음성과 음악 분류를 위한 특징 파라미터와 분류 방법의 성능비교

부산대학교 전자공학과
김수미, 김형순

Performance Comparison of Feature Parameters and Classifiers for Speech/Music Discrimination

Su Mi Kim, Hyung Soon Kim
Dept. of Electronics Engineering, Pusan National University
E-mail : {noise2, kimhs}@pusan.ac.kr

Abstract

In this paper, we present a performance comparison of feature parameters and classifiers for speech/music discrimination. Experiments were carried out on six feature parameters and three classifiers. It turns out that three classifiers shows similar performance. The feature set that captures the temporal and spectral structure of the signal yields good performance, while the phone-based feature set shows relatively inferior performance.

I. 서 론

오디오 데이터로부터 음성과 다른 오디오 신호(예를 들면 음악)를 분류하는 일은 멀티미디어 환경에서 다양하게 활용될 수 있다. 먼저 방송 뉴스 인식 시스템에서는 오디오 스트림 중에서 음악이나 주변 잡음을 인식기의 입력으로 사용되지 않게 하여 인식 성능을 향상시키는 역할을 한다. 다른 적용 분야로 저전송률 오디오 부호화기를 들 수 있다. 인터넷과 같은 멀티미디어 환경에서의 오디오 스트림은 대부분 음성과 음악이 함께 존재하게 되는데, 이 때 모든 오디오 스트림에 동일한 부호화 방식을 적용하지 않고 음성/음악 분류기를 이용하여 음성과 음악에 따라 적합하게 부호화하여 효율적인 압축을 할 수 있다. 이러한 접근 방식은 최근 MPEG4의 parametric coder 등에서 이용되고 있다. 그 외에도 멀티 미디어 정보 검색 시스템이나 FM 라디오 채널 기반의 오디오 콘텐츠의 음성인식을 등에도 이용될 수 있다[1][2].

신호를 분류하는 문제에 있어서 가장 중요한 문제는 적당한 특징 파라미터의 선택이다. 본 논문에서는 지금까지 음성/음악 자동분류 연구에서 비교적 성능이 좋은 것으로 알려진 특징 파라미터를 분류 방법에 따라 성능을 비교해 보았다.

본 논문의 구성은 다음과 같다. 2절에서는 본 논문에서 사용한 몇 가지 특징 파라미터에 대해 살펴보고, 3절에서는 음성/음악 분류를 위한 분류기에 대해 설명하였다. 그리고 4절에서는 실험 내용 및 결과에 대해 기술하고, 5절에서 결론을 맺는다.

II. 특징 파라미터

2.1 High Zero-Crossing Rate Ratio (HZCRR)

Zero-Crossing Rate(ZCR)는 시간 영역에서 영 교차 횟수로서, 간단한 계산으로 신호의 두드러진 스펙트럼을 잘 나타낸다. 따라서 ZCR은 음성/음악 분류 시스템에 많이 사용되어 왔다[3]. 음성 신호는 특성상 무성음과 유성음이 번갈아 나타나기 때문에, 1초 단위의 윈도우 내에서 스펙트럼의 변화가 크다. 따라서 ZCR을 그냥 사용하기 보다 변화량을 High Zero-Crossing

Rate Ratio로 정의하여 하나의 특징 파라미터로 제안되었다[4].

$$HZCRR=\frac{1}{2N}\sum_{n=0}^{N-1}[\,\mathrm{sgn}(ZCR(n)-1.5avZCR)+1] \quad (1)$$

여기서 $ZCR(n)$은 n 번째 프레임의 ZCR이고, N은 윈도우(본 논문에서는 1초로 사용)내의 총 프레임 수이다. 그리고 $avZCR$은 평균 ZCR값이고, $sgn[.]$은 sign 함수이다.

2.2 Low Short Time Energy Ratio

일반적으로 음성 구간은 음악에 비해 많은 휴지 구간을 포함하고 있다. 따라서 신호의 단구간 에너지의 변화 또한 하나의 특징 파라미터로 사용되었다[2]. LSTER (Low Short Time Energy Ratio)은 1초 윈도우 내의 평균 단구간 에너지의 0.5배 보다 작은 프레임의 개수의 비로 정의된다[4].

$$LSTER=\frac{1}{N}\sum_{n=0}^{N-1}[\,\mathrm{sgn}(0.5avSTE-STE(n))+1] \quad (2)$$

HZCRR과 마찬가지로 $STE(n)$는 n 번째 프레임의 단구간 에너지, $avSTE$는 평균 단구간 에너지값, N은 1초 윈도우 내의 총 프레임 수이다.

2.3 Spectrum Flux

Spectrum Flux는 인접한 프레임간의 스펙트럼 크기의 변화를 나타낸다[3].

$$SF=\frac{1}{(N-1)(K-1)}\sum_{n=1}^{N-1}\sum_{k=1}^{K-1}[\log(A(n,k)+\delta)-\log(A(n-1,k)+\delta)]^2 \quad (3)$$

여기서 δ는 계산상 log 함수에 0이 들어가는 것을 막기 위한 작은 상수이고, K는 DFT point, N은 총 프레임 수이다. 그리고 $A(n,k)$는 n 번째 입력 프레임에 대한 DFT(Discrete Fourier Transform) 값으로서, 윈도우 길이가 L인 입력 신호 $x(m)$에 대해 $A(n,k)$는 식 (4)와 같이 나타낸다.

$$A(n,k)=\left|\sum_{m=-\infty}^{\infty}x(m)w(nL-m)e^{j\frac{2\pi}{L}km}\right| \quad (4)$$

음성의 경우 프레임간 스펙트럼의 변화가 크기 때문에 음악에 비해 값이 큰 쪽에 분포하게 된다.

2.4 LSP 거리

신호의 스펙트럼 포락선(Envelope)을 나타내는 LPC(Linear predictive Coefficient)의 또 다른 표현으로 LSP(Line Spectrum Pairs)가 있다. LSP의 통계적 특성은 패턴인식에서 좋은 성능을 보이는 것으로 알려졌고, 이에 따라 음성/음악 분류 시스템에서 LSP를 기반으로 하는 특징 파라미터들이 사용되었다[1].

1초 입력 신호의 LSP 벡터로 추정한 확률밀도함수(pdf)를 p_{LSP}라 하면 훈련 음성 데이터로부터 구해진 p_{SP}와의 LSP 거리는 식 (5)와 같이 이들 분포사이의 Kullback-Leibler 거리로 정의된다. 이 값은 음성의 경우 음악보다 더 작은 값을 가지는 경향이 있다.

$$D=\int_x[p_{LSP}(\boldsymbol{x})-p_{SP}(\boldsymbol{x})]\ln\frac{p_{LSP}(\boldsymbol{x})}{p_{SP}(\boldsymbol{x})}d\boldsymbol{x} \quad (5)$$

만약 LSP의 분포가 정규 분포를 따른다고 가정하면 LSP Distance는 식(6)과 같이 나타낼 수 있다.

$$D=\frac{1}{2}tr[(\hat{C}_{LSP}-C_{SP})(C_{SP}^{-1}-\hat{C}_{LSP}^{-1})] + \frac{1}{2}tr[(C_{SP}^{-1}-\hat{C}_{LSP}^{-1})(\hat{\boldsymbol{u}}_{LSP}-\boldsymbol{u}_{SP})(\hat{\boldsymbol{u}}_{LSP}-\boldsymbol{u}_{SP})^T] \quad (6)$$

여기서 $\hat{C}_{LSP}$와 $\hat{\boldsymbol{u}}_{LSP}$는 입력 신호에 의해 추정된 공분산 행렬과 평균 벡터이다. 여기서 공분산 부분만 취하여 변형된 LSP 거리를 식 (7)과 같이 정의한다[4].

$$D=\frac{1}{2}tr[(\hat{C}_{LSP}-C_{SP})(C_{SP}^{-1}-\hat{C}_{LSP}^{-1})] \quad (7)$$

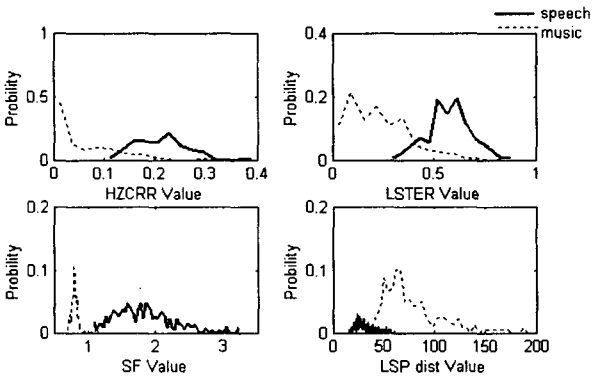

그림 1. 4가지 특징 파라미터의 분포

2.5 음소인식 기반 특징 파라미터[5]

음성/음악 분류하는 새로운 파라미터로 음소인식을 기반으로 하는 entropy와 dynamism이 제안되었다. Entropy는 주어진 분포에 대해 불확실성을 측정하는 척도로 사용된다. n 번째 프레임의 입력 벡터 x_n에 대해 K개의 클래스에 대한 사후 확률을 $P(q_k \mid x_n)$ 라 하면 entropy는 식 (8)과 같이 정의된다.

$$h_n=-\sum_{k=1}^{K}P(q_k \mid x_n)\log_2 P(q_k \mid x_n) \quad (8)$$

여기서 x_n을 음향 특징 벡터라고 하고, q_k를 K개의 음소라고 하면, 음성이 음악 보다 특정한 음소에 대해 높은 확률값을 가질 것이다. 이것은 불확실성이 작다

는 것을 뜻하며 entropy값은 0에 가까운 값을 가질 것이다. 반대로 음악의 경우 모든 음소에 대해 비교적 균일한 확률값을 가지고 큰 entropy값을 가질 것이다.

Dynamism은 프레임간 확률값의 변화의 정도를 측정한 값이다. n번째 프레임에서 dynamism은 식 (9)와 같이 정의된다.

$$d_n = -\sum_{k=1}^{K}[P(q_k \mid x_n) - P(q_k \mid x_{n+1})]^2 \qquad (9)$$

음성의 경우 프레임마다 나오는 음소의 종류가 달라지면서 확률값의 변화가 크지만 음악의 경우 변화의 정도가 작을 것이다. 따라서 dynamism은 음성이 음악보다 더 큰 값을 갖는다.

음성/음악 분류를 위해서는 식 (8)과 (9)을 N개의 프레임에 대해 smoothing한 값을 사용한다.

$$H_n = \frac{1}{N}\sum_{t=n-N/2}^{n+N/2} h_t, \quad D_n = \frac{1}{N}\sum_{t=n-N/2}^{n+N/2} d_t \qquad (10)$$

본 논문에서는 MFCC와 delta계수들을 음향 특징 벡터로 사용하였다. 그리고 사후 확률값을 추정하기 위해 묵음을 포함한 46개의 음소를 GMM으로 모델링하여 likelihood값 $P(x_n \mid q_k)$을 구하고 사전 확률 $P(q_k)$가 모든 음소에 대해 동일하다는 가정 아래 Bayesian rule에 의해 식 (11)과와 같이 추정하였다.

$$P(q_k \mid x_n) = \frac{P(x_n \mid q_k)P(q_k)}{P(x_n)} \approx \frac{P(x_n \mid q_k)}{\sum_{k=1}^{K} P(x_n \mid q_k)} \qquad (11)$$

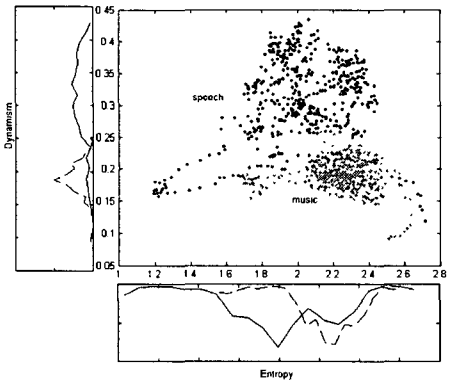

그림 2. Entropy와 dynamism의 분포

III. 분류방법

Gaussian Mixture Model(GMM)은 특징 벡터의 분포를 몇 개의 Gaussian 의 가중합으로 모델링 하는 것이다. GMM의 모델 파라미터들은 훈련 데이터로부터 EM Algorithm을 통해 추정된다. GMM을 이용한 분류 방법은 음악과 음성 모델로부터 추정된 likelihood값이 더 큰 모델을 선택하는 것이다.

Nearest-Neighbor(NN)추정 방법은 분포가 Gaussian이라는 가정 없이 새로운 특징 벡터가 떨어졌을 때 훈련 데이터의 특징 벡터 공간에서 거리를 비교하여 가장 가까운 훈련 데이터의 클래스로 분류한다. k-NN(k Nearest Neighbor) 분류 방법은 거리를 비교할 때 가장 가까운 하나의 거리를 구하는 대신 k개의 거리를 구한 후 평균 거리가 가장 작은 클래스로 분류한다. 본 논문에서는 특징 벡터들을 binary split 방법에 의해 벡터 양자화하여 훈련 데이터를 몇 개의 cell로 분류 한 다음 각 cell에 대해 Euclidean Distance를 구하였다.

HMM Classifier를 이용하여 음성과 음악을 분류하는 경우 그림 3처럼 2개의 상태를 가지는 HMM으로 모델링 하게 된다. 본 논문에서는 GMM으로 각 상태에서의 관측 확률을 구하고 Viterbi 알고리즘을 이용하여 최대 likelihood 음성과 음악의 상태열을 찾는다. 최종 상태열은 입력 오디오 파일의 끝에서 backtracking에 의해 결정된다. 따라서 긴 오디오 파일의 경우 적당한 길이에서 Viterbi 디코딩을 수행한다.

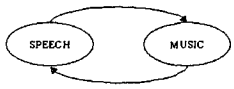

그림 3. HMM Topology

IV. 실험 내용 및 결과

4.1 실험 환경

훈련을 위해 사용된 음성 DB는 국어공학센터에서 구축한 PBS 589문장에 대한 남녀 50명분의 발성 데이터 약 13시간 분량을 사용하였다. 그리고 음악 DB은 클래식 음악 CD로부터 42곡을 16kHz로 다운 샘플링하고 16bit로 양자화 하여 음성 DB와 동일한 조건이 되도록 만들었다. 테스트 DB는 open test와 closed test의 2 set 1시간의 분량으로 구성하였는데 closed test는 훈련 시 사용된 음성과 음악이 15초씩 번갈아 가면서 나타나도록 구성하였고, open test는 훈련 시 사용되지 않고 clean 환경에서 녹음한 음성 데이타와 다양한 장르의 음악을 15초씩 번갈아 나타나도록 구성하였다. 여기서 사용한 음악은 클래식은 아니지만 사람의 목소리가 없는 연주곡이나 반주로 구성하였다.

LSP 거리의 기준 공분산 행렬은 훈련 음성 DB로부터 18차의 LSP를 구한 다음 벡터 양자화를 통하여 4개의 코드북을 생성하였다. 테스트 오디오 데이터에 대한 LSP 거리는 4개의 공분산 행렬로부터 각각 거리를 구한 다음 가장 작은 값을 사용하였다.

특징 벡터 추출을 위한 프레임의 크기는 25msec로 하였고 이동 크기도 25msec로 하였다. 그리고 40개의 프레임에 대해 2절에서 설명한 변화량과 평균값을 나타내는 특징 파라미터를 추출하였다.

Entropy와 dynamism(Hn와 Dn)을 추출하기 위한 음소 모델의 음향특징벡터로는 20msec의 프레임을 10msec씩 이동 시켜 MFCC와 delta, 총 24차를 구하였다. 그리고 묵음을 포함한 46개의 monophone을 모두 8개의 mixture를 가지는 GMM으로 모델링 하였다.

k-NN 분류기에서 k에 따른 성능 차이는 거의 없는 것으로 나타났다[3]. 따라서 본 논문에서는 k=2로 하였고 코드북 크기는 128개로 하여 실험하였다.

4.2 실험결과

HZCRR, LSTER, SF 및 LSP 거리를 각 분류기에 적용한 결과를 표 1, 2, 3에 나타내었다. 실험 결과 분류기에 따른 성능 차이는 거의 없었다. *k*-NN의 코드북 크기와 같은 128개의 mixture를 가질 때를 비교해 볼 때 closed test 경우 GMM 및 HMM 분류기가 더 좋은 성능을 보이지만 open test일 경우 *k*-NN 분류기가 약간 좋은 성능을 보인다. 그림 4에 음소 인식기에 바탕을 둔 Hn와 Dn을 HMM 분류기에 적용한 결과를 나타내었다. Hn과 Dn은 10msec 마다 특징 벡터들이 추출되기 때문에 1초 단위로 반올림하여 정확도를 계산하였다. 전반적으로 HZCRR, LSTER, SF 및 LSP 거리를 이용한 경우보다는 저조한 성능을 나타내었다.

표 1. *k*-NN 분류기에 의한 정확도(%)

Closed Test			Open Test		
Speech	Music	Ave.	Speech	Music	Ave.
99.72	97.78	98.75	99.17	85.56	92.36

표 2. GMM 분류기에 의한 정확도(%)

mix-ture	Closed Test			Open Test		
	Speech	Music	Ave.	Speech	Music	Ave.
4	99.50	99.72	99.61	98.17	87.67	92.92
8	99.83	99.72	99.78	99.39	85.89	92.64
16	100.00	99.78	99.89	99.67	85.28	92.47
32	100.00	99.00	99.50	99.56	84.56	92.06
64	100.00	99.06	99.53	99.78	82.83	91.31
128	**100.00**	**99.06**	**99.53**	**99.78**	**82.83**	**91.31**
256	100.00	99.06	99.53	99.89	80.94	90.42

표 3. HMM 분류기에 의한 정확도(%)

mix-ture	Closed Test			Open Test		
	Speech	Music	Ave.	Speech	Music	Ave.
4	99.44	99.44	99.44	95.50	88.39	91.94
8	99.89	99.44	99.67	99.11	85.50	92.31
16	99.94	99.78	99.86	99.44	85.06	92.25
32	99.94	99.72	99.83	99.39	84.94	92.17
64	99.94	99.61	99.78	99.56	82.89	91.22
128	**100.00**	**99.78**	**99.89**	**99.78**	**82.50**	**91.14**
256	100.00	99.67	99.83	99.78	81.78	90.78

V. 결론 및 향후 계획

실험결과 전체적으로 음악의 경우 open test시에 많은 성능저하를 보였다. 이는 음성에 비해 음악의 경우 다양한 특성이 존재하기 때문이다. 따라서 훈련 DB에 다양한 장르의 음악을 함께 포함시키는 것이 필요하다. 음소 인식기에 바탕을 둔 특징 파라미터는 음소 모델을 미리 잘 구성해야 할 뿐만 아니라, 입력 오디오 신호에 대해 1차적으로 MFCC를 계산하고 다시 2차 특징 벡터를 계산해야 하기 때문에 계산량이 앞에서 살펴본 4가지 특징 파라미터에 비해 많은 편이다. 그러나 2개의 파라미터만으로도 비교적 좋은 성능을 보인다. 향후 실험에서는 음소 기반 파라미터와 다른 파라미터를 함께 적용할 경우 정확도를 살펴 볼 계획이다.

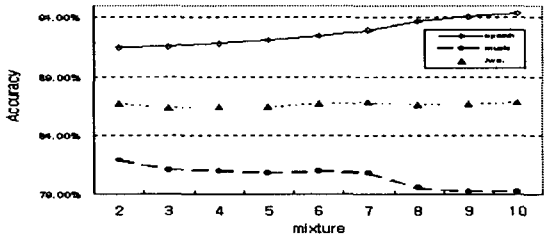

그림 4. Hn,Dn을 HMM 분류기로 분류한 결과

본 논문은 한국과학재단 특정기초연구(과제번호 2000-2-30300-002-3) 결과의 일부입니다.

참고문헌

[1] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal, "Speech/Music Discrimination for Multimedia Application," *Proc. ICASSP00*, vol.4, pp. 24455-2449, 2000.

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

[3] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Music/Speech Discriminator," *Proc. ICASSP97*, vol.II, pp.1331-1334, 1997.

[4] L. Lu, H. Jiang and H.J. Zhang,"A Robust Audio Classification and Segmentation Method," *Proc. of 9th ACM International Conference on Multimedia*, pp.203-211, 2001.

[5] J. Ajmera, I. McCowan, and H. Bourlard "Speech/Music Discrimination using Entropy and Dynamism Features in a HMM Classification Framework," *Speech Communication*, Vol40, Issue 3, pp. 259-430, 2003.