Proceedings of KFIS Autumn Conference 2007 Vol. 17, No. 2.

FM 방송 중 블록 단위 음성 음악 판별 시스템의 설계 및 구현

Design and Implementation of Speech Music Discrimination System per Block Unit on FM Radio Broadcast

장형종[1], 엄정권[2], 임준식[3]

[1,2,3] 경기도 성남시 수정구 경원대학교 전자계산학과

[1] E-mail: hjjang@kyungwon.ac.kr

[2] E-mail: iorikyo79@hotmail.com

[3] E-mail: jslim@kyungwon.ac.kr

요 약

본 논문은 FM 라디오 방송의 오디오 신호를 블록 단위로 음성 음악을 판별하는 시스템을 제안하는 논문이다. 본 논문에서는 음성 음악 판별 시스템을 구축하기 위해 다양한 특징 파라미터와 분류 알고리즘을 제안 한다. 특징 파라미터는 신호처리 분야(Centroid, Rolloff, Flux, ZCR, Low Energy), 음성 인식 분야(LPC, MFCC), 음악 분석 분야(MPitch, Beat)에서 각각 사용되는 파라미터를 사용하였으며, 분류 알고리즘으로는 패턴인식 분야(GMM, KNN, BP)와 퍼지 신경망(ANFIS)을 사용하였고, 거리 구현은 Mahalanobis 거리를 사용하였다.

Key Words : Music Speech Discrimination, Pattern Recognition, Neuro Fuzzy

1. 서 론

디지털 음향 매체의 발전과 더불어 오디오 신호를 저장하거나 다루는 일이 늘어남에 따라, 음성과 음악을 자동으로 구별하는 시스템은 여러 분야에서 유용하게 활용된다. 그 예로 오디오 데이터를 자동 인덱싱하거나 멀티미디어 정보를 검색하는 시스템, 그리고 데이터 전송시 음성과 음악에 적합한 압축 방식의 적용 등이 있다. 또는 방송 뉴스 인식 시스템에서는 배경 음악이나 주변 잡음을 인식기의 입력에서 제외시킴으로 인식 성능을 향상 시키는데 사용된다. 이를 위해 빠르면서도 성능이 뛰어난 음성 음악 판별 시스템이 요구된다[1][2].

본 연구에서는 음성 음악 판별 시스템을 구축하기 위해 다양한 특징 파라미터와 분류 알고리즘을 테스트 한다. 특징 파라미터는 기본 파라미터인 Centroid, Rolloff, Flux, ZCR, Low Energy, 음성 인식 분야에서 사용되는 LPC, MFCC, 음악 분석 분야에 사용되는 MPitch, Beat 를 사용하였으며, 분류 알고리즘으로는 GMM, KNN, ANFIS를 사용하였고, 거리 구현은 Mahalanobis 거리를 사용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로 본 논문에서 사용하는 특징 파라미터와 분류 알고리즘에 대해 개괄적으로 살펴보고, 3장에서는 설계와 구현, 4장에서는 실험 결과를 기술하였으며, 5장에서는 결론을 맺는다.

2. 관련 연구

방송 중의 오디오 신호는 매우 다양한 신호들이 존재한다. 오디오 신호는 악기음에 기반한 음악과 사람의 음성 구간으로 나눌 수 있으며, 사람의 음성 구간은 단독 화자에 의한 발성 구간과, 여러 화자에 의한 발성 구간, 배경음이 섞인 음성 구간 등으로 매우 다양한 특징을 갖는 패턴으로 표현된다. 음악 역시 다양한 장르별로 구분될 수 있으며, 각각의 장르마다 특징이 다양하게 나타나다.

오디오 신호에서 음성과 음악을 자동으로 판별하는 시스템은 멀티미디어 환경의 여러 분야에 유용하게 활용된다. 이런 음성 음악 판별 시스템은 입력된 오디오 신호로부터 효과적인 특징 파라미터를 추출하는 과정과 추출된 파라미터를 분류기를 사용하여 음성과 음악으로 분

류하는 과정으로 나누어진다.

특징 파라미터는 음성과 음악이 갖는 여러 특성으로부터 서로간의 변별력을 최대한 높여 주도록 파라미터를 구해야 한다. 이를 위해 음성과 음악의 특성을 시간 영역과 주파수 영역에서 추출한 다양한 특징 파라미터들이 제안되었다. 뿐만 아니라, 음성 인식 분야에서 연구되어온 다양한 특징 파라미터들이 제안되었다.

분류기는 특징 파라미터를 이용하여 음성과 음악, 또는 음성과 음악이 섞인 부분 등 두, 세 가지로 구분한다. 판별 시스템의 목적이 무엇이냐에 따라 음성과 음악이 섞인 부분은 둘 중 한 부분에 속하도록 구분한다. 기존의 음성 음악 판별 시스템에서 사용된 분류기로는 Gaussian Mixture(GMM), k-Nearest Neighbor(k-NN), Hidden Markov Model (HMM) 등이 있다[3].

2.1 음성과 음악의 특징을 이용한 파라미터

1) Centroid

$$C = \sum_{1}^{N} f M[f] / \sum_{1}^{N} M[f] \quad (1)$$

Centroid는 주파수 영역에서 스펙트럼의 균형점을 나타낸다. 주파수 전체의 에너지의 중간 즉 균형점 아래 주파수의 에너지와 균형점 위 주파수의 에너지가 동일한 주파수를 나타낸다.

2) Rolloff

$$\sum_{1}^{R} M[f] = 0.85 \sum_{1}^{N} M[f] \quad (2)$$

위 Centroid가 주파수 영역의 에너지의 50%라고 한다면, RollOff는 90% 영역이라 할 수 있다.

3) Flux

$$F = \| M[f] - M_{prev}[f] \| \quad (3)$$

두 연속된 프레임에서 측정된 STFT(Short Time Fourier Transform) 스펙트럼의 크기의 차이를 2-norm으로 계산한 것을 Flux라 한다.

4) Zero Crossing Rate

$$ZCR_w = \frac{1}{2} \sum_{m=0}^{N-1} |sgn[x(m)] - sgn[x(m-1)]| RECT \quad (4)$$

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

Zero Crossing Rate(ZCR, 영교차율)는 시간 영역에서 단위 구간당 영교차 횟수로서, 간단한 계산으로 신호의 두드러진 스펙트럼을 잘 나타내기 때문에, 음성 음악 판별 시스템에 사용되어져 왔다. 음성 신호는 특성상 무성음과 유성음이 번갈아 나타나기 때문에, 1초 단위의 윈도우 내에서 스펙트럼의 변화가 음악보다 크다.

5) Low Energy

1초 윈도우의 평균에너지 보다 적은 에너지를 갖는 윈도우의 퍼센트를 Low Energy라고 한다. 음성은 음악 보다 묵음 구간을 더 가지고 있는 경향이 있다.

6) LPC

LPC(Linear-prediction coefficients)는 음성 인식 분야에서 사용되는 특징으로, 음성 신호를 모델링하는데 용이하다.

7) MFCC

MFCC(Mel-Frequency cepstral coefficients)는 음성 인식 분야에서 사용되는 특징이다.

8) MPITCH

MPITCH(Multi Pitch)는 다중 피치 분석 알고리즘을 기반으로 화성 컨텐츠 표현을 위한 특징의 집합이다.

9) BEAT

BEAT는 이산 웨이블릿 변환을 기반으로한 비트 추출 알고리즘을 사용하여 음악의 비트 구조를 표현하는 특징의 집합이다.

2.2 분류 알고리즘

1) Gaussian Mixture Model (GMM)

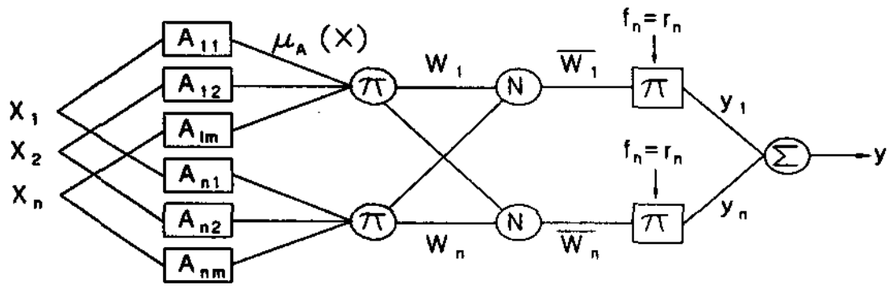
GMM은 특징 벡터의 분포를 몇 개의 Gaussian의 가중치합으로 모델링 하는 것이다. GMM의 모델 파라미터들은 훈련 데이터로부터 Expectation- Maximization(EM) 알고리즘을 통해 추정된다. GMM을 이용한 분류 방법은 음악과 음성 모델로부터 추정된 Likelihood 값이 더 큰 모델을 선택하는 것이다.

2) K-Nearest Neighbor (KNN)

KNN 추정 방법은 확률 분포에 대한 가정 없이 새로운 특징 벡터와 훈련 데이터의 특징 벡터 공간에서 거리를 비교하여 가장 가까운 훈련 데이터의 클래스로 분류한다. KNN 분류 방법은 거리를 비교할 때 가장 가까운 하나의 거리를 구하는 대신 k 개의 거리를 구한 후 평균 거리가 가장 작은 클래스로 분류한다.

3) ANFIS

본 연구에서는 퍼지신경망 모델로 <그림 1>과 같은 ANFIS(Adaptive Network-based Fuzzy Inference System)를 이용했으며 ANFIS는 최급강화법(Gradient Decent Method)을 이용하여 소속함수값들을 학습, 시스템에 전문가의 지식을 보다 정확하게 조정, 반영함으로써 결과 값을 최적화 시키게 된다.



<그림 1> ANFIS의 구조

<그림 1>에서

$$w_i = \prod_{j=1}^{m} \mu_{A_{ji}}(x_j) \quad (5)$$
$$= \mu_{A_{11}}(x_1) \cdot \mu_{A_{12}}(x_2), \mu_{A_{11}}(x_1) \cdot \mu_{A_{22}}(x_2) \cdot$$

$$\overline{w_i} = \frac{w_i}{\sum_{i=1}^{n} w_i} \quad (6)$$

$$y = \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \overline{w_i} f_i = \frac{\sum(w_i * f_i)}{\sum w_i} \quad (7)$$

이며, 여기서

x_j : 입력값,

A_{ji} : 전건부 소속함수 집합

$\mu_{A_{ji}}(x_j)$: 입력 x_j 의 집합 A에 대한 소속 함수값

w_i : i번째 규칙의 점화강도(firing strength)

$\overline{w_i}$: 정규화된 , w_i,

r_i : i번째 규칙 후건부의 소속함수값

y_i : i번째 규칙의 출력값

y : 시스템의 출력값

f_i : ANFIS에서 퍼지규칙의 후건부를 나타내는 다항식으로 zero order 시스템에서 $f_i = r_i$

N : 정규화된 점화강도(firing strength)

또한 결과의 소속함수값 r_i는 식(8)에 따라 학습하며 오차 E는 식(9)로 주어진다.

$$r_i(t'+1) = r_i(t') - lr \cdot \frac{\partial E}{\partial r_i} \quad (8)$$

$$E = \frac{1}{2}(y_i - y^t)^2 \quad (9)$$

여기서 y^t는 목표값, lr은 학습률(learning rate), t'은 학습횟수(time epoch)이며 E<Emax 때까지 학습을 수행한다.

4) Mahalanobis Distance

마할라노비스 거리는 군집분석에서 가장 많이 사용되는 거리개념으로서, 두 지점의 단순한 거리뿐만이 아니라, 변수의 특성을 나타내는 표준편차와 상관계수가 함께 고려된다는 특징을 가지고 있다.

3. 설계 및 구현

3.1 실험 데이터

실험 데이터는 라디오에서 임의로 녹음한 데이터로, 각 데이터는 30초를 기준으로 하였다. 음악은 Ballade, Dance, Rock, 연주곡으로 구성되며, 가수 성별이 균형이 되도록 하였다. 음성은 성별 균형을 고려하였고, 단독 음성 뿐 아니라, 공동의 목소리도 포함하였다. 기존 논문에는 클래식 음악과 사람들의 말소리를 녹음한 파일을 구분하였다.

실험데이터의 내용은 표 1과 표 2에 나타나 있다. 표 1은 실험에 사용된 음악데이터이고, 표 2는 실험에 사용된 음성데이터이다.

<표 1> 음악 데이터

장르	곡명
발라드	사랑2(윤도현밴드) 환생(이승환) 화장을고치고(왁스)
댄스	진실(쿨) 제비(김건모) 페스트발(엄정화)
Rock	어떤이의꿈(봄여름가을겨울) 진달래꽃(마야)
연주곡	간주곡(마스카니) 곡(재스오케스트라)

<표 2> 음성 데이터

성별	MC
남자	남궁연, 환희, 전영혁
여자	이상은, 이금희, 백승주
혼성	남여, 남혼성, 남다수, 여혼성

3.2 특징 파라미터

음성 음악 판별에 유용하다고 판단되는 특징 파라미터를 사용하였다. 기본 파라미터로는 Centroid, Rolloff, Flux, ZCR, Low Energy를 사용하였으며, Centroid, Rolloff, Flux, ZCR은 1초 구간의 평균과 표준편차를 사용하였다. 음

성 인식에서 사용하는 파라미터는 LPC(11차 중 5차), MFCC(5차)를 사용하였다. 음악 관련 파라미터로는 Pitch, Beat를 사용하였다.

3.3 분류 알고리즘

분류 알고리즘으로 ANFIS, GMM, KNN을 사용하였고, 거리 계산은 Mahalanobis Distance를 사용하였다.

3.4 구현 방식

실험용 음성 음악 데이터는 라디오 방송을 8KHz, 16bit로 샘플링하였고, 각각의 데이터 길이는 30초로 하였다. 특징 파라미터는 C로 구현하였고, 분류 알고리즘의 경우 ANFIS는 Matlab으로 그 외 알고리즘은 C로 구현 하였다.

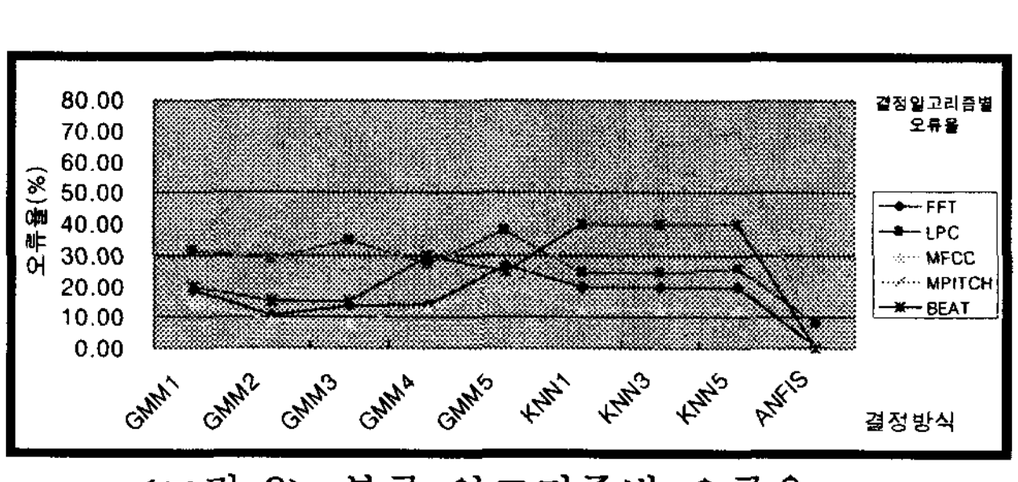
4. 실험 결과

실험 결과는 표 3과 같다. FFT는 기본 파라미터 9개를 나타내는 용어이다.

<표 3> 실험 결과표

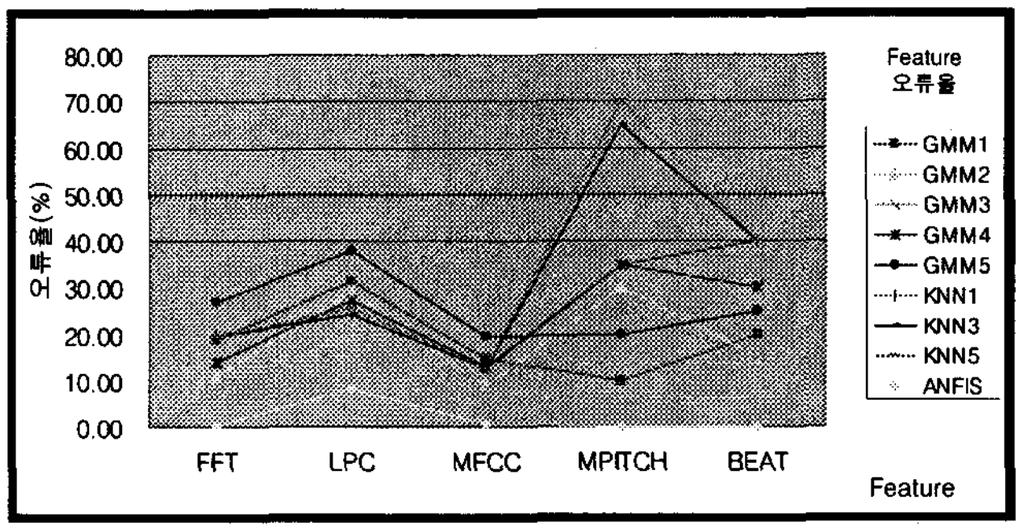
		GMM1	GMM2	GMM3	GMM4	GMM5	KNN1	KNN3	KNN5	ANFIS
FFT	오류	18.82	10.79	13.45	13.96	26.84	19.72	19.64	19.62	0.50
	성공	81.18	89.21	86.55	86.04	73.16	80.28	80.36	80.38	99.50
	오차	14.57	12.57	15.73	18.85	18.27	11.37	11.99	13.20	
LPC	오류	31.61	29.05	35.17	27.42	37.97	24.24	24.32	25.58	8.17
	성공	68.39	70.95	64.83	72.58	62.03	75.76	75.68	74.42	91.80
	오차	11.79	19.95	15.61	21.17	12.93	6.38	8.68	9.83	
MFCC	오류	14.67	8.90	8.18	12.93	19.51	12.59	12.67	12.71	0.83
	성공	85.33	91.10	91.82	87.07	80.49	87.41	87.33	87.29	99.16
	오차	18.50	15.77	17.89	19.31	21.65	10.24	10.27	10.57	
MPITCH	오류	10.00	30.00	25.00	35.00	20.00	35.00	65.00	70.00	0.00
	성공	90.00	70.00	75.00	65.00	80.00	65.00	35.00	30.00	100.00
	오차	20.00	24.49	25.00	22.91	24.49	32.02	32.02	33.17	0.00
BEAT	오류	20.00	15.00	15.00	30.00	25.00	40.00	40.00	40.00	0.00
	성공	80.00	85.00	85.00	70.00	75.00	60.00	60.00	60.00	100.00
	오차	24.49	22.91	22.91	24.49	25.00	30.00	20.00	20.00	0.00

분류 알고리즘에 대한 성능 평가는 그림 3과 같다. ANFIS가 가장 높은 성능을 나타냈으며, GMM, KNN 순으로 성능이 나타났다.



<그림 3> 분류 알고리즘별 오류율

특징 알고리즘에 대한 평가는 평균적으로 MFCC가 전체적으로 좋은 성능을 나타냈고, 음악 관련 파라미터들은 그 성능 편차가 크게 나타났다.



<그림 4> 특징 파라미터별 오류율

5. 결론

본 논문에서는 음성 음악 판별 시스템의 전반적인 특징을 알아보았다. 다양한 특징 파라미터와 분류 알고리즘을 테스트 하였다. 기존 연구는 클래식음악과 말소리를 비교하는 분야였는데, 실제 라디오에서 DJ 멘트와 사람의 목소리가 들어있는 음악을 분류했다는 점에서 새로운 연구 방향이라 할 수 있다[4].

향후 연구 방향으로는 실험 데이터베이스를 더 확대하여 다양한 분야에 응용할 예정이고, 라디오 등의 매체에서 음악만을 자동으로 추출하는 알고리즘을 구현 할 예정에 있으며, 장르 구분 및 내용기반 오디오 검색 시스템을 구현할 예정에 있다.

참 고 문 헌

[1] E. Scheirer, M. Slaney, "Construction and evaluation of a robust multifeature music/speech discrimination," *Proc. ICASSP97*, Vol.2, pp. 1331-1334, 1997.

[2] J. Saunders, "Real-time discrimination of broadcast speech/music," *Proc. ICASSP96*, Vol.2, pp.993-996, 1996.

[3] J. Ajmera, I. McCowan, H. Bourlard, "Speech/music discrimination using entropy and dynamism features in a HMM classification framework," *Speech Communication*, Vol.40, Issue 3, pp259-430, 2003.

[4] 김수미, 김형순, "음성/음악 판별을 위한 특징 파라미터와 분류기의 성능 비교", *말소리*, vol.46, pp. 37-50, 2003.