실시간 FM 방송중 음악/음성 검출에 관한 연구

황 진 만, 강 동 욱, 김 기 두
국민대학교 전자정보통신공학부
전화 : 02-910-4645 / 핸드폰 : 016-737-0521

A Study on Real-time Discrimination of FM Radio Broadcast Speech/Music

Jin-Man Hwang, Dong-Wook Kang, Ki-Doo Kim
Dept. of Electronics Engineering, Kookmin University
E-mail: choonjang2@kookmin.ac.kr

Abstract

본 논문은 FM 라디오 방송중의 오디오 신호를 블록단위로 음악 및 음성을 검출하는 알고리즘에 대한 것으로, 이를 기반으로 방송중의 노래(가요, 팝, 클래식...)만을 자동으로 인식하여 녹음하는 알고리즘을 개발한다. 본 논문에서는 기존에 제안되었던 것[1-4]과 같이 단지 음악과 음성을 구분함과 동시에 음악구간의 논리적 조합으로 이루어진 노래를 자동으로 인식하여 녹음하는 것을 알고리즘의 최종 목표로 한다. 알고리즘의 접근 역시 기존의 음소단위의 모델링을 거치는 GMM 기반의 접근이 아니기 때문에 모델링에 대한 훈련과정이 필요 없고, 시간영역에서의 오디오신호가 가지고 있는 직관적인 특징을 분석함으로써 비교적 적은 연산으로 실시간 구현이 가능하다.

1. 서론

사람의 귀는 방송중의 신호를 음악, 사람의 말, 잡음이 섞인 음악 등에 대하여 별 불편함 없이 인식할 수가 있다. 따라서 이를 구현하기 위해서는 보통의 인식 시스템에처럼 오디오 신호 중의 음악, 음성, 묵음 등을 특징별로 분류하고, 이에 알맞은 패턴을 수집하여 사람의 뇌와 비슷한 구조인 신경망을 이용하여 패턴을 학습시키는 인식과정을 거치게 되는데, 음성인식이 바로 이와 같은 구조이다. 그러나 이와 같은 방법으로 음악과 음성을 구분하기 위해서 음성 및 음악, 묵음 등의 클래스마다 특징을 추출하여 매우 다양한 패턴에 대한 수학적 모델링을 한다는 것은 쉽지 않다.

가공되지 않은 원래의 오디오 신호로부터 음성과 음악을 구분 짓는 특징 파라미터를 추출하는데 음성 벡터로 보통 LPC, MFCC, 켑스트럼 등이 널리 이용된다. 이런 1차적인 특징 벡터를 가지고 인식적 접근으로 음소 단위의 모델을 만든다. 보다 향상된 특징을 추출하기 위해서 모델에 적합한 대표 패턴을 찾기 위해 훈련 과정을 통해 모델링을 한다. 따라서 최종 입력으로 들어온 음악과 음성에 대해 모델링된 대표패턴과의 매칭을 통해 분류된다.

본 논문에서는 이러한 방식과는 달리 직관적이고 실험적이며 가장 일반적인 접근으로 음성과 음악 또는 묵음의 특징을 주로 시간영역에서 신호가 갖는 특징을 분석하고 이러한 특징을 조합하여 음성과 음악을 구분한다.

2. 음성, 음악, 묵음의 특징 및 파라미터

2.1 묵음(Silence)

일반적으로 묵음은 사람이 인식하기에는 불가능한 오디오 신호이다. 보통 묵음구간에서의 에너지는 다른 클래스의 신호에 비교하여 상대적으로 작으며, 따라서 이런 특성을 이용하면 우선 프레임 별 에너지의 임계값을 정해 이를 이용하여 오디오 클래스 상의 묵음을 구분해낼 수가 있다. 그러나 다른 클래스의 신호 즉, 음악이나 음성구간에서 역시 일시적이나마 상대적으로 작은 에너지 값이 존재할 수 있기 때문에 단순히 에너지 특성만을 가지고 묵음을 구분하기에는 정확성이 다소 떨어질 수가 있다. 따라서 본 논문에서는 이런 문제를 해결하기 위해서 ZCR(Zero Crossing Rate)을 이

용한다. ZCR의 응용은 실제적인 묵음구간의 검출의 정확성을 높여주는데, 음악이나 음성 구간에서의 묵음이 아닌 작은 에너지 부분에 대하여 ZCR을 실험적으로 살펴보면 실제의 묵음 구간에 비해 ZCR이 크다는 것을 관찰할 수 있다. 따라서 위에서 언급한 직관적인 두 가지 특징을 조합하여 묵음을 검출하면 에너지만으로 검출한 것에 비해 보다 정확한 묵음 검출이 이루어진다. 따라서 프레임 내의 정규화된 로그에너지와 ZCR의 곱으로 S_w를 식 (2.1)과 같이 정의한다.

$$S_w = E_w \times ZCR_w \tag{2.1}$$

여기서,

$$E_w = \frac{1}{N}\sum_{x=0}^{N} f(x)^2, \quad N: \text{프레임내 샘플수} \tag{2.2}$$

$$ZCR_w = \frac{1}{2}\sum_{m}|\,sgn[x(m)] - sgn[x(m-1)]| \tag{2.3}$$

$$sgn[x(m)] = \begin{pmatrix} x(n)=1, & x(n)>0 \\ x(n)=-1, & x(n)<0 \end{pmatrix}$$

여기서 E_w와 ZCR_w는 각각 30ms 구간의 프레임 에너지와 ZCR 을 나타낸다. 따라서 S_w 의 임계값 S_{th}를 정의하고 이를 이용하여 묵음을 검출한다.

2.2 일반음성(Speech)

본 논문에서의 일반적인 음성의 범위는 단어들이 연속적으로 발성되는 경우에 해당한다. 따라서 음성 중에서 고함, 괴성 등은 이에 속하지 않는다. 시간영역에서의 일반적인 음성신호의 파형을 살펴보면 발음이 될 경우 높은 에너지를 가지며, 발음과 발음 사이에 일시적으로 작은 에너지(묵음) 구간을 갖는다. 따라서 이런 특성을 이용하여 1초 블록 구간동안 10ms의 프레임 이동으로 총 100번에 대하여 S_w의 임계값(S_{th})을 기준으로 상하로 교차되는 빈도수를 계산한다. 이를 SCR(Silence Crossing Rate)라 정의하며, 실험적으로 일반적인 음성 구간에서는 보통 4~10의 SCR이 발생한다.

2.3 음악(Music)

음악 신호는 앞서 언급한 묵음 및 일반 음성신호의 특징과 상반되는 특성을 갖는다. 따라서 음악이란 클래스를 정의하면 우선 프레임별 에너지 크기의 포락선의 기울기가 완만하며, 묵음구간은 거의 발생하지 않는 특징을 갖는다. 또한 대상 신호를 분석하는 방법으로 앞서 설명한 시간 영역에서의 분석과 달리 주파수 영역에서의 분석으로 각 클래스를 분류하는 특성을 얻을 수 있다. 일반적으로 주파수 대역에서의 차이를 들 수가 있으며, 이를 바탕으로 경험적이면서 실험적인 특징 파라미터를 추출하였다. 이는 100Hz 미만의 파워 스펙트럼으로 음성 및 묵음 구간에서는 거의 발생하지 않는다. 장르에 영향을 받아서 발라드 음악 같은 경우에는 적용이 안되지만, 최근의 음악방송 중 많은 부분이 댄스음악, 강한 비트음 등의 음악에 의존하며, 이런 음악에 대해서는 매우 정확한 검출이 이루어진다. 따라서 본 논문에서는 파라미터 LFP(Low Frequency Power)를 식(2.4)와 같이 정의한다.

$$LFP_w = \begin{bmatrix} 1, & LFP_w > LFP_{th} \\ 0, & LFP_w < LFP_{th} \end{bmatrix} \tag{2.4}$$

2.4 엔트로피(Entropy)와 다이나미즘(Dynamism)

엔트로피와 다이나미즘은 2.1절에서 추출한 프레임별 에너지를 가지고 음성 및 음악 구분을 정규화하기 위해서 사용되는 파라미터이다.

엔트로피는 불규칙적이고 예측 불가능한 신호에 대한 정보량으로서 영상 압축 등에 널리 이용된다. 본 논문에서는 프레임별 로그에너지의 크기에 대한 확률값을 이용하여 엔트로피를 구한다. 프레임별 확률 값은 식 (2.5)과 같이 표현된다.

$$P(x_n) = \sum_{n=0}^{N}\frac{E_n}{E_{tot}} \tag{2.5}$$

여기서,

$$E_{tot} = \sum_{w=0}^{N} E_w \quad N: \text{블록내 프레임수} \tag{2.6}$$

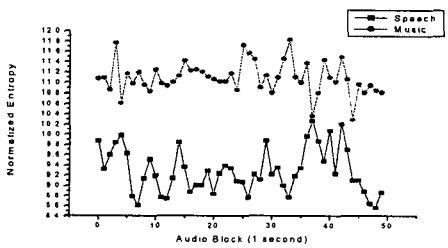

그림 1. 50초 동안의 음악과 음성에 대한 엔트로피

식 (2.7)에 나타낸 바와 같이 프레임별 확률 값이 규칙적인 경우에는 엔트로피 값이 크게 되며, 불규칙적이고 예측 불가능한 신호에 대해서는 엔트로피 값이 작아진다.

$$H = -\frac{1}{N}\sum_{n=0}^{N} P(x_n)\log_2 P(x_n) \tag{2.7}$$

따라서 그림 1에서 볼 수 있는바와 같이 음성에 대해서는 에너지의 분포가 불규칙 적이기 때문에 작은 엔

트로피 값이 나타나며, 음악에 대해서는 상대적으로 큰 값이 나타나게 된다.

다이나미즘은 프레임별 확률값의 차이를 제곱한 값으로 식 (2.8)과 같이 주어지며, 다이나미즘의 경우 엔트로피와 반대의 특성을 나타낸다.

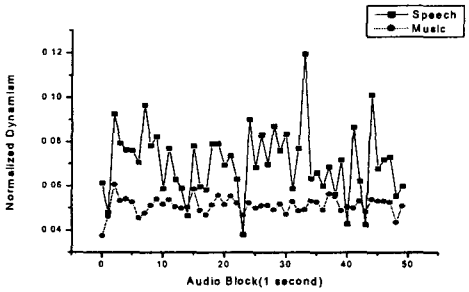

그림 2. 50초 동안의 음악과 음성에 대한 다이나미즘

$$D=\frac{1}{N}\sum_{n=0}^{N}(P(x_n)-P(x_{n+1}))^2 \qquad (2.8)$$

따라서 음성의 경우에는 높은 다이나미즘 값을, 음악에 대해서는 상대적으로 낮은 값을 나타난다.

3. 음악/음성 구분 알고리즘 및 구현

3.2 메인 알고리즘

그림 3은 앞서 설명한 알고리즘의 신호처리부를 나타낸다. 첫 번째 블록에서는 그림 4에서와 같이 1초간의 오디오 데이터를 프레임 쉬프트 M으로 (윈도우의 크기는 N) 겹침 윈도윙을 한다. 따라서 1초간의 블록에는 100개의 프레임 데이터가 존재하게 된다. 따라서 100개의 프레임에 대하여 2장에서 언급되었던 파라미터 ZCR_w , S_w , LFP_w 를 구한다

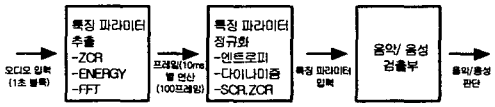

그림 3. 신호처리부의 블록 다이어그램

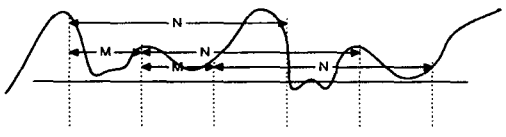

그림 4. 오디오 데이터를 겹치는 프레임들로 블록킹하는 과정

두 번째 블록에서는 음성 및 음악, 음악의 시작과 끝 부분의 특성을 찾기 위해 첫 번째 블록에서 추출된 파라미터를 바탕으로 클래스별 분류되는 특성을 얻기 위하여 SCR , 엔트로피 및 다이나미즘을 구한다. 이를 이용하여 마지막 블록에서는 1초간의 블록 데이터를 음악, 음성, 음악의 시작부 또는 음악의 끝부분 등으로 분류를 한다.

음악의 시작부(F): 식 (2.1)을 이용하여 $S_w<S_{th}$ 인 묵음의 연속적인 특성을 찾는다. 식 (3.1)과 같이 연속된 특성이 20 프레임 이상 지속될 경우 음악의 시작부로 판단한다.

$$F=\begin{bmatrix}1, & S_w<S_{th} \text{ 인 경우수가 연속적으로 20번 이상}\\ 0, & \text{그외}\end{bmatrix} \qquad (3.1)$$

음성(speech): 앞서 살펴본 SCR 과 LFP 의 특성의 조합으로 음성구간을 나타내는 S를 식 (3.2)와 같이 정의한다. 여기서 SCR 은 100개의 프레임 내에서 연속적인 S_w-S_{th} 값의 부호 변화의 빈도를 나타낸다. 또한 LFP 는 100개의 프레임 내에서 식 (2.4)를 만족하는 빈도수 값을 나타낸다.

$$S=g(SCR)+f(LFP) \qquad (3.2)$$

여기서 $g(\cdot)$ 와 $f(\cdot)$ 는 실험과정을 통하여 음성여부의 판단에 가장 정확성을 기하기 위해서 직관적으로 정의한 함수이다.

$$g(x)=\begin{bmatrix}1, & 4<x<10\\ 0, & \text{그외}\end{bmatrix},\quad f(x)=\begin{bmatrix}1, & x<50\\ 0, & \text{그외}\end{bmatrix} \qquad (3.3)$$

따라서 S = 2이면 오디오 블록은 음성구간으로 판단한다. 식 (2.7)과 식 (2.8)을 이용하여 E 와 D 를 구한다. 두 파라미터의 조합으로 음성구간을 나타내는 ED_S 를 식 (3.4)와 같이 정의한다.

$$ED_S=e(E)+d(D) \qquad (3.4)$$

여기서 $e(\cdot)$ 와 $d(\cdot)$ 는 식 (3.3)과 마찬가지로 실험과정을 통하여 음성여부의 판단에 가장 정확성을 기하기 위해서 직관적으로 정의한 함수이다. 이 함수의 정의는 아래와 같다.

$$e(E)=\begin{bmatrix}1, & E>E_{th}\\ 0, & E\le E_{th}\end{bmatrix},\quad d(D)=\begin{bmatrix}1, & D\le D_{th}\\ 0, & D>D_{th}\end{bmatrix} \qquad (3.5)$$

따라서 ED_S = 2이면 마찬가지로 음성구간으로 판단한다.

음악(music): 음악 구간은 음성 구간에서의 특성과 상반되는 특성을 지니기 때문에 위에서 마찬가지로 *SCR* 과 *LFP* 의 특성을 이용한다. 음악 구간을 나타내는 M을 식 (3.6)과 같이 정의한다.

$$M = g'(SCR) + f(LFP) \quad (3.6)$$

여기서 함수 $g'(\cdot)$ 와 $f(\cdot)$ 는 각각 식 (3.3)의 $g(\cdot)$ 와 $f(\cdot)$ 를 이용하여 정의한다.

$$g'(x) = 1 - g(x), \quad f(x) = 1 - f(x) \quad (3.7)$$

따라서 M = 2이면 1초간의 오디오 블록을 음악 구간으로 판단한다. 또한 음성 구간을 판단할 때 계산된 E, D 값의 조합으로 음악구간을 나타내는 ED_M 을 식 (3.8)과 같이 정의한다.

$$ED_M = e'(E) + d'(D) \quad (3.8)$$

여기서 함수 $e'(\cdot)$ 와 $d'(\cdot)$ 는 각각 식 (3.5)의 $e(\cdot)$ 와 $d(\cdot)$ 를 이용하여 정의한다.

$$e'(x) = 1 - e(x), \quad d'(x) = 1 - d(x) \quad (3.9)$$

앞서와 마찬가지로 ED_M = 2이면 마찬가지로 음악구간으로 판단한다.

4. 실험 결과

본 논문에서는 기존의 제시되었던 블록 단위의 오디오 데이터가 음악인지 음성인지 구분하는 것을 넘어서 그러한 블록들의 논리적인 조합인 음악만을 인식하는 알고리즘과 이를 이용하여 음악을 자동으로 녹음하는 시뮬레이터를 구현하였다. 이를 이용하여 FM 라디오 방송 107.7KHz 2002년 12월 17일 12시간 샘플을 기준으로 테스트 하였으며, 테스트 기준은 순수 음악만을 자동 인식 녹음하는 것을 기준으로 하였다. 구현한 알고리즘을 기반으로 FM 방송 샘플을 12시간 테스트한 결과 음악 녹음에서 81.25%의 정확성을 입증했다.

<table>
<tr><td colspan="5">음악구분 성공(SM)</td><td colspan="4">65곡</td></tr>
<tr><td rowspan="9">음악 구분 실패 (FM)</td><td colspan="4">음악시작 구분 성공(SS)</td><td colspan="3">73곡</td><td rowspan="5">7곡</td></tr>
<tr><td rowspan="3">음악 시작 구분 실패 (FS)</td><td rowspan="3">첨부(AS)</td><td>음악 첨부(MAS)</td><td>3곡</td><td rowspan="3">5곡</td><td rowspan="3">7곡</td></tr>
<tr><td>음성 첨부(SAS)</td><td>2곡</td></tr>
<tr><td colspan="2">손실(LS)</td><td colspan="2">2곡</td></tr>
<tr><td colspan="4">음악끝 구분 성공(SE)</td><td colspan="3">72곡</td><td rowspan="4">8곡</td></tr>
<tr><td rowspan="3">음악 끝 구분 실패 (FE)</td><td rowspan="3">첨부(AE)</td><td>음악 첨부(MAE)</td><td>2곡</td><td rowspan="3">3곡</td><td rowspan="3">8곡</td></tr>
<tr><td>음성 첨부(SAS)</td><td>1곡</td></tr>
<tr><td colspan="2">손실(LE)</td><td colspan="2">5곡</td></tr>
</table>

5. 결 론

음성 및 음악, 묵음 구간 등의 특징을 찾는 알고리즘과 이를 바탕으로 한 실시간 방송중의 음악만을 자동 인식하여 음악만을 녹음하는 시뮬레이터를 윈도우 환경에서 설계 구현 하였다. 현재 음악방송이 갖는 특성상 음악만을 녹음한다는 것은 매우 힘든 일이 아닐 수 없다. 사람이 인식하는 노래의 기준을 정의하기가 너무 엄격하기 때문이기도 하다. 예를 들어 잔잔한 배경음악이 존재하는 가운데 사회자의 진행은 논문에서 접근하는 노래의 기준에 부합하지 않는다. 또한 PCM 오디오 데이터를 1초 이상 버퍼링한다는 점은 하드웨어 메모리에 큰 부담을 주게 됨으로 알고리즘을 위한 최대한의 블록을 1초로 설계하였기에 음악중간의 상당히 긴 구간의 묵음 구간 등에 성능 저하의 문제가 있을 수 있다.

그러나 제안한 본 논문은 음악과 음성을 구분하는 접근 방식에서 주류를 이루고 있는 음소단위 모델링 과정, 또한 모델링을 위한 훈련 과정의 확률적 접근 방식과는 달리 시간영역에서의 본연의 오디오 신호를 처리하는 것을 기반으로 오디오 클래스 상의 특징 파라미터의 임계 값을 기준으로 실험적으로 접근하는 방법을 제시하였다. 이는 알고리즘의 수행 면에서 비교적 적은 연산으로써 추후 이를 바탕으로 한 하드웨어와 연계할 경우 실시간 응용이 가능해진다는 면에서 본 논문의 특수성을 찾을 수 있다.

6. 참고문헌

[1] J. Sounders, "Real-time discrimination of broadcast speech/music," Proc. IEEE Conf. on Acoustic, Speech and Signal Processing, pp. 993-996.

[2] G. Williams and D. Ellis, "Speech/music discrimination based on posterior probabilites," Proc. European Conf. on Speech Commun. and Technology, pp. 687-690, Sept. 1990.

[3] E. Scheirer and M. Slancy, "Construction and evaluation of a robust multi-feature speech/music discriminator," Proc. ICASSP' 97, pp. 1331-1334, 1997.

[4] Hadi Harb, Liming Chen, Jean-yves Auloge Ecole Centrale De Lyon, "Speech/Music/Silence and Gender Detection Algorithm," DMS'2001, Sept. 2001.

[5] Lawrence Rabiner and Biing-Hwang Juang *FUNDAMENTALS OF SPEECH RECOGNITINON*, PTR Prentice-Hall Inc, 1993.