자기상관을 이용한 음성 신호의 MIDI 변환

박상보°, 황인준
아주대학교 정보통신전문대학원 정보통신공학과, 고려대학교 전자공학과
bookcom°@ajou.ac.kr, ehwang04@korea.ac.kr

Speech-to-MIDI Conversion with Autocorrelation

Sangbo Park°, Eenjun Hwang
Graduate School of Information and Communication, Ajou University
Department of Electronics and Computer Engineering, Korea University

요 약

효율적인 멀티미디어 검색의 필요성이 증대됨에 따라 내용기반 멀티미디어의 검색에 대한 다양한 기법들이 소개되고 있다. 그 중에서 친숙한 멜로디를 가지고 사용자가 직접 마이크를 통해 생성한 음성 질의에 대한 분석에 대해 다루고자 한다. 음성 질의에 사용되는 음성 데이터를 분석함으로써 검색에 이용하는 것이다. 음성데이터를 분석하기 위한 방법으로 시간영역에서 가장 많이 쓰이는 기법 중의 하나인 자기상관함수를 사용한다. 자기상관함수를 이용하여 특정구간에서 발생하는 일정한 주기 즉 기본주기를 검출할 수 있다. 자기상관함수에 의해 분석된 결과를 가지고, 음의 높낮이를 구하기 위한 기본주파수 검출 알고리즘과 음의 길이, 음의 세기를 결정하기 위한 방법을 제안한다.

1. 서 론

최근 이미지, 오디오, 비디오 등 다양한 멀티미디어 데이터의 양이 빠른 속도로 증가함에 따라, 효율적인 멀티미디어 검색을 위한 다양한 방법들이 연구되고 있다. 타이틀, 저자명, 파일명 등 텍스트 기반 메타데이터를 이용하여 검색했던 전통적인 데이터베이스 검색 기법과는 달리 멀티미디어 데이터에 대한 내용기반 검색 기법의 필요성이 대두되었다. 단순히 텍스트 또는 SQL에 기반하고 있는 기존의 질의 형태는 인덱싱과 내용기반 검색 기법을 통하여 보완될 필요가 있다. 내용기반 검색 기법은 각 멀티미디어 데이터의 특성을 파악하여 검색에 활용한다. 대표적인 기법 중 하나는 친숙한 멜로디에 의한 허밍(humming)을 통하여 MIDI(Musical Instruments Digital Interface)로 저장된 음악 데이터베이스 시스템에 질의를 하는 것이다[1]. 이 방법은 음의 높이 차이에 대한 정보를 이용한 것으로 검색의 효율성을 위해서 MIDI를 사용한다. MIDI는 음 자체에 대한 파형 정보를 가지고 있는 것이 아니라 음을 연주하는 방법과 연주 시기 등에 대한 정보를 저장하고 있다. 마이크로 들어오는 디지털 음성 신호를 MIDI형식으로 바꾸어 데이터베이스 검색에 이용한다. 본 논문에서는 마이크로 입력되는 실시간 음성 신호를 분석하여 MIDI로 변환하는 효과적인 방법에 대해 제안한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서 음성신호 분석에 대한 관련 연구를 소개하고, 3장에서 디지털 음성신호에서의 특징추출 방법에 대해 설명한다. 4장에서 추출된 특징 정보를 가지고 MIDI 데이터로 변환하는 방법에 대해 다루고, 5장에서 구현 방법에 대해 설명하고, 6장에서 결론을 맺는다.

2. 관련 연구

실시간으로 들어오는 음성 데이터를 MIDI로 변환하기 위해서는 음성 신호에서 특징을 추출하는 작업이 선행되어야 한다. 음성 신호를 분석하여 피치(pitch)를 비롯한 음의 시작과 끝, 세기(strength) 등을 검출해야 한다. 피치는 유성음에서 관찰되는 주기 성분으로 피치 주파수는 그 주기의 역수로 음성의 기본주파수가 된다. 인간이 말을 할 때 성대가 진동하여 음성이 발생하는데 그 주기성은 완전하지는 않으나 (quasi-periodic) 보통 파형에서 눈으로 알아볼 수 있을 정도로 규칙적이다[2]. 피치 검출에 대한 알고리즘은 처리 영역에 따라 크게 시간 영역(time-domain)과 주파수 영역(frequency-domain)으로 나눌 수 있다. 시간 영역 피치 검출법은 시간 영역에서 직접 처리하기 때문에 다른 영역으로의 변환이 불필요하다. 시간 영역에서 분석할 수 있는 방법으로 영교차율(ZCR, Zero Crossing Rate), 자기상관함수(Autocorrelation function), 무음비율(Silence ratio) 등이 있다. 영교차율은 특정구간에서 진폭값이 '0'(Zero)이 되는 비율로, 잡음(noise)을 고려하여 '0' 근처의 값도 적당한 기준값(threshold)을 가지고 '0'으로 근사시켜서 처리한다. 자기상관함수는 특정구간에서 주기적인 정보를 찾는데 사용되는 함수로 이후에 자세히 설명한다. 주파수 영역에서의 분석 방법 중 대표적인 것으로 FFT(Fast Fourier Transform)가 있다. FFT는 모든 파형은 단순한 정현파(sine wave)의 합으로 표현 가능하다는 성질을 사용해서 시간영역의 신호를 주파수 영역으로 변환한다. 음악 데이터인 경우 하모닉(harmonic) 사운드를 검출함으로써 피치 검출할 수 있지만 FFT가 낮은 주파수 대역에서 신호를 처리하기 힘들고 표준 FFT의 빈(Bin)이 주파수에 대하여 선형적으로 위치함으로써 생기는 문제점을 보완시킨 constant Q-변환(transform)을 이용하는 방법 또한 연구된다[3]. MFCC는 FFT 변환 결과를 가지고 멜주파수 스케일(Mel-frequency scaling)을 사용하여 사람의 청각기관에 적합하도록 주파수 대역을 조정하는 방법으로 음성인식에도 많이 사용된다[4].

3. 특징검출

3.1 전처리

음성의 경우 매우 짧은 구간의 특성이 일정하게 유지되는

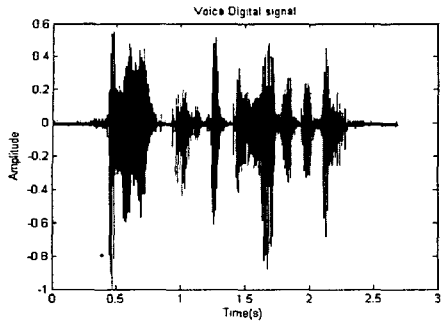


그림 1 음성신호의 시간영역 표현

성질이 있다. 즉 사람의 음성은 하나의 선율(monophonic)을 가지고 있으므로 한번에 한 개의 기본주파수(fundamental frequency)가 있다[5]. 긴 음성 신호를 분석하기에 앞서 신호를 프레임(frame) 단위로 분리하는 절차가 필요하다. 상대적으로 안정적인 주파수 성분을 얻기 위하여 보통 프레임의 길이를 20ms 정도로 한다. 이는 음성이 일반적으로 20ms 정도의 길이에서 특징이 크게 변화하지 않기 때문이다. 그림1은 전체 음성 신호의 그래프를 보여주고 있다. 음성신호의 한 단편인 20ms에 해당하는 프레임에 대한 확대된 그래프를 그림2에서 볼 수 있다. 프레임의 중복 구간의 비율을 보통 0~70%로 하지만 본 논문에서는 50%의 중복(overlapping) 프레임을 사용한다. 해당 프레임의 영교차율이 기준값보다 크고, 평균에너지가 기준값보다 작은 경우 무음구간으로 처리하며 그림2에서처럼 일정한 주기가 나타남을 알 수 있다.

$$ZCR=\sum_{n=1}^{N}\frac{|sgn(x(n)-sgnx(n-1)|}{2N} \quad \cdots (1)$$

영교차율을 구하기 위해서는 위의 식을 사용하고 평균에너지는 다음 절에서 다루어지는 자기상관함수에서 $s=0$ 일 때의 값을 이용한다. 그림3에서 영교차율과 평균에너지의 분포를 볼 수 있다.

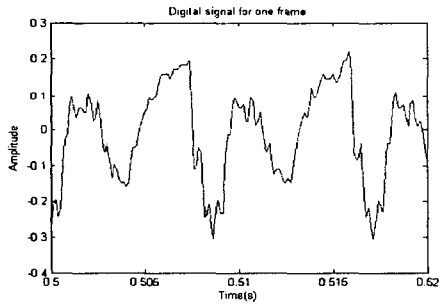


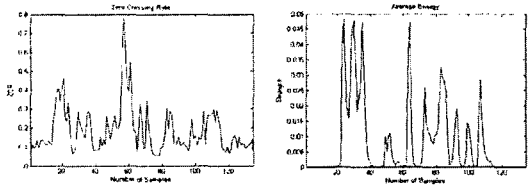
그림 2 한 프레임에 대한 시간 영역 표현

3.2 자기상관함수

프레임으로 구분된 각각의 음성신호에 대하여 자기상관함수를 사용한다. 자기상관함수는 두 신호의 시간 지연으로 유사성을 측정한다. 아래 식을 사용하여 음성 신호에서 발생하는 기본주기(Fundamental period)를 찾아낸다.

$$r(s)=\frac{\sum_{t=1}^{N-s}y(t+s)*y(t)}{N-s} \quad \cdots (2)$$

위 식에서 s 는 시간간격을 N 은 각 프레임에 있는 샘플의 개수, t 는 시간을 의미한다. 자기상관을 하기 전에 중앙 클리핑(clipping) 과정을 거쳐 무음(silence) 구간과 잡음(noise) 구간을 제거함으로써 보다 정확하게 피치를 검출할 수 있게 한다. 프레임 자체가 무음으로 구별된 경우는 자기상관함수를 적용하지 않으며, 무음이 아닌 프레임에 대하여 클리핑을 하는 것이다. 클리핑 레벨은 각 프레임의 최대 진폭의 1/3 정도로 정한다[2]. 각 프레임의 최대 진폭에 대하여 기준값을 정하여, 기준값 이하의 값은 '0' 으로 조정한다.



(a) 영교차율 (b) 평균 에너지

그림 3 영교차율과 평균에너지

3.3 기본주파수 검출

그림4의 그래프는 자기상관함수의 결과를 나타내는 것으로 샘플링율을 8000으로 하고 측정한 값으로 34개의 샘플간격 즉 4.25ms 정도에서 주기적인 성질이 나타남을 알 수 있다. 최소값이 되는 두 지점 사이가 주기가 된다. 기본주기를 구하기 위해 제안된 알고리즘은 표1과 같다.

프레임의 길이를 N이라고 하고, 검출된 극소점(local minimum)들의 수를 M, 극소값 특정한 범위에서 근사를 시키기 위한 상수를 K라고 한다. 상수 K는 일정한 주기를 나타내는 극소점을 구하기 위해 발생할 수 있는 오차를 허용하기 위한 것이다. 검출된 극소값을 근사적으로 환산한 후 극소값들 중 최소가 되는 지점간의 시간차를 가지고 기본주기를 검출한다. 기본주기를 구하기 위해 최소가 되는 최초의 두 지점 사이의 거리를 이용한다.

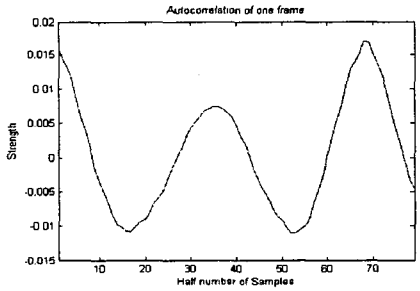


그림 4 한 프레임에 대한 자기상관함수

Algorithm

```
// Local Minimum Detection
for k=0..N-1
        if (x(k)>=x(k+1))
                Decrease_status=TRUE;
        else if (Decrease_status==TRUE
                        AND x(k)<x(k+1))
                min_array_value=x(k);
                min_array_time=k;
                Decrease_status=FALSE;
        else
                Decrease_status=FALSE;
end

// Approximation of min array value
for i=1..M
        min_array_value(i)=min_array_value(i)
                        -min_array_value(i)%K;
end

// Detection of period
min=min_array_value(1);
start_period=0;
for i=1..M
        if(start_period==0
        AND min_array_value(i)==MIN(min_array_value))
                start_period=min_array_time(i);
        else if(start_period!=0)
                period=min_array_time(i)-start_period;
end
```

표 1 기본주기 검출 알고리즘

4. MIDI로의 변환

자기상관함수에 의해 구해진 기본주기에 역수를 취함으로써 기본 주파수를 구한다. 연속한 여러 프레임이 같은 주파수를 가지고 있는 경우 하나의 음을 나타내므로 MIDI로 변환 시 하나의 피치에 해당되는 음을 할당한다. 그림5에서 보는 것과 같이 주파수의 값들이 2개 이상 모여 있는 경우를 하나의 음(Note)로 간주한다. 음의 길이는 연속된 프레임의 수에 따라 결정이 되며, 음의 세기는 각 프레임에서 얻어진 진폭의 평균값을 이용한다. 무음으로 분류된 프레임이거나 하나의 프레임이 떨어져서 존재할 경우 해당 구간의 음의 세기는 '0' (zero)으로 한다. 진폭의 평균값은 무음구간 판별을 위해 사용되었던 평균에너지를 의미하는 것으로 전에 구했던 값을 이용한다.

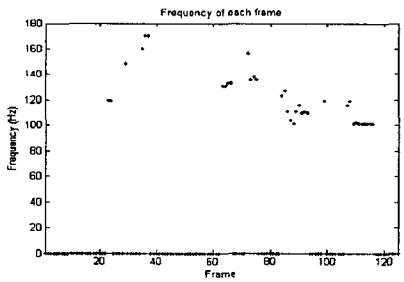


그림 5 주파수 분포

5. 구현

샘플링율이 커질수록 계산 시간이 오래 걸리기 때문에 초당 8000개의 샘플로 된 음성 신호를 가지고, 실험을 하였으며 실험 결과에 대한 통계는 다음 연구(Future work)로 남겨둔다. 구현에 사용되는 프레임의 크기는 20ms로 했고, 무음을 찾기 위한 영교차율과 평균에너지의 기준값은 각 프레임에서의 영교차율과 평균에너지의 평균값으로 했다. 평균에너지의 평균은 평균에너지의 분포를 나타내는 값들의 평균을 의미한다. 사람에 의해 발생할 수 있는 주파수의 범위가 보통 100~1200Hz이므로 이 구간을 벗어나는 주파수는 잡음으로 간주하여 제거하였다. 주파수를 이용하여 피치를 구하기 위해 기존에 알려진 MIDI 노트(note) 테이블을 사용하였다[6]. 표2는 표준 MIDI에 대한 노트 정보의 일부분을 보여주고 있다.

MIDI Note		Frequency	MIDI Note	Frequency	MIDI Note	Frequency
C	36	65.406	48	130.813	60	261.626
Db	37	69.296	49	138.591	61	277.183
D	38	73.416	50	146.832	62	293.665
Eb	39	77.782	51	155.563	63	311.127
E	40	82.407	52	164.814	64	329.628
F	41	87.307	53	174.614	65	349.229
Gb	42	92.499	54	184.997	66	369.995
G	43	97.999	55	195.998	67	391.995
Ab	44	103.826	56	207.652	68	415.305
A	45	110.000	57	220.000	69	440.000
Bb	46	116.541	58	233.082	70	466.163
B	47	123.471	59	246.942	71	493.883

표 2 미디 노트 테이블

6. 결론 및 향후과제

내용기반 음악검색에 사용하기 위한 음성질의를 분석하여 표준 MIDI 파일로 변환하기 위한 방법 중 시간영역에서 직접 처리할 수 있는 자기상관함수를 이용한 방법을 제안하였다. 주파수 영역으로 변환할 필요가 없어 변환시간의 단축은 있으나 여전히 계산하는데 많은 시간이 소요된다. 또한 기본주기를 찾는데 있어 자기상관함수의 그래프에 나타나는 극소점 사이의 길이를 구하는데 처음 2개의 값을 가지고 사용했지만, 주기가 여러 가지 나오거나 복잡한 상황을 고려한 보다 정교한 방법을 생각해 볼 수 있다. 무음구간을 검출하기 위해 사용되는 기준값 측정을 위한 실험 또한 병행되어야 한다.

7. 참고문헌

[1] S. Rho and E. Hwang, "Fast Melody Finding Based on Memorable Tunes," 1st International Symposium on Computer Music Modeling and Retrieval , Montpellier, France, pp. 227-239, May 26-27, 2003

[2] 박일서, 김대현, 조철우, "실시간 음성분석도구의 MatLab 구현," 말소리 제44호

[3] Johan Forberg, "Automatic conversion of sound to the MIDI-format," TMH-QPSR 1-2/1998

[4] Li Tan and Montri Karnjandecha, "Modified Mel-Frequency Cepstrum Coefficient"

[5] Jari Salo, "Pitch-to-MIDI conversion," Automatisoitu musiikin analysointi ja haku - seminaariesitys 17.11.2003

[6] http://tomscarff.tripod.com/midi_analyser/midi_note_frequency.htm