

# 음성/음악 판별을 위한 특징 파라미터와 분류기의 성능 비교\*

김형순(부산대), 김수미(부산대)

## <차 례>

- |                         |                       |
|-------------------------|-----------------------|
| 1. 서론                   | 2.4. LSP 거리           |
| 2. 음성/음악 판별을 위한 특징 파라미터 | 2.5. 음소 인식 기반 특징 파라미터 |
| 2.1. HZCRR              | 3. 음성/음악 판별을 위한 분류기   |
| 2.2. LSTER              | 4. 실험 내용 및 결과         |
| 2.3. Spectrum Flux      | 5. 결론                 |

## <Abstract>

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

**Hyung Soon Kim, Su Mi Kim**

In this paper, we evaluate and compare the performance of speech/music discrimination based on various feature parameters and classifiers. As for feature parameters, we consider High Zero Crossing Rate Ratio (HZCRR), Low Short Time Energy Ratio (LSTER), Spectral Flux (SF), Line Spectral Pair (LSP) distance, entropy and dynamism. We also examine three classifiers: k Nearest Neighbor (k-NN), Gaussian Mixture Model (GMM), and Hidden Markov Model (HMM). According to our experiments, LSP distance and phoneme-recognizer-based feature set (entropy and dynamism) show good performance, while performance differences due to different classifiers are not significant. When all the six feature parameters are employed, average speech/music discrimination accuracy up to 96.6% is achieved.

\* Keywords: speech/music discrimination, classifier, feature parameter

\* 본 논문은 한국과학재단 특정기초연구(과제번호 R01-2000-000-00275-0) 결과의 일부입니다.

## 1. 서 론

오디오 데이터로부터 음성과 음악을 판별하는 일은 멀티미디어 환경에서 다양하게 활용될 수 있다. 먼저 방송 뉴스 인식 시스템에서는 오디오 데이터 중에서 음악 구간을 인식기의 입력에서 미리 제외시킴으로써 인식 성능의 향상에 기여할 수 있다. 저전송률 오디오 부호화기에서도 전체 오디오 신호에 동일한 부호화 방식을 적용하는 대신에, 음성 구간과 음악 구간을 구분한 다음 이들 각각에 적합한 부호화 방식을 통해 효율적인 압축을 할 수 있다. 이러한 접근 방식은 최근 MPEG-4 parametric coder 등에서 이용되고 있다. 그 외에도 음성/음악 판별은 멀티미디어 정보 검색 시스템에도 유용한 수단이 된다[1][2].

이에 따라 음성과 음악을 판별하기 위한 여러 가지 접근 방식들이 제안되었으며[1]-[6], 이들 각각은 음성/음악 판별에 유용하다고 판단되는 특징 파라미터 분석 방법들과 이들 파라미터들을 이용한 분류 방법들을 제시하고 있다. 그러나 지금까지 제안된 방식들에 대해 동일한 오디오 데이터를 대상으로 충분한 성능 비교가 이루어지지 못했기 때문에, 본 논문에서는 기존 연구에서 비교적 성능이 우수하다고 보고된 대표적인 특징 파라미터들과 분류 방식에 대해 성능 비교를 수행하였다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문에서 검토한 음성/음악 판별용 특징 파라미터에 대해 살펴보고, 3장에서는 음성/음악 판별을 위한 분류기에 대해 기술한다. 그리고 4장에서 실험 방법 및 결과에 대해 서술한 다음, 5장에서 결론을 맺는다.

## 2. 음성/음악 판별을 위한 특징 파라미터

### 2.1. High Zero Crossing Rate Ratio (HZCRR)

Zero-Crossing Rate (ZCR)는 시간 영역에서의 단위 시간당 영 교차 횟수를 의미한다. ZCR은 간단한 계산으로 신호의 스펙트럼 특성을 나타낼 수 있다는 장점이 있어서, 음성/음악 판별 시스템에 많이 사용되어 왔다[3]. 음성 신호는 특성상 무성음과 유성음이 번갈아 나타나기 때문에, 1초 단위의 윈도우 내에서 스펙트럼의 변화가 크다. 따라서 프레임 별 ZCR을 그대로 사용하기 보다, 시간에 따른 ZCR의 변화량을 하나의 특징 파라미터로 사용하는 High Zero-Crossing Rate Ratio (HZCRR)가 다음 식과 같이 제안되었다[4].

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

여기서  $ZCR(n)$ 은  $n$  번째 프레임의  $ZCR$ 이고,  $N$ 은 윈도우(본 논문에서는 1초로 사용)내의 총 프레임 수이다. 그리고  $avZCR$ 은 윈도우 내의 평균  $ZCR$ 값이고,  $\text{sgn}(\cdot)$ 은 signum 함수로서 다음과 같이 정의된다.

$$(2) \quad \text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

## 2.2. Low Short Time Energy Ratio (LSTER)

일반적으로 음성 구간은 음악에 비해 많은 휴지 구간을 포함하는 경향이 있다. 이에 따라 신호의 단구간 에너지의 시간에 따른 변화 특성을 Low Short Time Energy Ratio (LSTER)라는 특징 파라미터로 제안되었다[2]. LSTER은 1초 윈도우 내의 평균 단구간 에너지의 0.5배 보다 작은 프레임의 개수의 비로 정의된다[4].

$$(3) \quad LSTER = \frac{1}{N} \sum_{n=0}^{N-1} [\text{sgn}(0.5avSTE - STE(n)) + 1]$$

HZCRR의 경우와 마찬가지로  $STE(n)$ 은  $n$ 번째 프레임의 단구간 에너지,  $avSTE$ 는 윈도우 내의 평균 단구간 에너지 값, 그리고  $N$ 은 윈도우 내의 총 프레임 수이다.

## 2.3. Spectrum Flux (SF)

Spectrum flux는 인접한 프레임간의 스펙트럼 변화의 크기를 나타내며, 다음 식과 같이 정의된다[3].

$$(4) \quad SF = \frac{1}{(N-1)(K-1)} \sum_{n=1}^{N-1} \sum_{k=1}^{K-1} [\log(A(n, k) + \delta) - \log(A(n-1, k) + \delta)]^2$$

여기서  $\delta$ 는 계산상  $\log$  함수에 영(0)이 들어가는 것을 막기 위한 작은 상수이고,  $K$ 는 DFT point 수,  $N$ 은 총 프레임 수이다. 그리고  $A(n, k)$ 은  $n$  번째 입력 프레

임에 대한 DFT (Discrete Fourier Transform) 값으로서, 입력 신호  $x(m)$ 에 대해  $A(n,k)$ 는 다음 식과 같이 나타낸다.

$$(5) \quad A(n, k) = \left| \sum_{m=-\infty}^{\infty} x(m)w(nL-m)e^{j\frac{2\pi}{L}km} \right|$$

여기서  $w(nL-m)$ 은 단구간 Fourier 변환을 하기 위한 길이가  $L$ 인 윈도우 함수이다.

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

## 2.4 LSP 거리

Line Spectrum Pairs (LSP)는 신호의 스펙트럼 포락선을 나타내는 선형예측부호화(Linear Predictive Coding) 계수의 또 다른 표현으로서, 음성 부호화에 널리 사용된다. LSP의 통계적 특성은 패턴인식에서 좋은 성능을 보이는 것으로 알려져 있으며, 이에 따라 음성/음악 판별 시스템에서 LSP를 기반으로 하는 특징 파라미터들이 사용되었다[1].

1초 동안의 입력 신호에 포함된 LSP 벡터들로부터 추정된 확률밀도함수(pdf)를  $p_{LSP}(\cdot)$ 라 하면, 훈련 음성 데이터로부터 구해진 pdf인  $p_{SP}(\cdot)$ 와의 LSP 거리는 이들 분포 사이의 Kullback-Leibler 거리로써 다음과 같이 정의될 수 있다.

$$(6) \quad D = \int_{\mathbf{x}} [p_{LSP}(\mathbf{x}) - p_{SP}(\mathbf{x})] \ln \frac{p_{LSP}(\mathbf{x})}{p_{SP}(\mathbf{x})} d\mathbf{x}$$

만약 LSP의 분포가 정규 분포를 따른다고 가정하면 LSP 거리는 다음 식과 같이 나타낼 수 있다.

$$(7) \quad D = \frac{1}{2} \text{tr} \left[ (\hat{\mathbf{C}}_{LSP} - \mathbf{C}_{SP}) (\mathbf{C}_{SP}^{-1} - \hat{\mathbf{C}}_{LSP}^{-1}) \right] \\ + \frac{1}{2} \text{tr} \left[ (\mathbf{C}_{SP}^{-1} - \hat{\mathbf{C}}_{LSP}^{-1}) (\hat{\mathbf{u}}_{LSP} - \mathbf{u}_{SP}) (\hat{\mathbf{u}}_{LSP} - \mathbf{u}_{SP})^T \right]$$

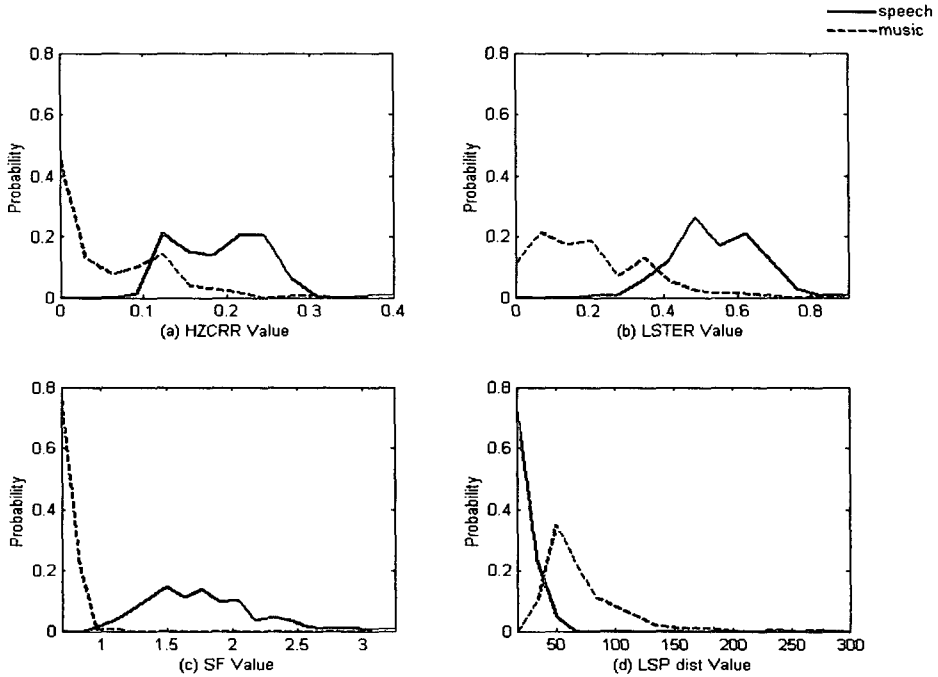
여기서  $\hat{\mathbf{C}}_{LSP}$ 와  $\hat{\mathbf{u}}_{LSP}$ 는 입력 신호에 의해 추정된 공분산 행렬과 평균 벡터이다. 식 (7)에서 공분산 부분만 취하여 변형된 LSP 거리를 식 (8)과 같이 정의한

다[4].

$$(8) \quad D_{LSP} = \frac{1}{2} \text{tr}[(\hat{C}_{LSP} - C_{SP})(C_{SP}^{-1} - \hat{C}_{LSP}^{-1})]$$

LSP 거리는 입력신호의 LSP 분포와 음성의 LSP 분포와의 거리이므로 음성에 비해 음악의 경우 상대적으로 큰 값을 가질 것이다.

<그림 1>은 지금까지 설명한 4가지 특징 파라미터들의 음성과 음악에 대한 분포 특성을 나타내고 있다.



<그림 1> 음성과 음악에 대한 4가지 특징 파라미터의 분포 특성  
(a) HZCRR (b) LSTER (c) SF (d) LSP 거리

## 2.5. 음소 인식 기반 특징 파라미터

최근 음성/음악 판별을 위한 새로운 파라미터로서 음소인식을 기반으로 하는 entropy와 dynamism이 제안되었다[5]. Entropy는 주어진 분포에 대해 불확실성을 측정하는 척도로 사용된다. n번째 프레임의 입력 벡터  $x_n$ 에 대해 K개의 클래스에

대한 사후(a posteriori) 확률을  $P(q_k | x_n)$  이라 하면 entropy는 식 (9)와 같이 정의된다.

$$(9) \quad h_n = - \sum_{k=1}^K P(q_k | x_n) \log_2 P(q_k | x_n)$$

여기서 K개의 클래스를 표현하는  $q_k$ 가 K개의 음소라고 하면, 음성의 경우 음악보다 특정한 음소에 대해 높은 확률값을 가질 것이다. 이는 불확실성이 작다는 의미가 되며, 따라서 entropy 값은 작아진다. 이에 반하여 음악의 경우 특정 음소와의 유사성이 크지 않기 때문에 상대적으로 모든 음소에 대해 비교적 균일한 확률값을 가지게 되며, 따라서 큰 entropy 값을 가진다.

Dynamism은 인접 프레임간 확률값의 변화의 정도를 측정한 값이다. n번째 프레임에서 dynamism은 식 (10)과 같이 정의된다.

$$(10) \quad d_n = - \sum_{k=1}^K [P(q_k | x_n) - P(q_k | x_{n+1})]^2$$

음성의 경우 발성내용의 변화에 따라 프레임별로 나오는 음소의 종류가 시간에 따라 달라지면서 특정 음소의 확률값의 변화가 크지만, 음악의 경우 상대적으로 변화의 정도가 작다. 따라서 dynamism은 음성이 음악보다 더 큰 값을 갖는다.

음성/음악 판별을 위해서는 식 (9)와 (10)을 N개의 프레임에 대해 smoothing한 값을 사용한다.

$$(11-a) \quad H_n = \frac{1}{N} \sum_{k=-N/2}^{-N/2} h_{n+k}$$

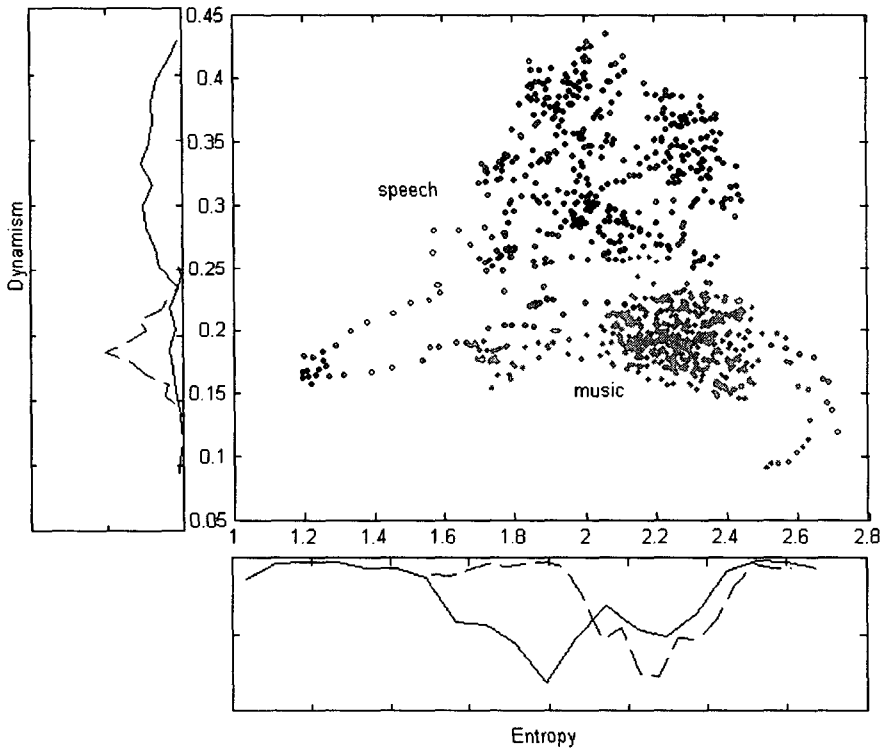
$$(11-b) \quad D_n = \frac{1}{N} \sum_{k=-N/2}^{N/2} d_{n+k}$$

본 논문에서는 12차 MFCC와 그 delta계수들을 음향 특징 벡터로 사용하였다. 그리고 사후확률  $P(q_k | x_n)$ 을 추정하는 방법으로는 다층 퍼셉트론 신경회로망이 사용되기도 하지만[5], 본 논문에서는 목음을 포함한 46개의 음소를 Gaussian Mixture Model (GMM)로 모델링하여 likelihood 값  $P(x_n | q_k)$ 을 구하고 사전(a priori) 확률  $P(q_k)$ 가 모든 음소에 대해 동일하다는 가정 아래 Bayesian rule에 의

해 다음 식과 같이 추정하였다.

$$(12) \quad P(q_k | x_n) = \frac{P(x_n | q_k)P(q_k)}{P(x_n)} \approx \frac{P(x_n | q_k)}{\sum_{k=1}^K P(x_n | q_k)}$$

<그림 2>에 음소인식 기반의 entropy 및 dynamism 파라미터들의 음성과 음악에 대한 분포 특성을 나타내었다.



<그림 2> 음소인식 기반의 entropy와 dynamism의 분포 특성

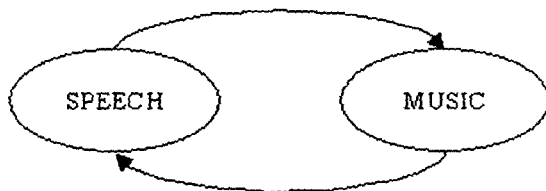
### 3. 음성/음악 판별을 위한 분류기

본 논문에서는 음성/음악 판별을 위해 Gaussian Mixture Model (GMM), k-Nearest Neighbor (k-NN), Hidden Markov Model (HMM)의 3가지 방법을 검토하였다. GMM은 특징 벡터의 분포를 몇 개의 Gaussian 분포들의 가중합으로 표현하는 것이다. GMM의 모델 파라미터들은 훈련용 데이터로부터 Expectation- Maximization

(EM) 알고리즘을 통해 추정된다. GMM을 이용한 분류 방법은 음악과 음성 모델로부터 추정된 likelihood 값이 더 큰 모델을 선택하는 것이다.

Nearest-Neighbor (NN) 분류 방법은 분포에 대한 특정한 가정 없이 새로운 특징 벡터에 대해 훈련 데이터의 특징 벡터 공간에서 거리를 비교하여 가장 가까운 훈련 데이터의 클래스로 분류한다. k-NN 분류 방법은 거리를 비교할 때 가장 가까운 하나의 거리를 구하는 대신, 가장 가까운 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 분류한 다음 각 cell의 centroid와의 Euclid 거리를 통해 k-NN 분류기를 구현하였다.

HMM 분류기를 이용하여 음성과 음악을 판별하는 방법에서는 오디오 신호를 <그림 3>과 같이 2개의 상태를 가지는 HMM으로 모델링한다. 본 논문에서는 GMM에 의해 각 상태에서의 관측 확률을 구하고 Viterbi 알고리즘에 의해 음성과 음악 구간의 최적 상태열을 찾는다. 실제로 최종 상태열은 입력 오디오 파일이 종료되는 시점에서 backtracking에 의해 결정되는데, 처리 지연을 막기 위해 긴 오디오 파일의 경우 적당한 길이 단위로 Viterbi 디코딩을 수행한다.



<그림 3> 음성/음악 판별을 위한 HMM Topology

## 4. 실험 내용 및 결과

### 4.1. 실험 환경

훈련을 위한 음성 DB로는 국어공학센터에서 구축한 음소 균형 문장 (phonetically balanced sentence, PBS) 589문장에 대한 남녀 50명분의 발성 데이터 약 13시간 분량을 사용하였다. 그리고 훈련용 음악 DB는 클래식 음악 CD로부터 42곡을 44.1kHz에서 16kHz로 다운 샘플링하여 음성 DB와 동일한 조건이 되도록 하였다. 테스트 DB는 open test와 closed test의 2 set 모두 1시간의 분량으로 구성하였는데, closed test DB는 훈련에 사용된 음성과 음악이 15초씩 번갈아 가면서 나타나도록 구성하였고, open test DB는 훈련에 사용되지 않은 clean 환경 음성 데이터와 다양한 장르의 음악을 15초씩 번갈아 나타나도록 구성하였다. Test에 사용한



다양한 장르의 음악은 클래식은 아니지만 사람의 목소리가 없는 연주곡이나 반주로 구성하였다.

LSP 거리를 구하기 위한 기준 공분산 행렬은 훈련용 음성 DB로부터 18차의 LSP를 구한 다음 벡터 양자화를 통하여 크기가 4인 코드북 형태로 생성하였다. 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 overlap 없이 25msec로 하였다. 그리고 40개의 프레임에 대해 2절에서 설명한 변화량과 평균값을 나타내는 특징 파라미터를 추출하였다.

Entropy와 dynamism ( $H_n$ 와  $D_n$ )을 추출하기 위한 음소 모델의 음향 특징 벡터로는 20msec의 프레임을 10msec씩 이동시켜 12차 MFCC와 그 delta 계수로 총 24차 벡터를 구하였다. 그리고 목음을 포함한 46개의 문맥독립음소(monophone)을 모두 8개의 mixture를 가지는 GMM으로 모델링 하였다.

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

#### 4.2. 실험 결과

먼저 분류기에 따른 성능 비교를 위해 HZCRR, LSTER, SF 및 LSP 거리의 4가지 파라미터를 함께 사용하여, k-NN, GMM 및 HMM 분류기에 각각 적용한 결과를 <표 1>에서 <표 3>까지에 나타내었다. 실험 결과 분류기에 따른 성능 차이는 별로 없었다. k-NN의 코드북 크기와 같은 128개의 mixture를 가질 때를 비교해 볼 때, closed test 경우 GMM 및 HMM 분류기가 조금 더 좋은 성능을 보이지만 open test일 경우 k-NN 분류기가 약간 좋은 성능을 보인다. 그러나 최적의 mixture를 사용했을 때의 의미있는 성능 차이는 보이지 않았다.

<표 1> k-NN 분류기에 의한 음성/음악 판별 정확도(%)

Closed Test			Open Test		
Speech	Music	Average	Speech	Music	Average
99.72	97.78	98.75	99.17	85.56	92.36

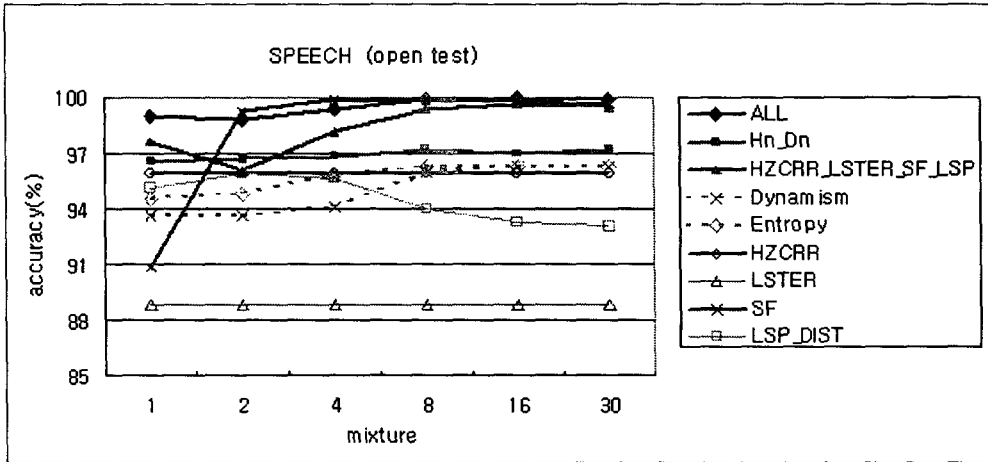
&lt;표 2&gt; GMM 분류기에 의한 음성/음악 판별 정확도(%)

Mixture	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

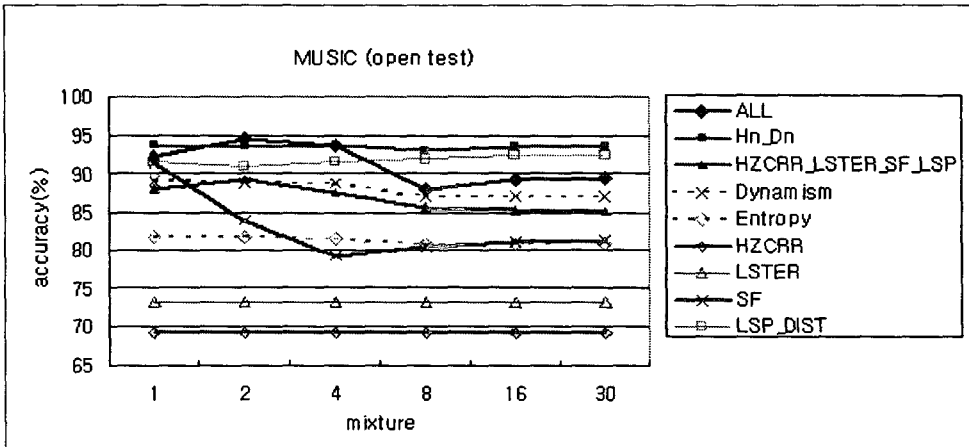
&lt;표 3&gt; HMM 분류기에 의한 음성/음악 판별 정확도(%)

Mixture	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

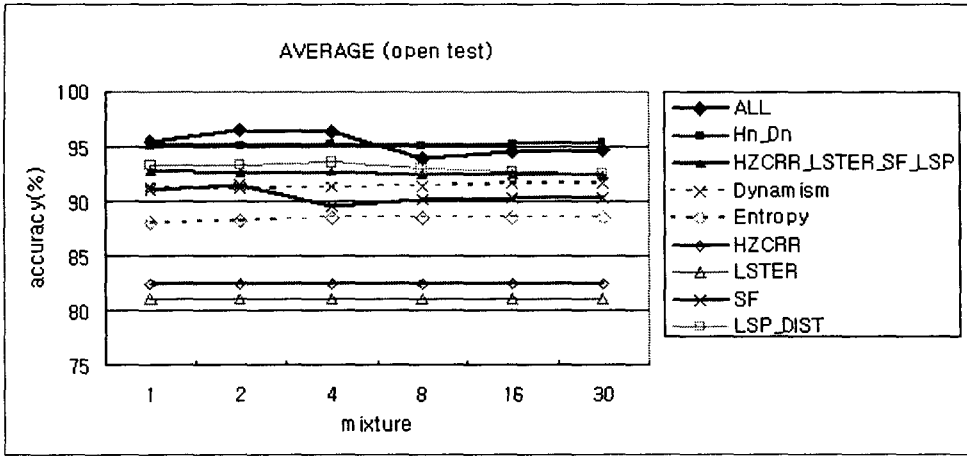
그 다음으로 본 논문에서 살펴본 여러 가지 특징 파라미터들에 따른 성능을 비교해 보았는데, 이 경우에는 GMM 분류기만을 사용하였다. 그리고 GMM에서의 mixture의 수는 1, 2, 4, 8, 16 및 30의 값을 고려하였다. <그림 4>에서 <그림 6>에 음소 인식기에 바탕을 둔 entropy와 dynamism, 그리고 이미 언급한 4가지 특징 파라미터(HZCRR, LSTER, SF, LSP 거리)를 GMM 분류기에 적용한 open test의 결과를 나타내었다. 그림에서 LSP\_DIST는 LSP 거리를 나타내며, HZCRR\_LSTER\_SF\_LSP는 HZCRR, LSTER, SF 및 LSP 거리를 함께 사용했을 때를 의미한다. 또한 Hn\_Dn은 entropy와 dynamism을 함께 사용한 결과이고, ALL은 본 논문에서 고려한 모든 파라미터들을 사용한 결과를 의미한다.



<그림 4> GMM 분류기에 의한 여러 특징 파라미터들의 음성 판별 정확도 비교 (%)



<그림 5> GMM 분류기에 의한 여러 특징 파라미터들의 음악 판별 정확도 비교 (%)



<그림 6> GMM 분류기에 의한 여러 특징 파라미터들의 전체 음성/음악 판별 정확도 비교 (%)

전체 성능을 나타내는 <그림 6>에서 볼 때 HZCRR과 LSTER은 시간 영역에서 간단하게 계산할 수 있기 때문에 계산량 면에서는 가장 유리하나 성능은 가장 떨어짐을 알 수 있다. 단일 파라미터로는 LSP 거리가 가장 우수한 성능을 보였다. (참고로 <그림 1>에서는 LSP 거리보다 SF가 변별력이 더 좋은 것으로 보이는데, 이는 <그림 1>의 분포가 closed test 데이터를 대상으로 구해진 것이기 때문이며, open test와 달리 closed test에서는 실제로 SF의 성능이 더 우수하였다.) 그리고, 동일하게 음소 인식기를 기반으로 하는 entropy와 dynamism 파라미터들은 함께 사용하였을 때 매우 뛰어난 성능을 보였다. 물론 모든 파라미터를 함께 사용하였을 때에 (적은 mixture 수를 사용하는 경우) 전체적으로 가장 우수한 성능이 얻어졌다.

## 5. 결 론

본 논문에서는 음악과 음성 구간의 판별을 위한 다양한 특징 파라미터와 분류 방법에 대해 성능 비교를 수행하였다. 실험 결과 분류 방법에 따른 성능 차이는 별로 없었으며, 특징 파라미터로는 LSP 거리를 사용하거나, 음소 인식 기반의 entropy와 dynamism 파라미터를 함께 사용하는 경우에 우수한 성능을 얻을 수 있었다. 실험 결과 전체적으로 음악의 경우 open test시에 많은 성능 저하를 보였다. 이는 음성에 비해 음악의 경우 다양한 특성이 존재하는데 반해서, 훈련 DB에 다양한 장르의 음악을 함께 포함시키지 않은 요인도 크게 작용했다고 판단되며, 앞으로 이 부분에 대한 추가 실험 및 보다 다양한 오디오 데이터를 이용한 실험이 수행되어야 할 것이다. 그리고 MPEG-7[7]에서 오디오 내용 기술을 위해 사용된

스펙트럼 유사성, 음색, 멜로디나 화음에 기반을 둔 파라미터들을 음성/음악 판별에 적용하는 방안도 검토할 예정이다.

## 참 고 문 헌

- [1] K. El-Maleh, M. Klein, G. Petrucci, P. Kabal, "Speech/music discrimination for multimedia application", *Proc. ICASSP00*, Vol. 4, pp.2455-2449, 2000.
- [2] J. Saunders, "Real-time discrimination of broadcast speech/music", *Proc. ICASSP96*, Vol. 2 pp.993-996, 1996.
- [3] E. Scheirer, M. Slaney, "Construction and evaluation of a robust multifeature music/speech discriminator", *Proc. ICASSP97*, Vol. 2, pp.1331-1334, 1997.
- [4] L. Lu, H. Jiang, H.J. Zhang, "A robust audio classification and segmentation method", *Proc. 9th ACM International Conference on Multimedia*, pp.203-211, 2001.
- [5] 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, pp.259-430, 2003.
- [6] 이경록, 서봉수, 김진영, "오디오 인덱싱을 위한 음성/음악 분류 특징 비교", *한국음향학회지*, 20권, 2호, pp.10-15, 2001.
- [7] ISO-IEC MPEG-7 Overview (ISO/IEC JTC1/SC29/WG11 N4980 document), <http://mpeg.telecomitalia.com/standards/mpeg-7/mpeg-7.htm>, July, 2002.

접수일자: 2003년 05월 31일

게재결정: 2003년 06월 12일

▶ 김형순(Hyung Soon Kim)

주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 전자공학과

소속: 부산대학교 전자공학과 음성통신연구실

전화: 051) 510-2452

FAX: 051) 515-5190

E-mail: kimhs@pusan.ac.kr

## ▶ 김수미(Soo Mi Kim)

주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 전자공학과

소속: 부산대학교 전자공학과 음성통신연구실

전화: 051) 510-1704

FAX: 051) 515-5190

E-mail: noise2@pusan.ac.kr