스마트폰환경에서 음성기반 감정인식 프레임워크

방재훈, 이승룡, 정태충
경희대학교 컴퓨터공학과
e-mail: jhb@oslab.khu.ac.kr, sylee@olsab.khu.ac.kr tcjung@khu.ac.kr

Speech Emotion Recognition Framework on Smartphone Environment

Jae Hun Bang, Sungyoung Lee , Taechung Jung
Dept of Computer Science, Kyung Hee University

요 약

기존의 음성기반 감정인식 기술은 충분한 컴퓨팅 파워를 가진 PC에서 수백개의 특징을 사용하여 감정을 인식하고 있다. 이러한 음성기반 감정인식 기술은 컴퓨팅 파워에 제약이 많은 스마트폰 환경을 고려하지 않은 방법이다. 본 논문에서는 제한된 스마트폰 컴퓨팅 파워를 고려한 음성의 특징 추출 기법과 서버 클라이언트 개념을 도입한 효율적인 음성기반 감정인식 프레임워크를 제안한다.

1. 서론

최근 스마트폰이 보급화 되면서 사용자 정보를 기반으로 다양한 개인화 서비스 연구가 활발히 진행 중이다. 사용자 정보로는 상황정보 및 감정정보 등이 있다. 특히 감정 정보는 사용자의 현재 감정 상태를 나타내는 정보로 감정 상태에 따라 달라지는 음악 추천과 같은 문화 콘텐츠 서비스 등에 사용되어지고 있다.

음성기반 감정인식이란 사용자의 음성신호를 분석하여 사용자의 감정을 자동으로 실시간 감정을 인식하는 기술이다. 최근 마이크로 폰 센서가 탑재된 스마트폰에서 사용자의 통화 음성 데이터 수집 및 처리가 용이해짐에 따라 감정인식 기술 연구가 활발히 이루어지고 있다.

스마트폰에서 감정기반 개인화 서비스를 제공하기 위해서는 최소한의 특징을 사용하여 정확한 감정인식 기술이 필요하다. 그러나 기존의 음성기반 감정인식 기술은 수초 정도의 음성 데이터를 해밍 윈도우 기법으로 잘게 분할한 뒤 수백개의 특징 벡터를 추출하여 기계학습 알고리즘을 사용하여 감정을 인식한다.[1-3] 이러한 방법론은 연산량이 많은 수백개의 특징 벡터를 사용하여 분류해야하는 기계학습 알고리즘을 제한된 스마트폰 자원에서 사용하기에는 적합하지 않다. 또한 연산량이 많은 기계학습 알고리즘을 서버로 분산하는 방법을 사용하더라도 수백개의 특징벡터 데이터를 전송하는 부분에서도 많은 프로세스가 발생하므로 특징 추출을 최소로 하고 처리량이 많은 기계학습 알고리즘을 컴퓨팅 파워가 높은 서버로 분산하는 음성기반 감정인식 프레임워크 기법이 필요하다.

본 논문에서는 스마트폰 환경에서 음성기반 감정인식을 위한 최소한의 특징을 추출하고 연산량이 많은 기계학습 알고리즘을 서버로 분산하는 프레임워크를 제안한다. 제안하는 기법은 13차 MFCC (Mel Frequency Cpestral Coefficient)를 활용하여 5초마다 60개의 특징을 추출하고 기계학습 알고리즘인 SVM (Support Vector Machine) 탑재된 서버로 전송하여 감정을 추출한다. 화남, 행복, 평범, 슬픔의 4가지 감정을 고려한 실험을 통하여 제안하는 감정인식 기술이 기존의 감정인식 기술보다 더 높은 정확도를 보임을 증명하였다.

2. 스마트폰환경 음성기반 감정인식 프레임워크

컴퓨팅 파워가 제한된 스마트폰에서 수백개의 특징을 사용하는 기존의 감정인식 기술을 적용하기에는 비효율적이다. 제안하는 스마트폰환경 음성기반 감정인식 프레임워크는 클라이언트가 되는 스마트폰에서는 감정인식을 위한 음성 특징 벡터 추출 부분을 처리하고 이를 서버로 전송하여 서버에서는 SVM을 사용하여 인식한다. 그림 1은 제안하는 스마트폰환경 음성기반 감정인식 프레임워크 전체 구조도 이다. 세부 모듈에 대한 설명은 다음과 같다.

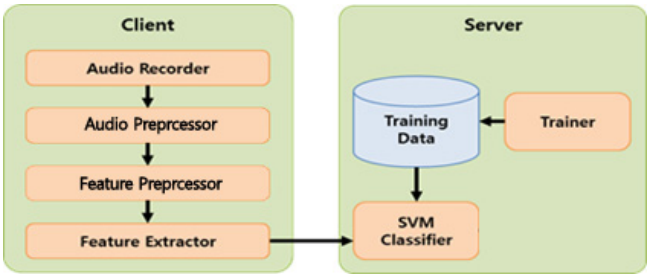

그림 1. 스마트폰환경 음성기반 감정인식 프레임워크 구조도

- Client (Smartphone)
  - Audio Recorder 모듈: 스마트폰의 마이크로 폰 센서

를 통해 사용자의 음성을 녹음하는 모듈이다.
- Audio Preprocessor 모듈: 녹음된 사용자의 음성을 전 처리하는 과정으로 사용자 음성만을 분리해내는 묵음을 제거하는 모듈이다.
- Feature Preprocessor 모듈: 사용자의 음성을 정제 가능한 데이터로 변환하기 위한 전 처리 과정으로 13차 MFCC Filter Bank 알고리즘을 이용하여 특징 벡터를 추출하는 모듈이다.
- Feature Extractor 모듈: 변환된 13차 MFCC 특징 벡터를 최소화하는 모듈이다. 추출된 특징 기계학습 알고리즘이 탑재되어 있는 서버로 전송한다.

● Client (Smartphone)
- Trainer 모듈: 미리 수집한 음성 훈련데이터를 SVM에 훈련시키기 위해 클라이언트에서 사용한 특징을 사용하여 특징 벡터를 추출하는 모듈이다.
- Training Data : 훈련을 위해 음성 훈련데이터의 특징벡터를 저장하고 있는 데이터 베이스다.
- SVM Classifier 모듈: SVM 분류 알고리즘을 통해 훈련을 위해 특징을 모아놓은 Training Database를 이용하여 훈련하고, 전송 받은 특징 벡터를 4개의 감정 화남, 행복, 평범, 슬픔으로 분류하는 모듈이다.

제안하는 감정인식프레임 워크에서의 기술적인 부분에서는 감정인식을 위한 음성 특징 벡터 추출과정과 감정인식 과정으로 나뉜다. 다음 절에서는 이 두 가지 부분을 설명한다.

2.1. 감정인식을 위한 음성 특징 벡터 추출 과정

스마트폰에서 음성의 특징 벡터를 추출하기 위해서는 사용자의 음성을 녹음하는 기능뿐만 아니라 사용자의 목소리만을 뽑아내기 위해 녹음된 데이터의 전처리 과정이 필요하다. 음성인식 분야는 다른 타 인식분야와 다르게 말하지 않는 공백데이터는 의미가 없는 데이터로 감정인식에 방해가 되는 요소이다. 따라서 이러한 묵음을 제거하는 것은 긴 단위의 음성기반 감정인식에 있어서 꼭 필요한 기술이다. 본 논문에서 사용하는 음성 묵음 제거 방법으로 소리의 크기에 임계 값을 두어 제거한다. 임계값은 일반적으로 사람이 소곤대는 소리인 15데시벨(db)로 설정한다.

그림 2는 통화음성데이터의 묵음을 제거하기 전 파형과 제거한 파형이다. 사용자가 말을 하지 않은 묵음이 효과적으로 제거되었음을 확인할 수 있다.

그림 2. 묵음 제거 전/후 파형

묵음이 제거된 데이터는 감정을 인식하기 위해 여러개의 서브 윈도우로 분할한다. 일반적으로 음성기반 감정인식은 3초에서 5초사이의 윈도우에서 높은 정확도를 갖는다. 본 논문에서는 감정 인식 및 음성 분할 주기를 5초 단위로 설정하고 분할한다.

5초단위로 분할된 음성은 감정인식에 적합한 특징을 추출하기 위해 데이터를 가공하는 필터뱅크 알고리즘을 통해 특징을 추출한다. 특징추출 과정에서는 필터뱅크 알고리즘으로 13차 MFCC를 사용하였다. MFCC는 인간의 청각 특성을 고려하는 필터뱅크 알고리즘으로 음성 인식 분야에서 널리 사용되고 있으며 인식 성능이 우수하다[4]. 13차 MFCC는 32ms를 프레임 단위로 윈도우를 분할하는 해밍윈도우 기법을 사용한다. 13차 MFCC의 첫 번째 특징은 전체 주파수의 음성의 세기를 나타내는 특징으로 이는 감정 음성에 대한 주파수별 특징이 아니므로 특징에서 제외한다.

본 논문에서는 5초크기의 타임 윈도우를 사용하므로 총 156개의 프레임이 생성되고 각 프레임마다 12개의 MFCC 값을 추출하여 총 1872개의 특징 벡터가 추출된다. 1872개의 특징은 스마트폰에서 서버로 전송하기 어려운 양이므로 31개의 프레임 단위로 각 프레임에서 추출된 13 MFCC값들의 평균을 특징으로 사용한다. 31개의 프레임은 음성데이터에서 1초의 데이터로 5초크기의 데이터를 1초마다 MFCC값의 평균을 추출하여 1초에 12개씩 총 60개의 특징 벡터를 추출한다.

2.2. 감정인식 과정

모든 인지 분야에서 인지 정확도는 특징 추출 부분뿐만 아니라 기계학습 알고리즘 선택에도 정확도에 영향이 있다. 그 중 SVM(Support Vector Machine)은 최근 등장하여 여러 가지 문제에 대하여 가장 우수한 해결 능력을 보여주는 알고리즘이다. 따라서 본 논문에서는 SVM을 사용하여 4가지 감정 기쁨, 화남, 보통, 슬픔을 인식한다. 또한 모든 상황에 높은 정확도를 가질 수 있도록 훈련용 데이터를 각 감정마다 모두 통합한 뒤 32ms 단위로 오버랩하여 훈련한다.

3. 실험 환경 및 결과

실험 환경은 안드로이드 스마트폰인 SHW-M250S에서 음성 데이터를 수집하였다. 음성은 샘플링 주파수 8000hz, 16비트, 모노 채널로 수집하였다. 데이터 수집 대상자는 남녀 2명씩 각 감정마다 5초식 25개를 녹음하여 훈련용 데이터를 수집하였다.

실험은 기존의 수 백개의 특징을 사용하는 기법[2]과 본 논문 에서 제안하는 기법을 비교 실험하였다. 정확도 측정은 10Fold-Cross Validation 기법을 사용하여 측정하였다. <표 1,2>는 인지 정확도를 비교한 Confusion Matrix 이다.

<표 1> 기존 기법의 정확도 - 평균 72% (단위 %)

	happy	angry	normal	sad
happy	81.20805	16.91275	1.610738	0.268456
angry	19.04762	70.47619	6.530612	3.945578
normal	2.755906	21.65354	58.66142	16.92913
sad	1.052632	10.07519	11.57895	77.29323

<표 2> 제안하는 기법의 정확도 - 평균 89% (단위 %)

	happy	angry	normal	sad
happy	91.4634	6.0976	1.2195	1.2195
angry	12.766	78.7234	0	8.5106
normal	0	0	99	1
sad	4.0541	6.0811	0	89.8649

3. 결론

본 논문에서는 컴퓨팅 파워에 제약이 있는 스마트폰 환경에서 효과적인 음성기반 감정인식 특징 추출 기법과 전체 프레임워크를 제안하였다. 제안한 프레임워크는 음성에서 특징 벡터를 최소화 하여 스마트폰에서 효율적인 감정인식이 가능한 프레임워크로 실험을 통하여 제안하는 특징벡터 최소화 기법이 기존의 기법이 높은 정확도가 보임을 입증하였다.

4.Acknoweledgement

본 연구성과는 중소기업청에서 지원하는 2011년도 산학연공동기술개발사업(No. 00048272)의 연구수행으로 인한 결과물임을 밝힙니다.

참고문헌

[1] A. B. Kandali, A. Routray, T. K. Basu, “Emotion recognition from Assamese speeches using MFCC features and GMM classifier,” TENCON 2008 - 2008 IEEE Region 10 Conference, pp.1-5, 19-21 Nov, 2008.

[2] Z. Xiao, Dellandrea, L. Chen, W. Dou "Recognition of emotions in speech by a hierarchical approach", ACII 2009. 3rd International Conference, 10-12, Sept, 2009

[3] D. Morrison, R. Wang, Liyanage C. D. Silva “Ensemble methods for spoken emotion recognition in call-centres", Speech Communication 49, pp.98 - 112 , 2007

[4] A. Klautau "The MFCC", [Online]. Available: http://www.cic.unb.br/~lamar/te073/Aulas/mfcc.pdf