Dynamic Time Warping 기법을 이용한 내용기반 디지털 오디오 검색

Contents based digital audio retrieval using the Dynamic Time Warping Technique

성보경, Bo-Kyung Sung, 고일주, Il-Ju Ko

숭실대학교 미디어학과

요약 최근 다양한 분야에서(웹 포털, 유료 음원서비스 등) 디지털 오디오의 검색이 사용되고 있다. 이러한 분야에서 디지털 오디오의 검색은 디지털 오디오 데이터가 가지고 있는 자체 메타 정보를 이용하여 이루어진다. 하지만 메타 정보가 다르게 작성 되었거나 작성되지 않은 경우 정확한 검색은 어렵다. 요즘 이러한 문제의 보완 방안으로 내용기반 정보 검색 기법을 이용한 검색이 이루어지고 있다. 본 논문에서는 내용 기반 디지털 오디오 검색 방법에 대해 논하고자 한다. 내용기반으로 디지털 오디오를 검색하기 위해 음성 인식 분야에서 유사도 측정에 사용하는 Dynamic Time Warping 기법을 활용하여 디지털 오디오 간의 유사도 측정을 하였다. 제안된 유사도 측정을 통한 내용기반 디지털 오디오검색 방법의 검증을 위해 같은 장르에서 무작위 추출된 100곡에서 시행한 90번의 검색은 모두 성공했다. 검색에 사용된 90개의 디지털 오디오는 10개의 디지털 오디오를 압축방식과 비트율을 다르게 조합하여 만들었다.

핵심어: Audio Retrieval, Contents Based, Dynamic Time Warping

1. 서론

최근 정보처리기술과 통신네트워크기술의 발달로 인하여 대용량의 멀티미디어 콘텐츠 데이터를 사용하는 범위가 인터넷 포털 사이트나 방송국 같은 콘텐츠 서비스 제공업체에서 일반 사용자로 확대 되었다. 이에 따라 많은 양의 멀티미디어 콘텐츠 데이터가 발생하게 되었고 필요한 콘텐츠를 편리하고 효율적으로 찾는 방법의 필요성이 증가 되었으며 이에 대한 여러 가지 방법론들이 제안 되고 있다. 멀티미디어 콘텐츠의 한 부분인 음악 콘텐츠 역시 사용 범위가 광범위해지고 그 양 또한 기하급수적으로 증가 하여 효율적으로 관리하거나 필요한 콘텐츠의 빠른 검색할 수 있는 방법을 요구하게 되었다.

현재 사용되어지고 있는 음악 콘텐츠들은 거의 모두 디지털화된 형태이다. 디지털 오디오 데이터는 텍스트 데이터와 다르게 내용자체를 명확하게 표현할 수 있는 방법이 없다. 그래서 그동안 오디오 데이터는 관련된 정보를 태그형태로 저장하였으며 태그 기반으로 필요한 콘텐츠를 검색하거나 관리 하였다.

오디오 데이터에 기록되어져 있는 태그 데이터를 이용한 검색은 크게 두 가지의 한계점을 가지고 있다. 첫째로 태그에 기록되어져 있는 정보 이외의 것에 대한 검색이 불가능하다. 태그로 오디오 데이터를 설명 할 수 있는 정보 양은 많지 않다. 그리고 사용자들은 태그의 정보 보다 실제 오디오 데이터의 내용에 더 관심을 가지며 기억한다. 한 노래를 기억할 때 몇 년도에 발매된 노래인지 보다 그 노래의 주요 선율이 어떤지를 기억한다. 그러나 실제로 이러한 오디오 데이터의 내용 자체는 태그 형태로 표현 할 수 없다. 즉 표면적 정보의 검색만 가능하다. 그리고 검색을 위해서는 필요한 오디오 데이터에 대한 규격화된 태그 정보를 알고 있어야 한다. 둘째로 모든 오디오 데이터들은 올바른 내용으로 태그 정보가 저장 되지는 않는다. 오디오 콘텐츠를 제공하는 서비스 업체의 경우는 올바른 태그 정보를 기록 하여 배포 하겠지만, 대다수의 개인 사용자들은 자신이 디지털 오디오를 생성할 경우 태그 정보를 기록하지 않는다. 또한 잘못 입력하는 경우도 있다. 이러한 두 가지 이유로 태그를 이용한 오디오 검색은 한계가 있다. 이러한 한계를 극복하기 위해 디지털 오디오 콘텐츠에서 내용기반검색 기법이 필요하다.

현재 사용되고 있는 내용기반 오디오 검색은 미디 데이터를 이용하거나 태그 데이터와 몇 부분의 오디오 데이터를

혼합하여 전처리 시켜놓은 메타 데이터를 이용한 경우가 많다.[1] 미디데이터를 이용한 경우는 미디 데이터 자체가 곡의 빠르기, 성조 등의 음악적 요소를 가지고 있기 때문에 이것을 이용한 검색은 용이하나 디지털 오디오 데이터에 이러한 방법을 적용시키기는 어렵다. 그리고 메타 데이터를 이용한 경우 역시 메타 데이터가 존재하는 오디오에 대해서만 검색이 가능 하므로 모든 오디오 데이터에 대한 검색이 불가능 하다. 위의 두 가지 내용기반 오디오 검색 방법은 태그기반 검색보다 좋은 성능을 보여주긴 하지만 위에서 언급했던 태그기반 검색이 가지고 있는 두 가지 한계점을 극복 하지는 못했다.

본 논문에서는 디지털 오디오 데이터의 파형 유사도 측정을 통한 내용기반 오디오 검색 방법에 대해 논하고자 한다. 일반적인 검색의 공통된 목적은 사용자가 주어진 자료를 기초로 가장 유사한 정보를 찾는 것이다. 디지털 오디오의 검색 또한 일반적 검색과 마찬가지로 주어진 디지털 오디오를 기초로 내용적으로 똑같거나 유사한 내용을 포함하는 디지털 오디오를 찾아내는 것이다. 하지만 디지털화된 오디오 데이터는 디지털화되는 과정에서 서로 다른 규격(인코딩 방식, 표본화 및 양자화간격, 초당 비트율 등)으로 인코딩이 되기 때문에 내용적으로 같은 것이라도 수치적으로 차이를 보이게 된다. 즉 디지털 오디오의 검색을 위해서는 검색을 위해 주어진 디지털 오디오와 비교해야할 디지털 오디오간의 파형적인 유사도를 측정 할 수 있어야 한다. 제안된 검색은 음성인식분야에서 목소리의 유사도를 측정하는 방법으로 사용되던 Dynamic Time Warping(DTW) [2] [3] 기법을 활용하여 디지털 오디오의 파형 유사도를 측정한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 본 논문에서 제안하는 내용기반 디지털 오디오 검색 방법의 전체적 구조에 대하여 간단히 설명 하며 3장에서는 유사도 측정을 위해 디지털 오디오 데이터를 단순화 시킨 Mel Frequency Cepstrum Coefficient(MFCC) [4] [5] 계수를 이용하여 프레임 단위로 변환하는 것에 대한 설명을 하고 4장에서는 Dynamic Time Warping(DTW) 기법을 활용하여 디지털 오디오 데이터의 내용적 유사도를 측정하는 방법에 대해 설명한다. 5장에서는 조합된 총 9가지의 경우에 대한 디지털 오디오 데이터의 검색 실험을 수행하였으며 마지막으로는 결론으로 끝을 맺는다.

2. 내용기반 음악 검색 구조

아래 (그림 1)은 제안된 내용기반 음악 검색 구조를 나타낸다. 내용기반 음악검색 구조의 흐름은 다음과 같다. 내용기반 음악 검색을 위해 찾기 원하는 음악의 일부분을 입력한다. 그리고 입력된 음악을 유사도 비교에 알맞은 규격으로 전처리를 한 다음 파형 데이터를 벡터 데이터 형태로 변환시킨다. 변환된 벡터 데이터는 유사도 측정 과정에서 오디오 데이터베이스에 벡터 형태로 저장되어 있는 모든 데이터와 각각의 유사도 정도가 측정된다. 마지막으로 측정된 데이터에서 유사도가 가장 높은 음악이 출력된다.

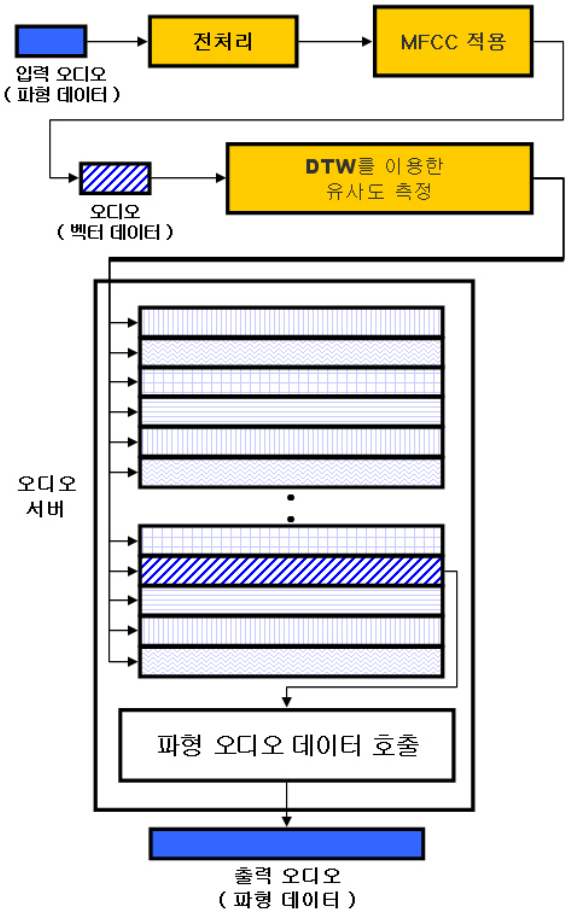


그림 1. 내용기반 음악 검색 구조

사용자들이 사용하는 음악 파일들은 모두 같은 규격으로 디지털화 되어 있지 않다. 그러므로 전처리 과정을 통해 음악 데이터베이스에 저장되어 있는 규격으로 바꿔야 한다. 일단 각기 다른 코덱으로 압축된 것을 비압축 파형 데이터 형태로 복원해야 한다. 그리고 음악 데이터베이스와 똑 같은 규격의 양자화 간격과 부호화 정도로 입력 오디오를 변환하는 과정이 필요하다.

파형 데이터는 순간순간의 소리의 크기 값 만을 가지고 있다. 이러한 값은 음악적 의미를 나타내주지 못하므로 내용기반 검색에 사용되기 힘들다. 그래서 시간에 따른 소리의 절대 값을 의미를 가질 수 있는 형태로 변환하여야 한다. 본 검색 구조에서는 여러 소리가 복합된 음악에 적합하게 간소화 시킨 MFCC를 이용하여 파형 데이터를 특징 벡터 형태로 변환 한다.

음악의 내용 기반 검색을 위해서는 음악 간의 유사성에 대한 측정이 필요하다. 입력된 음악의 특징 벡터들과 오디오 데이터베이스에 특징 벡터로 저장되어 있는 모든 음악 간의 유사도 측정을 위해 DTW 기법을 사용하였다.

3. 특징 벡터 추출

파형 데이터는 시간의 흐름에 따라 그 순간순간 에서의 소리 크기를 표현하는 값들의 연속이다. 이러한 데이터는 그 자체로 음악적 의미를 지니지 못한다. 그래서 파형 정보를 내용적인 의미를 가지는 형태로 변환 시켜줘야 한다. 소리의 유사성을 측정하는 음성 인식 분야에서도 파형 형태의 음성 데이터를 특징 벡터로 추출한 후 이것을 이용하여 유사성을 측정한다. 특징 벡터의 추출은 (그림 2)와 같이 입력 음악을 23ms 크기의 프레임으로 자르고, 각각의 프레임에서 특징 벡터를 구한다.

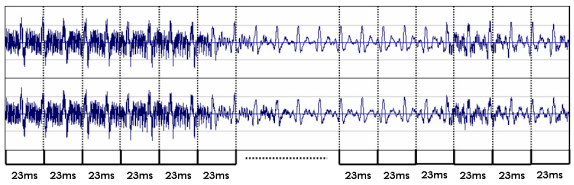
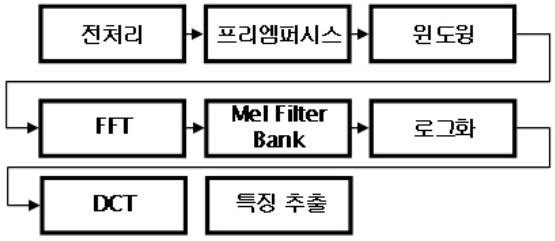


그림 2. 입력 음악의 프레임화

파형 데이터로부터 특징 벡터를 추출해주는 다양한 계수들이 있다. Short Time Fourier Transform(STFT)[6]값 스펙트럼의 중심을 뜻하며 스펙트럼의 형태를 측정하는 방법 중 하나인 Spectral Centroid[7], 오디오 신호의 리듬 정보를 수치적으로 산출하는 계수로서 Wavelet변환 후 대역 별 상관도를 구하여 Beat 히스토그램을 만들고 Beat정보를 추출하는 Beat Histogram[8], 인간의 발성 모델에 입각해서 음성 신호를 부호화 하는 방법으로 오디오 파형의 샘플값에서 필터 계수를 구하여 성대에서 입, 코까지 성도 특성을 8~12차의 전극형(All-pole) 필터에 근사 시키는 방법인 Linear Predictive Coding(LPC)[9][10] 및 인간 청각 특성을 모델링 하는 방법으로 오디오 신호의 절대값 스펙트럼을 log scale한 후 FFT bin을 그룹화 하여 인간의 청각 특성에 맞는 Mel-Frequency 스케일로 변환한 MFCC[4][5] 등이 있다.

MFCC 는 본래 음성인식 분야에서 입력된 사람 목소리로 부터 특징을 추출하기 위해 사용되었다. 이것은 인간 발성 모델을 기반으로 한 것이 아니라 청각 모델을 기반으로 만들어졌다. 그래서 원래 사용되던 단일 채널의 음성뿐만 아니라 여러 악기들이 복합된 음악 데이터 에서도 특징 벡터 추출이 가능하다고 가정하고 음악의 특징을 추출하는 것으로 사용하였다. 본 논문에서는 음성에 적합하게 설계된 것을 음악에 적합하게 사용하기 위해 몇 가지 단계를 축소하여 간소화 시킨 MFCC를 사용하였다. 그리고 그 중 13차 특징 벡터까지를 실험 데이터로 사용 하였다.

(그림 3)은 음성인식 분야에서 사용하던 일반적인 MFCC 과정과 음악에서 특징 벡터 추출을 위해 간소화 시킨 MFCC 과정을 보여준다. (그림 3)에서 (나)는 본 논문의 실험을 위해 기존의MFCC과정을 간소화 시킨 것이다. 일반적으로 사용되는 과정에서 전처리 과정, 프리-엠퍼시스 과정, 윈도우 과정이 생략 되었다. 이 과정은 소리의 파형이 사람 목소리로만 이루어진 경우 잡음과 불필요한 무음 부분을 제거하기 위해 사용 되었던 단계이다. 하지만 유사도 측정에 사용되는 음악은 여러 가지 소리들이 복합적으로 녹음된 상태이기 때문에 이 과정을 제거하고 단순화 시킨 MFCC 과정을 사용하였다.



(가) 일반적인 MFCC 과정

(나) 단순화시킨 MFCC 과정

그림 3. MFCC 과정

MFCC과정 에서 FFT처리는 시간 영역의 파형 신호를 주파수 영역으로 변환시키는 과정이다. 주파수 영역은 신호의 각 성분들을 알 수 있기 때문에 시간 영역에 비해 신호의 해석 및 처리가 용의한 장점이 있다. Mel Filter bank처리는 FFT처리된 결과를 가지고 멜 스케일 필터 값을 곱해주는 과정이다. 이 필터는 보통 사람은 1kHz 이하에서 잘 듣는다는 것을 이용 하여 1kHz 이하 부분은 촘촘히 분석하고 그 이상은 간격을 넓게 분석하여 좀 더 청각구조에 접근시킨 필터이다. 로그화 처리는 필터를 통과한 값에 로그를 취하는 것이다. 우리의 귀가 소리의 크기에 대해 로그 함수로 느끼기 때문에 로그화 과정을 거친다. Discrete Cosine Transform(DCT)처리는 필터 뱅크의 출력 간의 상관관계를 없애주고 파라미터의 특징을 모아주는 역할을 한다. 하나의 프레임에 MFCC를 적용할 경우 필터 뱅크의 개수 만큼 특징

벡터 값이 나온다. 특징 추출 처리는 필터 뱅크 수만큼 나온 특징 차수에서 작은 차수부터 필요한 개수 만큼 선택하는 것이다.

4. 유사도 측정

대부분의 사람은 동일한 단어를 발성할 경우 발성 시마다 발성시간 길이가 변화한다. (그림 4)는 동일단어에 대해 동일인의 발성을 비교한 것이다. 동일단어를 동일인이 발성 하더라도 파형의 길이 및 수치가 달라짐을 볼 수 있다. 이 두 파형은 일치하지 않지만 소리의 내용은 다르지 않다. 음악의 경우도 마찬가지다. 같은 파형을 다른 규격으로 디지털화 되거나 잡음이 섞일 경우 또는 악기 소리에서 적은 변화가 있을 경우 모두 파형이 달라진다. 하지만 이것을 내용적으로 완전히 다르다고 할 수 없다. 이러한 점을 고려하여 내용기반 음악 검색을 하기 위해서는 길이와 파형 값이 차이 나는 두 파형의 유사도를 측정할 수 있어야 한다.

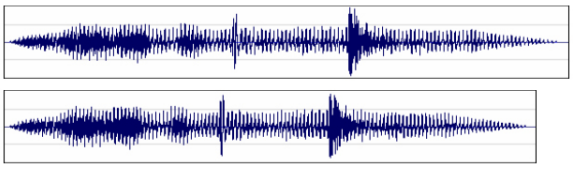


그림 4. 동일단어에 대한 동일인의 발성 파형 비교

본 논문의 실험에서는 추출된 특징 벡터들 간의 유사도를 수치적으로 측정하기 위한 방법으로 DTW[2][3]를 사용한다. DTW는 과거 음성인식을 위한 방법의 하나로 사용되던 것이다. 음성인식을 위해서는 입력되는 음성을 데이터베이스에 저장된 참조 데이터와 유사도를 측정해야 한다.

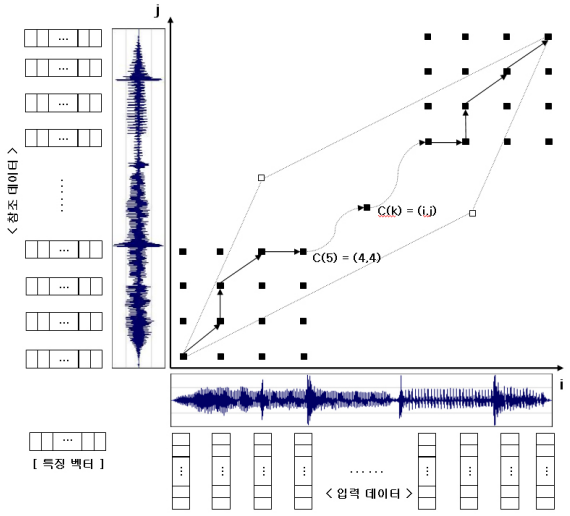


그림 5. DTW를 이용한 유사도 측정 방법

본 논문의 실험은 음성데이터 간의 비교가 아니라 복합 사운드 형태의 디지털 음악 간의 비교지만 DTW를 이용하여 유사도를 측정 하였다. 유사도는 두 음원이 얼마나 절대 거리차이를 보이는지 표현한 것이다.

위의 (그림 5)는 길이가 다른 입력패턴과 참조패턴이 정합되어지는 비선형 함수를 나타내고 있다. 이와 같이 서로 다른 두 개의 자료에서 최적의 정합 경로를 동적으로 찾아 두 데이터를 서로 비교할 수 있는 방법이 DTW 이다.

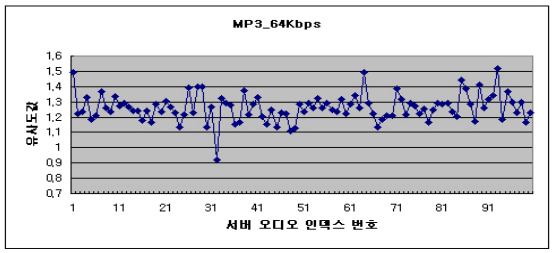
아래 식(1)은 DTW를 계산하는 식이다. D(A,B)에서 A,B는 입력되는 두 음원 이고, 거리차이를 구하는 것이 함수 D이며 DTW를 의미한다.

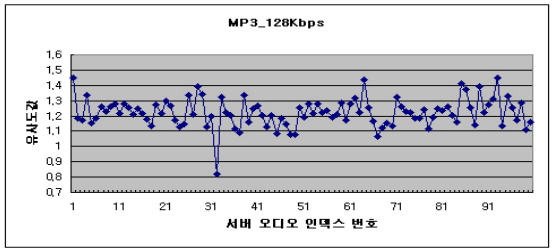
$$D(A,B) = \min_{F}\left[\frac{\sum_{k=1}^{K} d(c(k))w(k)}{\sum_{k=1}^{K} w(k)}\right] \quad (1)$$

유사도 측정을 위해 입력되는 두 음원은 특징벡터의 열로 표현할 수 있다. 두 음원의 총 길이가 다르다고 한다면 두 음원 을 하나의 시간 축 으로 정합 시켜야 한다. 이것을 해주는 함수를 와핑(Warping) 함수라 하고, c(k)가 이에 해당된다. k는 와핑 함수에서 열로 표현되는 음원의 포인트 갯수를 의미한다. w(k)는 가중치 계수로 와핑 함수의 탄력 있는 특성을 유도하는데 도입되며 적절한 와핑 함수를 찾는데 이용한다. Σw(k)는 와핑 함수들의 가중치 합의 거리를 의미하며, 함수에서 포인트 개수 k에 의한 영향을 보상하기 위해 사용된다.

5. 실험 및 결과

실험에서 사용하는 음악은 모두 22050Hz로 양자화 되고, 16bits로 부호화 되어 오디오 데이터베이스에 저장 되었다. 검색을 위해서 음악 서버와 검색 명령을 위한 음악들을 사용 하였다. 음악 서버는 동일 장르에서 무작위 추출된 100개의 곡으로 구성하였다. 검색 명령을 위해 사용한 곡은 10곡이며 이 곡은 음악 서버에 존재하는 곡이다. 그리고 10개의 곡은 각각 압축방식, 초당 비트율을 다르게 조합한 9가지의 다른 규격으로 변환하여 총 90개의 음악으로 만들었다. 사용한 압축 코덱은 mp3, ogg 그리고 wma로 3가지 이다. 그리고 초당 비트율은 각각의 코덱에서 저음질, CD음질, 고음질 로 분류하여 변환 하였다. Mp3 코덱의 경우는 64kbps, 128kbps, 320kbps 로, ogg 코덱의 경우는 96kbps, 128kbps, 350kbps 로, wma 코덱의 경우는 64kbps, 128kbps, 160kbps 로 비트 레이트를 변환 하였다.





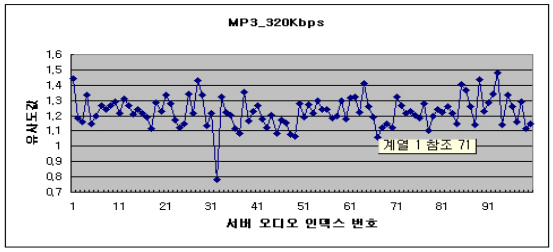
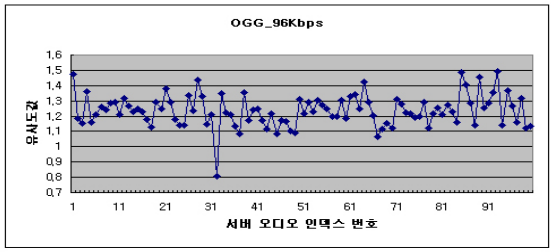
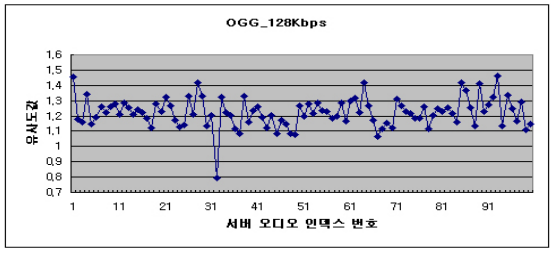


그림 6. MP3 파일의 유사도 그래프





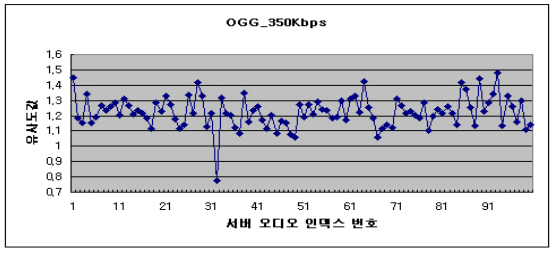
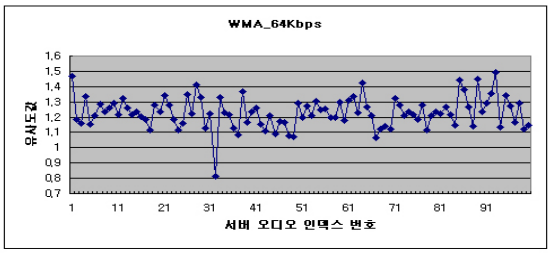
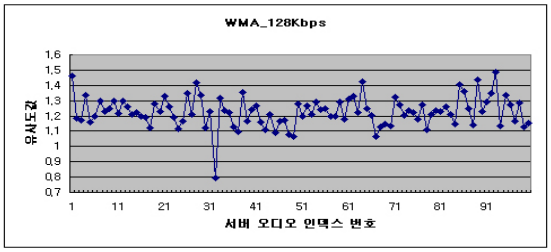


그림 7. OGG 파일의 유사도 그래프





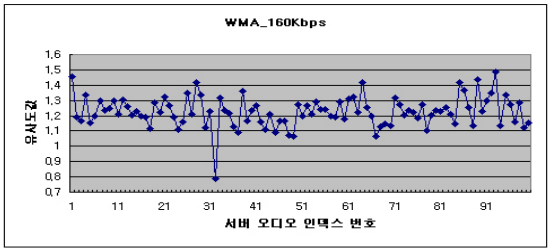


그림 8. WMA 파일의 유사도 그래프

내용기반 음악검색 구조에서의 검색은 동일한 규격으로 디지털화된 음악 간의 검색뿐만이 아니라 다른 규격으로 디지털화 되었어도 동일한 내용의 음악도 검색 하는 것이다. 실험은 한 곡을 9가지의 다른 규격으로 디지털화 하여 검색하여도 동일한 음악을 찾을 수 있다는 것을 증명하는 것이다. 위의 (그림 6)(그림 7)(그림 8)은 하나의 음악을 다른 규격을 가지는 9가지의 음악 파일로 변환한 후 검색을 한 결과이다. 위의 그림의 경우는 음악 서버의 32번째 음악과 동일한 노래를 9가지의 규격으로 변환하여 서버의 음악들과 측정한 유사도 값을 그래프로 표현한 것이다. 각 그래프의 세로축은 유사도 값이다. 유사도가 높을수록 유사도 수치는 0에 가까워진다. 각 그래프의 가로축은 서버에 저장된 음악의 인덱스이다. 모든 그래프의 32번째 부분에서 유사도 값이 다른 것 보다 0에 가깝게 나왔다. 유사도 그래프에서 0에 일치하지 않게 나오는 이유는 음악들 간에 조밀하게 비교하지 않았기 때문이다.

실험은 위와 같은 방식으로 10곡의 음악을 이용하여 총 90번의 검색을 실시하였다. 전체 실험에서 검색 결과는 100% 일치함을 보였다. 다른 규격으로 디지털화된 음악을 검색어로 사용 하여도 동일한 내용을 가지고 있는 음악을 찾을 수 있다는 것을 증명 하였다.

6. 결론

본 논문에서는 내용기반 음악 검색을 위한 디지털 음악간 유사도 측정 방법을 제안 하였다. 제안된 유사도 측정 방법은 DTW기법을 이용하였다. 입력된 디지털 음악은 간소화 시킨MFCC 를 이용하여 파형 데이터에서 특징 벡터들을 추출 하였고, 두 음악의 특징 벡터들의 유사도 측정을 통하여 음악 간의 유사도를 측정 하였다. 유사도 측정법을 통하여 다른 규격으로 디지털화된 음악 파일 사이에서 내용이 동일한 음악의 검색이 가능 하였다.

다른 규격으로 디지털화된 음악 간의 유사도 측정법은 내용기반 오디오 검색을 위해 필요한 기술이다. 대부분의 사용자가 음악 검색을 위해 입력하는 음악조각 들은 모두 같은 조건으로 디지털화 되어 있지 않을 가능성이 높다. 이러한 점을 고려한 내용기반 음악 검색을 위해서는 음악간 유사도 측정법이 필요하다.

향후 연구 과제로서 유사도 측정법을 발전시켜 비슷한 분위기의 음악을 검색하는 방법에 대한 연구가 필요하다.

참고문헌

[1] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search and retrieval of audio", IEEE Multimedia, 3(2), 1996.

[2] EJ. Keogh, MJ. Pazzani.: Derivative Dynamic Time Warping. First SIAM international Conference on Data Mining (2001)

[3] B-H. Juang.: Hidden Markov model and dynamic time warping for speech recognition—A unified view. AT&T BELL LAB.Tech.J.Vol.63,no.7, (1984)1213-1984

[4] M. Slaney.: A critique of pure audition. Computational Auditory Scene Analysis. (1997)

[5] Z. Jun, S. Kwong, W. Gang, Q.Hong.: Using Mel-Frequency Cepstral Coefficients in Missing Data Technique. EURASIP Journal on Applied Signal Processibng. (2004)

[6] SH. Nawab, TF. Quatieri.: Short-time Fourier transform. Prentice Hall Signal Processing Series. (1987)

[7] JJ. Burred, A. Lerch.: A hierarchical approach to automatic musical genre classification. Digital Audio Effects Conference. (2003)

[8] G. Tzanetakis, G. Essl, P. Cook.: Human perception and computer extraction of musical beat strength. Digital Audio Effect Conference. (2002)

[9] A. Harma, UK.Laine.: A comparison of warped and conventional linear predictive coding. IEEE Transactions on. Speech and Audio Processing. (2001)

[10] J. Makhoul.: Linear prediction--A tutorial overview. Proceeding of IEEE. (1975)