제 18 회 한국정보처리학회 추계학술발표대회 논문집 제 9 권 제 2 호 (2002.11)

오디오-비디오 정보 융합을 통한 멀티 모달 음성 인식 시스템

이상운*, 이연철**, 홍훈섭*, 윤보현*, 한문성*
*한국전자통신연구원
**㈜휴먼미디어테크
E-mail : lsu63479@etri.re.kr

Audio-Visual Integration based Multi-modal Speech Recognition System

Sahng-Woon Lee*, Yeon-Chul Lee**, Hun-Sop Hong*, Bo-Hyun Yun*, Mun-Sung Han*
*Electronics and Telecommunications Research Institute
**Human Media Tech. Inc.

요 약

본 논문은 오디오와 비디오 정보의 융합을 통한 멀티 모달 음성 인식 시스템을 제안한다. 음성 특징 정보와 영상 정보 특징의 융합을 통하여 잡음이 많은 환경에서 효율적으로 사람의 음성을 인식하는 시스템을 제안한다. 음성 특징 정보는 멜 필터 캡스트럼 계수(Mel Frequency Cepstrum Coefficents: MFCC)를 사용하며, 영상 특징 정보는 주성분 분석을 통해 얻어진 특징 벡터를 사용한다. 또한, 영상 정보 자체의 인식률 향상을 위해 피부 색깔 모델과 얼굴의 형태 정보를 이용하여 얼굴 영역을 찾은 후 강력한 입술 영역 추출 방법을 통해 입술 영역을 검출한다. 음성-영상 융합은 변형된 시간 지연 신경 회로망을 사용하여 초기 융합을 통해 이루어진다. 실험을 통해 음성과 영상의 정보 융합이 음성 정보만을 사용한 것 보다 대략 5%-20%의 성능 향상을 보여주고 있다.

1. 서론

사람의 음성 인지 능력은 화자의 소리를 듣는 것 뿐만 아니라 화자의 입 모양을 봄으로써 더욱더 높아진다 [1]. 자동 음성 인식 시스템(Automatic Speech Recognition: ASR)의 오디오-비디오 정보 융합에 대한 연구는 지금까지 널리 연구되어지고 있다. 특히, 잡음이 많은 환경에서 효율적으로 사용자의 음성을 인식하기 위해서는, 사용자의 영상 정보에 의존해야 할 것이다. 즉, 컴퓨터는 입술의 움직임과 같은 정보를 분석함으로써 음성 인식 성능을 향상시켜야 할 것이다 [2].

음성-영상 융합 시스템은 음성 특징 추출 모듈, 영상 특징 추출 모듈과 인식기 모듈로 3 부분으로 대개 구성되어져 있다. 그러나, 어떤 음성-영상 특징 벡터가 더 효과적이며, 음성 정보와 영상 정보가 어떻게 융합 되어져야 하는 것에 대한 분명한 통계 결과는 나오지 않은 상태이다. 지금까지 개발된 오디오-비디오 융합을 통한 인식 방법은 크게 은닉 마코프 모델 (Hidden Markov Models: HMM)과 시간 지연 신경회로망 (Time Delay Neural Network: TDNN)을 많이 사용하였으며,이 두 방법을 결합한 하이브리드 (Hybrid) 형태도 연구되어져 있는 상태이다 [3-5].

본 논문에서는 오디오-비디오 멀티 모달 융합 음성 인식 시스템을 제안한다. 제안된 시스템은 오디오 비디오의 특징 정보의 융합을 통해 잡음이 많은 환경에서 음성 인식률을 향상 시킨다. 변형된 시간 지연 신경회로망을 사용하여 음성과 영상 정보를 초기 융합을 통해 인식한다.

본 논문의 구성은 다음과 같다. 2 장에서 제안된 음성-영상 융합 시스템의 전체 구조도, 영상 특징 추출 방법과 음성 특징 추출 방법을 설명하고, 3 장에서 제안된 시스템의 실험 환경 및 결과에 관하여 설명하고, 4 장에서 결론 및 향후 과제에 관하여 살펴본다.

2. 제안된 음성/영상 융합 시스템

본 논문에서 제안된 음성/영상 융합 시스템은 그림 1 과 같이 크게 끝점 검출 모듈, 음성 특징 추출 모듈, 영상 특징 추출 모듈과 인식 모듈로 구성되어 있다. 끝점 검출 모듈은 마이크로 입력된 음성 신호에서 음성의 시작 부분과 끝 부분을 검출하는 모듈이며, 음성 및 영상 특징 추출 모듈은 인식에 용이한 각각의 특징을 추출하는 모듈로써, 본 시스템에서의 음성 특징 벡터는 멜 필터 캡스트럼 계수(Mel Frequency Cepstrum Coefficents: MFCC)를 사용하며, 영상 특징은 입술 영역에 대한 주성분 분석을 통한 구해진 가중치 벡터를 사용한다. 또한, 음성 정보와 영상 정보의 융합은 TDNN 을 통해 초기 융합을 통해 사람의 음성을 인식한다.

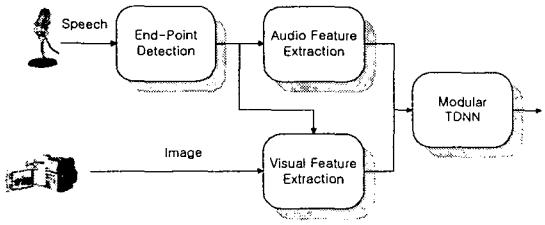


그림 1. 오디오/비디오 융합 시스템의 전체 구조도

2.1 음성 특징 추출

음성 특징 추출은 음성 끝점 검출기를 통해 전달된 음성 신호는 너무나 많은 양의 데이터를 포함하고 있어 실시간 인식 처리에 접합하지 않기 때문에 음성 신호를 인식에 용이한 적은 양의 데이터를 추출하는 것이다.

본 시스템에서 사용된 음성 특징 벡터는 음성인식에 가장 널리 사용되는 MFCC 를 사용하며, 16KHz 로 샘플링하고 16 비트로 양자화된 음성 데이터를 한 프레임을 30ms 로 하며, 매 프레임을 10ms 마다 이동하며 12 차의 MFCC 와 energy 와 1 차, 2 차 델타 값을 사용하여 총 39 차원의 벡터를 추출한다. 또한, pre-emphasis 상수는 0.97 이며 hamming 윈도우를 사용하며, 26 개의 필터뱅크 채널을 사용한다.

2.2 영상 특징 벡터 추출

영상 특징 벡터 추출은 입력 영상에서 얼굴 영역과 입술 영역을 검출하고 입술 영역의 정보를 추출한다. 얼굴 영역과 입술 영역 검출은 칼라와 형태 정보를 통하여 이루어지며, 입술 영역의 정보 추출은 주성분 분석을 통하여 이루어진다. 본 논문에서 제안된 영상 특징 벡터 추출은 그림 2 와 같이 구성되어진다. 우선, 입력 영상에서 얼굴 영역을 검출하고 검출된 얼굴 영역 내에서 입술 영역 만을 추출한 후, 주성분 분석을 통해 영상 특징 벡터를 추출한다.

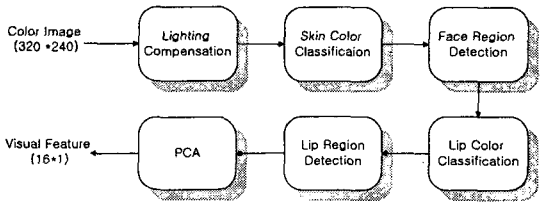


그림 2. 영상 특징 벡터 추출 구조도

입력 영상에서 얼굴 영역 검출은 3 단계를 거쳐 이루어 진다. 즉, 조명 보상, 피부 색깔 분류, 얼굴 영역 검출이다. 피부 색 칼라는 다양한 조명 조건에 따라 변할 수 있다. 이는 칼라에 기반 한 얼굴 영역 검출 성능의 저해를 가져온다. 이런 조명의 변화를 완화하기 위해서, 카메라로부터 입력된 칼라 이미지는 조명 보상 모듈을 거친다. 조명 보상 기술은 "참조-백색" (reference-white)을 사용하여 칼라를 정규화하는 방법이다 [6]. 조명이 보상된 칼라 이미지는 피부 색깔 분류 모듈을 거친다. 입력 영상에서 피부 색깔을 분류하기 위해 피부 색깔 모델(skin color model)을 이용한다. 피부 색깔 모델은 색채 칼라 공간(Chromatic color space)에서 얼굴 피부 색깔의 Cr, Cg 성분이 2D-가우시안 모델(2D-Gaussian Model)을 따른다고 가정하고 근사화 시킨 모델이다 [7]. 피부 색깔 모델을 통해 얻어진 이진화 영상에서 얼굴 영역을 검출하기 위해 얼굴의 형태 정보를 이용한다. 얼굴은 타원의 형태라는 가정을 통해 피부 색깔이 분류된 이진화 영상에서 몇 번의 반복을 통해 대략적인 타원을 찾아 얼굴 영역을 검출한다.

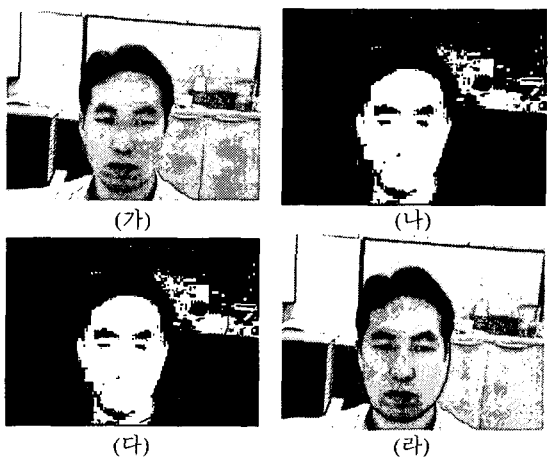


그림 3. 얼굴 영역 검출 과정

그림 3 는 얼굴 영역을 검출하는 과정을 보여준다. (가)는 입력 이미지이며, (나)는 피부 색깔 모델을 통해 얻어진 피부 색깔 분류 후의 이진화 영상이다. (다)는 초기 타원이며, (라)는 4 번의 반복 후에 얻어진 얼굴 영역 검출 결과이다. 사람 얼굴의 색깔 정보와 형태 정보의 결합을 통해 배경에 속한 피부 색깔을 제거할 수 있다.

타원 형태의 얼굴 영역이 검출되어지면 입술 영역 검출이 이루어진다. 입술 영역 검출은 사람의 입술 색은 채도(saturation) 공간에서 뚜렷하게 검출되어지는 성질을 이용한다. 본 논문에서는 사람의 입술은 타원의 아래 일정한 부분에 존재한다는 가정을 둔다. 이런 가정을 통해, 검출된 얼굴 영역의 아래 부분만을 추출하여 채도 공간으로 투영 한 후 히스토그램 평활화(histogram equalization) 과정을 채도 값의 분포가 일정하게 하게 한 후 일정한 범위의 채도 값을 갖는 입술 색을 분류 하고, 후처리 과정인 연결성 성분 분석(connected component analysis)을 통해 사각형 형태의 입술 영역을 검출한다.

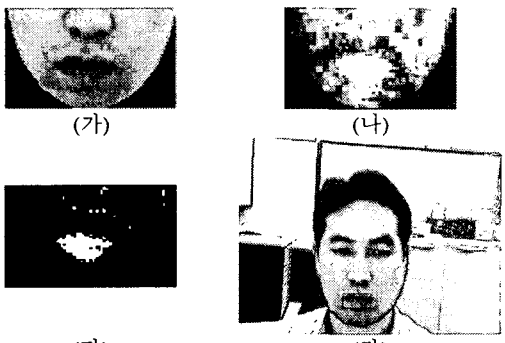


그림 4. 입술 영역 검출 과정

그림 4 는 입술 영역 검출 과정을 보여주고 있다. (가)는 타원의 아래 부분만을 추출한 영상이며, (나)는 (가)를 채도 공간에 투영한 후 히스토그램 평활화 과정를 수행 한 후의 영상이다. 채도 공간에서 입술 영역이 뚜렷하게 구별됨을 볼 수 있다. (다)는 입술 색깔 분류 후의 이진화 영상이며, (라)는 연결성 성분 분석을 통해 사각형 형태의 입술 영역을 검출한 결과 영상이다.

입술 영역이 검출되어지면 입술 영역의 크기를 일정한 크기로 정규화 시킨 후에 영상 특징 벡터 추출이 이루어진다. 영상 특징 벡터는 주성분 분석을 통해 입술 영역에 대한 고유공간(eigen space) 상에서 각 영역의 가중치 벡터(weight vector)를 사용한다 [8]. 주성분 분석은 데이터 집합의 공분산 행렬의 고유 벡터를 통해 심각한 정보 손실이 없이 낮은 차원으로 데이터를 표현하는 기술이다. M 개의 학습 입술 영역들로부터 M'개의 고유 벡터를 구하고, 새로운 입술 입력 영역은 직교공간에서 M'개의 고유 벡터들의 선형 조합을 구해 영상 특징 벡터로 사용한다.

2.3 음성 영상 융합

McGurk 효과는 인간의 음성 인지 과정에서 시각 정보에 상당한 영향을 받는다는 이론이다 [1]. 이런 이론을 바탕으로 오디오-비디오 융합 방법에 대한 많은 모델들이 제안되어지고 있다. 이런 모델은 크게 초기 융합(Early Integration)과 후기 융합(Late Integration)으로 나눌 수 있다. 초기 융합 모델에서, 음성과 영상 특징을 결합한 특징 벡터를 형성하기 위해 특징 공간(feature space)에서 융합이 이루어지며, 결합된 특징 벡터를 기반으로 인식을 수행한다. 또한, 후기 융합 모델에서는 음성과 영상 특징을 각각 먼저 인식한 후, 이 인식 결과를 융합함으로써 최종 인식을 수행하다. 초기 융합 모델이 후기 융합 모델보다 성능이 좀 더 우수함을 보인다.

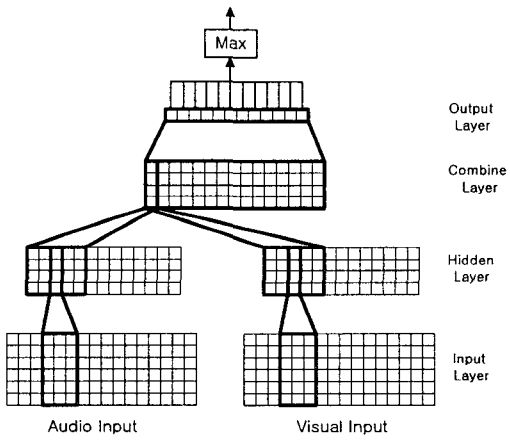


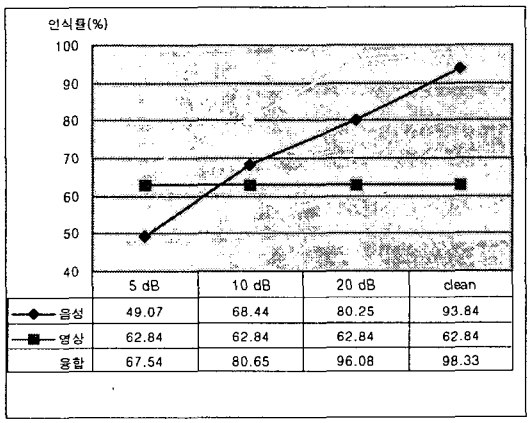
그림 5. 음성/영상 융합 TDNN

본 논문에서는 변형된 시간 지연 신경회로망을 사용하여 음성과 영상 정보를 초기 융합을 통해 인식한다. 즉, 음성과 영상 정보는 각각의 인식기를 통해 인식되지 않고, TDNN 의 결합층(combine layer)에서 음성과 영상 데이터의 융합이 이루어진다. 시간에 동기화 된 음성과 영상 특징 벡터는 각각의 입력층에 입력되어진다. 입력층과 은닉층(hidden layer)은 시간이 지연된 연결 구조를 가진 원도우를 사용한 기존 TDNN 의 방식을 그대로 사용하며, 결합층에서 동기화 된 음성 영상 데이터를 결합하여 출력층에서 인식 결과를 출력해 주는 구조이다. 그림 5 는 제안된 시스템에서 사용된 변형된 TDNN 을 기반 한 오디오-비디오 융합을 TDNN 인식기의 구조를 보여주고 있다. 기존의 TDNN 을 통한 음성 영상 융합 방식은 DTW 층을 거쳐 인식 결과를 출력해 주는 반면 [9], 본 논문에서는 이 DTW 층을 없애고 결합층에서 출력층으로 완전한 연결(fully connected)을 통해 인식을 수행하는 차이점을 가진다.

3. 실험 결과

제안된 시스템에 사용된 TDNN 의 입력은 음성, 영상 각각 정규화 된 16 차원의 64 개 프레임이며, 입력층의 윈도우 사이즈는 3 개의 프레임이다. 또한, 은닉층은 8 차원의 62 개 프레임이며, 윈도우 사이즈는 5 프레임이다. 결합층은 4 차원의 58 개 프레임과 윈도우 사이즈는 52 프레임이다.

TDNN 의 학습은 역전파 학습 알고리즘((back-propagation learning algorithm)을 사용하였으며, 훈련에 사용된 데이터는 남자 한 사람이 30 단어를 30 번 발음한 데이터를 TDNN 학습에 사용하였다. PC 카메라로부터 얻은 영상 데이터는 320 x 240 크기의 칼라 이미지를 초당 10 프레임 캡쳐 하였으며, PC 마이크를 사용하여 16KHz 로 샘플링하고 16 비트로 양자화 된 잡음이 없는 clean 한 음성 데이터를 사용하였다.



	5 dB	10 dB	20 dB	clean
음성	49.07	68.44	80.25	93.84
영상	62.84	62.84	62.84	62.84
융합	67.54	80.65	96.08	98.33

그림 6. 음성 영상 정보 융합 실험 결과

제안된 시스템의 성능을 평가하기 위해서 음성 데이터를 4 단계의 잡음 레벨(clean, 20dB, 10dB, 5dB)로 구성된 테스트 데이터를 사용하여 실험해 보았다. 그림 6 은 음성 데이터의 잡음 레벨에 따른 인식 결과를 나타내고 있다. 본 실험에서, 영상과 음성의 개별적인 인식 결과를 얻기 위해서 단일 TDNN 을 통하여 인식하였다. 그림 6 에서 보듯이, 음성과 영상의 정보 융합이 음성 정보만을 사용한 것 보다 대략 5%-20%의 성능 향상을 보여주고 있다. 또한, 잡음의 정도가 5dB 인 환경에서도 음성 영상 융합을 통한 인식률이 20%정도 향상 됨을 알 수가 있으며, 이는 잡음이 많은 환경에서 음성과 영상 정보 융합을 통해 음성 인식 성능이 향상 됨을 의미하는 것이다.

4. 결론

본 논문에서는 오디오와 비디오 정보의 융합을 통한 멀티 모달 음성 인식 시스템을 제안하였다. 제안된 방법은 변형된 시간 지연 신경회로망을 사용하여 음성과 영상 특징 벡터를 초기 융합을 통해 인식하였다. 음성 특징 벡터는 MFCC 를 사용하였으며, 효율적이고 정확한 입술 영역 검출을 위해 칼라와 형태 정보를 이용하여 얼굴 영역과 입술 영역을 검출하였다. 주성분 분석을 통해 검출된 입술 영역에 대한 고유 공간상에서 각 영역의 가중치 벡터를 영상 특징 벡터를 사용하였다. 제안된 방법을 통해 실험해 본 결과를 통해, 영상 자체의 인식률 향상 뿐만 아니라, 음성과 영상 정보 융합을 통해 잡음이 많은 환경에서도 음성 인식률이 20%정도 성능이 향상됨을 볼 수 있었다.

참고문헌

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices", Nature, pp. 746-748, 1976.

[2] T. Chen, "Audiovisual speech processing", IEEE Transactions on Signal Processing Magazine, pp. 9 -21, 2001.

[3] I.Meier, W. Hurst and P. Duchnowski, "Adaptive bimodal sensor fusion for automatic speechreading", Proceeding of ICASSP, pp. 833-837, 1996.

[4] S. Dupont and J. Leuttin, "Audio-Visual Speech Modeling for Continuous Speech Recognition", IEEE Transactions on Multimedia, pp. 141-151, 2000.

[5] A. Rogozan and P. Deléglise, "Visible Speech Modelling and Hybrid Hidden Markov Model_Neural Network Models Based Learning for Lipreading", Proceedings of the IEEE Symposia on Intelligence and Systems, pp. 336-342, 1998.

[6] R. L. Hsu, A. Mottaleb and Jain. A.K., "Face detection in color images", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 696-706, 2002.

[7] S. H. Park, E. Y. Kim, S. W. Hwang, Y. C. Lee and H. J. Kim, "Face detection for security system on the internet", Proceeding of IEEE Consumer Electronics, pp. 276-277, 2001.

[8] M. Turk and A. Pentland, "Face Recognition Using Eigenfaces", Proceeding of IEEE CVPR, pp. 586-591, 1991.

[9] U. Meler, R. Stiefelahagen, J. Yang and A. Waibel, "TOWARDS UNRESTRICTED LIP READING", International Journal of Pattern Recognition and Artificial Intelligence, pp. 571-585, 2000.