비디오 화자 인식 성능 향상을 위한 복합 신경망 모델*

이범진[1○], 장병탁[1]
{bjlee, btzhang} @ bi.snu.ac.kr
[1]서울대학교 컴퓨터공학부

A Hybrid Neural Network model for Enhancement of Speaker Recognition in Video Stream

Beom-Jin Lee[1○], Byoung-Tak Zhang[1]
[1]School of Computer Science and Engineering, Seoul National University

요 약

대부분의 실세계 데이터는 시간성을 띄고 있으므로 시간성을 지닌 데이터를 분석할 수 있는 기계 학습 방법론은 매우 중요하다. 이런 관점에서 비디오 데이터는 다양한 모달리티가 결합된 대표적인 시간 데이터이므로 비디오 데이터를 대상으로 하는 기계 학습 방법은 큰 의미를 갖는다. 본 논문에서는 음성 채널에 기반한 비디오 데이터 분석 방법의 예비 연구로 비디오 데이터에 등장하는 화자를 인식할 수 있는 간단한 방법을 소개한다. 제안 방법은 MFCC (Mel-frequency cepstrum coefficients)를 이용하여 인간 음성 특성의 분포를 분석한 후 분석 결과를 신경망에 입력하여 목표한 화자를 인식하는 복합 신경망 모델을 특징으로 한다. 실제 TV 드라마 데이터에서 가우시안 혼합모델, 가우시안 혼합 신경망 모델, 제안 방법의 화자 인식 성능을 비교한 결과 제안 방법이 가장 우수한 인식 성능을 보임을 확인하였다.

Key Words : 음성, 화자인식, 비디오

1. 서론

나날이 증가하는 TV 방송 채널들과 라디오 방송, 그에 더불어 발달되는 데이터저장 기술로 인하여 방대한 양의 디지털 비디오 데이터가 각 방송사 스토리지에 저장되고 있다. 예를 들면 프랑스의 Institut National de l'Audiovisuel (INA)에는 45년 기간 동안의 약 300,000 시간의 TV 방송 프로그램과 60년 기간 동안의 약 400,000시간의 라디오 프로그램이 저장되어있다[1]. 그러나 저장된 데이터를 수작업으로 검색하고 사용해야 한다면 큰 문제가 될 수 있다. 만일 사용자가 모든 데이터를 일일이 찾아야 한다면 시간 비용이 막대할 것이다. 그리고 검색기술이 발달되지 않았다면 동일 작업을 반복해야 하는 상황이 발생한다. 따라서 연구자들은 이와 같이 비효율적인 상황이 반복되는 것을 막기 위하여 텍스트, 이미지, 사운드와 같은 모달리티를 이용한 색인 작성을 통해 사용자가 원하는 데이터를 찾는 방법을 연구하여 왔다[2, 3, 4].

탐색 기법들을 비교해 보면 텍스트 및 이미지 탐색이 음성 이용 탐색보다 성능 면에서 뛰어나다[5]. 왜냐하면 다른 모달리티 탐색의 경우 검색 대상 항목이 거의 변하지 않는다는 특징을 이용할 수 있지만(예: 텍스트 검색에서 단어는 불변), 음성인식에서는 같은 단어라도 어휘, 억양, 단어가 제시된 배경에 따라 다른 특성을 보이기 때문이다[6, 7]. 따라서 음성 채널 측면에서 비디오를 분석하려면 잡음 등의 외부 환경이 존재하는 상황에서도 높은 성능으로 화자를 인식할 수 있는 방법이 필요하다.

본 논문에서는 음성 데이터의 중요 특징을 도출할 수 있는 MFCC를 이용, 음향벡터 (acoustic vector)를 추출하여 화자를 인식하는 방법을 개발하였다. 그리고 실제 TV 드라마 에피소드에 제안 방법을 적용하여 그 성능을 확인하였다.

2. 시스템 개요 및 실험 방법

그림 1은 제안된 시스템 구조로 제안된 화자 인식 방법론

* 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(No. 2012-0005643, videome), 정부(지식경제부)의 재원으로 한국산업기술평가관리원의 지원(No. 10035348, mLife) 및 교육과학기술부의 BK21-IT 프로그램에서 일부 지원되었음.

은 드라마 등장인물들의 화자인식을 위한 음성 특성추출, 음성인식을 위한 모델링 방법으로 구성된다. 본 논문에서는 20세기 폭스 텔레비전이 ABC를 위해 제작한 미국의 법률 드라마인 보스턴 리걸(Boston Legal) 시즌 1의 에피소드 5를 실험용 동영상으로 사용하였다. 그림 1에서 보이고 있듯 음성 스트림이 공급되면 전처리 과정을 통해 중요한 특성을 추출한 후 모델링을 수행한다. 이렇게 만들어진 모델에 기반하여 실험 데이터에서 목표한 화자를 인식한다.

2.1 전처리

음성 신호의 특징 파라미터 추출과정은 다음과 같다. 첫째, 전처리를 통하여 PCM signed 16비트, 스테레오 타입, 44.1khz의 샘플링 레이트의 신호를 추출하였다. 다음으로, 잡음에 대한 성능을 측정하기 위하여 잡음이 섞인 음성신호를 추출하였다. 마지막으로 음성 데이터는 녹음 시점의 설정 영향을 매우 크게 받기 때문에 동일 에피소드의 데이터를 훈련 데이터와 테스트 데이터로 구분하였으며 주어진 에피소드의 30%를 훈련 데이터로 설정하였다.

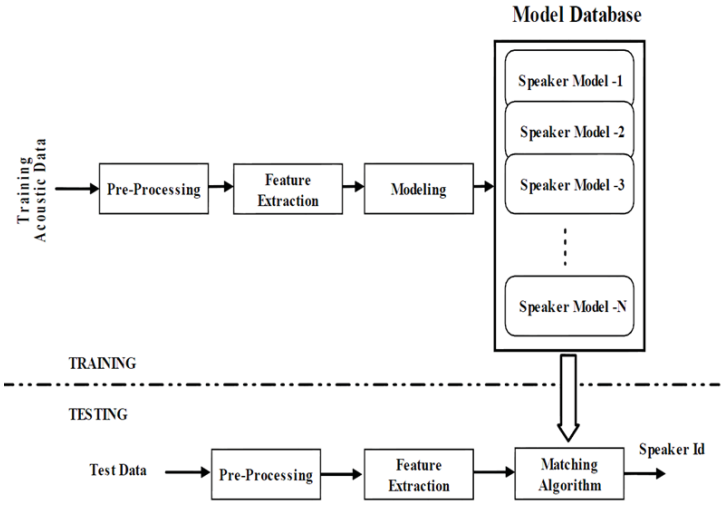

그림 1 시스템의 구조도. 음성인식은 훈련단계와 시험단계로 구분되며, 전처리 후 특성을 추출한 음성 데이터 모델링을 통하여 화자를 인식한다[9].

2.2 MFCC를 이용한 특성 추출

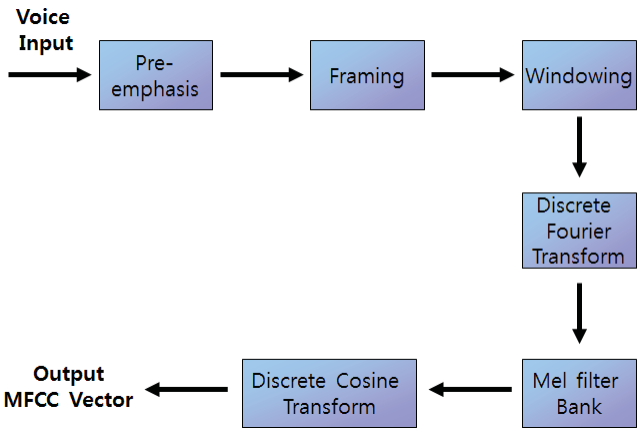

그림 2 MFCC 순서도. 음성 데이터가 들어오게 되면 높은 주파수역대를 증폭한 후 프레이밍과 해밍 윈도잉을 수행한다. 그 후 DFT (Discrete Fourier Transform)를 통하여 주파수역대로 전환 후 멜 필터 뱅크 (Mel filter Bank)를 통하여 로그 멜스펙트럼을 생성한다. 마지막으로 다시 시간대영역으로 전환하여 MFCC 음향벡터를 출력한다.

인간의 청각 시스템은 멜 주파수에 따라 반응하게 된다. 여기서 멜 (mel)이란 인간의 귀가 주파수에 따라 반응하는 정도를 측정하는 측정 단위이다. 실제 신호의 주파수와 멜 주파수 사이의 관계는 차이가 있다. 그러나 이러한 특성을 이용하여 실제 주파수 영역에서 멜 주파수 영역으로 변형시킨 후 멜-켑스트럼을 구할 수 있다.[9] 그림 2에서 MFCC의 음향벡터 추출과정을 설명하였다.

각각의 음성데이터들을 1초 단위로 조각내어 MFCC 생성과정을 수행 했으며, MFCC 생성과정 수행 중 필요한 분절화 단계의 파라미터는 20msec 단위로 설정하여 13개의 음향벡터를 추출하였다.

2.3 제안 방법을 이용한 모델링

표 1 제안 알고리즘

```
알고리즘 1. 복합 신경망 모델링
입력. 음성데이터(wav 포맷)
출력. 화자를 나타내는 이진값
방법. Mfcc(file, fs, time): 추출함수,
          file: 해당 음성 파일
          fs: 음성 파일의 샘플링비율,
          time: 음성 분절값
      TrainNN(input): 신경망 훈련
          input: 총 파일 길이 × 42 벡터
begin
      for i ← until 등장인물 수
          begin
              wav ←wav(1sec)
              values ←Mfcc(wav, fs, 2048)
              NN_input ←values
          end
end
begin
      TrainNN(NN_input)
end
```

표 1에서 제안된 음성 인식 알고리즘을 설명하고 있다. 출력된 음향벡터는 그림 3에서 표시된 신경망에 입력된다. 신경망 모듈을 위해서는 Matlab에서 제공되는 신경망 툴박스를 사용하였으며, Bayesian regularization을 사용하여 훈련 데이터를 모델링 하였다.

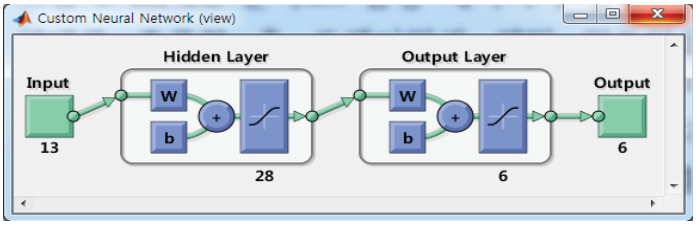

그림 3 Neural Network 구조도. Hidden neuron의 개수는 입력벡터 크기+15개로 설정하였고 Gradient 가 1.00e-05가 되었을 때 종료가 되도록 설정하였다.

2.4 실험 방법

실험용 비디오 음성데이터의 총 길이는 총 8분 44초이다. 이 데이터는 각 인물들이 최소한 1분 30초 이상 이야기를 하는 구간을 모아 재구성한 음성데이터이다. 유음과 무음으로 구분된 이 데이터를 훈련과 동일하게 1초 단위로 나눠 MFCC 음향벡터를 생성한다. 그 후 음향벡터에 에너지 기반 모델 (Energy-based model)[10]을 적용하여 음성데이터가 유음인지 무음인지 판단한다. 음성데이터가 놓인 환경에 따라 다르지만 약 95% 성능으로 무음과 유음을 구분함을 확인하였다. 그 후 유음 구간으로 판정된 데이터를 이용하여 6명의 화자에 대한 MFCC 음향벡터 값을 확보하였다. 각 화자별로 출력된 결과를 총 파일길이×42 크기의 행렬로 재구성하였다. 이 행렬은 신경망에 입력되어 화자인식을 수행하게 된다.

3. 실험 결과 및 토의

표 2 모델별 비교

모델	화자인식률
가우시안 혼합 모델	67%
가우시안 혼합 신경망 모델	71%
제안된 모델	80%

표 2는 각 모델별 성능을 비교한 표이다. 표에서 볼 수 잇듯이 제안된 모델에서는 80%의 화자인식률을 보여준다. 각 모델별 성능차이는 가우시안 혼합모델의 사용여부에 따라 달라짐을 볼 수 있다. 그 이유로는 가우시안 혼합모델을 이용하게 되면 실험데이터는 로그 우도 (likelihood) 값으로 출력이 되고 이는 본래 1초단위의 MFCC 음향벡터의 값들이 손상된 하나의 값으로 축소가 되기 때문이다. 이러한 데이터의 축소는 데이터 손실을 발생 시키고 손실에 대한 처리가 이루어지지 않는다면 화자인식률은 감소하게 되는 것이다.

우리는 기존 연구들과 달리 환경 잡음을 고려하지 않고 실험을 진행했음에도 80% 이상의 인식 성능을 달성하는데 성공하였다. 만일 비디오의 잡음을 제거 할 수 있다면 화자인식률을 더욱 높일 수 있을 것이다. 향후 연구에서는 독립성분분석 (ICA)를 통하여 노이즈를 제거하고, 환경에 영향을 받지 않는 화자인식 알고리즘으로 발전시킨 후 인물들의 대화 구간 추정을 통해 비디오를 설명할 수 있는 식별자를 생성하고자 한다. 이렇게 만들어진 비디오 식별자는 새로운 비디오 추천 알고리즘의 기반 구조로 사용될 것이다.

참고문헌

[1] P. Delacourt, and C.J. Wellekens, DISTBIC: A speaker-based segmentation for audio data indexing, *Speech Communication*, Vol. 32, No. 1-2, pp. 111-126, 2000

[2] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, Query by image and video content: The QBIC system, *Computer*, Vol. 28, No. 9, pp. 23-32, 1995.

[3] K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura, Video Handling with Music and Speech Detection, *Multimedia*, Vol. 5 No. 3, pp. 17-25, 1998

[4] H. Li, D. Doermann, and O. Kia, Automatic text detection and tracking in digital video, *Image Processing*, Vol. 9, No. 1, pp. 147-156 , 2000

[5] A. Hauptmann, R. Jin and T. D.Ng, Video Retrieval using Speech and Image Information, In Storage and Retrieval for Multimedia Databases 2003, EI'03 Electronic Imaging, p 148, 2003

[6] J.-P. Poli, An Automatic Television Stream Structuring System for Television Archives Holders, *Multimedia Systems*, Vol. 14, No. 5, pp.255-275, 2008

[7] G. Manson and S.-A. Berranim, Automatic TV Broadcast Structuring, *International Journal of Digital Multimedia Broadcasting*, vol. 2010, Article ID 153160, 2010

[8] S. Chakroborty, A. Roy and G. Saha, Improved Closed Set Text-Independent Speaker Identification by combining MFCC with Evidence from Flipped Filter Banks, *International Journal of Signal Processing*, Vol. 4, No. 2, pp.114-121, 2007

[9] J. W. Ko, W. J. Song, Analysis-by-synthesis Homomorphic Vocoder Using Mel-cepstrum, *대한전자공학회 학술대회 논문집*, 7권, 1호, pp. 284-287, 1994

[10] T. Kemp, M. Schmidt, Martin Westphal, and Alex Waibel, Strategies for Automatic Segmentation of Audio Data, *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Vol. 3, pp. 1423-1426, 2000