소셜 TV적용을 위한 사용자 반응 사운드 인식방식 비교

*류상현 **김형국
광운대학교
*rshfly@nate.com **hkim@kw.ac.kr

Comparison of User's Reaction Sound Recognition for Social TV

*Sang-Hyeon Ryu **Hyoun-Gook Kim
Kwangwoon University

요약

소셜 TV 사용 시 , 사용자들은 TV를 시청하면서 타 사용자와의 소통을 위해 리모컨을 이용해서 텍스트를 작성해야하는 불편함을 가지고 있다. 본 논문에서는 소셜 TV의 이러한 불편함을 해결하기 위해 사용자 반응 사운드를 자동으로 인식하여 상대방에게 이모티콘을 전달하기 위한 시스템을 제안하며, 사용자 반응 사운드 인식에 사용되는 분류방식들을 비교한다. 사용자 반응 사운드 인식을 위해 사용되는 분류 방식들 중에서, Gaussian Mixture Model(GMM), Gaussian Mixture Model - Universal Background Model(GMM-UBM), Hidden Markov Model(HMM), Support Vector Machine(SVM)의 성능을 비교하였다. 각 분류기의 성능을 비교하기 위하여 MFCC 특징값을 각 분류기에 적용하여 사용자 반응 사운드 인식에 가장 최적화된 분류기를 선택하였다.

1. 서론

최근 IT기술의 발전과 인터넷 통신망의 고도화와 스마트기기의 보급으로 언제 어디서나 소통이 가능한 환경이 구축되면서 다양한 소셜 미디어 서비스가 인기를 얻고 있다. 대표적인 예로 페이스북, 트위터 같은 소셜 네트워크 서비스를 들 수 있다. 최근에는 IP 기반의 네트워크를 이용하여 음성통화, 화상통화, 인터넷접속, 콘텐츠 시청 등을 할 수 있는 IPTV기기 및 서비스의 대중화가 이루어지면서 SNS와 IPTV서비스가 결합된 소셜TV서비스가 주목받고 있다[1]. 사용자는 TV시청과 동시에 인터렉션 서버를 통해 오디오 데이터, 텍스트 정보 등을 전송 및 수신하게 된다. 이를 통해 사용자들은 가상의 공유공간을 형성하여 의사소통을 할 수 있게 된다. 하지만 스포츠 경기와 같이 사용자의 감정상황 및 의사소통이 빠르게 이루어져야 하는 상황에서 경기를 TV로 시청하면서 텍스트를 작성하고 보내야하기 때문에 시청에 방해가 되며, 리모컨을 이용한 이모티콘을 수동을 선택해야하는 조작의 불편함이 발생한다. 이러한 불편함을 해결하기 위해서 스포츠 경기 시청 시, 실시간으로 사용자의 반응 사운드를 인식하여, 인식된 감정을 상대방에게 이모티콘으로 자동 전송하는 시스템이 필요하다.

본 논문에서는 사용자 반응 사운드를 자동 인식하는 방법들을 비교하고, 소셜 TV에 적용하여 사용자 반응 사운드를 자동 인식하고 상대방에게 이모티콘을 사용하여 실시간으로 감정을 전달하는 최적의 방법을 제안한다.

2. 사용자 반응 사운드 인식 시스템

사용자 반응 사운드 인식을 위한 오디오 신호의 특징값들을 사용하여 잡음환경에서 사용자 반응 사운드를 가장 잘 분류 할 수 있는 오디오 특징값을 선택한다. 선택된 특징값을 분류기를 이용하여 사용자 반응을 분류하고 사용자 반응 인식에 가장 최적화된 분류기를 선택하였다. 다음 그림 1은 사용자 반응 사운드 인식 시스템의 구조이다. 분류기의 작동은 크게 모델 학습단계, 분류단계로 나누어진다. 먼저, 모델링 단계에서는 분류할 각 클래스에 대한 충분한 오디오 데이터를 입력받아 특징값을 추출하고 추출된 특징값들을 이용하여 모델을 생성한다. 생성된 모델들은 모델 데이터 DB에 저장된다. 분류단계에서는 분류가 필요한 오디오 데이터를 입력받아 특징값을 추출한다. 추출된 특징값은 모델학습 단계에서 생성된 모델과 비교하여 가장 유사한 클래스로 분류한다. 본 논문에서는 사용자 반응 사운드 신호를 입력받아 박수, 환호, 웃음, 야유, 아쉬움, TV잡음으로 분류되고, 결과를 상대방의 디스플레이에 이모티콘으로 표현함으로써, 사용자의 반응 및 감정을 공유 할 수 있도록 한다.

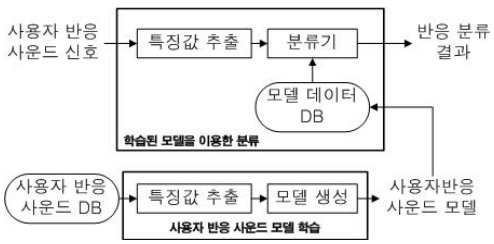

그림 1. 사용자 반응 사운드 인식 시스템

2.1 Gaussian Mixture Model(GMM)

GMM은 통계적 방법을 사용해 클러스터링 또는 밀도 추정을 생성하는 모델이다[2]. GMM은 가중치가 부여된 여러 개의 가우시안 확

률 밀도함수의 통계적 분포를 선형 결합하는 방법으로 최적 모델을 찾기 위해 Maximum Liklihood Estimate(MLE)를 사용한다. 그리고 MLE를 최적화하기 위한 방법으로 Expectation Maximization (EM) 알고리즘을 사용한다. GMM은 다음 식 1과 같이 가우시안 확률 $P(X|\lambda_c)$로 표현된다. 여기서 λ_c는 각 클래스에 해당하는 모델, $X=\{x_1,x_2,...,x_N\}$는 입력된 사운드 신호의 n차원 특징벡터, μ_i는 x의 평균 벡터, $\sum_i$는 x의 공분산이다.

$$P(X|\lambda_c)=\frac{1}{(2\pi)^{n/2}|\sum_i|^{1/2}}exp\left\{-\frac{1}{2}(x-\mu_i)^T\sum_i^{-1}(x-\mu_i)\right\} \quad (1)$$

2.2 Gaussian Mixture Model - Universal Background Model (GMM-UBM)

특징벡터 공간에서 구축된 UBM은 하나의 커다란 GMM으로 특징벡터들의 화자독립적인 특성을 나타내는 모델이다[3]. UBM 모델은 훈련된 모델과 테스트 사이의 불일치를 극복하기 위한 방법으로 사용되고 있다. GMM-UBM은 다음 식 2와 같이 GMM 가우시안 확률과 UBM 가우시안 확률의 비 $P(X)$로 표현된다. 여기서 λ_c는 각 클래스에 해당하는 모델, λ_{UBM}은 UBM 모델, $X=\{x_1,x_2,...,x_N\}$는 입력된 사운드 신호의 n차원 특징벡터이다.

$$P(X)=\frac{P(X|\lambda_c)}{P(X|\lambda_{UBM})} \quad (2)$$

2.3 Hidden Markov Model(HMM)

HMM은 은닉 관측열을 갖는 Markov Chain이다. 이는 관측 심볼과 관측 심볼을 구성하는 확률함수 및 상태와 상태간 천이확률로 구성할 수 있다. HMM은 관측 가능한 심볼 출력으로부터 관측 불가능한 프로세스를 확률로써 추정하는 방식으로 초기상태확률(π_i), 상태 a_i에서 a_j로의 상태천이확률(a_{ij}), 그 천이에서 심볼 k를 출력하는 관측확률$(b_{ij}(k))$ 등으로 HMM의 모델 $\lambda_c(A,B,\Pi)$를 정의하며, 관측열 $X=\{x_1,x_2,...,x_N\}$가 주어졌을 때, HMM은 다음 식 3과 같이 표현된다. 여기서 $S_{1,2,...N}$는 상태열을 나타낸다[4].

$$P(X|\lambda_c)=\sum_{s_1,s_2,...,s_N}\pi_{s1}b_{s1}(x_1)a_{s1s2}b_{s2}(x_2)...a_{sN-1sN}b_{sN}(x_N) \quad (3)$$

2.4 Support Vector Machine(SVM)

SVM은 높은 예측 정확성 때문에 많은 선형, 비선형 분류 문제에 폭 넓게 사용되고 있는 식별 모델로서, 두개의 클래스 사이에서 Margin이 최대가 되는 가장 좋은 결정 초평면을 찾는다. 이렇게 구해진 초평면은 결정 경계라고도 하며, 이 결정 경계에서 가장 가까운 클래스들의 데이터를 Support Vector라 한다[3,5]. 분류 문제가 비선형일 경우, 커널을 사용하여 데이터를 고차원으로 사상시킴으로써 선형분리가 되도록 한다. SVM은 다음 식 4로 정의 된다.

$$f(x)=\sum_{i=1}^{N}\alpha_i t_i K_i(x,x_i)+b. \quad (4)$$

여기서 t_i는 -1또는 1인 Ideal 출력으로 클래스 0 또는 클래스 1에 해당하는 것을 나타내고, x는 Support vector, b는 2차 분류 문제를 풀기위한 바이어스 상수이다. $K(\bullet,\bullet)$는 데이터를 고차원으로 사상시키는 커널함수이다. α_i는 조건부 최적화 문제를 해결하기 위해 사용하는 라그랑제 승수이다.

3. 실험결과

실험에는 실제 TV에서 스포츠 경기를 관람하며 스마트 TV에 설치된 캠코더(CY-STC110)를 사용하여 TV앞 4~5m 사이의 사용자 반응(박수, 함성, 웃음, 아쉬움, 야유, TV잡음)사운드를 15dB의 TV잡음 환경에서 녹음하였다. 오디오 데이터는 Stereo의 16Khz 샘플링레이트를 사용하였으며, 각 클래스에 대하여 7분길이의 오디오를 녹음하였다. 5분길이의 데이터는 모델링에 사용하였으며, 2분길이의 데이터는 테스트에 사용하였다. 본 논문에서는 특징값 MFCC를 이용하여 각 분류방식의 사용자 반응 분류를 실험하였다. 실험결과는 다음 표 1과 같으며, GMM이 평균 94.61%의 정확도로 가장 높은 인식률을 보였다.

	GMM	GMM-UBM	HMM	SVM
정확도(%)	94.61	92.85	90.38	90.45

표 1. 분류기 별 정확도 비교

4. 결론

본 논문에서는 소셜TV에서 사용자 반응 사운드 자동 인식 알고리즘을 제안하였다. 실험결과를 통해 MFCC를 특징값으로 사용하여, GMM을 통해 모델을 생성하고 클래스를 분류하는 것이 인터렉티브 소셜TV에서의 사용자 반응 사운드를 자동으로 인식 하여, 수동적인 조작 없이 자동으로 상대방의 디스플레이상에 이모티콘으로 감정을 전달해주는 시스템에 가장 적합하다고 판단되었다.

감사의 글

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(2012-0001941).

참고문헌

[1] K. Y. Lee, K. S. Cho and W. Ryu, “Social TV Service: A Case Study,” in Proc. of IEEE ICCE, pp. 287-288, Las Vegas, United States, Jan 2011.

[2] D. Reynolds, T. Quatieri and R. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, no. 1-3, pp. 19-41, Jan 2000.

[3] M. H. Liu, B. Q. Dai, Y. L. Xie and Z. Q. Yao, “Improved GMM-UBM/SVM for Speaker Verification,” in Proc. of IEEE ICASSP, pp. 925-928, Toulouse, France, May 2006.

[4] M. J. F. Gales and S. J. Young, "The Application of Hidden Markov Models in Speech Recognition," Foundations and Trends in Signal Processing, 2007.

[5] G. Mountrakis, J. Im and C. Ogole, “Support Vector Machines in Remote Sensing: A Review,” ISPRS J. Photogrammetry and Remote Sensing, vol. 66, no. 3, pp. 247-259, May 2011.