RUI용 음성신호기반의 감정분류를 위한 피치검출기에 관한 연구

* 변성우 ** 이석필
*, ** 상명대학교
*123234566@naver.com, ** esprit@smu.ac.kr

A study on pitch detection for RUI emotion classification based on voice

*Byun, Sung-Woo **,Lee, Seok-Pil
*, ** SangMyung University

요약

컴퓨터 기술이 발전하고 컴퓨터 사용이 일반화 되면서 휴먼 인터페이스에 대한 많은 연구들이 진행되어 왔다. 휴먼 인터페이스에서 감정을 인식하는 기술은 컴퓨터와 사람간의 상호작용을 위해 중요한 기술이다. 감정을 인식하는 기술에서 분류 정확도를 높이기 위해 특징벡터를 정확하게 추출하는 것이 중요하다. 본 논문에서는 정확한 피치검출을 위하여 음성신호에서 음성구간과 비 음성구간을 추출하였으며, Speech Processing 분야에서 사용되는 전 처리 기법인 저역 필터와 유성음 추출 기법, 후처리 기법인 Smoothing 기법을 사용하여 피치 검출을 수행하고 비교하였다. 그 결과, 전 처리 기법인 유성음 추출 기법과 후처리 기법인 Smoothing 기법은 피치 검출의 정확도를 높였고, 저역 필터를 사용한 경우는 피치 검출의 정확도가 떨어트렸다.

1. 서론

컴퓨터 기술이 발전하고, 컴퓨터 사용이 일반화 되면서 휴먼 인터페이스에 대한 많은 연구들이 진행되어 왔다. 휴먼 인터페이스에서 컴퓨터와 사람간의 상호작용은 필수적인 요소이다. 감정을 인식하는 기술은 인간과 컴퓨터간의 상호작용의 한 연구 분야로 컴퓨터가 사람의 감정을 통해 상황을 인지하고 그에 따른 인터랙션 작용을 하는데 있어 중요한 기술이다. 감정을 인식하는 기술들 중 음성신호를 사용하는 기술은 음성신호가 상대방과의 의사소통을 위한 신호이기 때문에 감정에 대한 정보를 많이 포함하고 있는 이유로 영상, 생체신호를 이용하는 기술과 더불어 활발하게 연구가 진행되고 있다[1][2][3].

한편, 피치검출은 음성인식, 감정인식, 멜로디와 허밍간의 톤 매칭을 이용한 Qbsh 등 많은 Speech Processing 분야에서 사용 된다. 특히, 음성신호기반의 감정인식분야에서는 음성신호에서 사람의 감정을 반영하는 특징으로 피치를 사용한 많은 연구가 진행되었다[4][5]. 이러한 연구들은 분류 정확도를 높이기 위해서 특징벡터를 정확하게 추출하는 것이 중요하다.

본 논문에서는 정확한 피치 검출을 위하여 Speech Processing 분야에서 사용되는 전 처리 기법들의 소개와 이 기법들을 사용하여 피치 검출을 수행하고 비교한다. 본 논문의 구성은 다음과 같다. 2장에서 음성구간과 비 음성구간 추출 기법을 소개하고, 3장에서 전처리 기법들에 대한 설명 4장에서는 실험결과 5장에서는 실험 결과를 통해 결론 및 향후과제로 끝을 맺도록 한다.

2. 음성구간/비 음성구간 추출

음성신호에서 음성구간을 추출하는 부분은 감정인식 시스템에서 불필요한 정보가 될 수 있는 비 음성 구간을 제거하는 이유로 감정인식에서 중요한 부분이다. 음성구간 추출하는 flow chart는 다음 그림1과 같다.

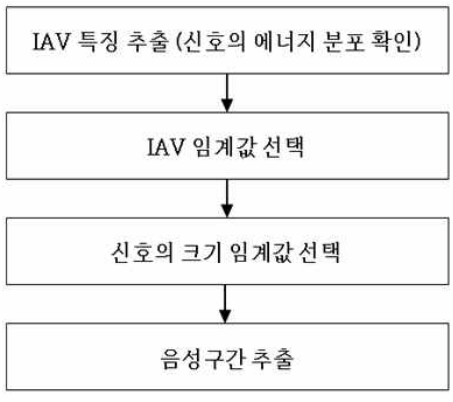

그림 1. 음성구간 추출 flow chart

음성구간 추출하는 flow chart는 다음 그림 1 과 같다. 음성신호 구간은 비 음성신호 구간에 비해 신호의 에너지 값이 크기 때문에 에너지 크기의 값을 반영하는 절대 적분치(Integral Absolute Value) 특징 벡터를 사용 하였으며 식은 다음과 같다.

$$\overline{X}=\sum_{i=1}^{N}|X(i\triangle t)|$$

여기에서,

X : 측정된 신호 ,

△t : 샘플링 시간 간격 ,

N : 샘플의 수 ,

i : 샘플의 순서

IAV 임계 값을 선택하는 과정은 신호에서의 IAV특징 벡터를 추출 한 후 최대 값 최소 값을 구한 후, 최대 값 최소 값 차의 10%만큼 최소 값의 위로 잡는다. 만약 최소 값이 최대 값의 70%보다 크면 임계 값은 최대 값의 20% 아래로 잡는다. 임계 값 선택하는 과정의 예시는 다음 그림 2와 같다.

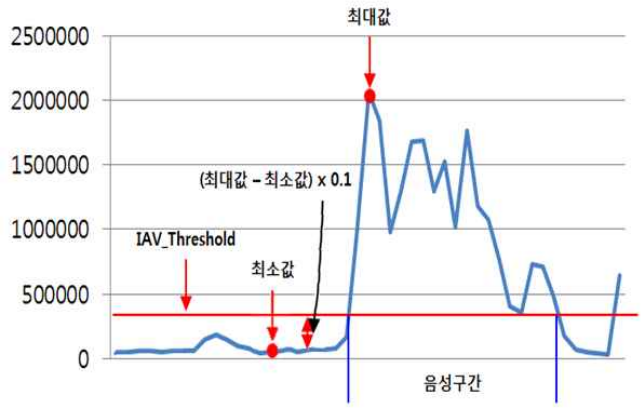

그림 2 IAV 임계 값 선택 예시

신호의 크기 임계 값은 IAV 임계 값에서 프레임 크기로 나눠주어서 구하게 된다. IAV 특징벡터가 프레임내의 모든 신호 값의 절대치를 더한 값이기 때문에 프레임 크기로 나눠주게 되면 프레임의 신호 평균 값이 나오게 된다. 따라서 이 값은 IAV 임계 값을 신호의 크기 임계 값으로 바꾼 값이 된다.

음성 구간을 추출하는 과정은 프레임 단위로 IAV 임계 값 보다 큰 구간이 나오면 해당 프레임 내에서 신호 에너지 임계 값 보다 커지는 지점을 시작 인덱스로 선정하고 시작 인덱스부터 IAV 임계치가 작아지는 구간이 나오면 그 지점을 끝 인덱스로 선정하게 된다. 위 방법을 사용하게 되면 정확하게 음성 구간을 추출할 수 있다.

3. 피치 검출

음성신호에서 피치 검출은 2가지의 전 처리 과정을 거친 후 검출하게 되는데 차단주파수가 700Hz인 Low Pass Filter 과정과 유성음구간 검출과정을 거치게 된다. 전 처리를 거친 신호는 자기상관함수를 사용하여 피치를 검출하게 되고, 검출된 피치는 후처리 과정인 Smoothing 과정을 거치게 된다.

3.1 전 처리 과정

일반적으로 피치는 80Hz~ 500Hz 사이에 존재하고, 대부분 100~ 200Hz에 범위에서 검출된다. 그러나 Fundamental frequency가 모든 하모닉 구조에서 가장 큰 값을 가지고 있는 것은 아니기 때문에 피치 검출 오류를 줄이기 위해 전처리 과정이 필요하다.

Fundamental frequency의 범위는 일반적으로 80Hz ~ 500Hz에 존재하는 이유로 500Hz이상의 주파수 영역은 피치 검출에 필요하지 않다. 따라서 차단주파수가 500Hz인 Low Pass Filter를 사용하면 피치검출의 성능을 높일 수가 있다. 본 연구에서는 차단주파수 700Hz인 Low Pass Filter를 사용하였다.

음성신호는 사람의 발음에서 나오는 파열음, 파찰음, 마찰음, 경음 등 피치와 관련 없는 고주파 성분인 무성음 구간이 존재하게 된다. 이는 피치 검출에서 반드시 제거해 주어야 정확도를 높일 수 있으며 이 부분은 전체 시스템 정확도를 떨어뜨릴 수 있는 부분이다. 일반적으로 음성신호에서 유성음은 준 주기성을 가지는 신호이고 무성음은 주기성이 없는 신호이다. 따라서 무성음은 Autocorrelation값을 Normalized한 값이 임계 값보다 작으면 주기성이 약한 신호이기 때문에 무성음이라 할 수 있다. 본 연구에서는 경험적으로 임계 값을 0.55로 정하였다. Normalized Autocorrelation 식은 다음과 같다.

$$Nomailized\,Autocorrelation = \frac{R_{s1s2}}{\sqrt{E_1 \times E_2}}$$

$$E_1 = \sum S_1^2$$

$$E_2 = \sum S_2^2$$

여기에서,

R : 자기상관함수

S : 시간영역 신호

E : 신호의 에너지 값

3.2 피치 검출 과정

피치는 주기신호의 기본주파수를 의미한다. 일반적으로 사람의 피치는 80Hz ~ 500Hz에 존재하게 된다. 따라서 신호에서 80Hz~ 500Hz까지 주기를 늘리면서 Autocorrelation값이 가장 큰 주기를 찾게 된다. 상관도가 가장 높은 주기는 그 신호의 기본 주파수가 된다.

3.3 후 처리 과정

위의 과정을 거쳐 피치를 검출하게 되면 잘못된 유성음 검출 과정과 같은 여러 가지 이유로 정확하지 않은 피치 값이 존재하게 된다. 이러한 값들은 분류 시스템 전체의 성능을 낮출 수 있기 때문에 노이즈를 제거하는 Soothing 작업이 필요하다. 본 연구에서는 일반적으로 많이 사용되는 Smoothing 기술인 median filter를 사용하였다.

4. 실험

4.1 실험데이터

본 연구를 위한 실험 데이터는 사람에 대한 음성 감정데이터로 감정은 보통, 슬픔, 기쁨, 화남 4가지로 분류하였다. 정확한 데이터를 위하여 방송매체를 통해서 데이터를 취득하였고, 각각의 감정에 대해 3~5개씩, 총 25명의 데이터를 수집하였다. 데이터 취득 대상은 여자/남자로 구분된다. 데이터를 취득할 때 데이터들은 음성만으로 감정구분이 명확한 데이터, 배경음이 없는 데이터들로 선정하였다. 이렇게 취득한 데이터는 16000Hz로 샘플링 하였고 Window 크기는 0.03초 간격으로 500샘플씩 분석하였다.

4.2 분리도 비교

본 연구에서는 하나의 감정에 대한 25명의 데이터를 하나의 클래스로 볼 때 클래스에 대하여 각각 피치의 구간 평균, 구간 분산을 특징벡터로 이용하게 된다. 추출된 특징벡터 비교는 각 클래스의 분포가 가우시안 형태를 가질 때 가장 좋은 평가 기준이 되는 바타캐리야 거리를 이용하여 각 클래스간의 거리를 구한다. 바타캐리야 거리의 식은 아래와 같다.

$$\mu(1/2) = \frac{1}{8}(M_2 - M_1)^T \left\{\frac{\Sigma_1 + \Sigma_2}{2}\right\}^{-1} (M_1 - M_2) + \frac{1}{2} ln \frac{\left|\frac{\Sigma_1 + \Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}}$$

여기에서,

M_1, M_2 : 클래스 1, 2의 평균,

Σ_1, Σ_2 : 클래스 1, 2의 공분산,

|| : 행렬식

4.3 실험 결과

실험 데이터에서 피치 추출 값의 정확도를 보기 위해 남성의 데이터만을 가지고 실험하였으며, 전 처리, 후처리 과정을 수행하여 피치 값을 추출한다. 분리도가 높은 전 처리, 후처리과정이 각 클래스간 거리가 떨어져 있기 때문에 패턴인식에 용이하다. 때문에 분리도가 높으면 특징이 잘 추출됐다고 할 수 있다.

특징벡터는 피치의 평균, 분산도로 하였으며, 실험 결과는 그림 3과 같다. Low Pass Filter와 Smoothing 과정을 거친 결과는 슬픔 기쁨 간의 거리 값이 20.10839로 가장 크게 나왔으며 전체 거리의 평균값은 9.871063833으로 나왔다.

	보통	슬픔	기쁨	화남
보통	0	0.573779	16.2278	14.126186
슬픔	0.573779	0	20.108389	6.362689
기쁨	16.2278	20.108389	0	1.82754
화남	14.126186	6.362689	1.82754	0

표 1. LPF+Smoothing과정의 분리도

유성음/무성음 구간 검출과 Smoothing 과정을 거친 결과는 보통 기쁨 간의 거리 값이 25.106256으로 가장 크게 나왔으며 전체 거리의 평균값은 12.880858으로 나왔다.

	보통	슬픔	기쁨	화남
보통	0	1.48738	25.106256	19.693692
슬픔	1.48738	0	23.679213	5.623253
기쁨	25.106256	23.679213	0	1.695354
화남	19.693692	5.623253	1.695354	0

표 2. U/V Detection+Smoothing과정의 분리도

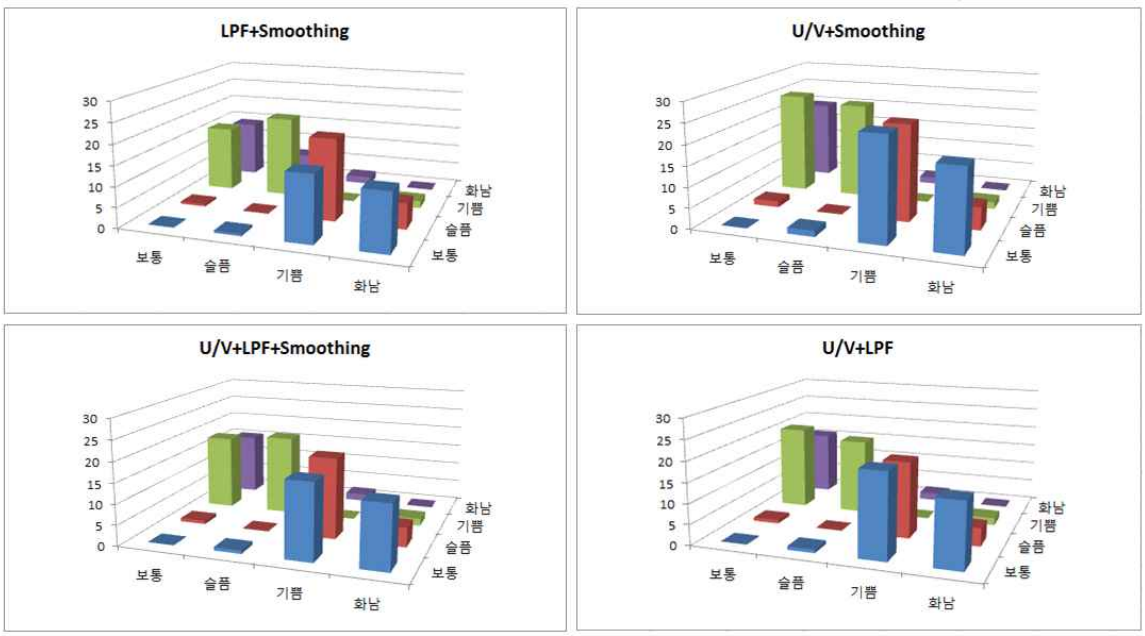

그림 3. 실험결과 오른쪽 위에서부터 (1) LPF + Smoothing (2) Unvoice/Voice Detection + Smoothing (3) Unvoice/Voice Detection + LPF (4) Unvoice/Voice Detection + LPF + Smoothing

유성음/무성음 구간 검출과 Low Pass Filter 과정을 거친 결과는 보통 기쁨 간의 거리 값이 20.593893으로 가장 크게 나왔으며 전체 거리의 평균 값은 10.10111667으로 나왔다.

	보통	슬픔	기쁨	화남
보통	0	0.867451	20.593893	15.774572
슬픔	0.867451	0	18.59258	4.465942
기쁨	20.593893	18.59258	0	1.814512
화남	15.774572	4.465942	1.814512	0

표 1. U/V Detection+LPF과정의 분리도

유성음/무성음 구간 검출과 Low Pass Filter, Smoothing 과정을 거친 결과는 슬픔 기쁨 간의 거리 값이 19.594427으로 가장 크게 나왔으며 전체 거리의 평균 값은 10.35149167으로 나왔다.

	보통	슬픔	기쁨	화남
보통	0	0.846748	18.344148	15.30574
슬픔	0.846748	0	19.594427	4.725327
기쁨	18.344148	19.594427	0	1.79031
화남	15.30574	4.725327	1.79031	0

표 1. U/V Detection+LPF+Smoothing 과정의 분리도

5. 결론

본 논문은 음성신호기반의 감정분류를 위한 피치 검출기에 대한 연구이다. 이를 위하여 음성신호에서 음성구간과 비 음성구간을 추출하였으며 피치검출에서의 전 처리 후처리 과정들에 대하여 설명하고 이를 사용하여 피치 검출에 대한 실험을 하였다. 그 결과 Low Pass Filter를 사용한 경우 각 클래스 간의 거리의 평균값이 약 10 정도로 비교 적 낮게 나왔고, 유성음/무성음 구간 검출 과정과 Smooting 과정을 거친 방법이 12.880858으로 크게 나왔다.

향후, 본 논문에서 제안한 피치 검출기를 이용하여 음성인식, 감정인식, 멜로디와 허밍간의 톤 매칭 기술에 특징으로 검출하여 사용할 수 있으며, 음성신호 기반의 감정 분류 시스템에 대한 연구도 필요하다고 판단된다.

감사의 글

본 연구는 산업통상자원부의 산업기술혁신사업의 일환으로 수행하였음. [10050499, 중국 국가전략산업 연계 상용화 R&D 지원]

참고문헌

[1] 변성우, 이소민, 이석필 "홍확률변수를 이용한 음악에 따른 감정분석에의 최적 EEG 채널 선택" 대한전기학회 논문지, 제62권, 제11호, pp.1598-1603, 2013

[2] 이소민, 변성우, 이석필 "음악에 따른 감정분류을 위한 EEG특징벡터 비교" 대한전기학회 논문지, 제63권, 제5호, pp.696-702, 2014

[3] 오지수, 강정진, 임명재, 이기영 "생체신호를 이용한 감정인식시스템의 설계 및 구현" 한국인터넷방송통신학회 논문지, 제 10권 제 1호 pp. 58-62, 2010

[4] 방재훈, 이승룡 "감성기반 서비스를 위한 통화 음성 감정인식 기법" 정보과학회 논문지 제 41권 제 3호 pp.208-213, 2014

[5] 정병욱, 천성표, 김연태, 김성신 "음성신호를 이용한 감정인식" 한국지능시스템학회 논문지 제 18권 제 4호 pp.494-500, 2008