음성신호 기반 감정인식을 위한 특징 파라메터 비교

김남균, 성우경, 하헌규, 김홍국
광주과학기술원 정보통신공학부
{skarbs001, wkseong, hnine, hongkook}@gist.ac.kr

Comparison of feature parameters for speech emotion recognition

Nam Kyun Kim, Woo Kyeong Seong, Hun Ku Ha, and Hong Kook Kim
of Information and Communications, Gwangju Institute of Science and Technology (GIST)

요 약

논문에서는 음성신호로 추출된 파라메터에 따른 감정인식 시스템의 인식성능을 비교 분석한다. 이를 위하여 여러 가지 감정 상태에 따라 분류된 한국어 음성 DB를 이용하여 얻은 음성신호의 특징 파라메터를 추출하고 감정인식기를 구현한다. 본 논문에 사용된 특징 파라메터로는 프레임 기반의 음성특징 파라메터와 발화 기반의 음성특징 파라메터를 이용하였고, 인식기로는 SVM (support vector machine)을 사용하였다. 6가지 감정음성 데이터로 이루어진 DB기반 감정인식 시스템에서 발화기반의 음성특징 파라메터를 사용한 결과, 69.2%의 감정인식률을 보였다.

. 서 론

음성 신호는 사람의 마음을 표현하기 위한 가장 기본적인 신호로, 성별, 나이 등의 여러 개인정보 뿐 아니라 사람의 감정 상태 또한 포함한다. 특히 스마트 폰을 비롯한 다양한 스마트 디바이스에 음성인식 시스템이 내재화됨에 따라, 음성신호 기반의 감정인식 기술은 로봇 등 HCI (human-computer interface)의 기능을 더욱 향상시킬 것으로 예상된다. 이 외에도 사람의 감정을 인식할 때, 영상신호 또는 생체신호를 사용한다. 예를 들어, 영상신호의 경우, 안면 영역 부분 특징 혹은 영상 전 영역을 기반으로 한 특징을 주로 사용해 왔고, 생체신호의 경우, 심전도, 뇌전도 등을 특징으로 사용하여 사람의 감정을 구분하는 척도로 사용한다[1].

본 논문에서는 음성신호에 대해 프레임 기반의 음성특징 파라메터와 발화 기반의 음성특징 파라메터를 이용하여 감정인식 시스템을 구축하고, 각 특징 파라메터들에 대해 인식 성능을 비교 분석한다. 프레임 기반의 음성특징 파라메터로는 MFCCs (mel-frequency cepstral coefficients)를, 발화기반의 음성특징 파라메터로는 LLDs (low-level descriptors)를 추출한 후 각 LLD의 통계치(statistical functional)를 사용하였다. 감정인식기로는 SVM (support vector machine)를 이용하였다.

Ⅱ. 프레임기반 및 발화기반 음성특징을 이용한 감정인식 시스템

본 절에서는 프레임 기반 및 발화기반 음성특징 파라메터를 이용한 감정인식 시스템에 대해 서술한다. 그림 1에서 보는 바와 같이, 입력 음성신호에 음성특징 파라메터를 추출하고 감정모델을 구성한다. 즉, 입력음성에서 다양한 acoustic feature를 추출하는데 가장 대표적인 것이 피치, RMS (root mean square) 에너지, 발화율, 발화 지속시간 등의 발화기반의 음성특징인 prosodic feature와 프레임 기반의 음성특징은 39차 MFCC 등의 spectral feature들이 있다[2]. 이때, 입력 음성은 10ms의 frame rate로 25ms 길이의 프레임으로 분할된다.

발화기반 음성특징 파라메터로는 표 1에 보는 바와 같이 openEAR

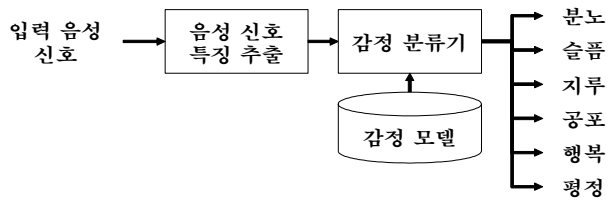

그림 1 음성신호 기반 감정인식 시스템

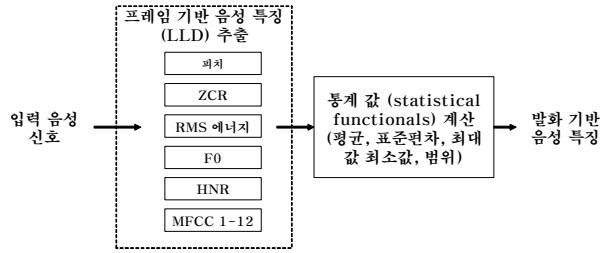

그림 2 발화기반 음성특징 추출

표 1 발화기반 음성특징 비교

	특징 차수	설명
emo_IS09	384	16 LLDs, Δ, 12 functionals
emo_base	950	25 LLDs, Δ, 19 functionals
emo_large	6,552	84 LLDs, Δ, 39 functionals

project[4]에서 사용된 음성특징들을 사용하였다. 즉, 프레임 기반으로 LLD들을 추출하였고, 그리고 나서 각각의 LLD의 통계치로 음성특징 파라메터로 사용하였다. 예를 들어, emo_IS09의 경우, LLD는 피치, 에너지, 주파수 변이(jitter), 진폭변위(shimmer), 발화율, HNR (harmonic-to-noise ratio), 포만트, 12차 MFCC이고, 이에 1차 미분을 적

표 2 음성특징 기반 감정인식률 비교

	음성 특징	감정인식률(%)
실험 1	프레임 기반 음성 특징 (13-MFCC+△+△△) [3]	62..5
실험 2	발화 기반 음성특징 (emo_IS90 - 384 차원)	48.3
실험 3	발화 기반 음성특징 (emobasse - 950 차원)	57.5
실험 4	발화 기반 음성특징 (emo_large - 6,552 차원)	69.2

. 통계치로는 평균, 표준편차, 최대값, 최소값, 범위, 왜도(kurtosis), 비대칭도(skewness) 등의 12개의 통계 값을 사용하여 384차원의 음성특징 파라메터를 추출하였다[4].

. 감정인식 실험 및 결과

발화기반의 음성특징 파라메터를 사용한 감정인식 시스템을 구축하기 위하여 openEAR project에서 제공하는 toolkit[4]을 사용하였다. OpenEAR project에서는 표 1에서 보는 바와 같이, emo_IS09, emobase 와 emo_large의 세가지 파라메터 셋을 정의하고 있으며, 각 셋은 384, 950 와 6,552 차원의 특징 파라메터로 구성된다. 본 논문에서 사용된 감정인식기인 SVM은 RBF (radial basis function)을 사용하였다[5].

구현된 시스템의 성능을 평가하기 위해서 6명의 남녀 배우가 10개의 문장을 6가지 감정으로 발성한 음성 DB[6]를 사용하였다. 총 360문장 중에 240문장 (남녀 4명)을 가지고 SVM을 학습시키고, 나머지 120문장 (남녀 2명)을 성능평가에 활용하였다. 이때, 훈련에 사용된 문장 중 듣기에 감정이 모호하다고 판단되는 31개의 문장은 제외시켰다.

표 2에 프레임 기반의 음성특징 파라메터와 발화기반의 음성특징 파라메터에 대한 감정인식 성능을 비교하였다. 표에서 보는 바와 같이, 실험 1에서는 프레임 기반의 음성특징 파라메터로 39차 MFCC를 사용하였고, 실험 2, 3, 4에서는 표 1에서 언급된 발화기반의 음성특징 파라메터를 각각 사용하였다. 그 결과, emo_large를 기반으로 한 음성특징 파라메터를 사용한 SVM이 69.2%의 성능으로 다른 발화기반의 음성특징 파라메터나 프레임 기반의 음성특징 파라메터를 사용하는 것과 비교하여 가장 좋은 성능을 보였다.

표 3은 사람의 감정인식 결과를 각 감정별 confusion matrix로 보여준다. 표의 결과는, 사람 (남녀 20명)이 평가에 사용된 음원을 직접 듣고 감정을 분류한 결과로써, 평균 87.6%의 인식률을 보였다[6]. 이에 반해, 표 4는 emo_large의 특징 파라메터를 사용하는 SVM의 성능에 대한 각 감정별 confusion matrix를 보여 준다. 표 3과 4를 비교한 결과, 공포 감정의 인식의 경우, 사람이 분류한 결과보다 15% 더 나은 인식 성능을 보였다.

Ⅳ. 결론

본 논문에서는 음성신호를 이용한 감정인식 시스템에서의 음성특징 파라메터에 따른 감정인식 시스템의 성능을 비교하였다. 구현된 시스템은 프레임 기반 음성특징 파라메터와 발화기반 음성특징 파라메터로써, 프레임 기반의 음성특징 파라메터는 39차 MFCC를 사용하였고, 발화기반의 음성특징 파라메터는 3가지 다른 종류의 low-level descriptor를 사용하였다. 실험 결과, 6552차의 발화기반 음성특징 파라메터를 사용하는 경우, 69.2%의 감정인식률을 보였다.

표 3 사람의 청취 판별에 의한 감정인식 혼동행렬[2]

		인식감정 (%)					
		화남	지루	공포	행복	평정	슬픔
입력 감정	화남	98.5	0.0	0.0	0.0	1.0	0.5
	지루	1.5	94.0	1.5	0.0	3.0	0.0
	공포	0.0	9.5	60.0	0.0	6.0	24.5
	행복	2.5	0.5	2.5	89.5	3.5	1.5
	평정	1.0	3.5	0.0	0.5	95.0	0.0
	슬픔	0.5	0.0	11	0.0	0.0	88.5

표 4 발화기반 특징 (emo_large) 기반 감정인식 시스템의 혼동행렬

		인식감정 (%)					
		화남	지루	공포	행복	평정	슬픔
입력 감정	화남	70	0	5	15	0	10
	지루	0	70	0	0	20	10
	공포	5	0	70	0	0	25
	행복	40	0	0	55	0	5
	평정	0	5	0	0	95	0
	슬픔	0	0	15	10	20	55

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2015년도 문화기술 연구개발지원사업의 연구결과로 수행되었음

참 고 문 헌

[1] 황구현, 신동규, 신동일. “복합 생체신호 기반의 감정인식 시스템 설계,” 한국통신학회 학술대회논문집, pp 782-783. 2013.

[2] 김남수, “감정인식 기술의 현황과 전망,” Telecommunications Review, 제 19권, 5호, May 2009.

[3] 강진아, 김홍국, “관객 반응 정보 수집을 위한 음성신호 기반 감정인식 시스템,” 한국방송공학회 하계학술대회 논문집, pp 1-2, 2013

[4] Eyben, F., Woollmer, M., and Schuller, B., “openEAR - introducing the Munich open-source emotion and affect recognition toolkit,” *Proc. ACII*, Amsterdam, Netherlands, pp. 576-581, 2009.

[5] Emotion01, Speech Information Technology & Industry Promotion Center, 2004.

[6] Scholkopf, B., Sung, K., Burges, C. J. C., Girosi, F., Niyogi, P., Poggio, T., and Vapnik, V., “Comparing support vector machines with Gaussian kernels to radial basis function classifiers,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2758-2765, Nov. 1997.