한국정밀공학회지 제 23 권 제 5 호 (2006 년 5 월)
Journal of the Korean Society of Precision Engineering, Vol. 23, No. 5, May 2006

음성의 특정 주파수 범위를 이용한 잡음환경에서의 감정인식

김은호*, 현경학*, 곽윤근#

Noise Robust Emotion Recognition Feature : Frequency Range of Meaningful Signal

Eun Ho Kim*, Kyung Hak Hyun* and Yoon Keun Kwak#

ABSTRACT

The ability to recognize human emotion is one of the hallmarks of human-robot interaction. Hence this paper describes the realization of emotion recognition. For emotion recognition from voice, we propose a new feature called frequency range of meaningful signal. With this feature, we reached average recognition rate of 76% in speaker-dependent. From the experimental results, we confirm the usefulness of the proposed feature. We also define the noise environment and conduct the noise-environment test. In contrast to other features, the proposed feature is robust in a noise-environment.

Key Words : Emotion recognition (감정인식), Voice (음성), Noise-environment (잡음환경), Noise robust feature (잡음에 강인한 특징), HRI (인간-로봇 상호작용)

기호설명

Δ = difference of two continuous time sequence dynamic feature values

1. 서론

지난 20 여 년간 로봇의 발전은 사람의 명령을 잘 따르는 로봇을 만드는 쪽으로 발전되어 왔다. 하지만 2000 년대에 들어와서는 소니의 AIBO[1], 혼다의 ASIMO[2], MIT 의 KISMET[3] 등 인간친화형 로봇에 산업계 및 학계에서 많은 관심을 보이기 시작했다. 이런 인간친화형 로봇의 특징은 로봇과 사람 사이에 모종의 인터렉션이 존재한다는 점이다. 다시 말하면 과거의 로봇은 사람으로부터 명령을 받고 이를 수행하는 일방적인 대화만을 고려하였으나 앞으로는 사람의 명령뿐만 아니라, 감정이나 건강 정보 등을 인식하여 명령을 수행하고 사람의 반응을 보며 자신의 행동을 수정할 수 있는 인간-로봇 상호작용이 가능한 로봇이 대두될 것이다.

본 연구는 이러한 사람과 로봇의 상호작용의 관점에서 인간친화형 로봇의 제작에서 빠질 수 없는 기술인 음성을 통한 감정인식에 관한 것이다.

☞ 접수일: 2005 년 11 월 1 일; 게재승인일: 2006 년 4 월 14 일
* KAIST 기계공학과
# 교신저자: KAIST 대학교 기계공학과
E-mail ykkwak@kaist.ac.kr Tel. (042) 869-3212

인간의 감정 정보는 심장 박동 수, 혈압, 체온, 얼굴 표정, 음성 등에서 다양하게 얻을 수 있지만 적용 분야에 따라 제한을 받을 수 있다. 특히 음성은 화자의 위치와 거리에 자유로우며 전화와 같은 시스템에서 사용될 때 유리하다. 따라서 본 연구는 인간의 감정인식 방법(modality)중 음성을 이용한 감정인식에 관한 것이다.

사람의 감정이란 주관적인 요소가 다분하고, 실제 사람도 모르는(혹은 친하지 않는) 사람의 감정을 알아내기란 쉽지 않으므로, 본 연구에서도 한 화자에 대하여 특별히 훈련된 화자 종속 시스템을 구성 하였다. 또한, 사람이 발화한 문장의 내용과는 무관하게 사람의 감정은 표현되므로 문맥적 내용을 고려하지 않은 문장 독립 형 시스템을 다루고자 한다.

논문의 본론에서는 사람과 로봇 사이의 인터페이스를 구성함에 있어서 사람에게 가장 기본적인 대화 채널인 음성에 초점을 맞추어 감정과 연관이 있다고 생각되는 새로운 특징, 음성의 특정 주파수 범위(Frequency Range of Meaningful Signal, FRMS)을 제안하였다. 그리고 FRMS 에 대한 설명과 함께 FRMS 를 이용한 음성으로부터 감정인식 실험 결과를 통해 FRMS 의 효용성을 보였다.

마지막으로 논문의 결론에서는 본문에서 언급한 FRMS 에 대하여 정리하고 향후 연구가 필요한 내용에 대해 언급하였다. 따라서 본 연구는 음성을 통한 사람의 감정인식에 관한 것으로 음성으로부터 사람의 감정을 추출할 수 있는 새로운 특징에 관한 것이다.

2. 관련 연구

2.1 감정과 연관이 있는 음성 특징

감정과 연관이 있는 음성 특징으로는 크게 운율적(prosody)특징과 음운론적(phonetic)특징으로 구분된다. 운율적인 특징이란 문장의 문맥적인 내용과는 상관없이 들어나게 되는 특징으로 음의 높낮이, 크기, 장단 등이 여기에 해당된다. 이미 음성에 관한 많은 연구에서 음성의 운율적인 특징이 사람의 감정과 관련이 있다는 것을 밝히고 있다.[4] 특히 Xiao Lin[5] 의 연구에서는 피치와 감정의 관계를 이용하여 감정 인식을 시도하였으며, V. Kostov[6]의 논문에서는 피치, 에너지, 템포를 이용하여 문장 독립적인 감정 인식 시스템에 적용한 예를 보이고 있다. 또한 2004 년에 Dimitrios Ververidis[7] 의 연구에서는 주파수, 피치, 에너지 동적 특징에 대하여 세부적으로 87 개의 정적 특징을 정의하여 각 정적 특징의 성능을 평가한 결과도 발표되었다. 음운론적인 특징이란 운율적 특징과 대비되는 것으로 성도의 모델링에 사용되는 선형 예측 계수(Linear Prediction Coefficient, LPC)가 이에 해당된다. 이외에도 멜 켑스트럼 계수(Mel-Frequency Cepstrum Coefficient) 등의 특징들이 존재 한다.

2.2 기존 연구

음성으로부터 감정인식에 관한 연구는 대부분 화자 종속방향으로 약 70~95%정도의 인식률을 가지고 진행되어 왔다. 화자 독립에 관한 연구는 2000 년 S. McGilloway[8] 에 의한 ASSESS 시스템으로 55%의 인식률을 가지고 있다. 이와 같이 화자 독립 시스템의 인식률이 급격히 떨어지는 것은 사람조차 모르는 사람의 감정인식에 있어서 60%의 낮은 인식률을 가지는 것을 보면 당연한 일이다.[9] 화자종속 시스템에 대한 연구는 1999 년 SpeakSoflty[10] 시스템이 신경 회로망(neural network)을 기반으로 5 가지 감정에 대해서 70%의 인식률을 보였으며, 2005 년 MEXI[11] 시스템은 퍼지규칙(fuzzy rule)을 기반으로 5 가지 감정에 대해서 화자종속 84% 화자독립 60%의 인식률을 얻었다.

3. 음성의 특정 주파수 범위

3.1 FRMS 의 의미

사람이 발화 시 긴 주파수 영역을 가진다. 그러나 실제로 음성으로의 의미가 있는 주파수 영역은 100~5000Hz[12] 로 일부분이며 그 영역은 사람의 감정의 따라 Fig. 1 과 같이 달라지게 된다.

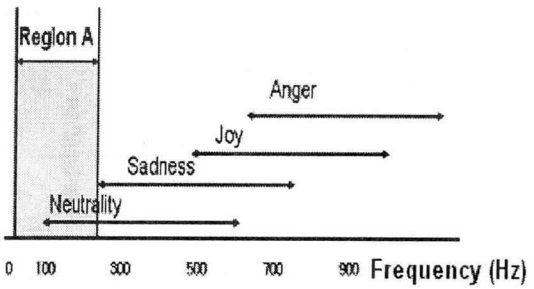

Fig. 1 Frequency range of meaningful signal

Table 1 Mean and standard deviation of the envelope and energy for each emotion

	Envelope		Energy (dB)	
	Mean	Std.	Mean	Std.
Neutrality	0.8496	0.0236	0.0089	0.0060
Joy	0.5991	0.0708	0.0200	0.0146
Sadness	0.6762	0.077	0.0065	0.0049
Anger	0.5252	0.0817	0.0189	0.0095

이것을 특징으로서 이용하게 되면 음성을 통해 사람의 감정을 인식할 수 있게 된다. 이것이 FRMS 라 부르는 새로운 특징이 된다.

Fig. 2 와 3 은 감정의 실제 신호와 160Hz 의 차단 주파수(cut-off frequency)를 가지는 저주파 필터를 통과한 후의 신호의 모양을 나타내고 있다. (여기서 160Hz 는 Fig. 1 에서 A 영역의 주파수) 저주파 필터를 통과한 후 4 개의 감정 모두 그 크기는 감소하였지만 평상 감정만 그 포락(envelope)이 살아 있는 것을 볼 수 있다. 이것은 평상 감정만이 차단 주파수 이하에 의미 있는 음성 신호가 남아 있는 것을 의미 한다. 우리는 차단 주파수를 바꿔가면서 실제 신호와 저주파 필터를 통과한 후의 신호의 포락 비교를 통해 감정에 따라서 의미 있는 주파수 대역이 다름을 알 수 있었다. 위와 같이 차단 주파수를 옮겨 가며 필터를 사용하게 되면 음성으로부터 인간의 감정 상태를 구분할 수 있게 된다.

여기서 의미 있는 신호라 함은 그 신호의 크기와 무관하게 Figs. 2 와 3 에서 보는 것과 같이 실제 음성 신호의 포락이 살아 있는 것을 의미 한다. 여기서 주목해야 할 점은 저주파 필터 후에도 에너지가 아닌 포락이 살아 있다는 점이다. Figs. 2 와 3 에서 얼핏 보면 평상이 저주파 영역에서 다른 감정 상태의 신호들보다 많은 에너지를 가지고 있는 것처럼 보인다. 하지만 tow sample Z test 를 이용하여 Table 1 로부터 우리는 에너지의 P-값을 0.0892, 포락의 P-값을 대략 0 으로 얻을 수 있다. 위 결과를 통해 우리는 0.05 유의 수준에서 '평상의 에너지가 슬픔과 같다'는 귀무가설이 받아 들여지는 반면에 '평상의 포락이 슬픔과 같다'는 귀무가설은 받아 들여지지 않음을 알 수 있다. 이것은 앞에서도 언급했듯이 저주파 필터 후에 각 감정들의 신호가 에너지는 비슷하나 포락의 남아있는 정도가 다름을 나타낸다. 즉 FRMS 특징이 에너지 특징과 다름을 나타낸다. 위 실험에서 남아 있는 포락의 정도를 나타내기 위해서 다음 장에 설명될 저주파 필터 전후의 상관관계(correlation)값을 이용하였다.

3.2 FRMS 추출 방법

FRMS 를 이용하기 위해서 우리는 저주파 필터 후 신호에 남아있는 의미 있는 신호의 양의 측정해야 한다. Figs. 2 와 3 에서 본 것과 같이 의미 있는 신호는 신호의 포락으로 나타낼 수 있다. 따라서 의미 있는 신호의 양은 실제 신호와 저주파 필터 후 신호 사이 포락의 상관관계 값을 통해서 얻을 수 있다.

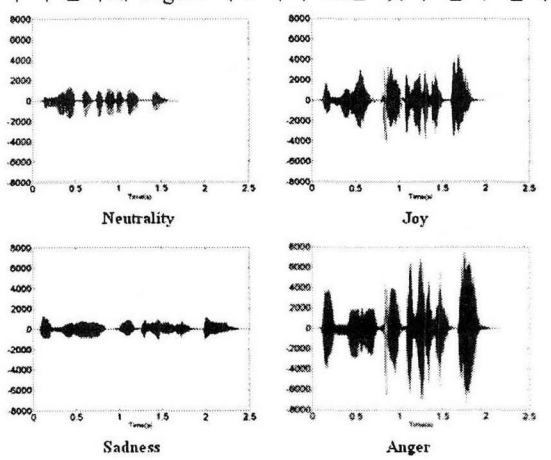

Fig. 2 Original speech signal for four emotions

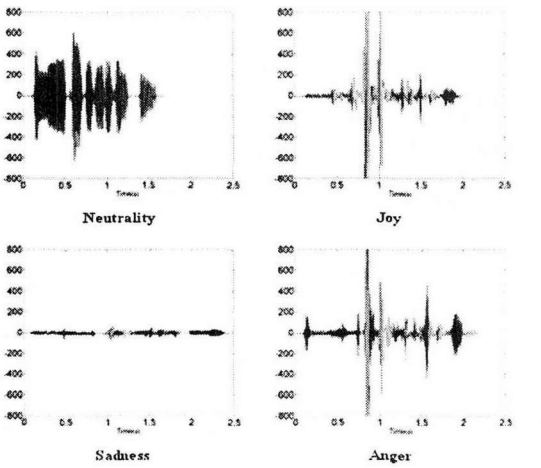

Fig. 3 Low-pass filtered speech signal for four emotions

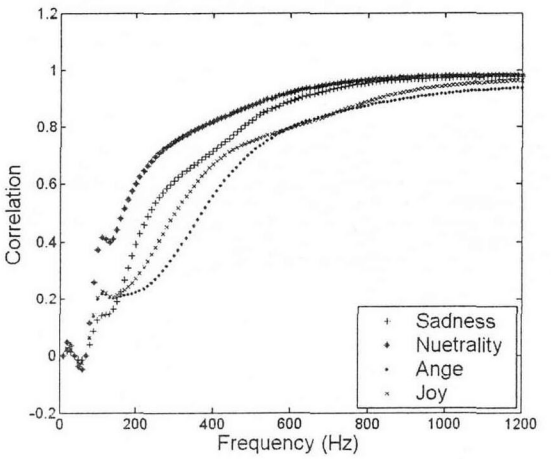

Fig. 4 Relation between frequency and correlation

Fig. 4 는 첫 번째 여자 화자의 135 개 문장에 대한 평균 값으로 각 감정 별 차단 주파수와 상관관계 값 사이의 관계를 나타낸다. Fig. 4 에서 각 감정 별로 그 특성이 뚜렷이 구분되는 것을 알 수 있다. 이것은 각 감정 별로 차단주파수에 따라 남아있는 의미 있는 신호의 양이 달라 상관관계 값이 다르고 상관관계 값이 급격히 증가하는 점이 다르기 때문이다. 또한 Fig. 5 에서, 모든 감정이 약 100Hz 부근에서 상관관계 값이 0 에서 급격이 증가하는 것을 볼 수 있다. 이것은 앞에서 언급한 의미 있는 신호의 영역 100~5000Hz 의 시작하는 부분과 일치하게 된다. 따라서, 우리는 저주파 필터 전후 신호 포락의 상관관계 값을 통해서 FRMS 를 측정할 수 있다

4. 데이터 베이스

4.1 음성 감정 데이터 베이스

음성 데이터 베이스를 구성하기 위해서는 데이터 베이스의 용도를 고려하여 감정 선정, 문장 선정, 녹음대상 선정, 녹음환경, 데이터 베이스 규모 등의 결정 작업이 필요하다.

실험에 사용된 음성 데이터 베이스는 G7 과제의 일환으로 제작된 것으로 다음과 같다. 대상 감정은 인간의 주요 감정인 기쁨, 슬픔, 분노의 3 가지 감정과 이들의 기준이 되는 평상 감정을 포함한 4 가지 감정으로 결정 하였다. 데이터 베이스

Table 2 Human performance of emotion recognition for the database

Recog. \ Human performance (%)	Neutrality	Joy	Sadness	Anger
Neutrality	83.9	3.1	8.9	4.1
Joy	26.6	57.8	3.5	12.0
Sadness	6.4	0.6	92.2	0.8
Anger	15.1	5.4	1.0	78.5
Overall	78.1			

제작시 화자독립-문장독립형 감정인식 시스템개발 및 테스트 용도의 데이터 베이스를 제작한다는 목적 하에 다음과 같은 기준을 고려하였다.[13]

1) 3 가지 감정(기쁨, 슬픔, 분노) 상태로 발음하기 용이한 문장

2) 자연스런 감정 표현이 담긴 대화체 문장

3) 전체적으로 한국어의 모든 음소를 고루 포함하도록 구성

4) 법, 높임 형 등 다양한 어법을 고려하도록 구성

제작된 데이터 베이스는 4 가지 감정 상태(기쁨, 슬픔, 분노, 평상)에 대해서 각각 3 회씩 45 문장을 이용하였으며 남녀 5 명씩 총 5400 개의 문장을 포함하고 있다.

4.2 주관적 평가

구축된 데이터 베이스가 화자의 감정을 어느 정도로 정확히 반영하는지를 판단하기 위해서 각 감정에 대한 주관적 평가를 실시 하였다. 감정 별 인식률 및 오 인식 패턴을 알 수 있는 주관적 평가에 대한 결과는 Table 2 와 같다.[13]

Table 2 의 결과를 보면 각 감정의 인식률이 균일하지 않음을 보여준다. 슬픔의 인식률이 92.2%으로 가장 높았고 기쁨의 인식률이 57.8%으로 가장 낮았으며 그 차이가 약 35%으로 크게 나타났다. 또한 오 인식 패턴에 있어서 모든 감정들이 평상 감정으로 오 인식하는 경우가 많을 것을 볼 수 있다. 이와 같은 결과는 평상시 감정 상태에 대한 주관적 평가자 들의 정신적인 기준치가 다르기 때문에 나타났고 볼 수 있다.

특히 주관적 평가에서 주목할 점은 다음과 같

다. 기존 음성을 통한 감정인식 결과에서 주로 나타난 쾌(valance)축에 대한 오 인식이 주관적 평가에는 나타나지 않는다는 것이다. 즉 분노와 기쁨 사이의 오 인식과 평상과 슬픔 사이에 오 인식이 주관적 평가에 현저히 나타나지 않는다는 것이다. 이것은 현재 음성을 통한 감정인식에서는 쾌 축에 관한 특징이 없는 반면 인간은 쾌 축에 대한 특징을 가지고 있는 것을 나타낸다.

5. 실험 결과

5.1 실험 방법

화자종속 시스템에 대한 실험을 위해 각각의 화자에 대해서 30 개의 문장(10 개 대화체에 3 번씩 반복), 약 20%의 데이터베이스를 이용해 학습 시켰으며 나머지 105 문장에 대해서 실험 하였다.

음성으로부터 FRMS 의 추출은 10~1200Hz 사이의 매 10Hz 총 120 개의 차단 주파수를 이용하여 저주파 필터 전후 신호 포락의 상관관계 값을 구하는 방법을 이용하였다. 신호 처리 시간을 줄이기 위해 120 개 차단주파수에 대한 상관관계 값을 모두 구하지 않고 주요 16 개의 차단주파수에 대한 상관관계 값을 구한 뒤 3 차 보간법(cubic interpolation)을 이용하여 120 개에 대한 값을 구하였다. 또한 120 개 차원의 특징벡터의 차원을 줄이기 위해서 주성분 분석(Principal Component Analysis)[14] 을 사용하였다. 주성분 분석을 통해 특징 벡터의 120 차원을 약 6~10 차원으로 감소 시켰다. FRMS 를 이용하여 음성으로부터 감정 인식을 위한 인식기로는 베이스 식별기(정규 분포를 가정함)를 사용하였다.

5.2 실험 결과

음성을 통한 감정인식은 크게 두 가지로 화자종속 시스템과 화자독립 시스템으로 나눌 수 있다. 화자종속 시스템은 감정을 인식하고자 하는 화자에 대한 정보, 즉 대상 화자의 학습 데이터베이스를 시스템이 알고 있는 경우로 이 시스템은 특정 화자에 특성화되어 있기 때문에 그 화자에 대해서만 올바른 감정 인식을 할 수 있게 된다. 반면 화자독립 시스템의 경우 감정 인식하고자 하는 화자에 대한 정보를 모르는 경우로 이 시스템은 일반 모든 화자에 대해서 감정 인식을 할 수 있기 때문에 감정인식 시스템의 실용화에 꼭 필요하게 된다. 화자종속 시스템에 대한 실험 결과는 Table 3 과 같다. Table 3 은 남자 여자 각각의 5 명의 화자에 대해서 4 개의 감정에 대한 평균 인식률과 오 인식률을 보여주고 있다. 실험 결과 남자의 경우 각 감정 별 인식률 차가 최대 약 4%로 주관적 인식률과 달리 균일하였다. 또한 여자의 경우 주관적 인식률과 같이 슬픔이 가장 좋은 인식률을 보였고 기쁨이 가장 낮은 인식률을 보였다. 그리고 그 차이는 약 23%로 주관적 인식률보다 낮은 것을 알 수 있다. 즉 FRMS 를 이용한 감정인식이 주관적 감정인식보다 감정 별 편차가 작은 것을 알 수 있다. 또한 전체 평균 인식률은 76%(남자 79%, 여자 72%)로 주관적 인식률 78%에 거의 근접한 결과를 얻었다.

Table 3 Results of recognition using the FRMS feature

Recog. \	Male (%)			
	Neutrality	Joy	Sadness	Anger
Neutrality	80.2	4.8	10.5	4.6
Joy	1.7	80.0	5.1	13.1
Sadness	8.4	8.0	80.6	3.1
Anger	2.7	20.8	0.2	76.4
Overall	79.3			
Recog. \	Female (%)			
	Neutrality	Joy	Sadness	Anger
Neutrality	69.1	5.0	21.1	4.8
Joy	8.2	61.1	4.8	25.9
Sadness	8.8	5.9	84.4	0.8
Anger	3.4	19.6	3.4	73.5
Overall	72.0			

오 인식 패턴을 보면 주관적 인식 결과에서 두드러지게 나타났던 모든 감정을 평상 감정으로 판단하는 경향은 보이지 않았다. 그 이유는 사람들의 평상시 감정 상태에 대한 정신적인 기준치가 다른 반면 구성된 시스템은 모든 감정 상태에 대한 기준치가 학습을 통하여 정립됐기 때문이다. 이와 같은 측면은 로봇을 이용한 감정 인식이 사람의 주관적 감정 인식보다 좋은 장점으로 생각할 수 있다. 그러나 앞에서 언급했듯 기존 연구결과와 같이 FRMS 를 이용한 오 인식 패턴을 보면 쾌 축에 대한 오 인식이 두드러지는 것을 볼 수 있다. 이것은 FRMS 특징이 쾌보다 기존 특징들과

Table 4 Comparison of the FRMS feature with energy, lpc and pitch

	FRMS (%)	Energy (%)	LPC (%)	Pitch (%)
Neutrality	74.7	65.8	80.5	65.0
Joy	70.6	67.5	48.1	61.7
Sadness	82.5	82.5	45.8	64.6
Anger	75.0	83.4	81.9	61.1
Overall	75.7	74.8	64.1	70.0

같이 각성(arousal)축에 가까운 특징이라는 것을 나타낸다.

6. FRMS 의 특성

6.1 다른 특징과의 비교

음성을 통한 감정인식에 관한 기존 연구 결과는 약 70~95%다. 그러나 각 연구마다 데이터베이스가 다르고 감정의 수 등 실험 환경이 다르기 때문에 단순히 인식률을 가지고 연구 결과를 평가하고 어느 특징이 좋은지를 판단하는 것은 무리이다. 따라서 본 연구에서는 기존 연구에서 가장 많이 사용되는 특징인 에너지, 피치, LPC 에 대해 위 실험과 같은 환경아래 감정인식 실험을 통하여 FRMS 를 비교 해보고자 한다.

$$E_n = \sum_{m=n-N+1}^{n} x^2(m) \quad w(m)=1 \;\; 0 \le n \le N-1$$
$$otherwise \;\; 0 \qquad (1)$$

에너지 특징은 식 (1)과 같이 얻을 수 있다. 에너지의 동적 특징으로부터 9 차원을 가지는 (평균, 최고 값, 표준편차, △평균, △최고 값, △표준편차, △△평균, △△최고 값, △△표준편차) 정적 특징을 얻을 수 있다. 음성의 피치는 자기상관함수 방법(Autocorrelation function method)과 선형예측계수 분석을 동시에 사용한 Simple inverse filtering tracking 방법을 사용하였다.[15] 피치 역시 에너지와 같이 동적인 특징으로부터 9 차원을 가지는 정적인 특징을 얻을 수 있다. 음성으로부터 LPC 는 자기상관(autocorrelation)법과 공분산(covariance)법 중에 자기상관법을 이용하여 음성신호의 20ms 구간을 한 프레임으로 잡고 10ms 씩 중첩하여 12 차수를 구하였다. LPC 역시 동적인 특징으로부터 9 차원을 가지는 정적인 특징을 얻어 사용하였다.

Table 4 는 같은 환경 아래서 에너지, 피치, LPC 의 화자종속 시스템의 대한 실험 결과와 FRMS 의 실험 결과이다. 실험 결과는 남자 여자 10 명에 대한 각 감정 별 평균 인식률을 나타내고 있다. 실험 결과 FRMS 가 가장 높은 인식률을 보이는 것을 알 수 있다. 뿐만 아니라 각 감정 별 인식률 편차에서도 FRMS 가 가장 작아 감정인식 특징으로서 좋은 특징임을 알 수 있다.

6.2 잡음환경에 대한 고찰

지금까지 음성을 통한 감정 인식에 관한 연구는 감정에 관련 있는 음성 특징을 찾는 것과 적절한 인식기 설계를 통하여 감정인식률 향상에 초점이 맞춰졌다. 그러나 실제 음성을 통한 감정인식 환경은 이제까지 연구해온 데이터 베이스상의 환경과 매우 다르다. 즉, 실제 환경에는 데이터 베이스상에 존재하지 않은 문제가 생기게 된다. 실제 환경에서는 음성잡음, 백색잡음과 같은 잡음이 존재 하게 된다. 또한 마이크와 화자 사이에 거리와 마이크의 설정 등에 따른 음성의 크기가 변화게 된다. 따라서 감정인식에 있어서 이와 같은 실제 환경에 대한 영향을 고려해주어야 한다. 음성인식 분야의 연구가 음성인식과 관련 있는 음성 특징을 찾는 연구로 시작하여 실제 환경에 강인한 인식기를 설계하는 연구로 진행 됐던 것처럼 음성을 통한 감정인식도 실제 환경에 강인한 감정 인식기에 관한 연구가 진행되어야 한다.

따라서 본 논문에서는 실제 환경의 음성잡음, 백색잡음, 마이크와 화자 사이의 거리와 마이크의 설정 등에 따른 음성의 크기 변화를 통틀어 잡음환경이라 정의한다. 그리고 이와 같은 잡음환경아래서 기존 특징들(에너지, 피치, LPC)과 FRMS 의 성능평가를 수행 하였다.

6.2.1 잡음환경에서 음성 감정 데이터베이스

음성 감정 데이터 베이스를 통하여 잡음환경(음성잡음, 백색잡음, 두 배 크기잡음, 반 배 크기잡음)은 다음과 같은 방법으로 생성하였다.

음성잡음(voice noise)은 원래 음성신호의 크기의 1/4~1/2 정도의 크기를 가지는 두 명의 다른 사람

의 다른 감정의 신호를 원래 음성신호에 중첩하여 생성하였다. 백색잡음(white noise)은 원래 음성신호의 크기의 1/4~1/2 정도의 백색잡음을 원래 음성신호에 중첩하여 생성하였다. 또한 원래 음성신호의 크기를 두 배 하여 두 배 크기 음성잡음(double noise)을 생성하였고 원래 음성신호의 크기를 반 배 하여 반 배 크기 음성잡음(half noise)을 생성하였다.

6.2.2 잡음환경 성능 평가

잡음환경에 대한 기존 특징들과 FRMS 의 성능 평가는 무 잡음환경에서 학습시킨 시스템으로 앞에서 구성한 4 가지 잡음환경에 대한 감정인식 실험을 수행하였다. 잡음환경에 대한 성능 평가 결과는 Fig. 5 와 같다. Fig. 5 에서 각각의 색은 4 개의 감정을 나타내며 'O', 'V', 'W', 'D', 'H'는 각각 무 잡음, 음성잡음, 백색잡음, 두 배 크기잡음, 반 배 크기 잡음환경을 나타낸다.

FRMS 의 인식률을 보면 4 가지 잡음환경에 대해서 무 잡음환경과 잡음환경의 인식률의 차이가 거의 없는 것을 볼 수 있다. 이것은 FRMS 가 잡음환경에서 강인한 특성을 가지는 것을 나타낸다. FRMS 가 잡음환경에서 강인한 특성을 가지는 반면, 에너지에 대한 잡음환경 성능평가 결과를 보면 모든 잡음환경 아래서 전체적 인식률뿐만 아니라 개별인식률 모두 급격히 감소하는 것을 알 수 있다. 특히 음성잡음 백색잡음, 두 배 크기잡음 환경에서는 분노, 기쁨처럼 높은 에너지를 가지는 감정으로 오 인식을 많이 하였으며 반 배 크기잡음환경 아래서는 슬픔, 평상 같은 낮은 에너지를 가지는 감정으로 오 인식을 많이 하였다. 이것은 에너지 특징이 신호의 세기와 관계가 있기 때문이다.

피치에 대한 잡음환경 성능평가를 보면 음성잡음을 제외하고 나머지 세 잡음환경에서 비교적 좋은 성능을 보이는 것을 볼 수 있다. 그러나 음성잡음환경에서 전체적 인식률뿐만 아니라 개별 인식률 모두 크게 감소하여 피치 역시 음성잡음환경에 민감한 것을 알 수 있다. 이 결과는 음성 인식 분야에서 일반적으로 알려진 피치의 잡음에 대한 특징과 비슷하다.

LPC 에 대한 잡음환경 성능평가를 보면 전체적인 인식률에서 대체적으로 잡음환경에서 인식률이 감소하는 것을 볼 수 있다. 특히 음성 크기 잡음환경에 대해서 큰 인식률 감소를 보이고 있다.

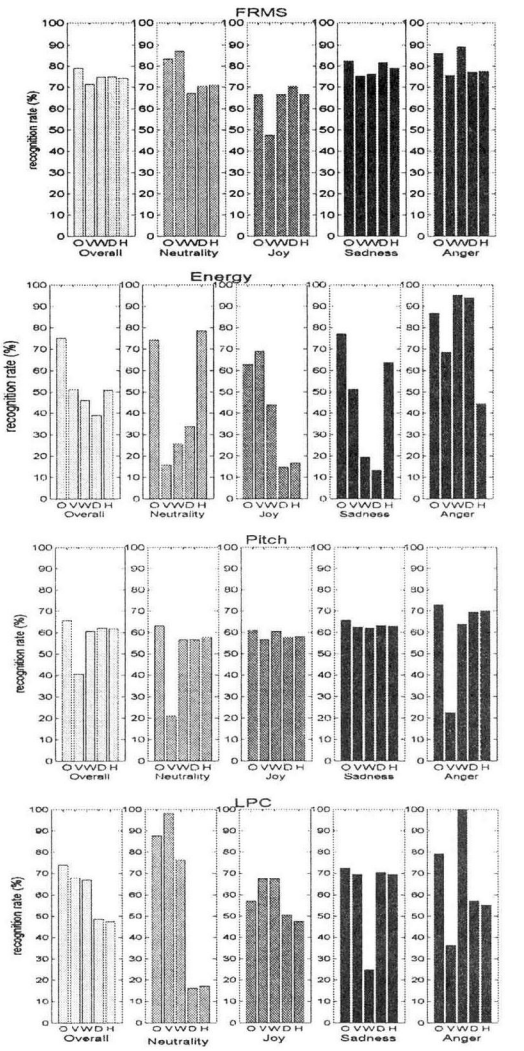

Fig. 5 Experiment results of the noise environment, and a comparison of the FRMS features with energy, pitch and LPC

전체적 인식률뿐만 아니라 개별 인식률에서도 잡음환경에 대해서 좋지 않은 결과를 보여주고 있다. 이와 같은 결과는 잡음환경이 성도 모델링에 영향을 미쳐 LPC 값이 변하기 때문이다. 이와 같이 LPC 특징 또한 잡음환경에 취약한 점을 나타내고 있다.

6.3 FRMS 의 장점

FRMS 의 가장 큰 장점은 위에 잡음환경에 대한 성능평가 결과에서 볼 수 있듯이 기존 특징들과 달리 화자와 마이크 사이의 거리 또는 화자의 발화 크기에 독립적이라는 것이다. 실제 감정 인식에서 화자와 마이크의 거리를 보정해 주는 것은 중요한 문제점 중 하나이다. 예를 들어, 에너지의 경우 위 실험결과에서 보듯이 화자와 마이크 사이의 거리 또는 화자의 발화 크기에 매우 민감하게 반응하게 된다. 따라서 에너지를 사용하기 위해서 화자와 마이크 사이의 거리의 보정이 불가피하게 된다. 그러나 FRMS 의 경우 에너지의 크기와 상관없이 그 포락을 보기 때문에 다시 말해 FRMS 의 경우 원래 음성신호와 저주파로 필터 된 신호 사이에 상관관계에만 관련되기 때문에 거리와 크기에 강인하여 실제 사용에 유용하게 된다. 이것은 FRMS 가 거리와 크기에 무관하며 거리와 크기에 관한 전처리 없이 사용가능 하다는 것을 나타낸다.

FRMS 의 또 다른 장점으로는 위 잡음환경 성능평가에서 볼 수 있듯이 음성잡음과 백색잡음환경에 강인하다는 점이다. FRMS 가 위와 같은 잡음환경에 강인한 이유는 대부분의 잡음환경은 차단주파수 이상의 고주파일 경우가 많다. 그러므로 대부분의 고주파의 잡음은 저주파 필터 후 사라지게 된다. 뿐만 아니라 만약 잡음의 크기가 인식하고자 하는 음성의 크기보다 약하게 될 경우 이러한 잡음은 대부분 음성신호의 포락에 영향을 주지 않게 된다. 이와 같은 이유로 FRMS 가 음성잡음과 백색잡음환경아래서 강인한 특징을 가지게 된다. 이러한 특징은 잡음제거를 위한 전처리 과정을 불 필요하게 해 잡음제거 과정의 어려움을 해결할 수 있게 하는 장점을 가지고 있다.

7. 결론

본 연구에서 우리는 음성을 통한 감정인식 분야에서 FRMS 라는 새로운 특징을 제안하였다. 그리고 화자종속 시스템에 대한 실험을 통해 제안한 특징의 효용성을 확인 하였다. 또한 우리는 실제 감정인식 환경과 데이터 베이스 환경상의 다른 4 가지 상황에 대해서 잡음환경을 정의하였다. 그리고 그 잡음환경에 대한 여러 특징들의 잡음환경 평가를 수행하여 Fig. 5 와 같은 결과를 얻었다. 잡음평가를 통해서 우리는 FRMS 가 잡음환경에 대해서 인식률 감소가 10% 미만인 것을 보여 잡음환경에서 강인한 것을 확인하였다. 이와 같은 FRMS 의 특징은 실제 사용에 있어서 잡음환경에 대한 전처리 없이 쉽게 사용할 수 있는 큰 장점이 된다.

앞으로 우리는 음성을 통한 감정인식 시스템의 실용화를 위해 화자독립 시스템에 대한 연구가 필요하다. 또한 보다 정확하고 보다 많은 감정의 인식을 위해 쾌 축에 관련된 특징을 찾는 연구가 필요하다.

후 기

이 연구(논문)는 산업자원부 지원으로 수행하는 21 세기 프론티어 연구개발사업(인간기능 생활지원 지능로봇 기술개발사업)의 일환으로 수행되었습니다.

참고문헌

1. Fujita, M., "On Activating Human Communications with Pet-type Robot AIBO," Proceeding of the IEEE, Vol. 92, No. 11, pp. 1804-1813, 2004.
2. Takenaka, T., "Honda Debuts New Humanoid Robt ASIMO," The Industrial Robot, Vol. 28, No. 2, pp. 2-2, 2001.
3. Srinivasan, S. and Tamkun, J. W., "Characterization of Kismet, a Trithorax-group Protein," Biochemistry and Cell Biology, Vol. 81, No. 3, pp. 253-560, 2003.
4. Plutchik, R., "Emotions and Life : Perspectives from Psychology, Biology and Evolution," American Psychological Association Press, 2003.
5. Lin, X., Chen, Y., Lin, S. and Lim, C., "Recognition of Emotional State from Spoken Sentences," IEEE Multimedia Signal Processing, pp. 469-473, 1999.
6. Kostov, V. and Fukuda, S., "Emotion in User Interface, Voice Interaction System," System Man and Cybernetics, IEEE Conf., Vol. 2, pp. 798-803, 2000.
7. Dellaert, F., "Recognition Emotion In Speech," ICSLP Proc. 4th International Conf., Vol. 3, pp. 1970-1973, 1996.

8. McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M. and Stroeve, S., "Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark," in ISCA Workshop on Speech and Emotion, Belfast, pp. 207-212, 2000.
9. Scherer, K., "Vocal Communication of Emotion: A Review of Research Paradigms," in Speech Communication, Vol. 40, pp. 227-256, 2003.
10. Petrushin, V. A., "Emotion in Speech: Recognition and Application to Call Centers," Proceeding of the 1999 Conference on Artificial Neural Networks in Engineering, Vol. 9, pp. 1085-1092, 1999.
11. Austermann, A., Esau, N., Kleinjohann, L. and Kleinjohann, B., "Fuzzy Emotion Recognition in Natural Speech Dialogue," Robot and Human Interactive Communication IEEE 14th workshop, pp. 317-322, 2005.
12. Borden, C. J., Harris, K. S. and Raphael, L. J., "Speech Science Primer," 3rd ed., Williams & Wilkins Press, p. 42, 2000.
13. Kang, B. S., "Text Independent Emotion Recognition Using Speech Signals," Yonsei Univ. pp. 35-40, 2000.
14. Kitter, J., "A Method for Determining Class Subspace," Information Processing Letters, Vol. 6, No. 3, pp. 77-79, 1977.
15. Markel, J. D., "The SIFT Algorithm for Fundamental Frequency Estimation," IEEE Trans., Vol. AU-20, No. 5, pp. 367-377, 1972.