard deviation and log energy mean. Based on the experimental threshold values of each feature, we developed a rule-based algorithm to classify music portion of the input audio signal. For the verification of the proposed algorithm, actual FM broadcasting signal was recorded for 24 hours and used as source input audio signal. From the experimental results, the proposed system can effectively recognize music section with the accuracy of 96% and non-music section with that of 87%, where the performance is good enough to be used as a pre-process module for the a sound source monitoring system.

Keywords: sound source monitoring system, music signal detection, speech/music discrimination.

I. 서 론

최근 인터넷, 컴퓨터 통신과 같은 네트워크의 급속한 발전과 방송용 디지털 오디오 콘텐츠의 증가로 FM 라디오나 TV 방송 신호로부터 음성, 음악, 효과음이나 광고음악 등 여러 유형의 오디오 신호를 자동으로 분류하는 연구가 활발히 진행되고 있다. 예를 들어, 입력 오디오 신호로부터 음악 구간만을 검출하여 해당 음원을 검색할 수 있는 오디오 인덱싱(audio indexing)이나, 음성이나 화자인식의 전처리 과정으로 음성 구간만을 검출하여 음성인식(Speech recognition)이나 화자인식(Speaker recognition)을 수행한다거나, 또는 대역폭이 제한된 멀티미디어 통신환경에서 음성과 음악을 자동 구분하여 각 유형에 맞는 압축 방식을 적용할 수 있는 전송기술 등 모두는 오디오 유형을 자동으로 분류해낼

* 정회원, 단국대학교 컴퓨터과학 및 통계학과
(Dept. of Computer Science and Statistics, Dankook University)

※ 본 논문은 2007년 정부(교육인적자원부)의 재원으로 한국학술진흥재단(KRF-2007-511-D00197)과 2010년도 단국대학교 대학연구비의 지원을 받아 수행된 연구임.

접수일자: 2010년6월30일, 수정완료일: 2010년8월9일

수 있는 기술을 전처리 과정으로 요구한다.

일반적으로 음성이나 음악 등의 오디오 유형을 분류하기위한 시스템은 크게 2단계로 구성된다. 첫째는 각 오디오 유형을 효과적으로 구분할 수 있는 특징 추출(feature extraction), 둘째는 특징 파라메터를 이용하여 오디오 유형을 판별할 수 있는 통계적 패턴 분류나 Rule-base 분류이다. 기존 연구로서 J. Saunders[1]는 라디오 방송에서 음성과 음악의 실시간 분류를 위해 Energy와 zero-crossing rate (ZCR)를 사용하였고, E. Scheirer[2]는 13개의 오디오 특징을 4가지 서로 다른 다차원 분류기에 적용하여 k-d spatial 분류기에서의 최적의 특징 벡터조합을 찾을 수 있었다. A. Pikrakis[3]는 Variable Duration Hidden Markov Model (VDHMM)과 Bayesian Network (BN)을 조합한 복합 구조를 적용하여 좋은 성능을 얻을 수 있었다. 국내 연구 역시 대부분 음성과 음악 신호 분류를 위한 최적의 특징벡터 조합이나 통계적 패턴 분류기의 성능 향상에 대한 연구가 주를 이루고 있으며 약 95% 가까운 좋은 성능을 나타내고 있다[4~9].

그러나 이상에서 살펴본 선행연구 대부분은 음성과 음악 또는 효과음만을 대상으로 하였으며, 연구 내용 또한 음성과 경음악(가수의 음성이 섞여있지 않은 음악)과의 구분을 위한 것이 주를 이루고 있다. 그러나 실제 FM 라디오나 TV 방송에서는 경음악이나 연주 위주의 클래식음악보다 가수의 목소리가 포함되어 있는 대중음악의 방송빈도가 훨씬 높기 때문에 기존 연구를 실제 방송 오디오 신호에 적용하기에는 한계가 있다. 또한, 기존 연구 대부분은 실제 방송 신호를 대상으로 하지 않았거나, 방송 신호를 대상으로 연구가 되었다고 하여도 실험데이터가 충분치 않고, 음성과 음악 구간을 수작업으로 분류한 짧은 오디오 클립(audio clip)들을 대상으로 하였기 때문에 24시간 연속적으로 방송되고 있는 라디오나 TV 방송을 대상으로 하기에는 무리가 있다.

본 논문은 실제 방송 환경에 적용 가능한 방송용 음원 모니터링 시스템을 구축하기 위한 사전연구로 방송 신호로부터 음악신호 구간만을 자동으로 검출할 수 있는 시스템에 대해 연구하였다. 음원 모니터링 시스템은 FM 라디오나 TV 신호로부터 음악신호 구간만을 검출하여 음원을 검색해 해당 음악 콘텐츠에 대한 저작권 보호와 지적 재산권을 행사할 수 있는 시스템으로 다양한 유형의 오디오 신호가 포함된 방송신호로부터 얼마나 정확히 음악신호 구간만을 검출할 수 있느냐가 전체 시스템의 성능을 좌우한다. 실제 방송 신호는 DJ나 게스트들의 음성과 음악만 존재하는 것이 아니라 방송에 도움을 주고 있는 기업들에 대한 광고 방송이나 교통정보 등 다양한 유형의 오디오 신호를 포함하기 때문에 일반적인 음성/음악 판별보다 훨씬 더 정교한 알고리즘을 요구한다.

본 논문에서는 사람의 음성 발화 특징, 즉 사람의 음성 신호에는 호흡이나 발언 간의 순간 등으로 음성신호가 존재하지 않는 정적(silence)구간이 많이 존재한다는 기본적인 사실을 이용하였다. 반면, 가수의 음성이 포함된 음악의 경우 가수의 음성뿐만 아니라 다른 악기들의 연주가 함께 존재하기 때문에 신호 정적 구간이 거의 없는 편이다. 즉, 이는 신호 에너지가 존재하는 구간과 존재하지 않는 구간의 출현이 빈번한 음성구간의 경우 에너지 편차가 많다는 것을 의미하고, 악기들의 연주에 가수의 음성이 더해지는 음악구간의 경우 상대적으로 에너지 편차가 작다는 것을 의미한다. 또한, 광고 신호의 경우에도 기본적인 배경음악(BGM: background music)에 광고 메시지를 음성으로 전달하는 것으로 가정한다면 신호 에너지 편차가 음성보다는 작고 음악보다는 클 것이라는 가정이 가능하다. 본 연구에서는 이러한 원리에 기반해 에너지 표준편차(standard deviation), Log 에너지 표준편차, Log 에너지 평균 등 3개의 간단한 시간-영역 특징 파라메터를 추출하였으며, 최종 음성신호 구간 판별은 각 에너지 한계값(threshold)을 이용한 Rule-base 분류를 기반으로 하였다. 또한 보다 정교한 신호 분류를 위해 1차적으로 확실한 음성신호 구간을 검출한 다음, 나머지 신호구간에 대한 최종 음악구간 판별은 2차 후처리 과정에서 수행하도록 하였다.

본 논문의 구성은 다음과 같다. 먼저 Ⅱ장에서는 제안된 시스템의 구조와 추출되는 특징에 대해 설명하고, Ⅲ장에서는 오디오 방송 신호에서 음악구간과 비-음악구간의 분류를 위한 특징 분석에 대한 설명을 한다. Ⅳ장에서는 실제 방송 신호에 대한 실험 결과를 설명하고, 끝으로 Ⅴ장에서 결론을 맺는다.

Ⅱ. 제안 시스템의 구조 및 특징 추출

본 논문에서 제안하는 시스템의 구조는 그림 1과 같다.

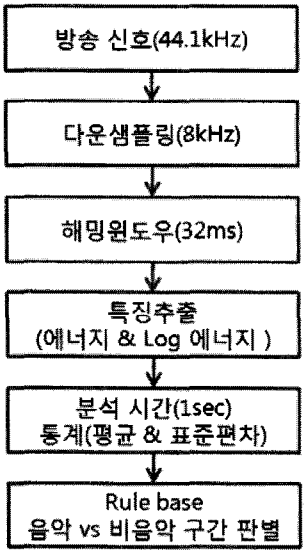


그림 1. 제안된 시스템의 구조
Fig. 1. Proposed system block diagram.

FM 라디오 신호는 각 방송사의 인터넷 라디오 청취 프로그램(미니, 고릴라, 콩 등)이나 홈페이지의 웹에서 듣기 서비스를 이용하여 고품질의 신호가 사운드 카드로 입력이 되지만, 연산량을 고려하여 8kHz로 다운 샘플링 된다. 다운 샘플된 신호는 32ms 해밍 윈도우를 중첩되지 않도록 적용하면서 에너지와 Log 에너지를 추출하고, 음악구간과 비-음악구간을 구분하기 위한 분석 시간인 약 1초(31프레임)간의 분석 구간 내에서의 통계값을 계산한다. 수식(1)은 위의 과정을 표현한 것이다.

$$E(l)=\frac{1}{N_{hop}}\sum_{n=0}^{N_{hop}-1}|s(n+lN_{hop})|^2$$

$$(0\le l\le L-1)$$

$$E_S(k)=\left(\frac{1}{M_{hop}}\sum_{m=0}^{M_{hop}-1}(E(m+kM_{hop})-\overline{E(k)})^2\right)^{\frac{1}{2}} \quad (1)$$

$$(0\le k\le \frac{L}{M_{hop}}-1)$$

$$\overline{E(k)}=\frac{1}{M_{hop}}\sum_{m=0}^{M_{hop}-1}E(m+kM_{hop})$$

위 수식에서 $E(l)$은 l번째 프레임의 에너지이고, $\overline{E(k)}$는 k번째 분석구간(1초) 내에서의 에너지의 평균을, $E_S(k)$는 k번째 분석구간 내에서의 에너지의 표준편차를 나타낸다. L은 신호의 총 프레임 수를 나타내고, N_{hop}은 1 프레임(32ms)에 포함되는 샘플의 수를, M_{hop}은 분석구간에 해당하는 프레임의 수를 나타낸다. Log 에너지에 대한 연산은 에너지를 구한 후 Log를 취한 다음, 식(1)과 동일하게 연산하면 되기 때문에 생략하였다. 분석 구간 내에서 추출된 통계 값을 이용하여 음악구간과 비-음악구간으로 구분을 하게 되는데, 자세한 설명은 다음 장에서 하도록 한다.

Ⅲ. 제안된 알고리즘

본 논문의 구간 검출 시스템은 1차적으로 FM 라디오 신호로부터 음성구간과 비-음성(음악과 광고)구간을 분류하고, 비-음성으로 분류된 구간에서 최종적으로 음악구간을 검출해주는 2단계로 구성이 되어있다. 이번 장에서는 이와 같은 분류구조를 갖게 된 이론적 배경을 실제 FM 라디오 방송 신호에 대한 파라미터 분석을 통해 설명하고자 한다.

1. 음성/비-음성(음악, 광고) 구간의 분류를 위한 특징 파라미터 분석

그림 2는 방송신호에서 음성과 음악 또는 광고 신호가 어떠한 차이점을 갖는지 분석하기 위한 것으로 실제 라디오 방송(FM 91.9MHz, 배철수의 음악캠프)을 녹음한 2시간(7,200초) 분량의 신호이다. 그림에서 S(Speech)는 음성구간, M(Music)은 음악구간, C(Commercial)는 광고구간을 나타내는 것으로 모두 수작업으로 인덱싱 되었다.

그림으로부터 가장먼저 주목할만한 점은 주로 S로 표시되어 있는 DJ나 게스트들의 음성에 대한 신호 에너지 크기가 음악이나 광고구간보다 크게 나타나고 있다는 점이다. 이는 생방송으로 진행되는 라디오 방송의 특성상 음악이나 광고 등은 미리 만들어진 음원들을 정해진 시간에 맞춰 음량을 설정하여 재생하는 반면, DJ 등의 음성은 마이크 음량을 설정해 놓는다 하여도 웃음소리나 기타 대화 환경 등의 음량을 통제할 수 없기 때문일 것이다. 그러나 단순히 신호의 에너지 크기 차이만을 이용한다면 그림 2의 첫 음성과 음악구간과 같이

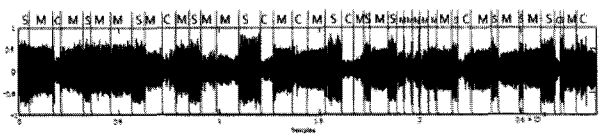


그림 2. FM 라디오 방송 신호의 파형
Fig. 2. Time domain waveform of FM radio broadcasting signal.

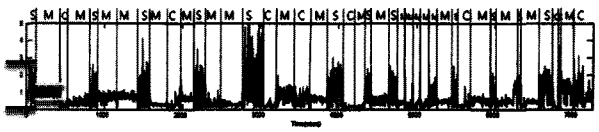


그림 3. 그림 2의 라디오 방송 신호에 대한 1초간의 에너지 표준편차

Fig. 3. Standard deviation of energy within 1 sec analysis window for radio broadcasting signal of Fig. 2.

신호 에너지 크기가 유사한 부분들을 구별할 수 없게 된다.

이에 본 논문에서는 사람의 음성 발화 특징, 즉 사람의 음성신호에는 호흡이나 발언간의 순간 등으로 음성신호가 존재하지 않는 정적구간이 많이 존재한다는 사실에 주목하였다. 반면, 가수의 음성이 포함된 음악의 경우에는 가수의 음성뿐만 아니라 다른 악기들의 연주가 함께 존재하기 때문에 신호 정적 구간은 거의 없게 된다. 즉, 에너지가 존재하는 구간과 존재하지 않는 구간의 출현이 빈번한 음성구간의 경우 에너지 편차가 크다는 것을 의미하고, 기본적으로 악기들의 연주에 가수의 음성이 더해지는 음악이나 배경음악에 음성이 더해지는 형태의 광고의 경우 에너지 편차가 상대적으로 적다는 것을 의미한다. 그림 3은 그림 2의 라디오 방송신호에 대해 분석구간(1초) 내에서의 에너지 표준편차를 나타낸 것이다.

그림 3에서 보듯이 음악과 광고를 포함한 비-음성 구간은 그림 2의 신호 에너지 크기에 상관없이 에너지 표준편차 값의 변화 폭이 좁은 반면, 음성구간은 상대적으로 에너지 표준편차 값의 변화 폭이 큰 것을 확인할 수 있다. 또한 그림 3의 약 5,000초 부분의 연속된 짧은 음악구간들은 DJ가 음악을 연속적으로 재생시키면서 음악차트의 순위를 알려주는 경우로 음악 재생 중간 약 10초 정도의 짧은 시간에 해당하는 DJ의 음성구간도 정확하게 찾아내고 있다.

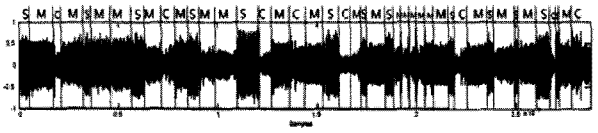
그러나 그림에서와 같이 광고와 음악, 즉 비-음성 구간이 연이어 재생되는 구간에서는 광고와 음악구간의 구분이 모호하기 때문에 에너지 표준편차 특징만으로는 광고와 음악을 구분하는 것이 어렵다는 것을 알 수 있다. 또한, 그림 3의 약 5,000초 부분의 짧은 음악구간들의 마지막부분처럼 다른 악기의 연주가 거의 없고 가수의 음성이 주를 이루는 Rap 음악의 경우 음성구간과 비슷한 특성을 나타내고 있어 음악구간을 음성구간으로 오 분류하는 문제가 발생할 수도 있다. 결론적으로 에너지 표준편차 특징은 음성과 비-음성(광고/음악) 구간을 효과적으로 분류할 수 있지만, 광고와 음악구간 분류에는 한계가 있다 할 수 있다.

2. 비-음성(광고, 음악) 구간에서 음악 신호 분류를 위한 특징 파라미터 분석

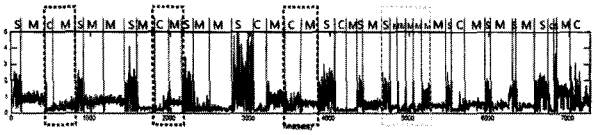
일반적으로 FM 라디오 방송 신호에는 DJ나 게스트들의 음성과 음악만 존재하는 것이 아니라 기업들에 대한 광고나 교통정보 등 다양한 유형의 오디오 신호를 포함하고 있기 때문에 정확한 구간 검출을 위해서는 광고신호와 음악신호의 구분이 필수적이다. 그러나 전절에서 설명한 바와 같이 광고신호는 음악신호의 에너지 표준편차 특성과 유사한 형태를 보이고 있기 때문에 광고와 음악이 연달아 재생되는 구간에서는 전체 (광고+음악) 구간을 음악구간으로 잘못인식하게 되는 문제가 발생하게 된다.

본 논문에서는 이러한 문제를 해결하기 위해 광고와 음악의 주된 의사전달 방법의 차이에 주목하였다. 광고는 전달하고자 하는 메시지를 주로 음성을 통해서 표현하는데 반해, 음악은 가사와 멜로디 모두를 통해서 표현한다는 것이다. 즉, 광고에서의 음악은 BGM으로의 성격이 강하고 정작 전달하고자 하는 내용은 광고 출연자들의 음성으로 강조하여 표현되고 있으며, 음악에서는 가수의 음성과 각 악기들의 연주가 어우러져서 표현된다는 것이다. 또한, 각 광고들 사이에는 이들을 구분하기 위한 정적 구간이 필연적으로 존재한다는 것 역시 광고와 음악을 구분하는데 중요한 척도가 될 수 있다. 이러한 특성들을 이용해 본 논문에서는 광고와 음악신호 구간의 구분을 위한 특징으로 프레임간의 에너지 편차와 에너지 존재 유무를 보다 세밀히 표현할 수 있는 Log 에너지 표준편차와 평균을 사용하였다. 그림 4는 그림 2의 라디오 방송신호에 대한 에너지 표준편차, Log 에너지 표준편차와 평균을 비교 도시한 것이다.

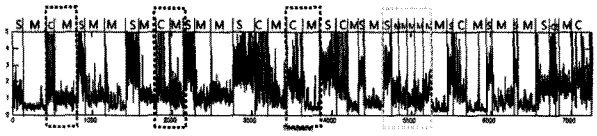
그림 4(c)와 (d)의 Log 에너지 표준편차와 평균에서, 광고와 음악이 연속되는 구간(처음 3개의 파선 박스)들을 살펴보면 광고의 경우 Log 에너지 표준편차가 음악구간과 달리 약 3 이상의 큰 값들을 포함하는 것을 확인할 수 있으며, Log 에너지 평균 역시 음성구간과 광고구간에서 -20에 근접한 반면, 음악구간의 경우 -12 이상의 일정한 값의 범위를 갖는 것을 확인할 수 있다. 그림에서 Log 에너지가 -20에 근접하다는 것은 프레임의 에너지가 거의 없는 정적구간을 의미하며 음성과 광



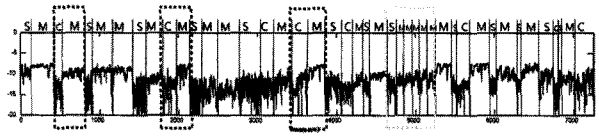
(a) 라디오 방송 신호의 파형



(b) 에너지 표준편차



(c) Log 에너지 표준편차



(d) Log 에너지 평균

그림 4. (a) 라디오 방송 신호, (b) 에너지 표준편차, (c) Log 에너지 표준편차, (d) Log 에너지 평균

Fig. 4. (a) radio broadcasting signal, (b) energy standard deviation, (c) Log energy standard deviation and (d) Log energy mean.

고구간에 이러한 특성들이 주로 나타나고 있다.

결론적으로 그림 4(b)의 에너지 표준편차 특징에서 구분이 힘들었던 광고와 음악 구간이 Log 에너지 표준편차와 평균을 이용해 구분이 가능해졌음을 확인할 수 있다. 또한 전절에서 언급했던 약 5,000초 부분의 짧은 음악구간들의 마지막부분의 Rap 음악(마지막 파선 박스)을 음성으로 오분류하는 문제도 해결할 수 있음을 확인할 수 있다. 따라서 본 논문에서는 분석구간 내에서의 에너지 표준편차, Log 에너지 표준편차 그리고 Log 에너지 평균의 3가지 특징의 한계값을 설정하였으며 이를 이용한 Rule-base 기법으로 음악/비음악 구간을 판별하였다.

3. 제안된 Rule-base 알고리즘

제안 시스템은 FM 라디오나 TV 방송에서 아나운서/MC/DJ/게스트들의 음성과 광고신호 등의 비-음악구간을 제외한 음악신호 구간만을 인식하여 음원 인식 모듈로 전달해주기 위한 전처리기 역할을 수행한다. 먼저 입력된 방송 신호로부터 1초간의 분석 구간을 설정하고 32msec 프레임 단위로 에너지와 Log 에너지 특징을 추출한 후 다양한 실험을 거쳐 에너지 표준편차, Log 에너지 표준편차, Log 에너지 평균 등 3개의 에너지 threshold 값을 설정하였으며 이를 이용해 Rule-base 기반으로 최종 음악신호 구간을 판별하였다.

시스템의 동작원리는 다음과 같다. 우선 1단계에서 방송 신호로부터 확실한 비-음악구간과 음악구간을 설정하고, 2단계에서는 음악구간으로 설정하기 모호한 구간(음성이 주를 이루는 음악들)에 대해서 최종적으로 음악/비-음악구간을 설정하게 된다. 이 때, 2장에서 추출한 특징인 에너지 표준편차, Log 에너지 평균 및 표준편차($E_S(k)$, $\overline{LogE(k)}$, $LogE_S(k)$)들이 음악구간을 효과적으로 인식할 수 있는 적절한 Rule을 설정해주는 것이 필요하며 본 논문에서는 실험을 통해 다음과 같은 Rule을 설정하였다.

먼저 비-음악구간과 음악구간의 구분을 위해 추출된 특징들의 4초간 변화를 살펴보고 해당 4초 구간이 비-음악구간인지 음악구간인지를 판단한다. 만약 음악구간이 1분 이상 지속될 경우 해당구간을 음악구간으로 최종 판단하게 된다. 이는 방송되는 음원의 시간을 확인해본 결과 1분 이하로 방송되는 음원이 거의 없었기 때문이며, 논문 [9]의 의견과도 일치한다. 결국 비-음악구간의 판단을 위해서는 4초간의 특징들의 변화를 확인하고, 음악구간의 판단을 위해서는 4초간의 특징의 변화를 1분 동안 지켜본 후 1분 동안 음악의 특성을 유지할 경우 음악구간으로 최종 판단하게 된다는 것이다. 한편 음악구간의 판단을 위해 1분을 4초씩 나누어서 판단하는 이유는 우선, 32ms를 1프레임으로 분석을 하다보면 1초마다 64샘플씩의 잔여 데이터들이 존재하게 되는데, 이러한 데이터들이 4초마다 1프레임의 분량으로 정확히 계산되기 때문이다. 또한, 시스템의 오인식률을 낮추기 위한 것으로, 1분 이내에 간헐적으로 비-음악구간으로 판단되는 구간이 존재하더라도 그 구간이 10%~20% 미만일 경우에는 음악구간일 확률이 높기 때문에 음원인식 모듈(해당 음악구간으로 판단이 을 통해 음원인식을 시도할 수 있도록 하기 위함이다. 그림 5는 본 논문에서 제안한 4초 단위의 음악구간 검출 시스템의 블록도이며, 각 단계의 임계값들은 실험을 통해 가장 적절하다고 판단된 값들이다.

처음 특징추출 단계에서 추출된 4초 분량의 특징들이 입력되면 그림 5와 같이 확실한 비-음악구간(1)과 확실한 음악구간(2)을 설정하고, 마지막으로 모호한 구

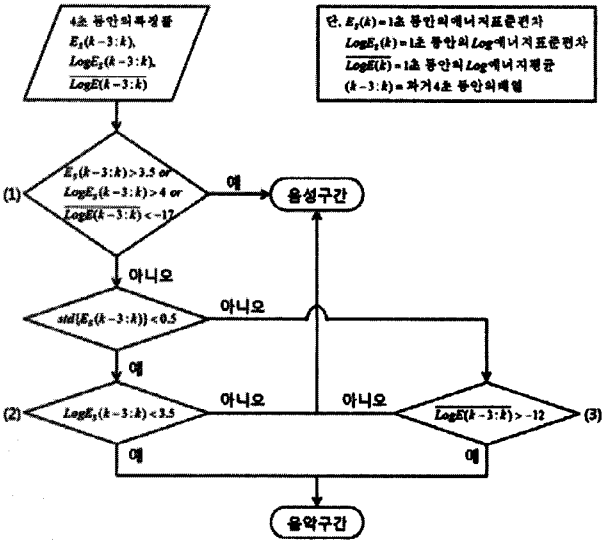


그림 5. 음악 구간 검출 시스템의 블록도
Fig. 5. System block diagram for music signal detection.

간(3)에 대한 최종 판단을 내리게 된다. 먼저, 음성 신호의 경우 음악에 비해 상대적으로 정적구간을 많이 포함하고 있기 때문에 4초 이내에 $E_S(k)$(에너지 표준편차)가 3.5 이상인 경우와 $LogE_S(k)$(로그 에너지 표준편차)가 4 이상인 경우 그리고 $\overline{LogE(k)}$(Log 에너지 평균)이 -17보다 작은 경우를 포함한다면 확실한 비-음악구간으로 설정을 한다. 다음으로는 음성을 제외한 신호구간에 대해 확실한 음악구간을 설정하게 되는데, 확실한 음악구간의 경우 악기들의 반주가 지속적으로 존재하기 때문에 정적구간이 거의 없다고 볼 수 있으며, 이에 따라 4초 이내에 추출된 $E_S(k)$들의 표준편차 값이 0.5 이하이면서 동시에 $LogE_S(k)$ 값이 모두 3.5를 넘지 않는 신호구간을 확실한 음악구간으로 설정한다. 이는 대부분의 음악에서 에너지 변화가 크지 않다는 특성을 이용한 것이다. 그러나 Hip-hop 등 Rap이 많은 음악이나 아카펠라(a cappella)처럼 사람의 음성이 주를 이루는 음악은 에너지 변화가 격렬하여 음성신호와 유사한 특성을 보이게 되는데, 이처럼 음성인지 음악인지 구분하기 모호한 구간에 대해서는 부가적으로 4초간의 $\overline{LogE(k)}$을 비교하여 그 값이 -12보다 클 때는 음악구간으로 설정하고 그렇지 않다면 최종적으로 비-음악구간으로 설정을 하게 된다.

IV. 실험 데이터 구성 및 실험 결과

본 논문에서는 제안 시스템의 성능평가를 위해 FM 라디오 방송 24시간 분량을 녹음하여 실험 데이터로 사용하였다. FM 라디오 방송은 MBC FM4U의 2009년 6월 6일 방송분으로 총 190개의 음악구간과 254개의 비-음악구간(175개의 음성구간, 79개의 광고구간)으로 구성되어 있다. 기존의 연구들에서는 각 방송사의 특정 음악장르들만을 대상으로 실험을 진행하거나 몇 시간 분량만을 대상으로 실험을 하였다. 그러나 실제 방송환경에 적용 가능한 음악/비음악 분류기에 대한 실험을 위해서는 모든 음악장르에 대한 실험이 필요하다고 판단하여 특정 방송사의 방송을 24시간 녹음하였다. 실제로 방송사의 음악전문채널의 편성을 살펴보면 새벽시간대(0시~6시)에는 클래식이나 발라드 위주의 잔잔한 음악들이 주로 편성되어 있고, 출근 시간대(6시~9시)에는 활기찬 하루의 시작을 위한 음악들이, 주부들의 청취율이 높은 시간대(9시~12시)에는 올드 팝이나 영화음악들이 편성된다. 이후로 정오부터 자정까지는 나른한 오후를 깨울 수 있는 음악들과 최신곡 위주의 음악들이 편성되어 다양한 장르의 음악들과 다양한 DJ, 게스트들의 음성에 대한 실험이 가능하다. 또한, 본 실험에서는 MBC 방송사에 대한 실험 결과만을 수록하였지만, 나머지 방송국의 방송 신호에 대해서도 유사한 실험 결과를 보이기 때문에, 이들을 제외하였다.

표 1은 FM 라디오 방송 신호에 대한 실험결과이며, 그림 6은 오인식 구간에 대한 분석 결과이다.

실험 결과 음악구간을 정확히 인식하고 저장한 경우(방송된 음원 길이의 98% 이상을 저장)는 190개 구간에 대하여 182개 구간으로써 96%의 정확도를 보였으며, 음성이나 광고구간으로 잘못 분류된 8개 구간은 그림 6의 중앙에 강조된 부분처럼 Rap 음악 원곡의 50%~60% 정도의 길이만을 음악으로 인식하였기 때문에 인

표 1. 자동 음악구간 인식 알고리즘의 실험결과
Table 1. Result of automatic musical interval recognition algorithms.

		음악구간 인식	비-음악구간 인식
음악구간		182(96%)	8(4%)
비-음악구간	음성구간	22(13%)	153(87%)
	광고구간	11(14%)	68(86%)

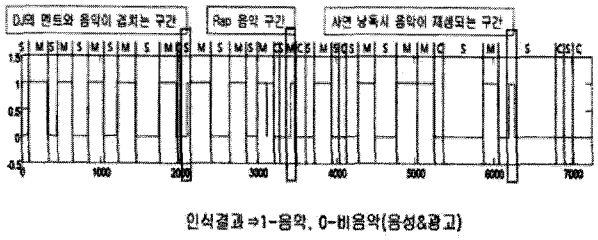


그림 6. 오인식 구간에 대한 분석
Fig. 6. Analysis for recognition error interval.

식실패로 구분된 것들이다. 하지만, 본 연구의 최종 목표인 방송용 음원 모니터링 시스템에서는 50% 정도만 음악구간으로 인식된다 하여도, 저장된 일부분의 음악 구간만으로도 음원인식모듈에서 해당 음원을 정확히 찾아낼 수 있기 때문에 본 연구의 제한 조건을 좀 더 완화할 수 시킬 수 있어 보다 높은 성능의 방송용 음원 모니터링 시스템의 구축이 가능할 것으로 기대된다.

한편, 비-음악구간(광고구간이나 음성구간)을 음악구간으로 잘못 인식한 구간의 성능은 전체 254개 중 33개 구간으로 87%의 성능을 나타내고 있다. 오인식된 구간들을 분석해 본 결과 32개의 구간이 그림 6의 앞, 뒤 박스로 강조된 부분처럼 DJ가 사연을 낭독하면서 BGM이 재생되는 구간이나 DJ가 음악을 미리 재생시켜놓은 상태에서 음악에 대한 설명을 하는 구간, 또는 광고에 이어서 프로그램의 로고송 등이 재생되는 구간으로 음악이 전혀 재생되지 않고 있는 구간을 음악구간으로 인식한 경우는 1구간 밖에 되지 않는다.

V. 결 론

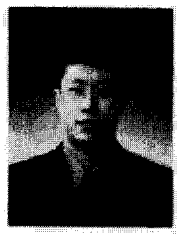
본 논문에서는 방송용 음원 모니터링 시스템의 사전 연구로 방송용 오디오 신호로부터 음악신호만을 자동 검출할 수 있는 알고리즘에 대한 연구를 진행하였다. 자동 음악구간 검출기는 TV나 라디오 방송에서 아나운서/MC/DJ/게스트들의 음성을 제외한 음악 구간들만을 검출하여 음원 인식 모듈로 전달하게 된다. 본 논문에서는 인간의 발화 특징과 광고와 음악의 근본적 차이에 근거하여 에너지 표준편차, Log 에너지 표준편차, 평균 등의 특징을 이용하여 실제 방송 라디오 신호로부터 음악구간과 비-음악구간을 효과적으로 분류할 수 있음을 입증하였다.

본 논문에서 사용된 특징들은 적은 연산량으로 계산할 수 있는 프레임 에너지와 Log 에너지로서 시스템의 복잡도를 낮출 수 있는 장점이 있다. 향후 연구로는 본 자동 음악구간 검출시스템과 연동할 수 있는 자동 음원 모니터링 시스템에 대한 연구를 진행하고자 한다.

참 고 문 헌

[1] J. Saunders, "Real-time discrimination of broadcast speech/music", in Proc. ICASSP 1996, vol 2, pages 993-996, Atlanta, May 1996.
[2] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator", in Proc. ICASSP 1997, pages 1331 - 1334, Munich, Germany
[3] C. Panagiotakis and G. Tziritas, "A Speech/Music Discriminator Based on RMS and Zero-Crossings", IEEE Trans. Multimedia, vol. 7(1), pp. 155 - 166, Feb. 2005.
[4] 이경록, 서봉수, 김진영, "오디오 인덱싱을 위한 음성/음악 분류 특징 비교", 한국음향학회지, 제 20권, 2호, pp. 10-15, 2001.
[5] 장형종, 엄정권, 인준식, "FM 방송 중 블록 단위 음성 음악 판별 시스템의 설계 및 구현", 한국퍼지 및 지능시스템학회 추계학술대회논문집, 제 17권, 2호, 2007.
[6] 금지수, 임성길, 이현수, "스펙트럼 분석과 신경망을 이용한 음성/음악 분류", 한국음향학회지, 제 26권, 5호, pp. 207-213, 2007.
[7] 김봉완, 최대림, 이용주, "멜 켑스트럼 모듈레이션 에너지를 이용한 음성/음악 판별", 말소리, 제 64호, pp. 89-103, 2007.
[8] 최무열, 김형순, "MFCC의 단구간 시간 평균을 이용한 음성/음악 판별 파라미터 성능 향상", 말소리, 제 64호, pp. 155-169, 2007.
[9] 강현우, "FM 라디오 환경에서의 실시간 음악 판별 시스템 구현", 정보처리학회논문지, 16권 B편, 2호, pp. 151-156, 2009.

저 자 소 개



윤 원 중(정회원)

2003년 상명대학교 정보통신학과 학사 졸업.
2005년 단국대학교 컴퓨터과학 및 통계학과 석사 졸업.
2010년 단국대학교 컴퓨터과학 및 통계학과 박사 졸업.

<주관심분야 : 음성 및 음향신호처리, 멀티미디어 신호처리, DSP 시스템 구현>

박 규 식(정회원)

1986년 Polytechnic University 전자공학과 학사 졸업.
1988년 Polytechnic University 전자공학과 석사 졸업.
1993년 Polytechnic University 전자공학과 박사 졸업.
1994년~1996년 삼성전자 마이크로사업부, 선임연구원.
1996년~2001년 상명대학교 컴퓨터·정보통신공학부 조교수.
2001년~현재 단국대학교 컴퓨터학부 교수.

<주관심분야 : 음성 및 음향신호처리, 멀티미디어 신호처리, DSP 시스템 구현>