비감독 학습과 감독학습의 결합을 통한 음성 감정 인식

배상호*, 이장훈*, 김현정**, 원일용*
*서울호서전문학교 사이버해킹보안과
**건국대학교 컴퓨터공학과
e-mail:nokozan2@gmail.com, ljh08888@naver.com, nygirl@konkuk.ac.kr, clccclcc@shoseo.ac.kr

Recognition of Emotional states in speech using combination of Unsupervised Learning with Supervised Learning

Sang-Ho, Bae*, Jang-Hoon, Lee*, Hyun-jung, Kim**, Il-Young, Won*
*Cyber Hacking Security Seoul Hoseo Technical College,
**Dept, of Computer Science and Engineering KonKuk University

요 약

사용자의 감정을 자동으로 인식하는 연구는 사용자 중심의 서비스를 제공할 때 중요한 요소이다. 인간은 하나의 감정을 다양하게 분류하여 인식한다. 그러나 기계학습을 통해 감정을 인식하려고 할 때 감정을 단일값으로 취급하는 방법만으로는 좋은 성능을 기대하기 어렵다. 따라서 본 논문에서는 비감독 학습과 감독학습을 결합한 감정인식 모델을 제시하였다. 제안된 모델의 핵심은 비감독 학습을 이용하여 인간처럼 한 개의 감정을 다양한 하부 감정으로 분류하고, 이렇게 분류된 감정을 감독학습을 통해 성능을 향상 시키는 것이다.

1. 서론

최근 IT 연구의 전체적인 방향이 PC중심에서, 네트워크 중심을 거쳐, 사용자 중심 흐름으로 가고 있다. 사용자 중심으로 서비스를 제공하기 위해서는 사용자의 행동은 물론 감정, 기호 등을 종합적으로 파악하여 맞춤형 서비스를 제공하는 것이 중요하다[1]. 그중에서도 특히, 음성은 사용자의 태도 및 감정을 인지하는 데 있어서 가장 간단하고 자연스러운 수단이다. 이러한 이유로 음성을 통해 사용자의 감정 상태를 자동으로 인식하는 연구가 진행되어왔다.

음성의 의미를 인식하는 연구에서 원시 음성데이터의 어떠한 특징들이 인식에 영향을 미치는지는 이미 많이 연구되어 왔지만, 감성인식에서 감정의 특징을 대변하는 요소는 많이 알려져 있지 않다. 이러한 이유로 신경망을 감정인식에 많이 사용한다.

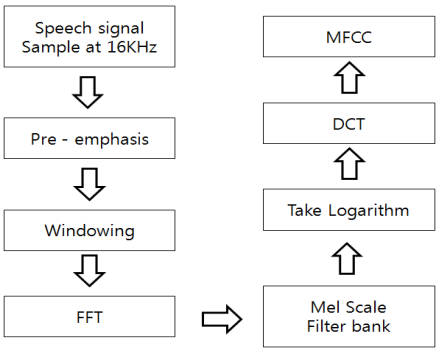
신경망의 학습에는 감독학습과 비감독 학습 방법이 있는데, 각각은 학습하고자 하는 데이터의 성질에 따라 또는 학습 환경의 조건에 따라 사용처가 다르다. 본 연구는 다양한 음성 데이터에서 화자의 감정 상태를 학습하고 자동으로 인식하는 방법에 대한 연구이다. 본 논문에서는 감독학습과 비감독 학습 신경망을 각각 실험하였으며, 성능의 향상을 위해 두 가지 학습 방법을 결합한 새로운 모델을 제시하였다. 제안된 방법의 유용성은 실험과 분석으로 증명하였다.

본 논문은 다음과 같이 구성된다. 2장에서는 관련 연구를 설명하고, 3장에서는 본 논문에서 제안하는 비감독 학습과 감독학습의 결합을 통한 감정인식 알고리즘을 제안한다. 4 장에서는 제안 알고리즘의 실험 및 결과를 언급하였다. 5장은 결론 및 향후 과제에 대하여 기술하였다

2. 관련 연구

2.1 Mel-Frequency Cepstral Coefficient (MFCC)

인간의 청각 시스템의 특성을 반영한 MFCC 알고리즘은 다른 특징들보다 좋은 성능을 제공한다고 보고되며, 일반적으로 Mel-cepstrum은 critical band filters를 사용하여 얻을 수 있다. 인간의 귀가 낮은 주파수 영역에서는 분해 능력이 높고 높은 주파수대에서 분해능력이 떨어지므로 1KHz 이하에서는 선형적으로 filter를 적용하고 그 이상에서는 log 스케일로 필터를 적용한다. (그림 1)은 Mel-cepstrum을 얻기 위한 MFCC의 개략적인 구조도이다[2].



(그림 1) MFCC의 처리과정

2.2 오류 역전파(Back-Propagation) 학습

오류 역전파 학습 알고리즘은 순방향 다층 신경망의 학습에 효과적으로 사용할 수 있어서 다양한 분야에 가장 널리 활용되는 학습 알고리즘이다. 출력층의 오차 신호를 이용하여 은닉층과 출력층 간의 연결강도를 변경하고, 또한 출력층의 오차신호를 은닉층에 역전파 하여 입력층과 은닉층과의 연결강도를 변경하는 학습 방법이다[3].

$$E=\sum_{p}E_{p'}\ (E_p=\frac{1}{2}\sum_{j}(d_{pj}-o_{pj})^2)$$

p는 p번째 학습 패턴을 말한다. E_p는 p번째 패턴에 대한 오차이며 d_{pj}는 p번째 패턴에 대한 j번째 요소이다. o_{pj}는 실제 출력의 j번째 요소이다.

2.3 자기조직화지도(Self Organizing Map) 학습

인간 뇌의 신경 생물학적 원리에 근거한 뇌 반응의 메커니즘을 모델링하여 고안되었다. SOM 알고리즘은 학습 단계에서 유사한 패턴끼리 2차원의 특징 지도를 조직화하여 영역 지도를 형성한다. 이후 인식 단계에서 이미 학습 단계에서 훈련된 연결 가중치 합에서 미지의 특징 벡터에 대하여 경쟁 층에서 반응이 일어나는 위치를 통하여 해당 클래스를 인식하는 알고리즘이다. SOM 알고리즘은 음성 인식, 이미지 패턴 분류에서 많이 사용되고 있다[4].

$$W_{new}=W_{old}+a(X-W_{old})$$

위의 식은 학습 시 뉴런들의 연결강도를 조정하는 방법이다. W_{old}는 조정되기 이전의 연결강도 벡터이며, W_{new}는 조정된 후의 새로운 연결강도 벡터이다. X는 입력패턴 벡터이며 a는 학습률이다[5].

3. 학습법의 결합을 통한 음성 감정 인식

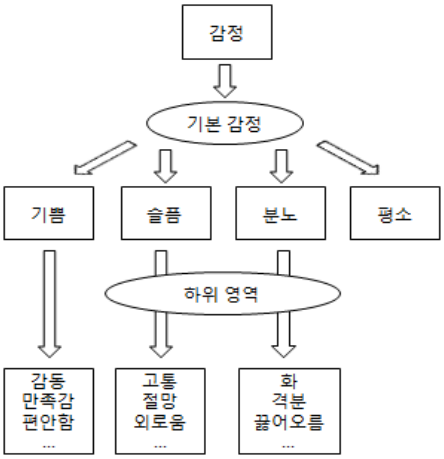
3.1 전처리

감정 인식을 위한 원시 음성 데이터의 전처리 방법에는 아직 표준적인 방법은 없다. 우리는 일반적으로 음성의 특징점만으로 데이터를 추출하는 MFCC를 사용하여 입력된 음성을 전처리하였다.

3.2 학습

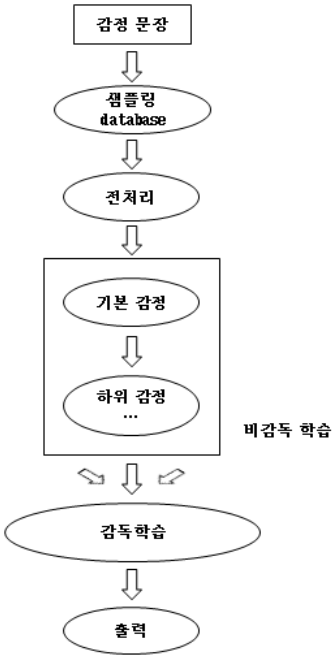
감정을 단순히 하나의 수치로 표현하기에는 무리가 따른다. 즉 감정에는 정성적인 면이 존재한다. 따라서 하나의 학습알고리즘을 이용하여 단순하게 감정을 인식한다면 좋은 성능을 기대하기 어렵다. 그러나 인간은 하나의 감정을 다양하게 분류하여 인식한다.

따라서 기계학습과 인간의 인식차이를 개선하기 위해서 본 논문에서는 비감독 학습과 감독학습을 결합한 모델을 제시한다. 비감독 학습을 이용하여 인간처럼 한 개의 감정을 다양한 하부 감정으로 분류하고, 이렇게 분류된 감정을 감독 학습하여 학습시킨다. 아래 (그림 2)는 이러한 제안을 보여준다.



(그림 2) 기본감정에서의 세분화된 감정 종류

제안된 음성 감정 분류 시스템의 구성은 아래 (그림 3)과 같다.



(그림 3) 음성 신호를 이용한 감정인식 순서도

감정이 포함된 문장을 전처리하고 전처리된 데이터를 비감독 학습으로 세분화한다. 이것을 감독학습의 입력으로 사용한다. 전체적인 과정은 학습 과정과 인식 과정으로 나누어진다. 핵심적인 알고리즘은 아래 (그림 4)와 같다.

학습 단계(Learning)

1. 학습할 모든 데이터를 전처리한다.
2. 비감독 학습엔진으로 학습한다.
3. 앞 단계의 출력을 이용하여 감독학습의 목표 값과 입력 노드를 설정한다.
4. 감독학습을 한다.
5. 감독 학습의 에러율이 기준에 도달하면 학습을 종료한다.

인식 단계(Recognition)

1. 인식하고자 하는 음성 데이터를 전처리한다.
2. 앞 단계의 데이터를 비감독 학습의 입력으로 하고, 그 출력 결과를 감독 학습의 입력으로 사용한다.
3. 감독학습에서 최종 결과를 출력하고 해석한다.

(그림 4) 결합된 신경망 알고리즘

4. 실험 및 결과

4.1 감성 데이터베이스 구축

음성 감정 인식에서 표준으로 사용되는 데이터베이스는 알려져 있지 않다. 몇몇 데이터들이 존재하기는 하지만 본 연구를 위한 실험데이터로 사용하기에는 문제가 있다. 이에 본 연구에서는 실험을 위해 자체 음성 데이터베이스를 구축하였다.

감정의 종류는 분노, 기쁨, 슬픔, 평소로 구분하고 한 개의 감정 당 10개의 예제 문장을 사용했다. 화자는 모두 15명으로 구성하였으며 총 600개의 원시데이터를 수집하였다. 본 연구에서 사용한 예제 문장은 논문[6]의 연구를 참고하여 결정하였다. 아래 <표 1>은 본 연구에서 사용한 예제 문장의 예시이다.

<표 1> 각 감정별 녹음 문장 예시

기본 감정	문장
분노	내가 그 정도 인격으로 밖에 안보여?
기쁨	모두들 합격의 기쁨을 누리고 있지요
슬픔	어머니의 뒷모습에 가슴이 아려옵니다.
평소	어리다는 사실은 제약이 아니라 무기입니다.
…	…

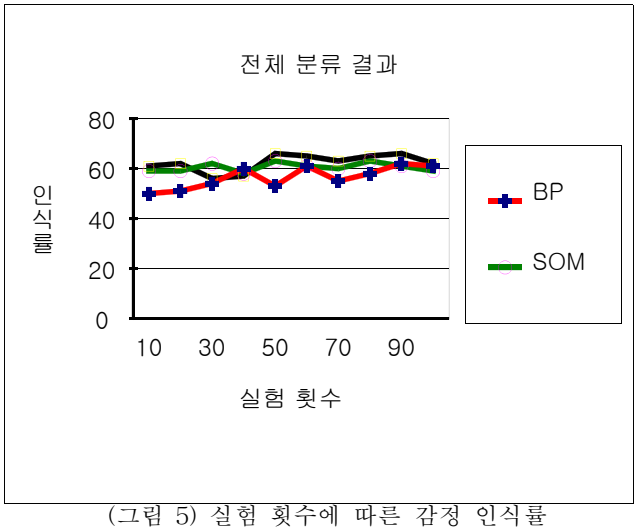
피험자 선정은 일반인을 대상으로 특정 감정의 영상을 보여줬을 때 몰입이 잘되는 남성 10명, 여성 5명을 선정하였다. 선정된 일반인은 21일간 하루 2시간씩 감정 표현에 대해 연습을 시켜서, 감정별로 화자가 감정에 몰입할 수 있는 시간에 데이터를 수집하였다.

녹음 환경은 조용한 환경의 강의실에서 마이크를 이용하여 15명의 피험자들에게 똑같은 환경을 구성하기 위해서 마이크와 피험자간의 거리는 10cm로 설정했다. 녹음된 파일 형태는 44.1KHz, 16bit이다.

4.2 실험 결과 및 분석

비감독 학습 알고리즘으로는 SOM을 사용하였으며, 감독학습으로는 BP를 사용하였다. 학습은 여러 번 반복하여 실시하였으며, BP의 에러율이 1% 미만이 될 때 학습 단계를 종료하고 성능을 실험하였다. BP는 3계층을 사용하였으며, 입력층의 노드 수는 SOM의 출력층 개수를 사용하였다. 수집된 원시데이터의 70%를 사용하여 학습하고, 나머지 30%로 테스트하였다.

실험결과 실험 횟수에 따른 전체 감정 인식률은 (그림 5)를 통해 알 수 있다. 비감독 학습 알고리즘, 감독학습 알고리즘으로 각각 실험했을 때보다 결합한 알고리즘 모델로 실험했을 때 성능이 좋았지만, 그 인식률은 2% 정도로 미미했고, 결합한 알고리즘의 성능이 안정적이지 못하고 인식률의 격차가 9%~11% 정도로 큰 편이었다.



(그림 5) 실험 횟수에 따른 감정 인식률

5. 결론 및 향후 과제

본 논문에서는 인간의 감정인지와 기계학습을 통한 감정인식의 차이를 최소화하기 위한 감정인식 모델을 제안하였다. 비감독 학습 알고리즘을 이용하여 감정을 세분화하고, 이것을 다시 감독학습 기법으로 학습하였다. 제안된 방법은 한가지만의 기계학습방법을 사용하여 인식했을 때보다 크게 우수한 성능은 얻지 못하였다. 불안정한 결합 알고리즘의 개선이 필요하고, 개선 방안으로는 감성 데이터베이스의 확장과 다양한 비감독 학습법, 그리고 전처리 기에서의 특징추출을 보완한다면 알고리즘의 성능이 개선될 것이라고 제안한다. 하부 감정 분류에서 추상적이고 주관적인 면이 있으며 향후 이를 보완 하고 개선하기 위해

서 여러 알고리즘을 이용한 다양한 실험이 필요하다.

참고문헌

[1] 신동일, “감정인식 기술 동향”, 주간기술동향, 통권 1283호, 2007.
[2] 정영규, 한문성, 이상조, “성대신호 기반의 명령어인식기를 위한 특징벡터 연구”, 정보과학회논문지: 소프트웨어 및 응용, 제34권 제3호, pp.230, 2007.
[3] 한국과학기술원, “영상인식을 위한 신경회로망의 광학적 구현 기술에 관한 연구, 제2차년도 최종보고서”, 과천: 과학기술처, 1992.
[4] 주종태, 박창현, 심귀보, “자기 조직화 신경망을 이용한 음성 신호의 감정 특징 패턴 분류 알고리즘”, 한국퍼지 및 지능시스템학회, 2006년도 추계학술대회 학술발표 논문집, 제16권 제2호, 2006.
[5] 최성기, “코호넨의 자기 조직화 신경회로망의 성능에 관한 연구”, pp.26, 1999.
[6] 조윤호, “화자 독립 음성 감성인식 시스템의 구현 및 응용에 대한 연구”, 2009.