한국음향학회지 제26권 제5호 pp. 207~213 (2007)

스펙트럼 분석과 신경망을 이용한 음성/음악 분류

Speech/Music Discrimination Using Spectrum Analysis and Neural Network

금 지 수*, 임 성 길*, 이 현 수*
(Ji-Soo Keum*, Sung-Kil Lim*, Hyon-Soo Lee*)
*경희대학교 컴퓨터공학과
(접수일자: 2007년 6월 7일, 채택일자: 2007년 7월 10일)

본 연구에서는 스펙트럼 분석과 신경망을 이용한 효과적인 음성/음악 분류 방법을 제안한다. 제안하는 방법은 스펙트럼을 분석하여 스펙트럴 피크 트랙에서 지속성 특징 파라미터인 MSDF (Maximum Spectral Duration Feature)를 추출하고 기존의 특징 파라미터인 MFSC (Mel Frequency Spectral Coefficients)와 결합하여 음성/음악 분류기의 특징으로 사용한다. 그리고 신경망을 음성/음악 분류기로 사용하였으며, 제안하는 방법의 성능 평가를 위해 학습 패턴 선별과 양, 신경망 구성에 따른 다양한 성능 평가를 수행하였다. 음성/음악 분류 결과 기존의 방법에 비해 성능 향상과 학습 패턴의 선별과 모델 구성에 따른 안정성을 확인할 수 있었다. MSDF와 MFSC를 특징 파라미터로 사용하고 50초 이상의 학습 패턴을 사용할 때 음성에 대해서는 94.97%, 음악에 대해서는 92.38%의 분류율을 얻었으며, MFSC만 사용할 때보다 음성은 1.25%, 음악은 1.69%의 향상된 성능을 얻었다.

핵심용어: 음성/음악 분류, 스펙트럼 분석, 스펙트럴 피크 트랙, 화자 인덱싱

투고분야: 음성처리분야 (2.4)

In this research, we propose an efficient Speech/Music discrimination method that uses spectrum analysis and neural network. The proposed method extracts the duration feature parameter (MSDF) from a spectral peak track by analyzing the spectrum, and it was used as a feature for Speech/Music discriminator combined with the MFSC. The neural network was used as a Speech/Music discriminator, and we have performed various experiments to evaluate the proposed method according to the training pattern selection, size and neural network architecture. From the results of Speech/Music discrimination, we found performance improvement and stability according to the training pattern selection and model composition in comparison to previous method. The MSDF and MFSC are used as a feature parameter which is over 50 seconds of training pattern, a discrimination rate of 94.97% for speech and 92.38% for music. Finally, we have achieved performance improvement 1.25% for speech and 1.69% for music compares to the use of MFSC.

Key words: Speech/Music discrimination, Spectrum analysis, Spectral peak track, Speaker indexing

ASK subject classification: Speech Signal Processing (2.4)

I. 서 론

최근 멀티미디어 데이터의 인덱싱 (Indexing)과 검색 (Retrieval), 이해 (Understanding) 그리고 이를 결합한 내용 기반 분석의 연구가 활발하게 진행되고 있다 [1-4]. 특히, 방송 뉴스와 영화 그리고 스포츠 중계를 대상으로 다양한 연구가 진행되고 있으며, 연구의 접근 방법으로는 오디오 또는 영상 처리에 기반을 둔 접근 방법과 두 가지 접근 방법을 결합한 연구도 활발하게 진행되고 있다. 이러한 접근 방법들에 있어서 오디오 처리에 기반을 둔 대표적인 연구 분야로 화자 인덱싱 (Speaker Indexing)이 있다. 화자 인덱싱은 발성 화자의 수와 특성 등의 사전 지식이 없는 상태에서 화자의 변화 구간을 찾고, 동일 화자 구간을 묶는 기술로 방송 뉴스, 토론, 드라마 등에 적용할 수 있다 [5]. 화자 인덱싱을 효과적으로 수행하기

책임저자: 이 현 수 (leehs@khu.ac.kr)
경기도 용인시 기흥구 서천동 1 경희대학교 컴퓨터공학과
(전화: 031-201-2947; 팩스: 031-202-1723)

위해서는 전처리 과정으로 오디오 데이터가 음성인지 아니면 음악적 성향을 포함한 것인지의 분류가 필요하다. 이러한 분류는 음성 구간에 대해서만 화자 변화 검출과 클러스터링 (Clustering)을 수행하도록 하며, 더 나아가 음악과 효과음 등의 분류로 구조적인 데이터 인덱싱이 가능하다.

음성/음악 분류에 관한 연구는 크게 특징 파라미터 연구와 학습 모델에 관한 연구로 나눌 수 있다. 음성/음악 분류에 사용되는 파라미터로는 HZCRR (High Zero-Crossing Rate Ratio)과 LSTER (Low Short-Time Energy Ratio) 그리고 멜-켑스트럼 (Mel-cepstrum)과 스펙트럼 (Spectrum) 분석을 기반으로 하는 다양한 특징 파라미터들이 있다 [2, 6-11]. 그리고 대표적으로 사용되는 음성/음악 분류기로는 GMM (Gaussian Mixture Model)과 HMM (Hidden Markov Model), NN (Neural Network) 등이 있다 [12, 13]. 현재까지 발표된 연구 결과들에 의하면 음성/음악 분류기의 성능은 학습 모델에 의한 영향보다 사용되는 특징 파라미터의 종류와 특징간의 상호조합 그리고 학습 데이터의 양에 의한 영향이 큰 것으로 알려져 있다. 이러한 결과들은 음성/음악 분류에서 특징 파라미터의 선정이 중요하다는 것을 나타내며, 이와 관련하여 다양한 특징 파라미터를 사용한 연구 결과에서도 음성/음악 분류에 좋은 성능을 보이는 몇 종류의 특징 파라미터를 선별 했을 때 높은 결과를 보이고 있다 [7, 8]. 이러한 기존 연구 결과를 볼 때 효과적인 음성/음악 분류를 위한 특징 파라미터의 선별에 관한 연구가 필요하다.

음성/음악 분류기로 신경망을 사용하는 기존의 연구 중에서 [13]의 방법은 스펙트럼을 분석하여 MFSC (Mel Frequency Spectral Coefficients)를 특징 파라미터로 사용하는 방법이다. 이 방법은 적은양의 학습데이터를 사용하고도 높은 성능을 갖으며, 특징 파라미터로는 멜-스케일링 (Mel-scaling)된 주파수 채널별 MFSC를 구한 후 정규화 과정을 거친 평균과 분산을 사용한다. 이렇게 구한 특징 파라미터는 음성은 음악에 비해서 MFSC의 변화가 크다. 그러나 이 방법은 동일한 주파수 채널에서의 변화가 크게 나타나는 음악이나, 변화가 적은 음성에 대해서는 오분류가 발생한다. 따라서 본 연구에서는 기존의 MFSC를 특징으로 사용하는 방법이 적은양의 학습데이터를 사용하고도 높은 성능을 갖는 장점을 이용하면서, 채널별 변화에 민감하게 반응하는 단점을 줄이고자 한다.

본 연구에서 제안하는 방법은 음성/음악의 분류에 사용될 수 있는 다양한 특징 중, 스펙트럴 피크 트랙 (Spectral Peak Track)에서 지속성을 추출하여 특징으로 사용하고자 한다. 스펙트럴 피크 트랙은 [1, 4, 9]의 연구에도 기술된 바와 같이 음성이 음악의 경우보다 짧은 지속성을 가지며 피치 (pitch)가 존재하는 낮은 주파수 대역에 주로 나타난다. 그리고 음악의 경우 음성에 비해서 피크의 지속성이 길다. 본 연구에서는 1kHz 이하의 구간에서 나타나는 스펙트럴 피크 트랙 중에서 모음의 내재적 길이의 평균인 약 92ms의 지속성을 갖는 피크들을 주파수 채널별로 분석한다. 그리고 분석된 주파수 채널별 지속성에서 최대의 지속성을 음성/음악 분류의 특징으로 사용한다. 이러한 최대 지속성 특징인 MSDF (Maximum Spectral Duration Feature)는 스펙트럴 피크가 일정 시간 지속하고 있지만 음성으로 오분류 되거나, 짧은 시간 동안 스펙트럴 피크가 지속되고 있는 음성이 음악으로 오분류 되는 MFSC의 단점을 줄일 수 있다.

본 논문의 구성은 다음과 같다. 제 II장에서는 음성/음악 분류를 위한 스펙트럼 분석에 대해 기술하고, 제 III장에서는 제안하는 신경망 기반의 음성/음악 분류기를 기술한다. 그리고 제 IV장에서는 제안한 방법의 성능 평가를 위한 실험 데이터 구성과 다각적인 실험에 의한 결과를 분석 고찰한다. 마지막으로 제 V장에서는 결론과 향후 연구 방향을 기술한다.

II. 음성/음악 분류를 위한 스펙트럼 분석

2.1. 스펙트럴 피크 트랙 분석

스펙트럴 피크 트랙은 동일한 주파수 채널에서 일정시간 지속성을 가지고 나타나며 이를 분석하면 지속성에 기반을 둔 음성/음악 분류뿐만 아니라 오디오 데이터의 배음구조도 알 수 있다. 이러한 스펙트럴 피크 트랙은 음성의 경우 피치가 존재하는 저주파 대역에 주로 나타나며 음악의 피크 트랙과 비교하여 짧은 지속성을 가지고 있다. 본 연구에서는 이러한 스펙트럴 피크 트랙의 특성을 이용하여 음성/음악을 분류할 수 있는 특징 파라미터를 추출한다.

오디오 데이터에서 스펙트럴 피크 트랙을 추출하는 과정은 다음과 같다. 먼저 1초 길이의 오디오 세그먼트에 대해 FFT (Fast Fourier Transform) 스펙트럼을 계산한다. 그리고 스펙트럼의 크기가 임계값 (θ_s) 보다 크면 대

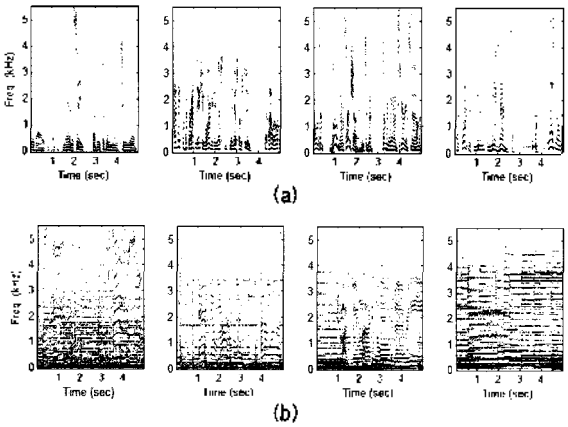


그림 1. 음성/음악에 대한 스펙트럼 피크 트랙 (a) 음성, (b) 음악
Fig. 1. Spectral peak tracks for Speech and Music (a) Speech, (b) Music.

응하는 스펙트럼의 값을 1로 그렇지 않으면 0으로 설정한다. 식 (1)에서 $X_n[k]$와 $CX_n[k]$는 각각 1초 길이의 세그먼트에서 n번째 프레임의 k번째 bin에 대한 스펙트럼과 임계값에 의해 절단된 스펙트럼이다. 여기서 임계값은 세그먼트 내에서 최대값을 갖는 스펙트럼의 크기에 임계비율을 적용하여 피크 트랙이 적절히 추출될 수 있도록 결정한다. 그림 1을 살펴보면 음성의 경우에는 1kHz 이하의 범위에 피크 트랙이 주로 나타나고, 음악은 음성에 비해서 넓은 범위에서 피크 트랙이 나타나며 음성보다 지속성이 길게 나타나는 것을 볼 수 있다.

$$CX_n[k] = \begin{cases} 1, & \text{if} |X_n[k]|^2 \geq threshold(\theta_s) \\ 0, & otherwise \end{cases} \quad (1)$$

2.2. 스펙트럼의 지속성 추출

화자가 발성하는 음성에서 모음의 길이는 화자의 신체적인 특성에 따른 요인과 발성 환경과 교육 수준 그리고 발성 속도와 습관에 영향을 받는다. 본 연구에서 추출하는 지속성은 모음이 내재적인 길이를 가지고 있으며 발성 시 길이가 변화하기는 하지만 일정 시간 지속되는 특성을 이용하고자 한다. [14]의 연구를 살펴보면 우리말 모음의 길이와 발화 속도에 따른 말소리의 길이 변화가 나타나있다. 가장 짧게 발음되는 모음 /ㅡ/의 길이는 약 70ms이고, 가장 길게 발음되는 모음 /ㅔ/와 /ㅏ/는 약 110ms이다. 따라서 본 연구에서는 기존 연구 결과를 바탕으로 이들 모음의 중간 길이인 92ms를 지속성 특징 추출을 위한 피크 트랙의 지속 시간으로 한다.

스펙트럼 피크 트랙에서 지속성을 추출하는 과정은 1kHz 이하의 각 주파수 bin에 대해서 이웃하는 프레임과의 지속성이 92ms인 히스토그램을 구한다. 이렇게 얻어진 각 주파수 bin의 지속성에서 최대값을 세그먼트의 지속성 특징으로 사용한다. 다음의 식 (2)는 각 주파수 bin에 대해 지속성 히스토그램을 구하고 세그먼트내의 프레임수로 정규화 시킨 값이다. 식 (3)의 MSDF는 분석 세그먼트의 최대 지속성으로 음성/음악 분류를 위한 특징으로 사용한다. 식 (2)에서 N은 세그먼트내의 전체 프레임수를 나타내고, 식 (2)와 식 (3)에서 k는 주파수 bin을 나타낸다.

$$SDF[k] = \frac{1}{N-1}\sum_{n=1}^{N-1}\{CX_{n-1}[k] \cdot CX_n[k] \cdot CX_{n+1}[k]\} \quad (2)$$

$$MSDF = Max(SDF[k]) \quad (3)$$

그림 2의 (a)는 음성과 음악에 대해 제안하는 방법으로 추출한 MSDF와 기존의 특징 파라미터 중에서 HZCRR과 LSTER의 분포를 나타낸다. 그리고 그림 2의 (b)와 (c)는 특징 파라미터들 간의 상호 관계를 나타낸다. 그림 2에서 볼 수 있듯이 MSDF가 기존의 특징 파라미터와 비교하여 음성/음악의 중첩된 영역이 적기 때문에 분류에 보다 효과적이다. 그리고 MSDF는 음성은 0.7 보다 작은 쪽으로, 음악은 0.7보다 큰 쪽으로 분포하고 있는 것을 알 수 있다. 또한 MSDF가 HZCRR이나 LSTER과 상호 결합한 관계에서도 HZCRR과 LSTER을 결합한 관계보다 중첩되는 영역이 적어 음성/음악 분별력이 높은 것을 알 수 있다.

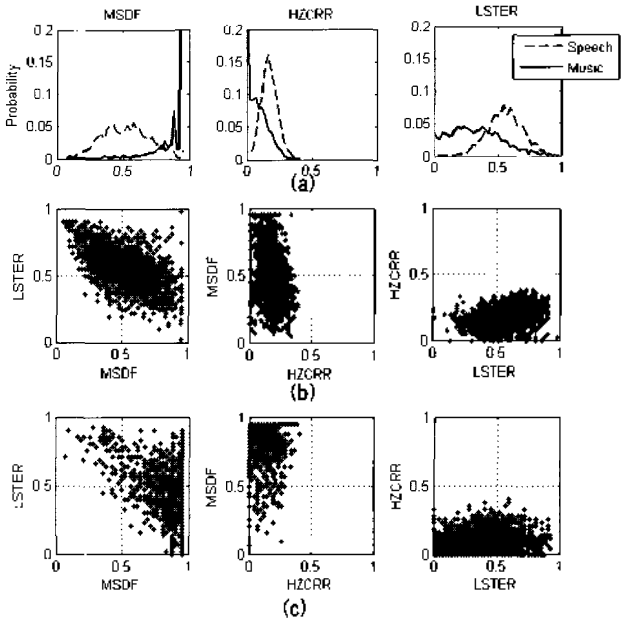


그림 2. 음성/음악에 대한 특징 파라미터 분포
(a) MSDF, LSTER, HZCRR의 분포,
(b) 음성에 대한 특징들의 상호 관계,
(c) 음악에 대한 특징들의 상호 관계
Fig. 2. Distribution of feature parameter for Speech and Music
(a) Distribution of MSDF, LSTER and HZCRR,
(b) Correlation between features for Speech,
(c) Correlation between features for Music.

III. 신경망을 이용한 음성/음악 분류

3.1. 신경망의 구성 및 학습

본 연구에서는 다양한 신경망 모델들 중에서 패턴 분류에 적합하며 뛰어난 일반화 능력을 가지고 있는 에러 역전파 (Error Back-propagation) 신경망을 음성/음악 분류기로 사용한다.

신경망의 구성을 살펴보면, 신경망의 입력으로는 본 연구에서 제안한 1차원의 MSDF와 기존의 특징 파라미터인 40차원의 MFSC로 전체 41차원의 특징이 사용된다. 이때 MSDF와 MFSC는 모두 1초 길이의 세그먼트를 분석 단위로 하였으며, MFSC는 [13]의 연구에 적용된 방법과 같이 세그먼트 내의 각 프레임에 대해 FFT를 수행하고 멜-스케일링과 정규화 과정을 수행한 평균과 분산을 사용하였다. 그리고 중간 뉴런의 개수 결정을 위한 다양한 방법이 있으나, 5개에서 20개까지 5개씩 추가하였다. 출력 뉴런은 음성과 음악에 대응하는 것으로 2개로 구성하였으며, 최대 출력을 갖는 뉴런을 분류 결과로 한다. 그림 3은 음성/음악 분류를 위한 신경망의 구조를 나타낸다. 신경망의 학습에 영향을 미치는 다양한 학습 파라미터 중에서 초기 연결강도는 -1.0~1.0 범위에서 임의의 값으로 설정하였고, 각 층간의 연결강도는 모멘텀 (momentum)을 적용하여 조절하였다. 그리고 학습률은 학습 단계별 에러의 변화를 고려하여 적응적으로 변경하였다.

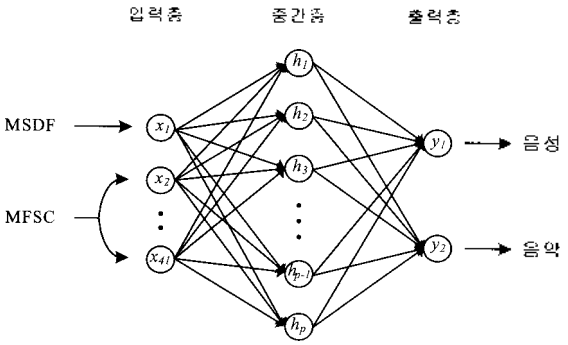


그림 3. 음성/음악 분류의 신경망 구조
Fig. 3. Neural Network architecture for Speech/Music discrimination.

3.2. 분류 결과의 후처리

신경망에 의한 음성/음악 분류 결과는 이웃하는 세그먼트의 결과를 바탕으로 그림 4와 같은 후처리 과정을 적용하여 현재 세그먼트의 결과를 보정한다. 후처리는 $n-1$번째와 $n+1$번째 세그먼트의 결과가 음성일 때 n번째의 결과가 음악이면 n번째의 결과는 음성으로 보정한다. 이

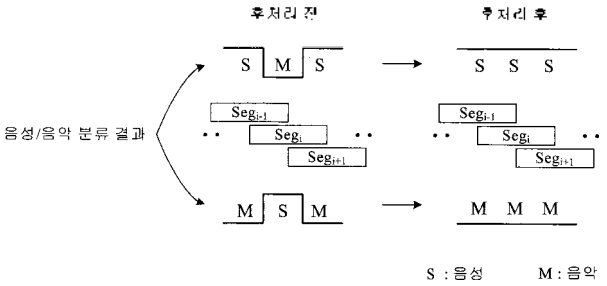


그림 4. 음성/음악 분류 결과의 후처리
Fig. 4. Post-processing for Speech/Music discrimination results.

와 같은 후처리는 음성이나 음악이 일정 시간 계속하여 나타나고 있을 때 분류기에 의한 오분류를 줄일 수 있다.

IV. 실험 및 결과

4.1. 실험 데이터의 구성

음성/음악 분류를 위한 실험으로는 [7]의 연구에서 사용된 데이터를 사용하였다. 이 데이터는 전체 245개의 파일로 구성되어 있으며, 각 파일의 길이는 15초 이다. 실험에 사용한 파일은 음성과 음악이 섞여있는 데이터와 효과음 파일 64개를 제외한 181개 파일로 약 45분 분량이다.

학습 패턴의 선별과 양에 따른 음성/음악 분류기의 성능을 평가하기 위해 다음과 같은 학습 패턴 집합을 구성하였다. 먼저, 학습 패턴 집합의 기본이 되는 집합 A는 [7]의 연구에서 사용되지 않은 데이터로 25초의 방송 뉴스에서 앵커와 기자의 음성을 사용하였고, 음악에는 보컬이 포함된 가요와 팝을 사용하였다. 그리고 기본 집합 A에 [7]의 연구에 사용된 데이터에서 임의적으로 5개의 파일씩 선택하여 추가적으로 새로운 5개의 학습 패턴 집합 B~F를 구성하였다. 추가적으로 구성한 집합들은 각 파일에서 25초에서 125초까지 25초씩 데이터를 추가하면서 패턴의 양을 다르게 구성하였다. 따라서 실험에 사용하는 학습 패턴 집합의 선별에 따른 종류는 6개 (A~F) 이고, 학습 패턴의 양에 따라서는 25초에서 125초까지 25초씩의 차이로 5종류를 구성하여 총 30개의 학습 패턴 집합을 구성하였다. 그리고 음성/음악 분류 실험으로는 [7]의 데이터에서 학습 패턴 집합(A~F)에 사용되지 않았던 파일에 대해 성능 평가를 수행하였다.

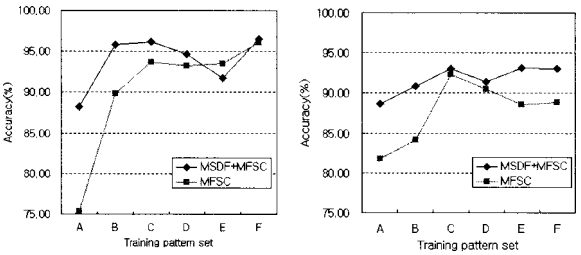
신경망의 입력은 1차원의 MSDF와 40차원의 MFSC로 총 41차원이고, 중간 뉴런의 수에 의한 영향을 평가하기 위하여 5개에서 20개까지 5개씩 중간 뉴런을 추가하면서

실험하였다. 그리고 초기 연결 강도에 의한 영향을 분석하기 위해 구성한 학습 패턴 선별과 중간 뉴런수의 변화에 대해 각각 5회씩 연결 강도를 다르게 초기화 하였다.

4.2. 학습 패턴 선별에 따른 성능 평가

학습 패턴을 다르게 구성한 집합 A~F에 대해 음성/음악 분류기의 성능을 평가하였다. 그림 5는 학습 패턴의 선별에 따른 음성/음악 결과를 나타낸다. 그림 5에 나타난 결과는 학습 패턴으로 집합 A는 25초를 사용하였고, 나머지 집합 B~F는 25초씩 데이터를 추가하여 50초를 사용한 경우의 결과이다. 학습 패턴의 양이 50초 일 때의 실험 결과를 분석해보면 패턴 집합 E를 제외한 모든 경우에 대해 MFSC만 단독으로 사용했을 때보다 MSDF를 추가적으로 사용했을 때 성능 향상이 이루어졌다. 집합 B~F에 대해 성능 변화를 살펴보면 음성에 대해 MFSC만 사용했을 때 최대-최소 차이가 6.22%이고 MSDF를 함께 사용한 경우 4.78%이다. 그리고 음악에 대해서는 MFSC의 경우 8.16%이고 MSDF를 함께 사용한 경우에는 2.34%이다. 이러한 성능 향상은 MSDF가 MFSC만 사용할 때의 오분류를 줄일 수 있다는 것을 보여주며, 학습 패턴 선별 과정에서 발생할 수 있는 문제를 줄일 수 있음을 알 수 있다. 또한 패턴 선별에 따른 성능 차이는 음성/음악 분류에서 어떤 특징 파라미터를 사용할지의 결정도 중요하지만, 학습 패턴의 선별 또한 중요함을 나타낸다.

그림 6은 음성/음악 분류 특징 파라미터로 MFSC만 사용한 경우와 MSDF를 추가적으로 사용했을 때 패턴 집합 C의 분류 결과로 오분류된 패턴 분포를 나타낸다. 실험에서 MSDF를 추가함으로써 음성에 대해서는 지속성이 짧은 오분류 패턴을 줄일 수 있었고, 음악에 대해서는 1에 가까운 지속성을 갖는 오분류 패턴을 줄일 수 있었다. 기존의 특징 파라미터인 HZCRR과 LSTER의 분포로 오분류 감소를 분석하면 음성과 음악 모두 분류가 어려운 영역들에 대해서 MSDF의 사용으로 오분류가 줄었음을 확인할 수 있다. 실험에서 MSDF를 MFSC와 함께 사용함으로써 전체적인 성능 향상을 보였으나, 음성에 대해서는 1에 가까운 MSDF에서, 그리고 음악에 대해서는 0에 가까운 MSDF에 대해서 오분류 패턴이 증가한 경우도 발생하였다.



(a) (b)

그림 5. 학습 패턴 선별에 대한 성능 (a) 음성, (b) 음악
Fig. 5. Performance for training pattern selection (a) Speech, (b) Music.

4.3. 학습 패턴의 양에 따른 성능 평가

그림 7은 학습 패턴의 양에 따른 음성/음악 분류기의 성능이다. 음성의 경우 학습 패턴의 양이 증가함에 따라 75초까지는 성능도 같이 향상되었으나, 그 이상의 양에 대해서는 많은 성능 변화가 나타나지 않았다. 그리고 음악의 경우에는 50초 이상의 데이터에 대해서 큰 성능 변화가 나타나지 않았으며, MSDF를 추가함으로써 음악이 음성에 비해 더 높은 성능 향상이 이루어졌다. 음성/음악 분류의 성능에 있어서 학습 패턴의 양과 관련한 결론으로 MSDF와 MFSC를 모두 사용할 때 분류에 더욱 효과적이며, 학습 패턴의 양을 증가시켜도 MSDF를 추가했을 때보다 높은 결과가 나타나지 않음을 알 수 있다.

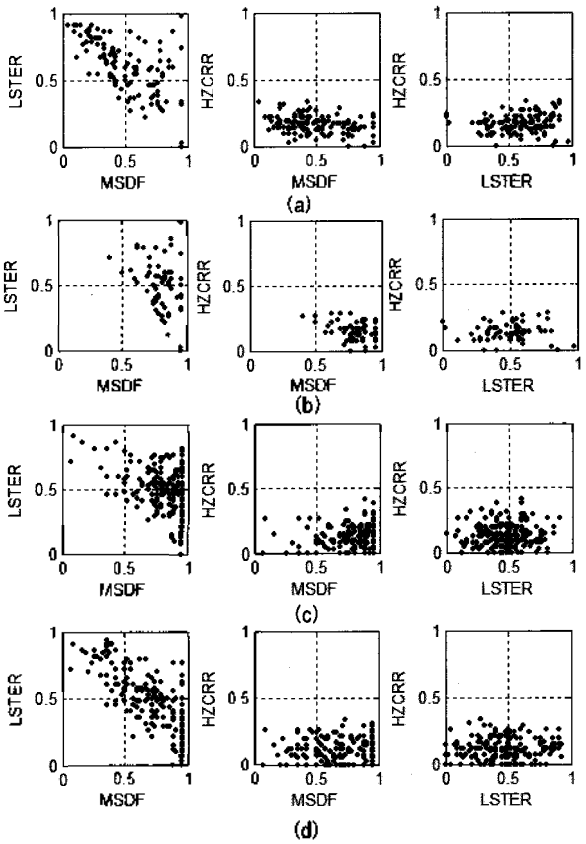


그림 6. MSDF 추가에 따른 성능 개선
(a) 음성 (MFSC), (b) 음성 (MSDF+MFSC)
(c) 음악 (MFSC), (d) 음악 (MSDF+MFSC)
Fig. 6. Performance improvement for MSDF addition
(a) Speech (MFSC), (b) Speech (MSDF+MFSC)
(c) Music (MFSC), (d) Music (MSDF+MFSC).

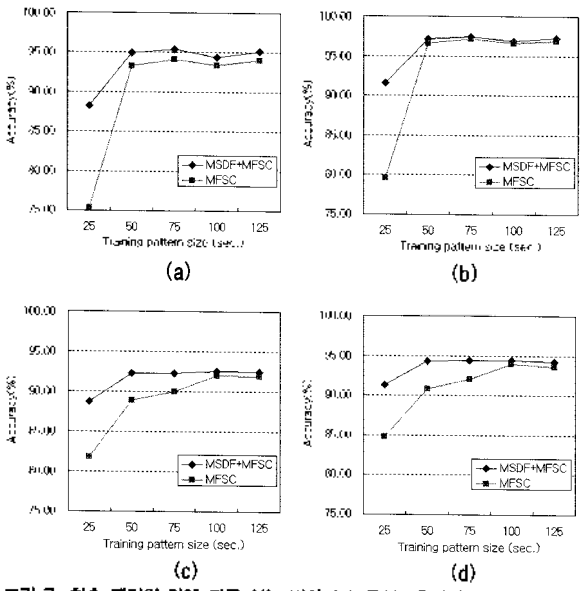


그림 7. 학습 패턴의 양에 따른 성능 변화 (a) 음성 (후처리 과정 전), (b) 음성 (후처리 과정 후) (c) 음악 (후처리 과정 전), (d) 음악 (후처리 과정 후)

Fig. 7. Performance variation for training pattern size (a) Speech(without post-processing), (b) Speech (with post-processing), (c) Music (without post-processing), (d) Music (with post-processing).

그리고 후처리 과정의 영향을 보면 MFSC만 사용한 경우 음성은 3.1%, 음악은 1.92%의 성능 향상이 있었고, MSDF를 함께 사용한 경우 음성은 2.24%, 음악은 1.97%의 성능 향상이 있었다. 이러한 성능 향상은 이웃하는 세그먼트가 동일한 음성이거나 음악임에도 불구하고 불명확한 결정 경계에 의해 MFSC만 사용했을 때 더 많은 오분류가 있었음을 나타낸다.

4.4. 신경망 구성에 따른 성능 평가

음성/음악 분류기의 성능은 특징 파라미터의 종류와 학습 패턴의 선별에 영향을 받는다는 것을 확인하였다. 이번에는 신경망 모델의 구성에 따른 음성/음악 분류기의 성능을 분석하고자 한다.

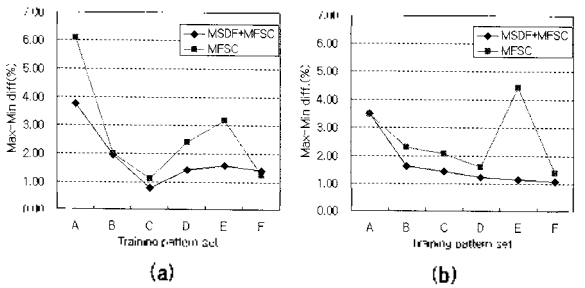


그림 8. 신경망 구성에 따른 성능 평가 (a) 음성, (b) 음악

Fig. 8. Performance evaluation for neural network architecture (a) Speech, (b) Music.

그림 8은 각 패턴 집합에 대해 5개부터 20개까지 5개씩 중간 뉴런을 추가하면서, 5회씩 각기 다른 연결 강도로 초기화한 실험 결과로 최대 분류율과 최소 분류율의 차이를 나타낸다. 그림 8은 최대-최소 분류율의 차이를 나타내기 때문에 이 값의 차이가 적을수록 신경망의 구성이 음성/음악 분류 성능에 미치는 영향이 적다고 할 수 있다. 집합 B~F에 대해 MFSC만 사용하는 경우 최대-최소 차이는 음성은 2.05%, 음악은 3.04%이고 MSDF를 함께 사용하는 경우 음성은 1.17%, 음악은 0.53%이다. 이 결과는 각 패턴 집합별로 MFSC만 사용하는 경우보다 MSDF를 추가한 경우가 분류 성능도 안정적이고, 신경망 구성에 따른 영향도 적음을 나타낸다.

V. 결 론

본 연구에서는 멀티미디어 인덱싱의 전처리 단계로 사용할 수 있는 음성/음악 분류에 관한 연구를 수행하였다. 음성/음악 분류를 위하여 적은 학습 데이터를 사용하면서도 높은 성능을 발휘하는 기존의 특징 파라미터에 스펙트럴 피크 트랙 분석을 기반으로 하는 새로운 특징 파라미터를 추가하여 성능을 향상시켰다. 다양한 실험 방법으로 제안한 음성/음악 분류기의 성능을 평가한 결과 기존의 연구에서도 밝혀진 바와 같이 특징 파라미터의 선별과 학습 패턴의 선별이 중요함을 확인할 수 있었다. 그러나 음성에 비해서 음악은 장르가 다양하고 범위가 넓기 때문에 이에 대한 효과적인 학습 패턴의 선별이 필요하다.

따라서 향후 연구 방향으로는 본 연구에서 제안한 방법의 장점을 살리면서, 학습 패턴 선별을 어떻게 하는 것이 음성/음악 분류기의 성능에 중요한 영향을 미치는지 연구할 계획이다.

참 고 문 헌

1. Tong Zhang, C.-C. Jay Kuo, "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification," IEEE Trans. Speech and Audio Proc. 9 (4) 441-457, 2001.
2. Lie Lu, Hong-Jiang Zhang, Hao Jiang, "Contents Analysis for Audio Classification and Segmentation," IEEE Trans. Speech and Audio Proc. 10 (7) 504-516, 2002.
3. Serkan Kiranyza, Ahmad Farooq Qureshi, Moncef Gabbouj, "A Generic Audio Classification and Segmentation Approach for Multimedia Indexing and Retrieval," IEEE Trans. Speech and Audio Proc. 14 (3) 1062-1081, 2006.

4. 한학용, 김수훈, 허강인, "오디오 데이터의 특징 파라미터 구성에 따른 내용 기반 분석," 한국음향학회지, 21 (2) 182-189, 2002.
5. Soonil Kwon, Narayanan, S., "Unsupervised Speaker Indexing Using Generic Models," IEEE Trans. Speech and Audio Proc. 13 (5) 1004-1013, 2005
6. John Saunders, "Real-Time Discrimination of Broadcast Speech/Music," in Proc. ICASSP, 2 993-996, 1996.
7. Eric Scheirer, Malcolm Slaney, "Construction and Evaluation of A Robust Multifeature Speech/Music Discriminator," in Proc. ICASSP, 2 1331-1334, 1997.
8. Carey, M.J., Parris, E.S., Lloyd-Thomas, H., "A Comparison of Features for Speech, Music Discrimination," in Proc. ICASSP, 1 149-152, 1999.
9. Ji-Soo Keum, Hyon-Soo Lee, "Speech/Music Discrimination using Spectral Peak Feature for Speaker Indexing," in Proc. ISPACS, 323-326, 2006.
10. Balaji Thoshkahna, V. Sudha, K.R. Ramakrishnan, "A Speech-Music Discriminator using HILN Model Based Features," in Proc. ICASSP, 5 425-428, 2006.
11. 박슬한, 최무열, 김형순, "켑스트럼 거리 기반의 음성/음악 판별 성능 향상," 대한음성학회 말소리, 56 195-206, 2005.
12. 이경록, 서봉수, 김진영, "오디오 인덱싱을 위한 음성/음악 분류 특징 비교," 한국음향학회지, 20 (2) 10-15, 2001.
13. Hard Harb, Liming Chen, "Robust Speech Music Discrimination using Spectrum's First Order Statistics and Neural Networks," in Proc. ISSPA, 125-128, 2003.
14. 지민제, "음성의 음향 스펙트로그램 분석," 대한음성언어의학회 학술대회 심포지움 및 워크샵, 111-127, 1995.

저자 약력

• **금 지 수 (Ji-Soo Keum)**

1998년 2월 : 강남대학교 전자계산학과 (공학사)
2000년 2월 : 경희대학교 전자계산공학과 (공학석사)
2000년 3월~현재 : 경희대학교 컴퓨터공학과 (박사과정)
※ 주관심분야 : 멀티모달 화자인덱싱, 패턴인식, 신경망

• **임 성 길 (Sung-Kil Lim)**

1997년 2월 : 경희대학교 수학과 (이학사)
1999년 2월 : 경희대학교 전자계산공학과 (공학석사)
1999년 3월~현재 : 경희대학교 컴퓨터공학과 (박사과정)
※ 주관심분야 : ASA(Auditory Scene Analysis), 패턴인식, 신경망

• **이 현 수 (Hyon-Soo Lee)**

1979년 2월 : 경희대학교 전자공학과 (공학사)
1982년 4월 : 일본 게이오대학원 전기공학과 (공학석사)
1985년 4월 : 일본 게이오대학원 전기공학과 (공학박사)
1999년 9월--2000년 8월 : 미국 오레곤 주립대학교 전기 및 컴퓨터공학과 방문연구원, 미국캘리포니아대학교 (U.C.I) 전기 및 컴퓨터공학과 방문연구원
1985년~현재 : 경희대학교 컴퓨터공학과 교수
2005년~현재 : 경희대학교 정보통신대학원 원장 겸 전자정보대학 학장
※ 주관심분야 : 컴퓨터구조 및 VLSI, 병렬처리, 패턴인식, 신경망, 음성처리