신경망을 이용한 이중모달 음성 인식 모델링

류정우[0] 성지애 이순신 김명원
숭실대학교 컴퓨터학부
ryu0914[0]@orgio.net, mkim@comp.ssu.ac.kr

Bimodal Speech Recognition Modeling Using Neural Networks

Joung-Woo Ryu[0] ,Ji-Ae Sung, Sunshine Lee, Myoung-Won Kim
School of Computing, Soonsil University

요 약

최근 잡음환경에서 강인한 음성인식을 위해 음성 잡음에 영향을 받지 않은 영상정보를 이용한 이중모달 음성인식 연구가 활발히 진행되고 있다. 기존 음성인식기로 좋은 성능을 보이는 HMM은 이질적인 정보를 융합하는데 있어 많은 제약과 어려움을 가지고 있다. 하지만 신경망은 이질적인 정보를 효율적으로 융합할 수 있는 장점을 가지고 있으며 그에 대한 많은 연구가 수행되고 있다. 따라서 본 논문에서는 잡음환경에 강인한 이중모달 음성 인식 모델로 이중모달 신경망(BN-NN)[1]을 제안한다. 이중모달 신경망은 특징융합 방법으로 음성정보와 영상정보를 융합하고 있으며, 입력정보의 특성을 고려하가 위해 윈도우와 중복영역의 개념을 적용하여 시제위치를 고려하도록 설계되어있다. 제안된 모델은 잡음환경에서 음성인식기와 성능을 비교하고, 화자독립 고립단어 인식에서 기존 융합방법인 CHMM과 비교하여 그 가능성을 확인한다.

1. 서 론

최근 들어 사회가 점차 멀티미디어화함에 따라 인간과 기계의 인터페이스를 좀 더 간편하고 정확하게 실현하기 위하여 얼굴표정이나 방향, 입술모양, 응시추적, 손동작 그리고 음성 등을 이용한 멀티모달(multimodal)형태의 인식연구가 활발히 진행되고 있다. 이러한 연구의 궁극적인 목적은 잡음에 강인한 음성인식기 개발에 있다. 즉 음성 잡음에 영향을 받지 않은 정보들을 찾아내어 융합함으로서 실생활에 효율적으로 이용할 수 있는 음성인식기를 개발하는 것이다.

특히, 음성만으로 인식할 수 없는 잡음 환경에서 사람들은 일반적으로 상대방의 입술 모양을 보고 음성을 인식한다. 이와 같이 최근 음성정보와 영상정보를 융합한 이중모달(bimodal) 음성 인식 방법이 활발하게 연구되고 있다. 이러한 이중모달 음성 인식 방법에서 가장 중요한 연구주제는 음성정보와 서로 보완적인 형태를 이루고 있는 영상정보를 얼마나 잘 추출하는 것과 이질적인 두 정보를 얼마나 효율적으로 융합하는 것이다.

융합방법으로는 기존 음성인식기로 좋은 성능을 보이는 HMM(Hidden Markov Model)을 이용한 방법들이 많이 연구되고 있다. 하지만 융합방법에 있어 HMM은 입력특징들이 확률적 독립성이란 조건을 만족해야 하는 제약사항들이 있어 적용하기가 쉽지 않으며 학습 변수 예를 들면, 상태(state) 수와 가우시안 혼합(Gaussian mixture) 수를 결정하기가 어렵다는 문제점을 가지고 있다[1]. 반면, 신경망(Neural Network)은 이질적인 정보를 효율적으로 융합할 수 있는 장점을 가지고 있다[2].

따라서 본 논문에서는 신경망을 이용하여 단어 인식을 위한 이중모달 음성 인식 모델링인 이중모달 신경망(BM-NN : BiModal Neural Network)을 제안한다.

본 논문의 구성을 살펴보면, 2절에서는 기존의 이중모달 음성 인식 모델인 MS-TDNN과 HMM을 이용한 모델을 살펴보고, 3절에서는 제안한 BM-NN에 대해 기술한다. 4절에서는 잡음환경에서 음성인식기와 융합인식기의 성능을 비교분석하고 HMM을 이용한 융합모델인 CHMM(Coupled Hidden Markov modal)과 비교하여 그 가능성을 보인다. 마지막으로 5절에서는 결론 및 향후연구에 대해 검토한다.

2. 관련연구

이중모달 음성 인식을 위한 융합 방법은 이질적인 정보를 융합하는 시점에 따라 특징융합(feature fusion)과 결정융합(decision fusion)으로 나누어진다. 특징융합은 인식하기 전에 정보를 융합하는 방법을 의미하고 결정융합은 인식된 결과를 융합하여 최종 인식을 수행하는 방법을 의미한다. [3]에서는 특징융합이 결정융합보다 더 높은 인식률을 보여주고 있다. 이러한 방법에서 사용되는 융합모델은 HMM과 신경망이 일반적이다. 특히, 신경망 중 TDNN(Time-Delay Neural Network)은 음소의 지속시간 및 음성신호내의 시제 위치 등의 다양한 조건에서도 상당히 정확하게 음소를 인식할 수 있는 구조로 음성인식에서 많이 사용되고 있는 모델이다.

기존 이중모달 음성인식 모델을 살펴보면 다음과 같다.

2.1 MS-TDNN

MS-TDNN(Multi State TDNN)[4]은 결합융합 방법을 사용하여 음성정보와 영상정보를 융합하고 있다. MS-TDNN은 두 단계 학습과정을 통해 단어를 인식하게

1) 본 연구는 정보통신 선도기반기술개발사업에 의하여 수행되었습니다.

된다. 첫 번째 학습과정은 음소단위로 이루어지며 음성정보와 영상정보 각각에 대해 독립적인 TDNN 인식기를 생성한다. 두 번째 학습과정은 단어단위로 DTW(Dynamic Time Wrapping)에서 가장 적합한 단어에서부터 각 TDNN 출력층 까지 역전파(backpropagation) 알고리즘을 통해 학습이 이루어진다. 이처럼 MS-TDNN은 음소레벨에서 단어를 인식해야 함으로 시간 축 변화(time axis variation) 문제를 해결하기 위해 DTW알고리즘이 요구됨으로 보다 복잡한 모델이 생성될 뿐만 아니라 음소인식에서 어려운 점인, 잡음에 민감하고 음소간의 구분이 어렵다는 문제점을 가지고 있다.

2.2 HMM을 이용한 융합.

[1]에서는 HMM으로 특징융합 방법을 사용하여 음성정보와 영상정보를 융합하고 있다. 특징융합은 음성정보와 영상정보의 표본비율(sampling rate)이 다르기 때문에 융합하기가 결정융합 보다 어렵다. 이러한 동기화 문제를 [1]에서는 보간법(low-pass interpolation)을 사용하여 표본을 추출하고 새로운 특징은 10msec가 중복된 25msec 윈도우로부터 생성함으로써 해결하고 있다. 생성된 새로운 특징들을 이용하여 HMM을 학습한다.
이와 같이 HMM을 이용한 이질적인 정보융합은 HMM의 학습변수, 가우시한 혼합 수와 상태 수를 결정하기가 어려운 문제점을 가지고 있다.

2.3 CHMM

[3]에서는 HMM을 이용하여 특징융합이나 결정융합을 하였을 경우 문제점을 기술하고 있다. 즉, 효율적인 결정융합을 위해 음성정보와 영상정보가 서로 독립적이어야 하며 연관성이 결여되어야 하는 가정을 만족해야만 한다. 또한 특징융합은 음성정보와 영상정보의 동기화 문제점을 해결해야만 한다. 따라서 이와 같이 다중 시계열 융합모델에 대해 특징융합이나 결정융합은 적합하지 않으므로 CHMM을 제안하고 있다.

3. BM-NN

본 논문에서는 효율적으로 단어를 인식할 수 있는 이중 모달 음성 인식 모델을 신경망을 이용하여 제안한다. 제안된 BM-NN의 구조는 <그림.1>과 같다.

BM-NN은 네 계층 즉, 입력층, 은닉층, 융합층, 출력층으로 구성되어 있으며 전방향 신경망 구조를 가지고 있다. 역전파 알고리즘을 사용하여 학습이 이루어지고 학습단위와 인식단위는 고립단어이다. 따라서 고립단어 인식에 대해 인식률이 높은 중복 영역(overlap zone) 구조[5]를 사용한다. 또한 입력층과 은닉층은 윈도우(window) 개념을 사용함으로써 입력정보를 시제위치를 고려하여 함축하고 이를 상위계층으로 전파한다. 입력정보는 시간에 따라 샘플(sample)한 프레임(frame)별로 특징을 추출한다. 따라서 윈도우는 한개 이상의 프레임들로 구성되어 있다.

연결구조를 살펴보면 윈도우에 포함된 모든 프레임들의 노드들과 대응되는 상위계층의 프레임의 노드들이 완전연결(fully connect)로 연결되어 있으며 융합층은 윈도우 개념이 없기 때문에 출력층과 완전연결로 연결되어 있다. 그러므로 윈도우 크기와 중복영역의 크기가 결정되면 자동으로 각 층의 프레임 개수가 결정된다.

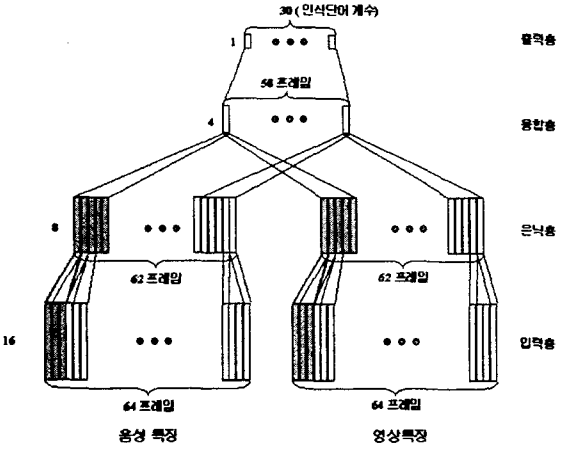

그림 1 . BM-NN(BiModal Neural Network) 구조

BM-NN은 인식단위가 단어이기 때문에 시간 축 변화 문제를 해결할 필요가 없음으로 음소단위인 MS-TDNN보다 학습방법과 구조가 간단하다. 또한 이질적인 정보를 융합하기 위해 HMM은 입력정보가 만족해야 하는 제약사항들이 있어 적용하기 어려운 문제점을 본 논문에서는 이질적인 정보를 효율적으로 융합할 수 있는 신경망을 이용하여 해결하고 있다. 하지만 특징융합 방법을 적용하고 있는 BM-NN 역시, 음성정보와 영상정보를 동기화해야 하는 문제점을 가지고 있다. 따라서 본 논문에서는 이미지를 버퍼에 시간정보(system tick)와 함께 저장하고 동시에 음성신호가 입력되며, 입력된 음성 신호는 끝점 검출 과정을 통해 단어로 세그멘테이션 된다. 이때, 끝점 검출 시간과 같은 시간을 나타내는 시점을 계산하고 영상 버퍼로부터 동일한 시간(시작시간 ~ 끝 시간)에 입력된 영상을 버퍼로부터 읽어 들임으로써 동기화 문제를 해결한다.

4. 실험

실험에 사용한 데이터는 ETRI에서 제작한 데이터를 사용한다. 남성 28명이 1회 발음한 78개의 고립단어에 대한 데이터이다. 단어들은 이동단말기에서 사용될 수 있는 명령어들로 구성되어 있다. 예를 들어 "앞으로", "뒤로","정지","선택", "1번항목", "2번리스트" 등이다.

모델을 형성하기 위해 21명의 데이터를 학습데이터로 사용하고 7명의 데이터를 테스트데이터로 사용한다.

음성정보는 ZCPA (Zero-Crossings with Peak-Amplitudes)[6]에 의해, 영상정보는 PCA(Principle Components Analysis)에 의해 특징이 추출된다.

본 실험에서 고립단어 인식을 위해 입력 프레임을 64프레임(:프레임당 10ms)으로 설정하고 각 프레임에서 16차원의 특징을 추출한다. 입력층의 윈도우 크기는 음소를 표현하기에 충분한 30ms인 3프레임으로 설정하고 중

복영역 크기는 2프레임으로 설정한다. 은닉층의 윈도우 크기는 더 넓은 시제 영역을 학습할 수 있도록 입력층 윈도우 크기보다 크게 5프레임으로 설정하고 중복영역 크기는 4프레임으로 설정한다. 따라서 각 층의 프레임의 개수는 <그림.1>과 같이 설정된다. 단, 출력층의 노드는 78개로 인식할 단어의 개수와 같다.

본 실험에서 제안한 융합인식기 BM-NN이 잡음에 강인한 인식기인지 알아보기 위해 가우시안 잡음(20db, 10db, 5db SNR)을 발생시켜 음성신호와 혼합하고, 융합인식기(BM-NN)와 음성인식기의 성능을 비교한다. 실험에 사용된 음성인식기는 BM-NN에서 영상특징을 0으로 입력하여 모델을 생성하고 테스트 한다.

잡음환경에서 음성인식기와 융합인식기의 성능비교 실험은 <그림.2>와 같이 30db에서는 융합 결과(73.99%)와 음성(73.44%) 결과 차이가 없지만 잡음이 증가할수록 음성결과보다 융합결과가 완만하게 감소하여 5db인 경우 15.2% 정도 융합인식기가 영상정보의 도움을 받아 더 나은 결과를 보이는 것을 알 수 있다.

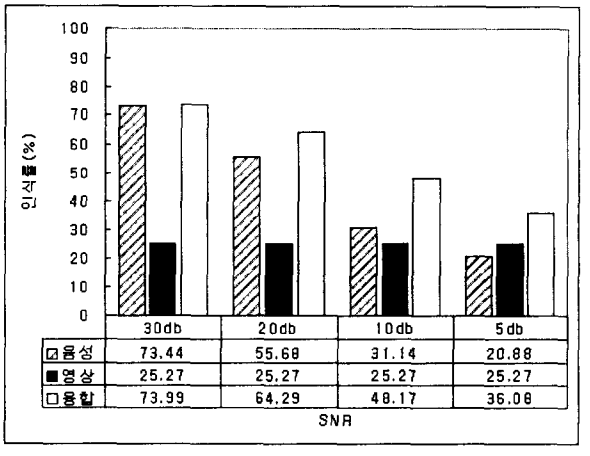

그림 2 . 음성인식과 융합인식 비교

<표.1>에서는 [3]에서 제안한 HMM을 이용한 융합인식기인 CHMM과 성능을 비교한 결과를 나타낸다. 각각의 인식기에 사용한 데이터가 틀리기 때문에 정확한 비교분석은 이루어 질 수 없지만, 인식단어의 개수를 같게 하고 화자독립 실험을 하였기 때문에 본 논문에서 제안한 신경망을 이용한 융합인신기인 BM-NN의 가능성을 확인할 수 있다.

표 1. BM-NN과 CHMM 비교

SNR	BM-NN	CHMM[3]
10db	48.17 %	38.43 %
30db	73.99 %	66.17 %

5. 결론 및 향후연구 계획

본 논문에서는 잡음환경에 강인한 음성 인식 모델로 이중모달 신경망, BM-NN을 제안한다. 제안한 모델은 특징융합 방법으로 음성정보와 영상정보를 융합하며, 이질적인 정보들을 효율적으로 융합할 수 있는 신경망을 이용한다. BM-NN은 MLP(Multi-Layer Perceptron)에서 이질적인 정보를 융합하기 위해 융합층을 추가하였으며, 입력정보의 특성을 고려하기 위하여 입력층과 은닉층에 윈도우와 중복영역의 개념을 적용하여 시제위치를 고려한다. 상위 계층일수록 더 넓은 시제 영역을 학습할 수 있도록 윈도우와 중복영역의 크기를 더 크게 설정한다. 제안된 융합모델의 타당성을 검증하기 위해 잡음환경에서 음성인식기와 융합인식기의 성능을 비교하였고 또한, 화자독립 실험에서 인식단어 개수를 같게 하여 CHMM 모델과 성능을 비교하고 그 가능성을 확인하였다.

향후연구로는 같은 데이터를 이용하여 기존의 모델과 비교 분석함으로써 모델의 타당성을 확인하고 잡음환경에 영향을 받지 않는 영상정보이외의 정보를 찾아 같이 BM-NN과 융합할 수 있는 멀티융합방법을 연구할 계획이다.

6. 참고문헌

[1] Kaynak, M.N.; Qi Zhi; Cheok, A.D.; Sengupta, K.; Ko Chi Chung; "Audio-visual modeling for bimodal speech recognition", Systems, Man, and Cybernetics, 2001 IEEE International Conference on , Vol. 1, 2001, pp.181-186

[2] Gemello, R.; Albesano, D.; Mana, F.; Moisa, L.; "Multi-source neural networks for speech recognition: a review of recent results", Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on , Vol. 5, 2000, pp. 265-270

[3] Xiaozheng Zhang; Merserratt, R.M.; Clements, M.; , " *Bimodal fusion in audio-visual speech recognition* ", Image Processing. 2002. Proceedings. 2002 International Conference on ,Vol.1,2002, pp.964-967

[4] C.Bregler, S.Manke, H.Hild and A. Waibel, "Bimodal sensor integration on the example of "speech-reading", Proc. of IEEE Int. Conf. on Neural Networks, San Francisco, 1993

[5] Mary Jo Creaney-Stockton, Beng., MSc., "Isolated Word Recognition Using Reduced Connectivity Neural Networks With Non-Linear Time Alignment Methods", Dept. of Electrical and Electronic Engineering Univ. of Newcastle-Upon-Tyne, August 1996

[6] Doh-Suk Kim,Soo-Young Lee, Rhee M. Kil, "Auditory Processing of Speech Signals for Robust Speech Recognition *in Real-World Noisy Environments*", IEEE Trans. on Speech and Audio Processing, Vol.7, No.1, January 1999, pp.55-69

[7] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. J. Lang, "Phoneme Recognition Using Time-Delay Neural Networks", IEEE Trans. on Acoustics, Speech and Signal Processing. Vol.37, NO.3, March 1989. pp.328-339