1999년도 한국음향학회 학술발표대회 논문집 제18권 제2(s)호

음성 신호를 이용한 화자의 5가지 감성 인식

강봉석*, 한철희**, 우경호*, 양태영*, 이충용*, 윤대희*
*연세대학교 전기·컴퓨터공학과
**연세대학교 신호처리 연구 센터

Recognizing Five Emotional States Using Speech Signals

Bong-Seok Kang*, Chul-Hee Han**, Kyoung-Ho Woo*, Tae-Young Yang*,
Chungyong Lee*, Dae-Hee Youn*
*Department of Electrical Computer Engineering, Yonsei Univ.
**Center for Signal Processing Research, Yonsei Univ.
e-mail : kbs@radar.yonsei.ac.kr

<본 연구는 한국표준과학연구원의 연구비 지원에 의해서 이루어졌습니다.>

요 약

본 논문에서는 음성 신호를 이용해서 화자의 감정을 인식하기 위해 3가지 시스템을 구축하고 이들의 성능을 비교해 보았다. 인식 대상으로 하는 감정은 기쁨, 슬픔, 화남, 두려움, 지루함, 평상시의 감정이고, 각 감정에 대한 감정 음성 데이터베이스를 직접 구축하였다.

피치와 에너지 정보를 감성 인식의 특징으로 이용하였고, 인식 알고리듬은 MLB(Maximum-Likelihood Bayes)분류기, NN(Nearest Neighbor)분류기 및 HMM (Hidden Markov Model)분류기를 이용하였다. 이 중 MLB 분류기와 NN 분류기에서는 특징벡터로 피치와 에너지의 평균과 표준편차, 최대값 등 통계적인 정보를 이용하였고, HMM 분류기에서는 각 프레임에서의 델타 피치와 델타델타 피치, 델타 에너지와 델타델타 에너지 등 시간적 정보를 이용하였다.

실험은 화자종속, 문장독립형 방식으로 하였고, 인식 실험 결과는 MLB를 이용해서 68.9%, NN을 이용해서 66.7%를 얻었고, HMM 분류기를 이용해서 89.30%를 얻었다.

I. 서 론

음성은 인간의 가장 편리한 정보 전달 수단의 하나로써, 최근 음성 인식, 음성 합성을 이용한 휴먼-컴퓨터 인터페이스에 대한 관심이 높아지고 있으며, 일부는 이미 실용화되었다. 감성 인식 기술은 음성 신호 분석을 통해 기쁨, 슬픔, 화남, 두려움 등 화자의 감정 상태를 판별하는 기술로써, 이를 이용해서 음성에 의한 휴먼-컴퓨터 인터페이스의 수준을 한 단계 높일 수 있고, 감정에 반응하는 보다 인간적이고, 고차원적인 컴퓨팅 기술을 구현할 수 있다[1].

감성 인식에 관련된 연구는 현재 미국 MIT대 Media Lab의 *Afective Computing Group*을 중심으로 음성뿐만 아니라 얼굴표정 및 인간의 생체신호를 이용한 감성적 착용 컴퓨터를 구현하기 위한 연구가 활발히 진행되고 있다[1].

감성 인식 시스템을 구현하기 위해서는, 우선 어떤 분야에 적용될 것인지를 고려하고, 해당 분야에서 발생 가능하거나, 필요로 하는 화자의 대상감정을 선정하고, 화자 종속 시스템인지 화자 독립 시스템인지의 여부를 결정해야 한다.

본 논문에서는 PC 환경에서 사이버 캐릭터를 키우는 게임 시나리오를 가정하고, 대상 감정으로는 인간의 주요 감정인 기쁨, 슬픔, 화남, 두려움의 감정과, 게임진행 시에 사용자에게 발생하기 쉬운 지루한 감정, 타 감정과의 비교기준으로 삼을 평상시 감정 등 6가지 감정을 판별하는 시스템을 제작하였다[2].

II. 감정 음성 데이터베이스 구축

음성을 이용한 감성 인식을 수행하기 위해서는 기쁨, 슬픔, 화남, 두려움 등 대표적 감정에 따라 분류된 감정 음성 데이터베이스 구축이 필수적이지만, 현재 한국어로 된 표준 데이터베이스가 없는 실정이다. 이러한 감정 음성용 표준 데이터베이스는 적용 어플리케이션이 미리 고려된 상태에서 그에 맞게 구축되어야 할 것이다.

본 논문에서는 감성 인식 실험을 위해서 시스템 제작에 필요한 데이터베이스를 직접 구축하였는데, 음성 입출력 및 감정 인터페이스의 적용이 용이한 PC 환경에서 사이버 캐릭터를 키우는 게임 시나리오를 가정하고, 이 사이버 캐릭터와 상호 작용하는 PC 사용자를 인식 대상으로 하였다. 대상 감정은 평상시, 기쁨, 슬픔, 화남, 두려움, 지루함의 6가지로 정하였고, 데이터베이스에 포함될 대상 어휘는 표1과 같이 사이버 캐릭터를 키우는데 필요한 5개의 명령어로 구성하였다.

표 1. 인식 대상 단어

index	1	2	3	4	5
command	밥먹어	운동해	청소해	세수해	잠자

데이터베이스를 위한 녹음은 20대의 교내 아나운서 부원 남/여 각 4명을 대상으로 하였다. 발음 횟수는 개인별, 감정별, 문장별로 13회씩으로 하였고, 총 3120개의 데이터를 제작하였다.

녹음은 감정별로 하였고, 한 감정에 대해서 녹음 시에 대상 단어 5가지를 단어별 휴지기를 1초 정도 둔 상태에서 연속해서 발음하도록 요구하였고, 이 과정을 5초 정도의 휴지기를 두면서 13회 반복하였다. 한 감정을 녹음한 후, 다음 감정을 녹음하기 위한 휴지기는 3분으로 하였고, 각 감정에 대한 녹음을 시작하기 전에 보조 영상물의 사용 요청 시 1회 상영하였다.

보조 영상물은 데이터베이스 제작에 참여하는 화자의 기쁨, 슬픔, 화남, 두려움의 4개 감정에 대한 감정몰입을 유도하기 위해 대해서 편집 비디오 테이프를 사용하여 제작하였다. 이 편집 비디오 테이프는 평소에 영화관람을 즐기는 20 명을 대상으로 한 설문조사 결과를 참조하여 선정한 영화 중 대상 감정이 가장 고조된 장면을 감정 당 약 1분 씩 녹화하는 방식으로 제작하였다.

녹음환경은 방음이 잘 되는 대학 방송국 녹음실에서 하였고, 녹음실외부에 설치된 콘솔과 DAT(Digital Audio Tape)를 연결하여 사용하였다.

III. 특징벡터의 추출 방법

3.1 감정 음성의 분석

음성을 통한 감성 인식을 위해서는 각각의 감정이 음성에 어떠한 변화를 만들어 내는가를 정확히 규명하여야 한다. 음성에 포함된 감정은 주로 운율 정보에 의해서 표현되고, 운율 정보에는 피치(F0)의 변화, 에너지의 변화, 발음속도 등이 있다[3]. 감성 인식을 위해서는 음성에서 이러한 운율 정보를 잘 반영하는 특징을 찾아내어 모델링을 해야 된다.

운율정보가 어떻게 감정 음성과 결부되는지를 살펴보면, 통계적인 관점에서는 기쁘거나 화난 음성인 경우에는 전체적으로 에너지와 피치가 높고, 발음 속도가 빠른 반면, 슬프거나 지루한 음성은 전체적으로 에너지와 피치가 낮고 발음 속도가 느리다는 것을 알 수 있다[2][3]. 그림1에서는 음성신호의 일부구간에 대한 에너지를 크기를 나타내었다. 이와 같이 음성 전반에 걸친 피치와 에너지의 높낮이는 발음된 음성 전구간의 피치 평균과 에너지 평균 등의 통계적 정보를 이용하면 모델링이 가능하다. 본 논문에서는 피치와 에너지의 평균, 표준편차, 최대값을 이용하였다.

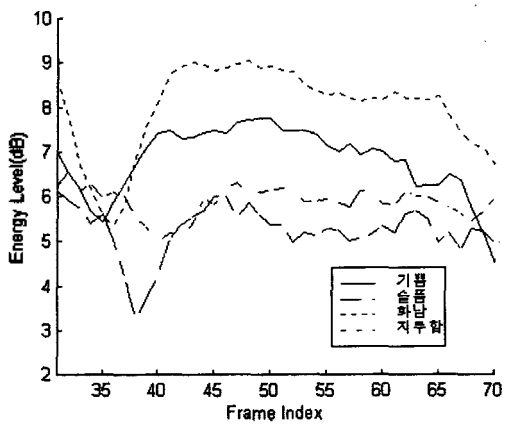

그림 1. '밥먹어'의 '어'에서 에너지의 크기 비교

한편, 시간적인 관점에서 보면, 동일한 문장이라도 감정에 따라 운율적 요소를 변화시켜 다르게 발음하는 것을 알 수 있다. 이것은 감정별로 피치와 에너지의 시간적인 궤적이 다르다는 것을 나타내고, 이를 모델링하기 위해서는 시간의 진행에 따른 피치 변화와 에너지 변화를 이용해야 한다. 모델링을 위한 구체적인 특징벡터로, 델타 피치와 델타델타 피치의 2차원 벡터, 델타 에너지와 델타델타 에너지의 2차원 벡터를 이용하였다. 이와 같은 시간적 정보는, 비슷한 정도의 여기 상태에 있는 감정, 예를 들어 기쁘거나 화난 상태의 감정, 지루하거나 두려운 상태의 감정에 대한 피치와 에너지의 통계적 값이 근접할 경우에 감정 구분을 위한 더 효과적인 특징이 될 수 있다.

3.2 에너지와 피치를 구하는 방법

에너지는 로그 에너지를 구하였고, 델타 에너지와 델타델타 에너지는 구해진 에너지에 대한 프레임간의 차분에 의해서 식 (1)과 같이 구했다.

$$\triangle E_k(t) = E_k(t+\delta) - E_k(t-\delta) \quad (1)$$

피치는 성대가 진동함으로써 발생하게 되는 유성음의 진동 주기인데, 자기상관함수나 AMDF방법을 사용하면 구할 수 있지만, 음성 신호의 비정적 특성, 성대진동의 불규칙성, 잡음 환경하에서의 신호 왜곡 등 때문에 정확하게 구하기는 어렵다. 감성 인식에 있어서 피치값의 정확한 측정은 절대적으로 중요하다. 따라서 현재 음성 부화기에서 우수한 성능을 보이는 EVRC (IS-127)에서 사용한 피치 추출 알고리듬 이용하여 구하였고, 델타 피치와 델타델타 피치는 구해진 피치에 대한 프레임간의 차분에 의해서 식 (1)과 같이 구했다[4].

IV. 인식 알고리듬

감성 인식을 위한 인식 알고리듬은 선택한 특징의 양식에 따라 다르게 선택했는데, 통계적인 특징을 이용할 경우에는 MLB와 NN을 이용했고이고 시간적 변화에 따른 특징을 이용할 경우에는 HMM을 이용하였다.

4.1 MLB(Maximum-Likelihood Bayes) 알고리듬

MLB는 패턴공간내의 어느 점에서 특정 클래스에 속하는 패턴이 출현하는 확률이 알려져 있는 경우, 패턴분포에 관한 정보로 식별함수를 계산해서, 미지패턴이 어느 클래스에 속하는가를 최적으로 결정하는 통계적 결정방법이다. 입력패턴의 확률분포가 알려져 있지 않은 경우에는 보통 가우시안 분포라고 가정한다[5].

4.2 NN(Nearest Neighbor) 알고리듬

NN은 패턴분포에 관한 정보로 식별함수를 계산하는 대신에 미리 저장해 놓은 기준패턴과의 거리를 계산하여 최소거리를 갖는 기준패턴의 클래스를 미지패턴의 클래스로 결정하는 방법이다[6]. 기준패턴을 생성하는 방법으로는 LBG 군집화 알고리듬을 많이 사용한다[7].

4.3 HMM(Hidden Markov Model) 알고리듬

HMM은 음성 신호의 시간적 변화를 Markov 프로세스로 모델링하며, 이중 확률 과정(doubly stochastic process)으로 음성의 단구간 정보와 시간적 변화에 따른 정보를 각각 모델링하는 구조를 갖는다. 따라서, 학습 데이터를 사용하여 모델 파라메터를 추정한 후, 미지의 입력 데이터가 어떤 모델에 해당되는가 하는 것을 확률적으로 판단하는 것이다[8].

V. 인식 실험 결과

5.1 데이터 처리

음성 신호는 16kHz 16비트로 샘플링되어, 20ms(320 샘플)의 길이를 갖는 해밍윈도우를 사용하여 10ms씩 이동하면서 분석하였다. 각 프레임마다 피치와 에너지를 구해서, 전프레임에 대한 피치와 에너지의 평균과 분산 및 최대값을 구해서 통계적인 정보를 표현하는 특징으로 이용하고, 현재 프레임을 중심으로 2프레임 간격으로 델타 피치, 델타 에너지, 델타델타 피치,델타델타 에너지를 구하여 시간적 정보를 나타내는 특징 벡터로 사용하였다.

5.2 주관적 평가 결과

제작된 테스트용 감정 음성 데이터베이스가 화자의 감정을 어느 정도로 정확히 반영하는지를 판단하고, 인식 시스템의 인식 결과와 비교를 위해서 주관적 평가를 실시하였다. 23명의 평가자들에게 13회 발음 중 랜덤하게 각 1개씩 선택해서, 2회씩 들려 주었고, 6감정 중 가장 근접한 1가지 감정을 선택하도록 요구하였다. 총 56%의 주관적 평가 인식률을 얻을 수 있었다. 감정별 인식률은 표2와 같다.

표 2. 주관적 평가 결과

	보통	기쁨	슬픔	화남	두려움	지루함	인식률
보통	65.57	0.61	0.78	28.26	0.43	4.35	65.57
기쁨	28.35	39.04	2.17	23.57	1.83	5.04	39.04
슬픔	18.35	0.61	44.00	4.52	13.74	18.78	44.00
화남	6.17	8.43	1.22	81.83	0.70	1.65	81.83
두려움	18.09	2.70	14.43	5.91	53.04	5.83	53.04
지루함	12.61	0.87	17.30	11.30	5.30	52.61	52.61
종합							56.00

5.3 MLB 분류기를 이용한 인식 실험 결과

통계적 특징에 대한 모든 조합을 이용해서 가장 인식률이 좋은 경우를 찾아내는 특징선별과정을 거쳐, 최종 특징으로 피치의 평균, 표준편차, 최대값, 에너지의 표준편차, 최대값을 사용하였다. 입력패턴의 분포는 가우시안이라고 가정하였으며 인식 결과, 총 인식률은 68.9%였다. 감정별 인식 결과는 표3과 같다.

표 3. MLB 분류기에 의한 인식 결과

	보통	기쁨	슬픔	화남	두려움	지루함	인식률
보통	73.79	2.071	5.52	3.45	0.69	14.48	73.79
기쁨	2.67	61.33	2.00	15.33	6.00	12.67	61.33
슬픔	4.00	2.67	52.67	4.00	12.67	24.00	52.67
화남	2.00	4.67	1.33	89.33	0.67	2.00	89.33
두려움	2.00	4.00	9.33	6.00	70.67	8.00	70.67
지루함	3.73	3.73	11.94	5.22	9.70	65.67	65.67
종합							68.90

5.4 NN 분류기를 이용한 경우

통계적 특징에 대한 모든 조합을 이용해서 가장 인식률이 좋은 경우를 찾아내는 특징선별과정을 거쳐, 최

종 특징으로 피치 평균, 피치 표준편차, 피치 최대값, 에너지 평균, 에너지 최대값을 사용했다. 클러스터 수는 10개로 하였고, 최소거리 계산은 유클리디안 디스턴스를 이용하였다[6]. 총 인식률은 66.7%였고, 감정별 인식 결과는 표4와 같다.

표 4. NN 분류기에 의한 인식 결과

	보통	기쁨	슬픔	화남	두려움	지루함	인식률
보통	72.41	6.21	10.34	2.76	1.38	6.90	72.41
기쁨	3.33	64.00	4.00	18.00	8.67	2.00	64.00
슬픔	4.00	5.33	62.67	2.00	11.33	14.67	62.67
화남	2.00	7.33	2.67	81.33	4.00	2.67	81.33
두려움	2.00	9.33	9.33	8.00	62.67	10.67	62.67
지루함	2.99	3.73	17.91	1.49	17.91	55.97	55.97
종합							66.70

5.5 HMM시스템의 인식 결과

시간적 정보를 나타내는 특징 벡터를 LBG 알고리즘을 사용하여, 에너지의 경우는 64개의 코드를 갖는 코드북을 구하였고, 피치의 경우는 코드북 크기를 256개를 사용하여 인식 실험을 하였다. 이렇게 구해진 코드북은 반연속 HMM의 가우시안 분포의 초기치로 사용되었다.

각 단어의 모델은 8개의 상태 개수를 가지며, HMM의 학습은 Baum-Welch 알고리즘을 사용하였고, 인식에는 Viterbi 알고리즘을 사용하였다[8].

위와 같은 방법으로 인식실험을 수행한 결과 총 인식률은 89.30%였다. 이는 통계적 특징을 이용한 MLB 분류기로 인식실험을 한 경우보다 20.4%의 인식률 향상을 보였고, NN 분류기를 이용한 경우보다 22.7% 인식률의 향상을 보인 결과이다. 감정별 인식률은 표5와 같다.

표 5. HMM 분류기에 의한 인식 결과

	보통	기쁨	슬픔	화남	두려움	지루함	인식률
보통	96.67	1.67	1.67	0.00	0.00	0.00	96.67
기쁨	0.00	85.83	0.83	9.17	2.50	1.67	85.83
슬픔	0.83	3.33	82.50	0.00	3.33	10.00	82.50
화남	0.00	4.17	0.83	92.50	1.67	0.83	92.50
두려움	0.00	1.67	0.83	0.00	93.33	4.17	93.33
지루함	0.00	0.83	11.67	0.00	2.50	85.00	85.00
종합							89.30

IV. 결론

본 논문에서는 음성 신호를 이용해서 화자의 기쁨, 슬픔, 화남, 두려움, 지루함, 평상시 등의 감정을 인식하기 위하여 MLB, NN, HMM 등의 알고리듬을 사용해 인식 시스템을 제작하고, 이들의 성능을 비교하였다. 음성신호에 포함된 운율적 요소 중 피치와 에너지의 통계적 정보를 구해서 MLB와 NN의 특징으로 사용했고, 시간적 변화에 대한 정보를 구해서 HMM의 특징으로 사용하였다. 각 감정에 대한 감정 음성 데이터베이스는 직접 제작하였고, 감정 유발을 위한 보조도구도 필요에 따라 사용하였다.

인식 실험은 화자종속, 문장독립 방식으로 하였고, 인식 실험 결과는 MLB를 이용해서 68.9%, NN을 이용해서 66.7%, HMM 분류기를 이용해서 89.30%를 얻었고, HMM을 이용한 경우 인식률이 가장 우수하였다.

결국 시간적 정보를 이용하는 것이 통계적 정보를 이용했을 경우보다 전체적인 에러가 줄었고, 기쁜 감정과 화난 감정사이의 에러율과 슬픈 감정과 지루한 감정 사이의 에러율이 급격히 감소하였다.

추후 과제로는 화자의 감정이 정확히 반영된 충분한 분량의 음성 데이터베이스 구축하는 것이고, 이를 바탕으로 통계적인 정보와 시간적인 정보를 동시에 고려할 수 있는 시스템을 제작하여야 할 것이다. 또 운율적 정보 중 발음속도를 모델링 하는 특징을 부가적으로 감성 인식에 사용할 수도 있을 것이다.

참 고 문 헌

[1] Rosalind W. Picard, *Affective Computing*, The MIT Press, 1997.

[2] Iain R. Murray and John L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", in *J. Accoust. Soc. Am*, vol. 93, no. 2, pp. 1097-1108, February. 1993.

[3] Janet E. Cahn, *Generating expression in synthesized speech*, Masters thesis, MIT Media Laboratory, May, 1989.

[4] QUALCOMM Inc., Proposed TIA/EIA/PN-3292 Standard - Enhanced Variable Rate Codec, *Speech Service Option 3 for Wideband Spread Spectrum Digital Systems*, Official Ballot Version, April, 1996.

[5] R.O. Duda, and P.E. Hart, *Pattern classification and scene anlaysis*, John Wiley & Sons Inc., 1973.

[6] Earl Gose, Richard Johnsonbaugh, and Steve Jost, *Pattern Recognition and Image Analysis*, Prentice Hall PTR, 1996.

[7] Y. Linde, A. Buzo and R. M. Gray, "An Algorithm for Vector Quantization," *IEEE Trans. Commun.* vol. COM-28, no. 1, pp. 84-95, Jan. 1980.

[8] Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, no. 2, Feb. 1989.