한국음향학회지 제28권 제3호 pp. 273~278 (2009)

주파수 변화율을 이용한 음성과 음악의 구분

Speech and Music Discrimination Using Spectral Transition Rate

양 경 철*, 방 용 찬*, 조 선 호*, 육 동 석*
(Kyong-Chul Yang*, Yong-Chan Bang*, Sun-Ho Cho*, Dongsuk Yook*)

*고려대학교 컴퓨터학과

(접수일자: 2009년 1월 16일; 수정일자: 2009년 3월 12일; 채택일자: 2009년 3월 27일)

주파수 분석을 통해 음성과 음악의 특성을 살펴보면, 대부분 악기는 특정 주파수 소리를 지속적으로 내도록 고안되어 있다는 것을 알 수 있고, 음성은 조음 현상에 의해서 점차적인 주파수 변화가 발생하는 것을 알 수 있다. 본 논문에서는 이러한 음성과 음악이 갖고 있는 주파수 변화 특성을 이용하여 음성과 음악을 구별하는 방법을 제안한다. 즉, 음성과 음악을 구분해 주는 특성 값으로서 주파수 변화율을 사용하고자 한다. 제안한 주파수 변화율인 STR (spectral transition rate) 기반의 SMD (speech music discrimination) 실험 결과, 기존의 알고리즘보다 빠른 응답 속도에서 상대적으로 높은 성능을 보임을 알 수 있었다.

핵심용어: 음성과 음악의 구분, 주파수 변화율
투고분야: 음성처리 분야 (2,5)

In this paper, we propose the spectral transition rate (STR) as a novel feature for speech and music discrimination (SMD). We observed that the spectral peaks of speech signal are gradually changing due to coarticulation effect. However, the sound of musical instruments in general tends to keep the peak frequencies and energies unchanged for relatively long period of time compared to speech. The STR of speech is much higher than that of music. The experimental results show that the STR based SMD method outperforms a conventional method. Especially, the STR based SMD gives relatively fast output without any performance degradation.

Keywords: Speech and music discrimination, Spectral transition rate
ASK subject classification: Speech Signal Processing (2,5)

I. 서 론

최근 음성 인식 시스템의 응용 분야가 넓어지면서, 실제 생활 환경에서도 좋은 성능을 얻기 위한 전처리 방법이 많은 관심을 받고 있다. 전처리 응용 분야도 다시 세분화 되면서 방송과 같은 음악 환경에서 음성을 음악으로부터 구분해 내는 방법에 대한 연구가 진행되고 있다.

기존의 SMD (speech and music discrimination) 방법을 살펴보면 음악의 주요 특성이라고 생각될 수 있는 시간에 따라 변하는 리듬을 이용하여 음성과 음악을 구분하는 방법들이 제안되었다 [1-3]. 이러한 방법들은 대체로 음악은 음성의 변화에 비해 상대적으로 느리며 비교적 일정한 간격으로 변한다는 원리를 사용하였기 때문에, 음악의 종류에 따라 템포가 빨라지거나 사용하는 악기가 변화하면 그 성능이 크게 변할 수밖에 없다. 주파수가 발생하는 빈도를 사용한 방법은 음성과 비슷한 주파수 영역을 가지는 짧고 빠른 음악에는 강인하지 못한 단점이 있다 [4]. 주파수 성분의 분석 방법인 spectral temporal response field (STRF)를 사용하는 방법은 시간에 따른 주파수 변화율을 고려하지 않는다는 단점이 있다 [5]. 최근에 발표된 MMCD (mean of minimum cepstral distance)의 경우 일정한 프레임 사이의 cepstrum 거리의 최소값의 평균이 작으면 음성으로 크면 음악으로 구분하였고 [6], spectral flux (delta spectrum magnitude)의 경우 프레임 사이의 스펙트럼 에너지 차이를 이용하여 음성과 음악을 구분하였다 [7]. 이러한 방법들은 비교적 좋은 성능을 보였지만, 성능을 유지하면서 빠른 응답을 얻지 못하는 단점을 갖고 있다.

주파수 분석을 통해 음성과 음악의 특성을 살펴보면, 대부분 악기는 특정 주파수 소리를 내도록 고안되어 있다

책임저자: 육 동 석 (yook@voice.korea.ac.kr)
136-701 서울특별시 성북구 안암동 5가 1번지 고려대학교 컴퓨터·통신공학부 음성정보처리연구실
(전화: 02-3290-3202; 팩스: 02-3290-3641)

는 것을 알 수 있고, 음성은 발성 과정에서 조음 기관이 서서히 움직여 가면서 소리를 내고 그 과정에서 점차적인 주파수 변화가 발생하는 것을 알 수 있다. 본 논문에서는 이러한 음성과 음악이 갖고 있는 주파수 변화 특성을 이용하여 음성과 음악을 구별하고자 한다. 즉, 음성과 음악을 구분해 주는 특징 값으로서 주파수 변화율을 사용하고자 한다. 제안한 주파수 변화율인 STR (spectral transition rate) 기반의 SMD 실험 결과, MMCD 기반 방법보다 전체적으로 좋은 성능을 보였으며, 특히 음성이 발성되는 과정에서 조금씩 변하는 주파수 변화율을 사용함으로써 빠른 응답 속도에서 상대적으로 높은 성능을 보였다.

제 2 장에서 음성과 음악의 주파수 특성에 대해 살펴보고, 제 3 장에서는 새로운 특징 값인 STR을 제안하고, 제 4 장에서는 STR과 MMCD의 성능을 실험적으로 비교한다.

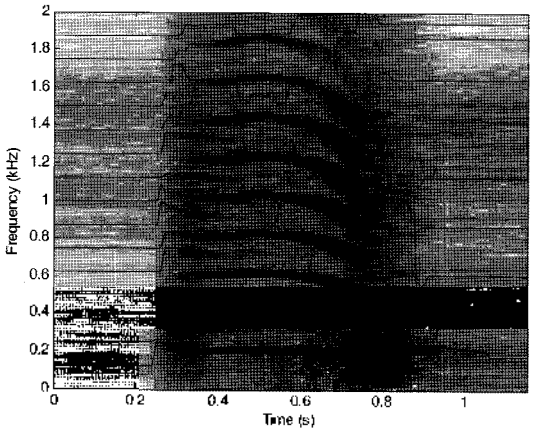
II. 음성과 음악의 주파수 변화 특성

이 장에서는 음성과 음악의 스펙트로그램 상에서의 특성 차이를 비교하여 주파수 도메인에서 음성과 음악을 구분해주는 특성을 분석한다.

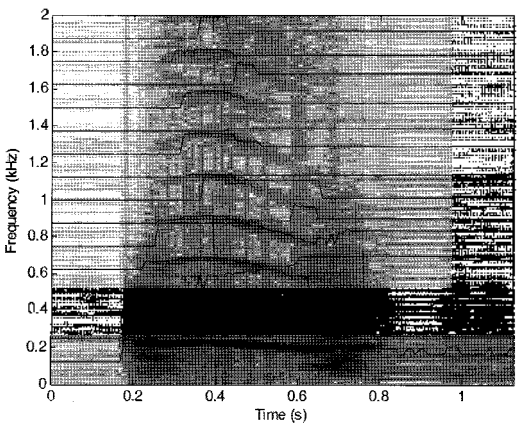
2.1. 음성의 주파수 변화 특성

그림 1은 단모음 '아' (aa)와 '이' (iy)의 스펙트로그램에 각 밴드에서 에너지 피크를 가지는 주파수 변화 추이를 실선으로 표시한 것이다. (125 Hz 간격으로 주파수 대역을 묶은 것을 밴드라고 한다.) 각 밴드에서 최고 에너지 값을 스펙트럼 피크 (spectral peak)라고 한다. 만약 다수의 스펙트럴 피크가 발생하면 최대값을 취한다. 그림 1의 '아'와 '이'는 어느 정도 일정한 주파수 간격으로 스펙트럼 피크가 발생한다. 그림 1-(a)의 '아'의 경우 음소가 지속되는 중간 부분은 상대적으로 피크의 변화가 작고 음소가 시작되는 부분과 특히 끝나는 부분에서는 변화가 크게 발생하는 것을 볼 수 있다. 이 경우는 화자가 음소의 중간 부분은 일정한 주파수의 소리를 냈지만 발성하는 전후 과정에서 조음 기관을 움직여 다른 주파수로 전이하는 발성을 했기 때문이다. 그림 1-(b)의 '이'의 경우에는 단모음이지만 음소가 발성되는 전 과정에서 피크의 변화가 발생한다. 많은 경우 화자가 의도적으로 해당 주파수를 유지하지 않으면 단모음이 발성되는 동안에도 높은 주파수 또는 낮은 주파수로 점차적으로 움직여 가는 주파수 변화가 쉽게 발생한다. 이와 같이 각각의 음소를 발성하기 위해 발성 기관을 움직이는 순간마다, 밴드 별 에너지가 최고인 지점의 주파수가 점차적으로 변화하는 것을 알 수 있다. 즉, 음성은 하모닉스의 변화가 매 순간 점차적으로 발생한다. 그 이유는 음성은 성도를 통해서 조음 기관이 변형될 때 주파수 변화가 발생하는데 [8], 사람은 조음 기관을 움직이며 소리를 발성하므로 악기에서처럼 기계적으로 단절된 주파수의 소리를 순간적으로 내는 것이 아니라 소리를 변화시킬 때마다 주파수 대역이 연속적으로 변화하기 때문이다.

그림 2는 연속 문장 "She had your dark suit in greasy wash water all year"의 스펙트로그램에 밴드별 에너지 피크를 가지는 주파수 변화 추이를 표시한 것이다. 연속 문장에서는 다양한 주파수 변화를 볼 수 있다. 즉, 현재 음소에서 다음 음소로 변화해 가면서 주파수 변화가 점차적으로 발생하고, 이미 살펴 본 단모음의 경우와 같이 음소가 시작하거나 끝나는 부분에서도 주파수 변화가 발생한다.



(a) '아' (aa)



(b) '이' (iy)

그림 1. 모음의 spectral peak
Fig. 1. The spectral peaks of vowels.

2.2. 음악의 주파수 변화 특성

그림 3-(a)는 기타 연주곡의 스펙트로그램에 스펙트럼 피크를 표시한 것이다. 이 경우, 밴드 별 스펙트럼 피크주파수가 일정하게 유지되다가 순간적으로 변화되는 것을 볼 수 있다. 일반적으로 악기들은 인성한 주파수의 소리를 내도록 고안되어 있어, 연주된 소리는 특정 주파수를 일정한 시간 동안 지속되다가 새로운 음이 발생할 때 다른 주파수로 순간적으로 변화한다. 가끔 음악에서 음성과 같은 점진적인 주파수의 변화도 있으나 많지 않아 실험 결과에 큰 영향을 미치지는 못하는 것을 발견하였다.

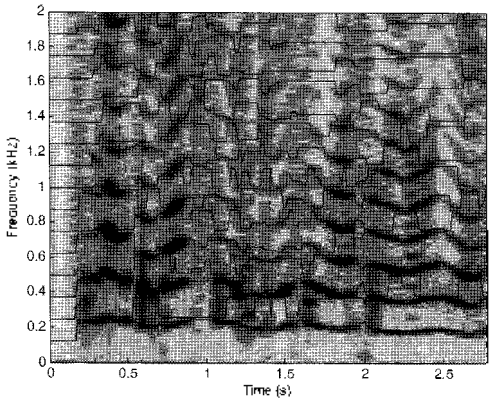
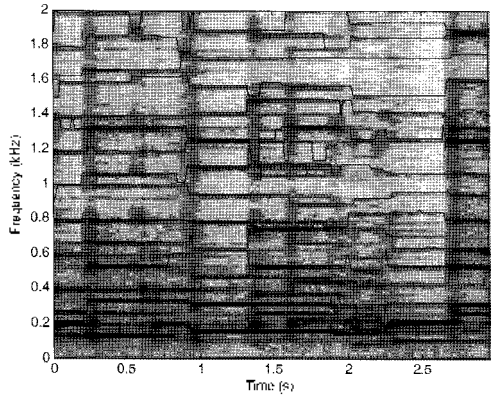
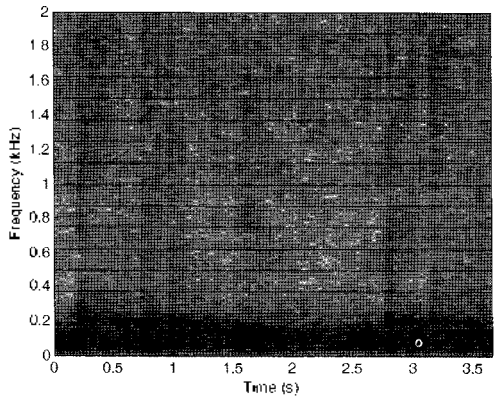


그림 2. 연속음의 spectral peak
Fig. 2. The spectral peaks of an utterance.



(a) 기타 연주곡



(b) 드럼 연주곡

그림 3. 음악의 spectral peak
Fig. 3. The spectral peaks of music.

그림 3-(b)는 비교적 빠르고 강한음의 드럼 연주곡이다. 드럼과 같은 타악기의 경우 주파수 변화 현상은 거의 없다는 것을 알 수 있다.

이러한 분석을 통해서 음악은 특정 주파수에서 시작하여 일정한 시간 동안 같은 주파수를 유지하는 반면, 음성의 경우 발성하는 매 순간마다 지속적으로 조금씩 변하는 것을 관찰할 수 있었다. 다음 장에서는 이러한 특성을 이용하여 음성과 음악을 구분하는 방법을 제안한다.

III. STR 기반 SMD

이 장에서는 주파수 변화율을 이용하여 음성과 음악을 구분하는 STR 특징 기반 SMD 알고리즘을 제안한다.

3.1. Spectral Transition Rate

피크 주파수의 변화량은 아래 식 (1)과 같이 계산한다.

$$d(t,b)=\begin{cases}0, & \text{if } |f(t,b)-f(t-1,b)|>f_{max}\\ |f(t,b)-f(t-1,b)|, & otherwise\end{cases} \quad (1)$$

여기서 $d(t,b)$는 시간 t에서 밴드 b의 주파수 변화량이며, $f(t,b)$는 시간 t에 밴드 b의 피크 주파수이고, f_{max}는 변화 제한 폭이다. 순간 주파수 변화량은 시간 t와 $t-1$ 사이의 주파수 변화량이다. 이때, 주파수 변화량이 f_{max} 이상이면 새로운 소리가 다른 주파수에 발생한 것으로 간주하고 $d(t,b)$를 0으로 계산한다. 또한, 밴드의 평균 에너지가 전체의 평균 에너지에 비해 일정한 비율 이하인 경우에도 순간 변화량 $d(t,b)$를 0으로 계산하여 상대적으로 낮은 에너지를 갖는 주파수 대역의 변화를 제외 하도록 하였다.

식 (2)은 일정한 기간 동안의 피크 주파수 변화량이다.

$$STR(t)=\sum_{b=start}^{end}\left(\sum_{\tau=0}^{T}d(t+\tau,b)\right)^2 \quad (2)$$

여기서 $STR(t)$은 입력된 소리가 T시간 동안 점차적으로 높은 주파수 대역으로 또는 낮은 주파수 대역으로 움직여가는 변화량이다. 순간 주파수 변화량인 $d(t,b)$를 각 밴드 별로 T까지 더한 후 그 제곱 값을 유효 밴드까지 합한 값이다. 여기서 $start$와 end는 주파수 변화 현상을

관찰하는 유효 밴드 대역이다. 음성과 음악을 구분하는 특징 값을 만들기 위해서, 하모닉스의 변화를 일정한 밴드로 나누어 그 추이를 추적하는데, 평균 270 Hz부터 3,010 Hz 사이에 포만트 (formant) 주파수와 평균 피치를 고려하여 [8], 밴드별로 스펙트럴 피크 값을 추적한다.

음성의 $STR(t)$은 소리가 변화하는 순간에 그 크기가 크게 나타나지만 소리가 유지되는 구간에서는 작게 나타날 수 있다. 이런 경우를 보상하기 위해 일정한 구간의 평균값을 사용한다. 식 (3)은 SMD 알고리즘에서 사용하는 최종 STR 값을 구하기 위한 식이다.

$$MSTR(t) = \left(\sum_{t=0}^{W} STR(t) \right) / W \tag{3}$$

여기서 W는 평균을 구하는 윈도우의 크기이다.

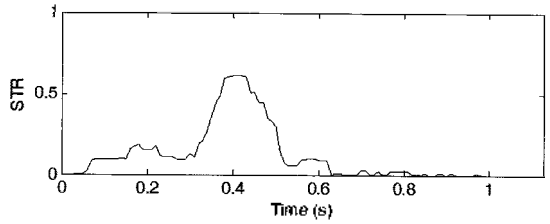
3.2. 음성과 음악의 STR

그림 4는 그림 1의 단모음 '아' (aa)와 '이' (iy)의 STR 값을 나타냈다. 단모음 '아'의 경우에는 주파수 변화 현상이 발성 시작 부분과 끝 부분에서 크게 발생하였으므로 시작하는 시점과 끝나는 시점에서 STR 값이 크게 나타났다. 단모음 '이'의 경우는 발성하는 과정에서 점차적으로 낮은 주파수로 변화한 경우다. 주파수 변화 현상은 발성 초기에 크게 나타나므로 STR 값도 전반부에서 크게 나타났다.

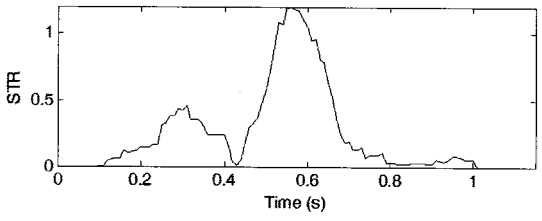
그림 5는 연속 문장의 STR 값을 나타내었다. 음성의 경우 다양한 주파수 변이 현상으로 인해 STR 값이 매 순간 크게 나타난다.

그림 6은 그림 3의 기타와 드럼 연주곡의 STR이다. 기타와 드럼 연주곡 모두 음성과 비교하면 주파수 변화 현상이 적게 발생하므로 STR 값이 음성에 비해 상대적으로 작게 나타난다.

STR 기반의 SMD에서는 학습 데이터를 이용하여 음성과 음악의 STR 분포를 구하고 오차율이 최소가 되는 STR threshold 값을 구한 후, 테스트 데이터의 STR 값이 threshold 보다 크면 음성으로 작으면 음악으로 구분한다.



(b) 단모음 '이' (iy)



(a) 단모음 '아' (aa)

그림 4. 모음의 STR
Fig. 4. The STR of vowels.

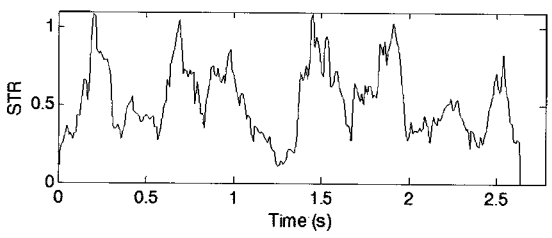
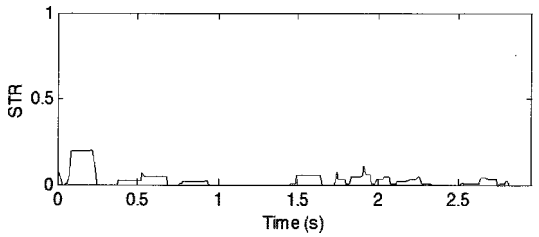


그림 5. 연속 문장의 STR
Fig. 5. The STR of a sentence.

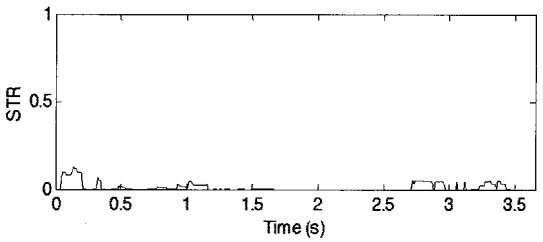
IV. 실험 및 결과 분석

4.1. 실험 환경

실험을 위하여 음성 데이터는 노이즈가 없는 TIMIT 데이터베이스를 사용하였으며, 음악 데이터는 사람들에게 친숙한 여러 가지 음악을 장르별로 분류하여 사용하였다. 학습 데이터 세그먼트 (segment) 는 총 177 개의 곡을 사용하였으며 장르별로는 시네마 38 개, 키보드 26 개,



(a) 기타 연주곡의 STR



(b) 드럼 연주곡의 STR

그림 6 음악의 STR
Fig. 6 The STR of music.

오케스트라 19 개, 그리고 현악기, 관악기, 재즈는 각각 16 개, 팝은 46 개를 사용하였다. 학습 시간은 평균적으로 처음부터 약 80초 정도의 구간을 사용하였고 분류 실험에 사용된 데이터 세그먼트는 장르별로 각 6 개씩 총 42 개 데이터 세그먼트를 사용하였다. 음악 데이터는 CD원본에서 16,000 Hz 샘플링 하였고, Fourier Transform 윈도우의 크기는 128 ms, 10 ms 간격으로 전진하며 SMD를 수행하였다. STR 계산에 사용된 유효 음성 주파수 대역으로는 125 Hz에서 2,000 Hz까지로 정하였다.

식 (1)에서 f_{max}는 40.3 Hz으로 하였으며, 식 (2)에서 주파수 변화 계산을 위한 시간 T는 실험을 통해 최적한 값인 200 ms를 사용하였다. STR의 평균 윈도우 W에 따라서 threshold를 구하여 실험하였다. 또한, STR의 빠른 응답에 대한 성능을 알아보기 위해 MMCD의 경우 cepstral distance 계산을 위한 시간 보다 작은 T=150 ms, W=150 ms으로도 실험을 진행하였다.

4.2. 실험 결과

그림 7은 평균 윈도우 W가 250 ms인 경우의 MMCD와 STR의 음성과 음악의 장르별 SMD 성능 (음성이든 음악이든 정확하게 분류한 것의 백분율)을 비교한 것이다. MMCD는 음악의 종류에 따라서 성능의 변화가 크다. 반면, STR은 음악의 종류에 무관하게 상대적으로 안정적인 성능을 보인다. STR과 MMCD의 평균 SMD 성능은 그림 8과 같다. STR을 이용한 SMD는 평균 윈도우 W의 크기가 큰 경우 MMCD와 유사한 성능을 나타낸다. 윈도우 W가 크기가 작은 경우, 즉, 빠른 응답에서는 MMCD에 비해 높은 성능을 보이는 것을 알 수 있다.

V. 결 론

본 논문에서 음성은 음악과 달리 점차적인 주파수 변이 현상이 있음을 보이고 이를 이용하여 음성과 음악을 구분하는 특성 값인 STR을 제안하였다. 실험을 통해서 STR은 MMCD보다 우수한 성능을 보이는 것과, 고전적인 패턴인식 방법보다 빠른 성능을 보이는 것도 알 수 있었다 [5]. 특히 음악에 장르 변화에 강인하며, 비교적 빠른 응답에서도 우수한 SMD 성능을 보였다.

이번 연구에서 고려하지 못한 인접 구간의 판정 결과를 참조하는 후처리기와, 피치 하모닉스의 인접밴드로의 이동 문제는 추후에 연구할 계획이다.

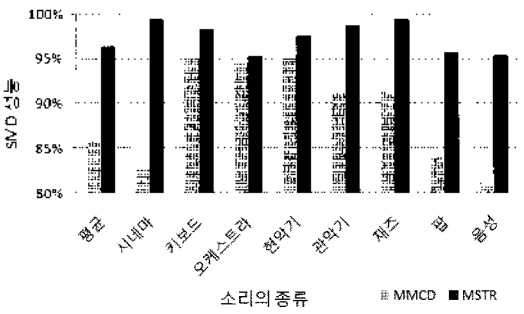


그림 7. MMCD와 STR의 장르별 SMD 성능 비교 (W=250 ms, T=200 ms)

Fig. 7. SMD performances of MMCD and STR for each genre (W=250 ms, T=200 ms).

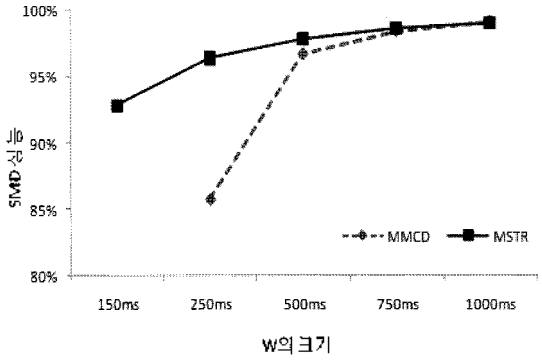


그림 8. W 변화에 따른 MMCD와 STR의 성능 비교

Fig. 8. Performance comparison between MMCD and STR for various W.

감사의 글

이 논문은 2008년도 정부 교육과학기술부)의 재원으로 한국학술진흥재단의 지원을 받아 (KRF-2006-311-00822) 수행된 연구임. 본 논문의 초고는 한국음향학회 추계학술대회에서 발표 되었음 [9].

참 고 문 헌

1. R. Jarina, N. O'Connor, and S. Marlow, "Rhythm detection for speech-music discrimination in MPEG compressed domain," *IEEE Conference on Digital Signal Processing*, vol. 1, pp. 129-132, Jul. 2002.
2. O. M. Mubarak, E. Ambikairajah, and J. Epps, "Novel features for effective speech and music discrimination," *IEEE International Conference on Engineering of Intelligent Systems*, pp. 1-5, Sep. 2006.
3. M. J. Carey, E. S. Parris, and H. Lloyd-Thomas, "A comparison of feature for Speech, music discrimination," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 149-152, Mar. 1999.

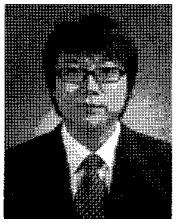
4. Ji – Soo Keum, and Hyon – Soo Lee, "Speech/music discrimination based on spectral peak analysis and multi –layer perceptron," *International Conference on Hybrid Information Technology (ICHIT'06)*, vol. 2, pp. 56–61, 2006.
5. Nima Mesgarani, Malcolm Slaney, and Shihab A. Shamma, "Discrimination of speech from nonspeech based on multi–scale spectro–temporal modulation," *IEEE Trans on Audio, Speech, and Language Processing*, vol. 14, No. 3, pp. 920 –930, May. 2006.
6. M. Y. Choi, H. J. Song, and H. S. Kim, "Speech/music discrimination for robust speech recognition in robots," *IEEE International Symposium on Robot and Human Interactive Communication*, pp. 118–121, Aug. 2007.
7. E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1331–1334, Apr. 1997.
8. L. Rabiner and B–H. Juang, *Fundamentals of Speech Recognition* Prentice Hall, New Jersey, pp. 20–28, 1993.
9. 양경철, 육동석, "주파수 변화율을 이용한 음성과 음악의 구분," 한국음향학회 2008년도 추계학술발표대회 논문집, 37–41쪽, 2008.

저자 약력

•양 경 철 (Kyong–Chul Yang)

1995년 : 숭실대학교 전기공학과 학사
2009년 : 고려대학교 컴퓨터학과 석사

•방 용 찬 (Yong–Chan Bang)



2008년 : 건양대학교 전산게임학과 학사
2009년 : 고려대학교 컴퓨터학과 석사과정

•조 선 호 (Sun–Ho Cho)

2009년 : 고려대학교 컴퓨터학과 학사
2009년 : 고려대학교 컴퓨터학과 석사과정

•육 동 석 (Dongsuk Yook)

1990년 : 고려대학교 컴퓨터학과 학사
1990년 : 고려대학교 컴퓨터학과 석사
1999년 : Ph.D. in Computer Science, Rutgers – The State University of New Jersey
1999년~2001년 : Senior Software Engineer, IBM T. J. Watson Research Center, New York, USA
2001년~현재 : 고려대학교 컴퓨터·통신공학부 교수