관객 반응정보 수집을 위한 음성신호 기반 감정인식 시스템

강진아 김홍국
광주과학기술원 정보통신공학부
{jinari, hongkook}@gist.ac.kr

A Speech Emotion Recognition System for Audience Response Collection

Jin Ah Kang Hong Kook Kim
School of Information and Communications, Gwangju Institute of Science and Technology

요약

본 논문에서는 연극공연을 관람하는 관객의 반응정보를 수집하기 위하여, 청각센서를 통해 관객의 음성을 획득하고 획득된 음성에 대한 감정을 예측하여 관객 반응정보 관리시스템에 전송하는 음성신호 기반 감정인식 시스템을 구현한다. 이를 위해, 관객용 헤드셋 마이크와 다채널 녹음장치를 이용하여 관객음성을 획득하는 인터페이스와 음성신호의 특징벡터를 추출하여 SVM (support vector machine) 분류기에 의해 감정을 예측하는 시스템을 구현하고, 이를 관객 반응정보 수집 시스템에 적용한다. 실험결과, 구현된 시스템은 6가지 감정음성 데이터를 활용한 성능평가에서 62.5%의 인식률을 보였고, 실제 연극공연 환경에서 획득된 관객음성과 감정인식 결과를 관객 반응정보 수집 시스템에 전송함을 확인하였다.

1. 서론

음성은 사람의 가장 기본적인 의사소통 수단으로써, 음성을 통해 감정을 표현하고 상대방의 감정 또한 예측할 수 있다. 특히 최근들어 음성인식 기능이 스마트폰을 비롯한 다양한 기기들에 성공적으로 탑재됨에 따라, 발화된 음성신호의 감정을 인식하는 기술은 음성 기반 사용자 인터페이스 기능을 더욱 향상시키는 역할을 할 것으로 기대된다. 뿐만 아니라, 인터랙티브 미디어아트 등의 문화기술 분야에서도 감정인식은 관객반응을 인지하기 위한 중요한 기술로써 부각되고 있다[1].

본 논문에서는 연극공연을 관람하는 관객의 반응정보를 수집하기 위하여, 청각센서를 통해 관객의 음성을 획득하고, 이에 대한 감정을 예측하여 관객 반응정보 관리시스템에 전송하는 음성신호 기반 감정인식 시스템을 구현한다. 이를 위해, 관객용 헤드셋 마이크와 다채널 녹음장치를 이용하여 관객음성을 획득하는 인터페이스를 구현하고, 음성신호의 MFCCs (mel-frequency cepstral coefficients) 성분을 추출하여 SVM (support vector machine) 분류기에 의해 감정을 예측하는 시스템을 구현한다. 또한 구현된 시스템을 관객 반응정보 수집 시스템에 적용하여, 관리자가 관객의 음성신호 및 그로부터 예측된 감정인식 결과를 확인할 수 있는 기능을 제공한다.

2. 관객 반응정보 수집을 위한 음성신호 기반 감정인식 시스템

그림 1은 본 논문에서 사용된 관객 반응정보 수집 시스템의 구성도이다. 이 관객 반응정보 수집 시스템은 센서를 통해 관심관객의 음성, 표정, 동작, 생체 신호 raw data를 획득하여 감정인식을 수행하는 센서단, 획득된 raw data 및 인식결과를 받아서 DB (database)에 저장하는 관리 시스템, 그리고 저장된 데이터로부터 원하는 정보를 제공하는 관리자 인터페이스로 구성된다.

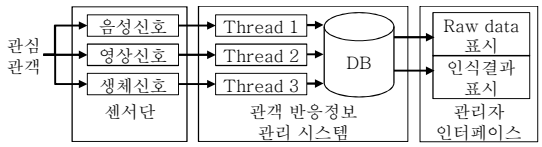

그림 1. 관객 반응정보 수집 시스템 구성도.

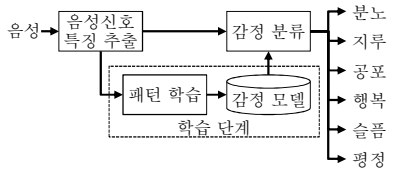

그림 2. 음성신호 기반 감정인식 시스템 구성도.

그림 2는 그림 1에서 음성신호 기반 감정인식 시스템의 구성도를 보여준다. 그림에서 보는 바와 같이, 감정인식의 학습단계에서는 분노, 지루, 공포, 행복, 슬픔, 평정의 6가지 감정으로 녹음된 음성 DB로부터 음성신호 특징을 추출하여 패턴학습을 통해 감정모델을 생성한다. 이후에는 입력된 음성신호의 특징을 추출하여 감정 분류기를 통해 예측된 감정을 출력한다.

3. 구현 및 성능평가

음성신호 기반 감정인식 시스템의 구현을 위해서는 MFCCs 특징 파라메터와 SVM 분류기를 이용하였다. 그림 3은 구현된 알고리즘의 흐름도를 나타낸 것이다. 즉, 입력 음성 파일로부터 20ms 단위로 읽어들인 프레임 데이터에 대해 VAD (voice activity detection)를 수행한

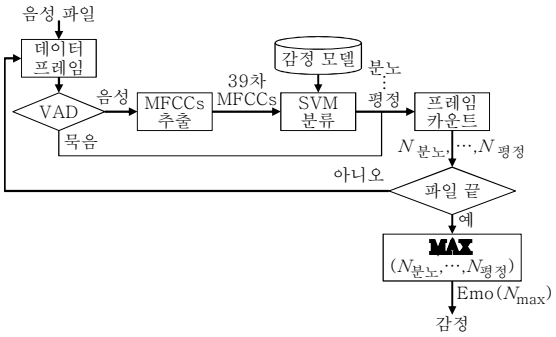

그림 3. 구현된 음성신호 기반 감정인식 알고리즘 흐름도.

표 1. 사람의 청취 판별에 의한 감정인식 결과(%)[2].

		인식 감정					
		분노	지루	공포	행복	평정	슬픔
입력 감정	분노	98.5	0.0	0.0	0.0	1.0	0.5
	지루	1.5	94.0	1.5	0.0	3.0	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 2. 구현된 음성신호 기반 SVM 기반 감정인식 결과(%).

		인식 감정					
		분노	지루	공포	행복	평정	슬픔
입력 감정	분노	60.0	0.0	15.0	15.0	5.0	5.0
	지루	0.0	85.0	10.0	0.0	0.0	5.0
	공포	0.0	0.0	90.0	0.0	0.0	10.0
	행복	20.0	0.0	20.0	55.0	5.0	0.0
	평정	0.0	55.0	0.0	0.0	40.0	5.0
	슬픔	0.0	10.0	30.0	10.0	5.0	45.0

다. 만약 해당 프레임이 묵음으로 검출된 경우, 감정예측 결과를 묵음으로 분류한다. 하지만 VAD 결과가 음성인 경우에는 음성신호로부터 39차 MFCCs를 추출하여 SVM 분류기에 의해 해당 프레임의 감정을 분노에서 평정까지의 6가지 중의 하나로 분류한다. 이때 음성 파일 전체에 해당하는 감정은 가장 많은 프레임 수 (N_{max})를 갖는 감정으로 분류한다.

구현된 시스템의 성능검증을 위해서는 6명의 남녀 배우가 10개의 문장을 6가지 감정으로 발성한 음성 DB[2]를 사용하였다. 총 360문장에서 SVM 학습과 인식률 평가를 위해 각각 240문장 (남녀 4명)과 120문장 (남녀 2명)을 할당한 후, 듣기에 감정이 모호하다고 판단되는 31개의 음원은 제외시켰다. SVM 커널은 사전실험을 통해 polynomial kernel[3]과 RBF (radial basis function)[4] 중 RBF를 사용하였다.

표 1은 사람 (남녀 20명)이 평가에 사용된 음원을 직접 듣고 감정을 분류한 결과로써, 평균 87.6%의 인식률을 보였다[2]. 또한 표 2는 구현된 시스템에 의해 분류된 결과로써, 평균 62.5%의 감정 인식률을 보였다.

실제 연극공연 환경에서의 동작검증을 위해서는 그림 4(a), (b)와 같이 다채널 녹음장치를 이용한 관객음성 획득 인터페이스를 구현한 후, 이를 앞서 구현한 음성신호 기반 감정인식 시스템에 연동되도록 하였다. 획득된 음성신호와 그에 대한 감정인식 결과는 SQL (structured query language)로 작성하여 TCP/IP 프로토콜을 통해 관객 반응정보 관리 시스템에 전달하도록 하였다. 실험은 90분의 연극공연 시간 동안 관객 5명에 대한 정보를 수집하도록 하였다. 그 결과, 구현된 음성신호 기반 감정인식 시스템은 정상적으로 동작하였으며, 그림 4(c)와 같이

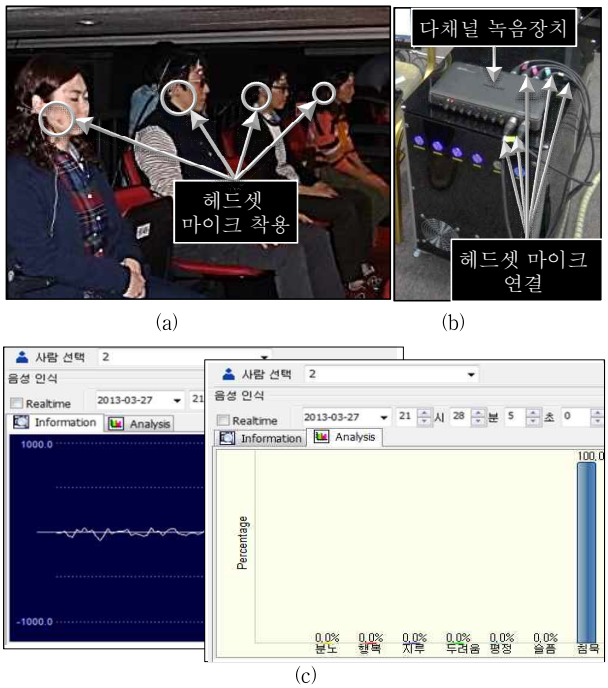

그림 4. 연극공연 환경에서의 동작검증: (a)헤드셋 마이크를 착용한 관객, (b)다채널 녹음장치 기반 H/W 구성, (c)관리자 인터페이스에 표시된 음성신호 파형 및 감정인식 결과.

관객 반응정보 수집 시스템의 관리자 인터페이스를 통해 특정 관객 및 특정 구간 설정에 따라 수집된 음성신호 및 감정인식 결과를 확인할 수 있었다.

4. 결론

본 논문에서는 연극공연을 관람하는 관객의 반응정보를 수집하기 위한 음성신호 기반 감정인식 시스템을 구현하였다. 구현된 시스템은 관객용 헤드셋 마이크와 다채널 녹음장치를 이용하여 관객의 음성신호를 획득하고, 이에 대한 MFCCs 성분을 추출하여 SVM 분류기를 통해 감정을 예측하였다. 실험을 통해, 구현된 시스템은 6가지 감정음성 데이터를 활용한 성능평가에서 62.5%의 인식률을 보였고, 실제 연극공연 환경에 적용되어 관객의 음성신호 획득 및 감정인식을 수행하여 관객 반응정보 수집 시스템에 전송함을 확인하였다.

감사의 글

본 논문은 문화체육관광부 및 한국콘텐츠진흥원의 2012년도 콘텐츠산업기술지원사업 및 2013년도 미래창조과학부의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2012-010636).

참고문헌

[1] 신동일, "감성인식 기술동향," *IITA 주간기술동향*, 1283호, pp. 1-9, 2007년 2월.
[2] Emotion01, Speech Information Technology & Industry Promotion Center, 2004.
[3] G. F. Smits and E. M. Jordaan, "Improved SVM regression using mixtures of kernels," in *Proc. of International Joint Conference on Neural Networks, Honolulu, Hawaii*, vol. 3, pp. 2785-2790, May 2002.
[4] B. Scholkopf, K. Sung, C. J. C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, "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.