논문 2004-41CI-3-8

신경망 기반 음성, 영상 및 문맥 통합 음성인식

(Speech Recognition by Integrating Audio, Visual and Contextual Features Based on Neural Networks)

김 명 원*, 한 문 성**, 이 순 신***, 류 정 우*

(Myung Won Kim, Mun-Sung Han, Sunshine Lee, and Joung Woo Ryu)

요 약

최근 잡음환경에서 신뢰도 높은 음성인식을 위해 음성정보와 영상정보를 융합하는 방법이 활발히 연구되고 있다. 본 논문에서는 이절적인 정보의 융합에 적합한 신경망 모델을 기반으로 음성, 영상 및 문맥 정보 등 다양한 정보를 융합하여 잡음 환경에서 고립단어를 인식하는 음성인식 기법에 대하여 기술한다. 음성과 영상 특징을 이용한 이중 모드 신경망 BMNN(BiModal Neural Network)을 제안한다. BMNN은 4개 층으로 이루어진 다층퍼셉트론의 구조를 가지며 각 층은 입력 특징의 추상화 기능을 수행한다. BMNN에서는 제 3층이 잡음에 의한 음성 정보의 손실을 보상하기 위하여 음성과 영상 특징을 통합하는 기능을 수행한다. 또한, 잡음환경에서 음성 인식률을 향상시키기 위해 사용자가 말한 단어들의 순차 패턴을 나타내는 문맥정보를 이용한 후처리 방법을 제안한다. 잡음환경에서 BMNN은 단순히 음성만을 사용한 것 보다 높은 성능을 보임으로써 그 타당성을 확인할 수 있을 뿐 아니라, 특히 문맥을 이용한 후처리를 하였을 경우 잡음 환경에서 90%이상의 인식률을 달성하였다. 본 연구는 잡음환경에서 강인한 음성인식을 위해 다양한 추가 정보를 사용함으로써 성능을 향상시킬 수 있음을 제시한다.

Abstract

The recent research has been focused on fusion of audio and visual features for reliable speech recognition in noisy environments. In this paper, we propose a neural network based model of robust speech recognition by integrating audio, visual, and contextual information. Bimodal Neural Network(BMNN) is a multi-layer perceptron of 4 layers, each of which performs a certain level of abstraction of input features. In BMNN the third layer combines audio and visual features of speech to compensate loss of audio information caused by noise. In order to improve the accuracy of speech recognition in noisy environments, we also propose a post-processing based on contextual information which are sequential patterns of words spoken by a user. Our experimental results show that our model outperforms any single mode models. Particularly, when we use the contextual information, we can obtain over 90% recognition accuracy even in noisy environments, which is a significant improvement compared with the state of art in speech recognition. Our research demonstrates that diverse sources of information need to be integrated to improve the accuracy of speech recognition particularly in noisy environments.

Keywords: 신경망(neural network), 이중모드(bimodal), 문맥 정보(context information), 융합방법(fusion method), 후처리(post-processing)

I. 서 론

최근 들어 사회가 점차 멀티미디어화 됨에 따라 인간과 기계의 인터페이스를 좀 더 간편하고 명확하게 실현하기 위하여 얼굴표정이나 방향, 입술모양, 응시추적, 손동작 그리고 음성 등을 이용한 다중모드(multi-modal) 형태의 인식 연구가 활발히 진행되고 있다.

* 정회원, 숭실대학교 컴퓨터학부
(School of computing, Soongsil University)
** 정회원, 한국전자통신연구원 디지털 홈 연구단
(Digital home research division, ETRI)
*** 정회원, LG-CNS 기술연구부문 솔루션센터 DW/BI팀
(DW/BI Team, Solution Center, IT Group Div. , LG-CNS)
※ 본 연구는 숭실대학교 교내연구비 지원으로 이루어졌음
접수일자: 2003년6월21일, 수정완료일: 2004년5월6일

특히, 이러한 연구는 최근 이동 단말기의 기술이 발전함에 따라 잡음환경에 강인한 음성인식 방법인 이중모드(bimodal) 음성인식 방법으로 활발히 연구되고 있다. 이중모드 음성인식 방법이란 잡음환경에 민감한 음성정보를 보완할 수 있는 영상정보를 동시에 고려함으로써 음성인식률을 향상시키는 방법이다. 예를 들어, 공장과 같은 시끄러운 환경에서 대화할 때 사람들은 서로의 음성뿐만 아니라 입모양 혹은 제스쳐와 같은 영상정보를 이용하여 음성을 잘 인식하는 경우를 생각할 수 있다.

이중모드 음성 인식 방법에서 가장 중요한 연구 주제는 음성정보와 서로 보완적인 형태를 이루고 있는 영상정보를 얼마나 잘 추출하느냐와, 이질적인 두 정보를 얼마나 효율적으로 융합하느냐에 있다. 본 논문에서는 두 번째 주제인 정보 융합방법에 대해서만 기술한다.

기존 융합 방법은 이질적인 정보를 융합하는 시점에 따라 특징 융합(feature fusion)과 결정 융합(decision fusion)으로 나누어진다[1]. 특징 융합은 인식하기 전에 특징 정보를 융합하는 방법을 말하고 결정 융합은 인식된 결과를 융합하여 최종 인식을 수행하는 방법을 말한다. 효율적인 융합을 위해 특징 융합은 음성정보와 영상정보의 동기화 문제를 해결해야 하는 반면 결정 융합은 음성정보와 영상정보가 서로 독립적일 때 융합의 효과가 크다. 따라서 결정 융합은 특징 융합에 비해 적용은 쉬우나 성능이 떨어지고 특징 융합은 성능은 좋으나 입력정보의 동기화를 고려하여야 하므로 적용하기가 어렵다. 이러한 융합 방법들로는 HMM(Hidden Markov Model)과 신경망이 일반적으로 많이 사용된다.

[2]에서는 HMM을 이용한 특징 융합 방법으로 음성과 영상정보를 융합하였다. 특징 융합은 음성과 영상정보의 표본비율(sampling rate)이 다르기 때문에 결정 융합에 비해 융합하기가 어렵다. 이러한 동기화 문제를 해결하기 위해 저주파 통과 보간법(low-pass interpola-tion)을 사용하여 표본을 추출하였고, 새로운 특징은 10msec가 중복된 25msec윈도우로부터 생성하였다. 그러나 HMM을 이용한 융합 방법에 있어 결과에 민감한 반응을 주는 학습 변수인 상태(state) 수와 가우시안 혼합(Gaussian mixture) 수를 결정하기 어렵고, 특히 일반적으로 사용되는 CDMM(Continuous Density Hidden Markov Model)은 입력특징들이 확률적 독립성 조건을 만족해야 하는 제약사항들이 있어 적용하기 어렵다[3][4].

신경망 중 TDNN(Time-Delay Neural Network)은 음소의 지속시간 및 음성 신호 내의 시제 위치 등 다양한 조건에서도 상당히 정확하게 음소를 인식할 수 있는 신경망 모델이다[5]. MS-TDNN(Multi State TDNN)은 DTW(Dynamic Time Warpping)층을 추가하여 연속 단어를 인식할 수 있도록 TDNN을 확장한 모델이다[6][7]. 이러한 MS-TDNN을 이용하여 음성정보와 영상정보를 융합한 이중모드 MS-TDNN이 개발되었다[8].

이중모드 MS-TDNN은 두 단계 학습과정을 통해 모델이 형성된다. 첫 번째 학습과정은 음소단위로 이루어지며 음성정보와 영상정보 각각에 대해 독립적인 TDNN 인식기를 생성한다. 두 번째 학습과정은 고립단어 단위로 DTW에서 가장 적합한 단어에서부터 각 TDNN 출력층까지 역전파(backpropagation) 알고리즘을 통해 학습이 이루어진다.

이와 같이 이중모드 MS-TDNN은 음소레벨에서 단어를 인식해야 하므로 시간 축 변화(time axis varia-tion) 문제를 해결하기 위한 DTW 알고리즘이 요구된다. 그러므로 보다 복잡한 모델이 생성될 뿐만 아니라 잡음에 민감하고 음소간의 구분이 어렵다는 음소인식의 문제점을 그대로 가지게 된다.

또한, [9]에서는 잡음환경에서 숫자음을 인식하기 위해 음성과 영상정보를 융합한 이중모드 인식기를 적용하고 있다. 적용된 이중모드 인식기는 두 정보를 특징 융합방법으로 융합하기 위해 단지 입력층에 영상정보를 위한 노드를 추가한 다층퍼셉트론으로 설계되었다. 따라서 모델에 대한 견고성(robustness)은 좋으나 모델 크기가 커짐에 따라 계산량이 증가하는 비효율적인 문제점을 가진다.

따라서 본 논문에서는 이질적인 정보들을 효율적으로 융합할 수 있는 신경망을 이용하고, 효율적으로 모델을 생성할 수 있는 고립단어 인식 모델인 이중모드 신경망(BMNN: BiModal Neural Network)을 제안한다. 또한, 잡음환경에서 음성인식률을 향상시키기 위해 사용자가 말한 단어들의 순차 패턴을 나타내는 문맥정보를 이용한 후처리 방법을 제안한다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 본 논문에서 음성과 영상 특징 추출 방법으로 사용한 기존 방법을 서술한다. Ⅲ장에서는 제안한 이중모드 신경망 모델에 대해 서술하고 Ⅳ장에서는 잡음환경에서 음성인식률을 향상시키기 위해 제안한 문맥정보를 이용한 후처리 방법에 대해 서술한다. Ⅴ장에서는 본 논문에서 제안한 방법에 대한 실험 및 분석 결과를 서술하며, Ⅵ장에서는 결론 및 향후 연구에 대해 기술한다.

Ⅱ. 음성과 영상 특징 추출

본 논문에서 사용하고 있는 음성 특징 추출방법과 영상 특징 추출 방법은 기존의 방법들로서 음성 특징 추출방법에 ZCPA(Zero Crossing with Peak Ampli-tude)[10] 방법을, 영상 특징 추출 방법에는 PCA(Princi-ple Component Analysis)[11] 방법을 사용하였다. 본 절에서는 각각의 방법에 대해서 간략하게 서술한다.

1. 음성 특징 추출

ZCPA는 청각 시스템에서 청각신경까지를 모델링한 것이며, 대역 통과 달팽이관 필터 뱅크와 각 달팽이관 필터의 출력단에 연결되어 있는 비선형 변환단으로 구성되어 있다. 일반적인 청각 모델처럼 와우각 필터 뱅크는 basilar 막을 모델링한 것이며, 비선형 변환단은 basilar 막의 기계적인 진동이 신경세포를 자극하는 과정을 모델링한 것으로 선형 필터들과 직렬 연결되어 있다.

ZCPA는 16개의 채널을 갖는 필터 뱅크 블록, 영 교차점 검출 블록과 비선형 변환 블록, 특징 추출 블럭으로 구성되어 있다. 필터 뱅크는 2의 제곱 계수를 갖는 FIR 필터로 구성되었으며 이등분을 반복적으로 사용하여 정밀도가 높은 주파수 계산이 가능하도록 하였다. 비선형 변환 블록에서는 이등분 방법과 이진 탐색 방법을 이용하여 메모리 크기와 계산 속도를 빠르게 하였다. 마지막으로 특징 추출 블록에서는 각 필터 뱅크의 프레임 크기만큼 해당 주파수 대역에 비선형화 된 최대값을 누적시키며 특징벡터를 추출하였다.

2. 영상 특징 추출

영상 기반 접근 방법 중에 가장 널리 쓰이는 방법은 통계적 분석에 기반한 영상 변환 방법인 PCA이다. PCA는 통계적인 분석을 통해 입력 영상의 차원을 줄여주며, 차원이 줄어들더라도 영상을 표현하는 중요한 정보는 보존되는 특성을 가지고 있다. 따라서 PCA를 통해 영상을 표현하기 위한 기저(basis)를 추출하였다. 주어진 16×16 크기의 입술 영상은 <그림 1>과 같이 기저와 기저의 가중치인 c_1, c_2 , ..., c_n 로 표현될 수 있다. 이 때 (c_1, c_2 , ..., c_n)를 영상의 표현 값이라 하고 이것이 곧 입술 모양의 특징이 된다.

영상이 M프레임일 경우, n차원 벡터 M개를 계산하여 음성의 시각 특징으로 제공하게 된다.

그러나 추출한 특징은 화자 간에 상이하므로, 이를

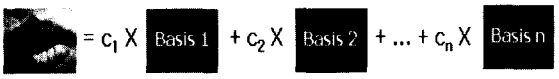


그림 1. 입술 영상 표현
Fig. 1. Lip image representation.

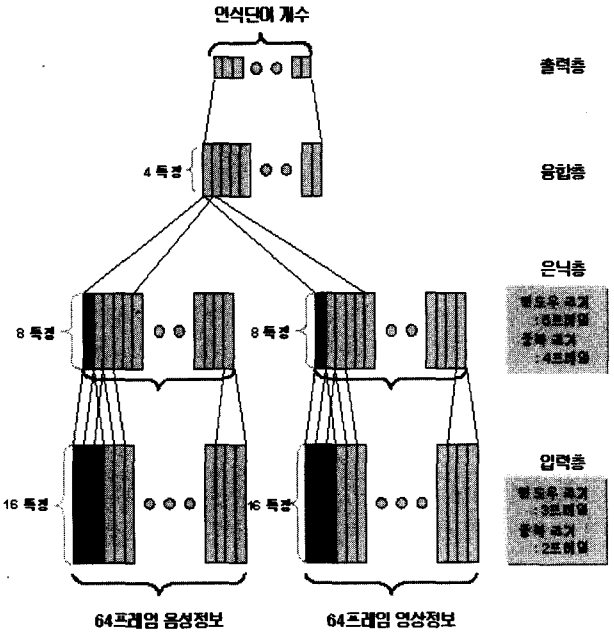


그림 2. BMNN 구조
Fig. 2. BMNN architecture.

어느 정도 상쇄해 주기 위해서 M 프레임의 영상에 대해 평균 벡터를 계산한 뒤, 이로부터의 차이로 입술 특징을 표현하였다. 따라서 입술 특징 v는 (식 1)과 같이 계산된다.

$$\bar{u} = \frac{1}{M}\sum_{i=1}^{M} u^i$$
$$v^k = u^k - \bar{u} \ , \ k = 1,2,...,M \tag{1}$$

여기서 u^i는 i번째 프레임의 특징 벡터를 의미하고, v^k는 새로 계산된 k번째 프레임의 특징 벡터를 의미한다.

본 논문에서는 벡터의 길이를 16으로 하였으며, 인식기의 특성상 입력 벡터의 개수가 고정되어야 하기 때문에 보간법을 이용하여 벡터의 개수를 64 프레임으로 만들어 인식 및 융합 모듈에 제공하였다.

Ⅲ. BMNN (BiModal Neural Network)

본 논문에서는 신경망을 이용하여 잡음환경에서 강

인한 이중 모드 음성 인식 모델을 제안한다. 제안한 BMNN의 구조는 <그림 2>와 같다.

BMNN은 4개 층(입력층, 은닉층, 융합층, 출력층)으로 구성되어 있으며 전방향 신경망 구조로 설계하였고 학습 알고리즘은 역전파 알고리즘을 사용하였다. 또한 고립단어로 학습 및 인식이 이루어지기 때문에 고립단어 인식에서 성능이 높은 중복 영역(overlap zone)구조[12]를 사용하였고 입력정보에 대한 함축적 시제위치를 고려하기 위해 윈도우 개념을 적용한 계층적 구조로 설계하였다. 특히, 융합층에서는 잡음에 의한 음성정보의 손실을 보상하기 위하여 음성과 영상 특징을 통합하는 기능을 수행한다.

모델의 연결구조와 각 계층의 프레임 개수 및 노드 개수를 살펴보면 윈도우에 포함된 모든 프레임들의 노드들과 대응되는 상위계층 프레임의 노드들이 완전 연결(fully connect)로 연결되고 융합층은 윈도우가 없기 때문에 출력층과 완전연결로 이루어진다. 그러므로 하위계층의 프레임 개수와 윈도우 크기 그리고 중복 영역의 크기가 결정되면 (식 2)에 의해 자동으로 각 층의 프레임 개수가 결정된다. 본 논문에서 중복영역의 크기는 (식 2)의 값이 상수가 되도록 설정하였고, 각 층의 프레임별 특징 개수는 실험을 통해 하위 계층의 프레임별 특징 개수의 1/2씩 감소하여 설정하였다. 마지막으로 출력층의 노드 개수는 인식할 고립단어의 개수로 설정하였다. 이와 같은 구조는 [9]에서 사용된 다층퍼셉트론 보다 모델 크기, 즉 연결 개수가 적어짐으로써 효율적으로 모델을 생성할 수 있다.

$$HF = \frac{LF - O}{W - O} \tag{2}$$

여기서 HF: 상위 계층 프레임 수, LF: 하위 계층 프레임 수, W: 윈도우 크기, 그리고 O: 중복 영역 크기를 의미한다.

Ⅳ. 문맥정보를 이용한 후처리 방법

최근 이동단말기 기술의 발전으로 잡음 환경에 강인한 음성인식 연구의 필요성이 대두되고 있다. 그러므로 잡음 환경에서 보다 향상된 음성인식률을 위해 사용자의 명령어(고립단어) 사용 패턴과 같은 순차 패턴을 나타내는 문맥정보를 이용하는 방법을 제안한다.

사용자의 명령어 사용 패턴과 같은 문맥정보는 영상, 음성 정보와는 다르게 신호가 아닌 추상적인 정보로써

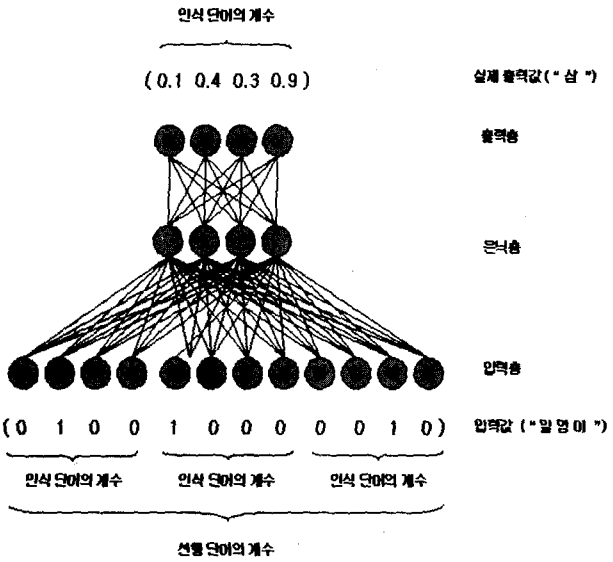


그림 3. 문맥 정보 인식기 구조
Fig. 3 Context recognition architecture.

특히, 잡음 환경에 영향을 받지 않는다. 따라서 문맥정보도 BMNN을 통해 융합한다면 전체적으로 인식률이 향상될 수 있을지 모르지만 추상적인 문맥정보와 영상, 음성 정보를 융합하기 위해서는 제안된 BMNN보다 더 복잡한 모델이 요구될 것이다.

따라서 본 논문에서는 문맥정보를 이용하기 위해 순차 패턴을 인식할 수 있는 문맥 정보 인식기를 제안하고 이를 통해 최종으로 인식할 수 있는 후처리 방법을 제안한다.

1. 문맥정보 인식기

사용자 명령어 사용패턴과 같은 순차 패턴을 인식하기 위해 <그림 3>과 같은 문맥정보 인식기(context recognition)를 제안한다. 일반 신경망은 입력노드에 순차 정보를 가지고 있지 않으므로 순차 패턴을 인식할 수 없다. 따라서 제안된 문맥정보 인식기는 다층퍼셉트론으로 설계하였으며 입력 층에 순차적 정보를 주기 위해 입력 값을 이진형(0과 1)으로 표현하였다.

예를 들어 인식할 명령어(단어)가 "영","일","이","삼" 네 개라고 가정할 경우, 문맥정보 인식기는 <그림 3>과 같이 설계될 수 있다. 여기서 입력노드 개수는 인식할 명령어(단어) 개수에 선행 단어 개수를 곱한 것과 같고 출력노드 개수는 인식할 명령어(단어) 개수와 같게 설정한다. 선행단어 개수란 현재 단어를 예측하기 위해 고려되는 단어의 개수로써, 앞서 인식한 명령어(단어)들 중 순차적으로 가장 최근에 인식한 단어의 개수를 말한다. 만약 입력값이 "일영이 : 0100 1000 0010"

으로 입력되면 출력노드에는 "일영이" 다음에 가장 많이 사용되는 해당 명령어(단어) 노드에 가장 높은 값이 출력된다. <그림 3>에서는 "삼"이 가장 높기 때문에 "일영이" 다음에 사용될 명령어(단어)를 "삼"으로 예측하게 된다. 따라서 선행단어 개수를 설정할 때, 너무 크게 설정하게 되면 패턴에 따라 발생빈도가 낮기 때문에 예측 값이 낮은 경향을 보이는 반면, 너무 작게 설정하게 되면 예측 값이 극소의 패턴으로 편중되는 경향을 보이게 된다.

2. 문맥 정보를 이용한 후처리

잡음 환경에서 보다 강인한 음성인식기를 개발하기 위해 본 논문에서는 문맥 정보를 이용한 후처리 방법을 제안한다. 후처리 방법이 적용된 음성인식기 구조는 <그림 4>와 같다. 여기서 최종 인식 결과는 BMNN인식기의 출력 값과 문맥정보 인식기의 출력 값을 결합함으로써 나타나게 된다. 따라서 BMNN 인식기와 문맥정보 인식기는 독립적으로 학습을 수행하여 모델을 생성하였다.

두 인식기 결과를 효율적으로 결합하기 위한 방법으로 <그림 5>와 같은 순차 결합(sequential combina-tion) 방법을 제안한다.

제안한 결합방법은 BMNN의 인식결과가 사용자가 설정한 임계값(θ) 보다 작을 경우 문맥정보 인식 결과를 고려하는 방법이다. 만약 두 인식기의 결과가 모두 임계값(θ)보다 작을 경우 입력 정보에 대해 정확한 인식이 이루어지지 않았다고 보고 두 인식기의 출력 값을 곱함으로써 출력 값의 차이가 적은 것을 선택할 수 있도록 한다. 각 인식기의 출력 값은 [0,1]의 값을 갖으며, 1에 가까울수록 출력 값의 신뢰성이 높다는 것을 의미한다. 임계값(θ)은 사용자에 의해 결정되며 사용자가 인식기의 결과를 신뢰할 수 있는 최소 한계 값을 의미한다. 따라서 실험에 사용된 임계값(θ)의 범위는 [0.8,1.0]으로 정하였다.

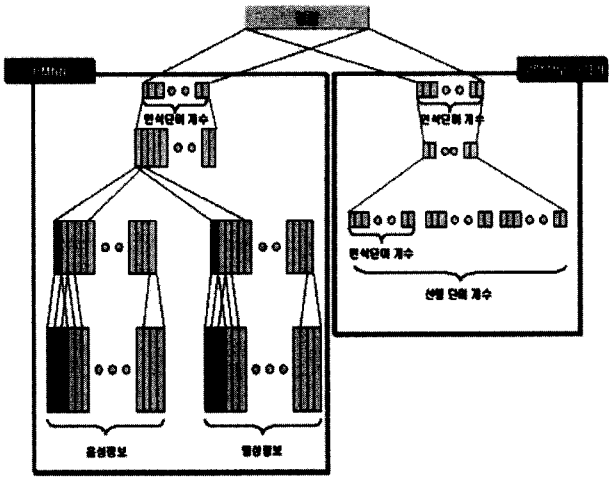


그림 4. 문맥정보를 이용한 후처리
Fig. 4. Post-processing using context information.

$BMNN(O_i)$: BMNN에서 i번째 출력노드의 출력 값
$Con(O'_i)$: 문맥정보 인식기에서 i번째 출력노드의 출력 값
θ: 임계값

if ($\theta < BMNN(O_i)$)
  $i = \max_i(BMNN(O_i))$ 번째 고립단어로 인식
else if ($\theta < Con(O'_i)$)
  $i = \max_i(Con(O'_i))$ 번째 고립단어로 인식
else if (($BMNN(O_i) \le \theta$) and ($Con(O'_i) \le \theta$))
  $i = \max_i(BMNN(O_i) \cdot Con(O'_i))$ 번째 고립단어로 인식

그림 5. 순차 결합 알고리즘
Fig. 5. Sequential Combine algorithm.

V. 실 험

실험에 사용된 데이터는 한국전자통신연구원(ETRI)에서 제작한 화자독립과 화자종속 데이터로 실험하였다. 화자독립 데이터는 62개의 고립단어를 남성 80명이 1회 발음한 데이터이며, 화자종속 데이터는 35개의 고립단어를 27회 발음한 데이터이다. 단어들은 별지.1에서와 같이 이동 단말기에서 사용될 수 있는 명령어들로 구성되었다. 잡음 환경에서의 음성신호를 생성시키기 위해 가우시안 잡음(20db, 10db, 5db)을 인위적으로 추가하여 잡음 데이터를 생성하였다.

본 실험에서 사용될 BMNN의 모델 구조는 고립단어 인식을 위해 입력 프레임을 64프레임(프레임당 10ms)으로 설정하였고 각 프레임에서 16차원의 특징을 추출하였다. 입력층의 윈도우 크기는 음소를 표현하기에 충분한 30ms인 3프레임으로 설정하였고 중복 영역 크기는 2프레임으로 설정하였다. 은닉층의 윈도우 크기는 더 넓은 시제 영역을 학습할 수 있도록 5프레임으로 설정하였고 중복 영역 크기는 4프레임으로 설정하였다. 따라서 (식 2)에 의거하여 은닉층의 프레임 수는 62프레임이 되고 융합층의 프레임 수는 58프레임이 된다.

잡음환경에서 제안한 이중모드 음성인식기인 BMNN의 타당성을 검증하기 위해 단일 음성인식기와 성능을 비교 분석하였다. 실험에 사용된 단일 음성인식기는 BMNN에서 영상특징을 "0"으로 입력하여 모델을 생성하였고 단일 영상인식기 역시 음성특징을 "0"으로 입력하여 모델을 생성하였다.

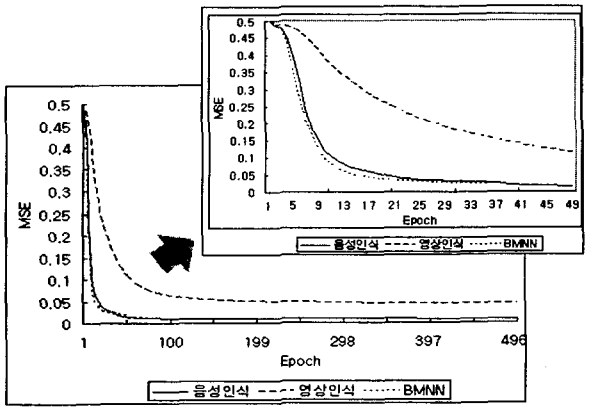
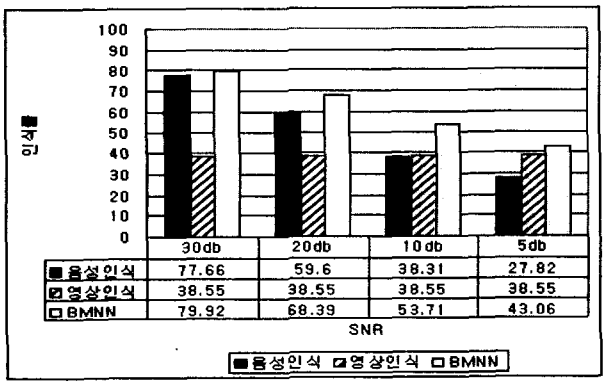


그림 6. 화자독립에 대한 학습 그래프
Fig. 6. Learning curves for speaker-independent.



	30db	20db	10db	5db
음성인식	77.66	59.6	38.31	27.82
영상인식	38.55	38.55	38.55	38.55
BMNN	79.92	68.39	53.71	43.06

그림 7. 화자 독립 실험 결과
Fig. 7. Experimental result of speaker-independent.

1. 화자 독립 실험

화자 독립에서 제안한 BNMM의 타당성을 검증하기 위해 62개의 고립단어를 60명이 1회 발음한 3720개 데이터를 학습데이터로 하고 테스트 데이터는 20명이 발음한 1240개 데이터를 사용하였다. 각 모델을 학습할 때 확인할 수 있는 에러 변화률은 <그림 6>과 같으며 테스트 데이터에 대한 실험 결과는 <그림 7>과 같다.

SNR이 30db일 경우 단일 음성인식 결과(77.66%)와 BMNN 인식 결과(79.92%)가 크게 차이는 없지만 잡음 증가에 대한 음성인식의 평균 감소률이 16.61%인 반면, BMNN 인식의 평균 감소률은 약 4%가 낮은 12.28%를 보인다. 따라서 잡음 신호가 많을수록 영상정보의 도움을 더 많이 받는다는 것을 알 수 있다.

<그림 8>은 5db 잡음환경에서 단일 음성인식과 BMNN 인식 결과로서 단어별 인식한 개수의 차이가 7개 이상인 단어들을 보여준다.

2. 화자 종속 실험

화자 종속에서 BMNN의 타당성을 검증하기 위해 35

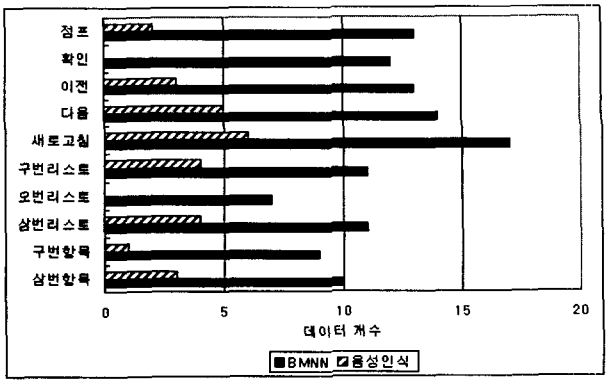


그림 8. 단어별 인식 개수
Fig. 8. Number of recognition words for each word.

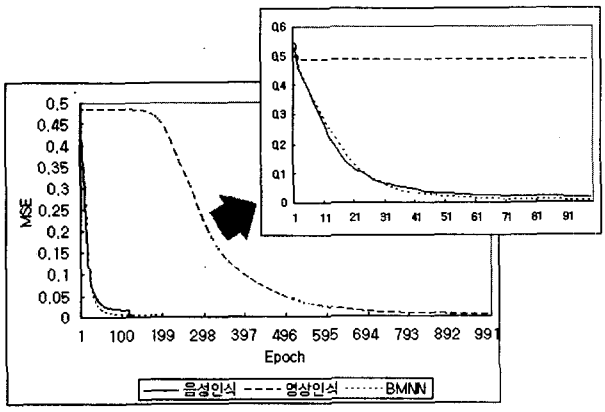


그림 9 화자 종속에 대한 학습 그래프
Fig. 9. Learning curves for speaker-dependent.

개 고립단어를 10회 발음한 350개의 학습데이터로 모델을 생성하였고 생성된 모델을 평가하기 위해 17회 발음한 595개를 테스트 데이터로 사용하였다. 각 모델에 대한 학습시 에러 변화률은 <그림 9>과 같으며 테스트 데이터에 대한 실험 결과는 <그림 10>과 같다.

SNR이 30db일 경우 단일 음성인식 결과(94.43%)와 BMNN 인식 결과(95.49%)가 크게 차이는 없지만 잡음이 증가할수록 음성인식의 평균 감소률은 13.16%인 반면 BMNN 인식의 평균 감소률은 약 3.96%가 낮은 8.2%임을 알 수 있다. 화자 종속 역시 화자 독립과 마찬가지로 잡음 환경에서 영상 정보의 도움을 받는 다는 것을 알 수 있다.

[13]에서는 실시간 임베디드 음성 인식 시스템을 위한 음성 인식기 설계 방법을 제안하고 있다. 따라서 본 논문과 목적은 다르지만 이동 단말기에서 사용될 수 있는 명령어들로 이용하여 시스템의 타당성을 검증한 면에서는 같다. 그 결과 [13]에서는 명령어에 대한 인식률은 96%, 숫자음에 대한 인식률은 94%로써 전체 인식률은 95%임을 보이고 있다. 제안한 방법이 [13]의 인식률

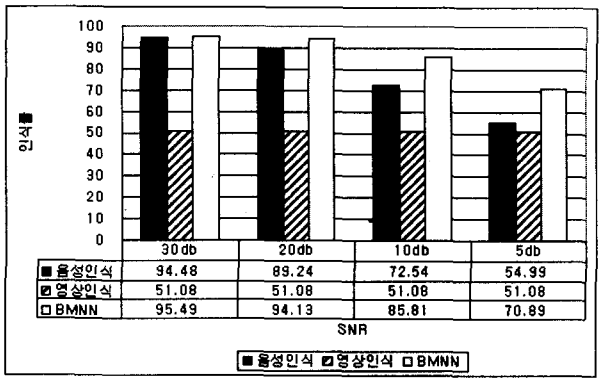


그림 10. 화자 종속 실험 결과
Fig. 10. Experimental result of speaker-dependent.

보다 다소 낮은 이유는 (별지.1)과 같이 실험에서 사용하고 있는 이동 단말기 명령어들 중 "일번리스트 ~ 십번리스트", "일번항목 ~ 십번항목"과 같이 혼동되기 쉬운 명령어들이 포함되어 있기 때문이다.

3. 문맥 정보를 이용한 후처리 실험

본 실험은 잡음 환경에서 음성 인식 성능을 보다 향상시키기 위해 이동 단말기 상에서 사용자가 사용하는 순차적 명령어 패턴과 같은 정보를 문맥 정보라 정의하고 이러한 정보를 이용한 후처리 방법에 대해 타당성을 검증하기 위한 실험이다.

먼저 후처리 방법에 사용할 문맥정보 인식기를 생성하기 전에, 이동 단말기 상에서 사용자가 <그림 11>와 같이 사용하는 순차적 명령어 패턴이 존재한다고 가정한다. 이때 <그림 11>-(a)와 같은 순차적 명령어 패턴이 사용자가 사용하는 전체 명령어 패턴들 중에서 발생되는 비율이 70%, 50%, 30%등으로 다르게 하여 각각의 학습데이터들을 생성하였다. 예를 들어 <그림 11>-(a)에서처럼 "브라우저 시작, 즐겨찾기, 오번항목"이라고 명령한 다음 "선택"이라는 명령어를 제시할 경우를 7회 발생시키고 3회는 선행단어들을 임의로 선택하여 생성함으로써 70%의 규칙성을 갖는 학습데이터를 생성하였다. 이와 같이 비율을 다르게 하여 학습데이터를 생성한 이유는 문맥정보 인식기가 학습데이터에 포함하고 있는 특정 패턴의 비율에 민감하게 반응하는지 알아보기 위해서이다.

본 실험에서 사용할 문맥정보 인식기는 명령어(단어)를 예측하기 위한 선행단어 개수를 3개로 설정하였다. 따라서 각 층의 노드의 개수는 입력층 105개, 은닉층 52개, 출력층 35개로 설정하였다. 입력층은 인식할 단어 개수가 35개이고 선행단어 개수가 3개 이므로 105개의

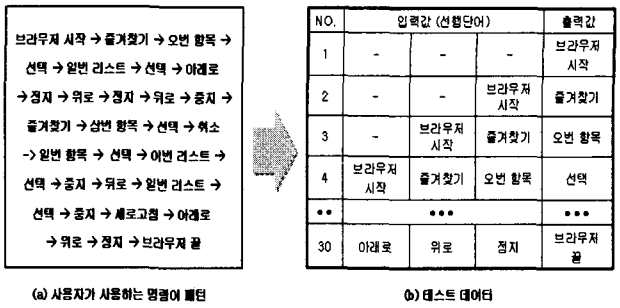


그림 11. 사용자의 명령어 사용 패턴과 테스트 데이터
Fig. 11. Sequence pattern of user commands and test data.

노드로 설정되었다. 반면, 은닉층의 노드 개수는 실험을 통해 얻어졌다.

생성된 문맥정보 인식기의 성능을 확인하기 위한 테스트 데이터는 <그림 11>-(a)패턴을 사용하여 생성하였다. 테스트 데이터는 <그림 11>-(b)와 같이 사용자가 처음 사용하는 명령어 "브라우저 시작"부터 마지막으로 사용하는 명령어 "브라우저 끝"까지 총 30개의 데이터로 구성되었다. 이렇게 생성된 테스트 데이터로 실험한 결과 70%, 50%, 30% 모델에 대한 인식률이 83.33%, 83.33%, 86.67%임을 확인할 수 있었다. 여기서 "브라우저시작", "즐겨찾기", "오번 항목" 같이 선행단어가 전부 존재하지 않는 단어들에 대해서는 모든 모델들이 올바르게 인식하지 못하였다. 또한 앞에서 기술한 바와 같이 학습데이터를 생성할 때 임의적으로 생성된 패턴에 따라 성능의 차이가 다소 발생하였다. 그 이유는 임의적으로 발생한 패턴들 때문에 선행단어는 같으나 예측할 단어가 틀린 경우가 발생함으로써 올바른 학습이 이루어지지 않았기 때문이다. 그러나 비율에 상관없이 비슷한 성능을 보이고 있음을 확인할 수 있다. 따라서 문맥정보 인식기는 일정 이상의 비율을 갖는 패턴들을 학습한다는 것을 알 수 있다.

이와 같이 학습된 문맥정보 인식기를 <그림 4>와 같이 후처리 방법에 적용하였다. 문맥정보 인식기는 특정 사용자의 명령어 사용패턴을 학습하기 때문에 화자 종속에만 적용된다. 따라서 앞서 기술한 화자 종속 실험에 후처리 방법을 적용한 결과는 <그림 12>과 같다. 결과를 살펴보면 잡음환경에서 음성 평균인식률은 69.51%이고 BMNN의 평균인식률은 81.84%인 반면, 문맥 정보를 이용한 후처리 방법을 적용하였을 경우 평균 인식률은 93.57%로 가장 높은 결과를 보인다. 또한 잡음 증가에 따른 인식에 대한 평균 감소율을 살펴보면 음성인식 경우 13.36% 평균 감소율을 보이고 BMNN 경우 9.24% 평균 감소율을 보이고 있으나, 문맥 정보를

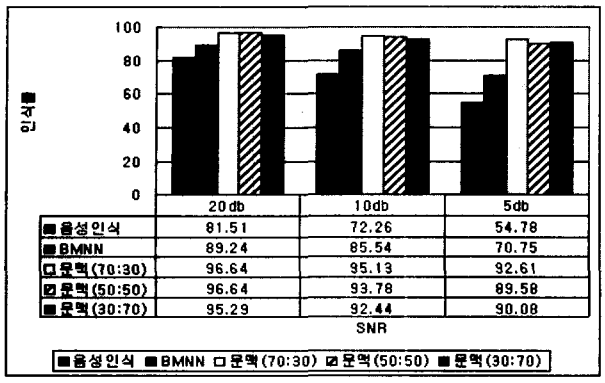


그림 12. 문맥정보를 이용한 후처리 실험 결과
Fig. 12. Experimental result of post-processing using context information.

이용한 후처리 방법을 적용할 경우 2.72% 평균 감소율을 보임으로써 보다 잡음에 영향을 받지 않는다는 것을 확인할 수 있다. 이처럼 만약 사용자의 순차적 명령 패턴이 존재한다면 이러한 정보들을 문맥정보 인식기로 학습함으로써 잡음환경에서 보다 우수한 성능을 보일 수 있다는 가능성을 제시한다. 순차적 명령 패턴이 존재한다는 가정 하에서 제안한 방법에 의해 성능이 향상된 이유는 음성과 영상정보만으로 구별될 수 없었던 패턴들을 구별할 수 있었기 때문이다. 그러나 순차적 명령 패턴이 존재하지 않는다면 문맥정보가 성능을 향상시켜준다고 보장 할 수는 없다. 따라서 본 논문에서는 잡음환경에서 명확한 음성인식을 위해 문맥정보와 같은 사용자 행동패턴이 새로운 정보로 이용될 수 있다는 가능성을 제시한다.

VI. 결론 및 향후 연구

본 논문에서는 잡음환경에서 강인한 음성인식을 위해 음성과 영상정보를 효율적으로 융합할 수 있는 이중 모드 신경망인 BMNN을 제안한다. BMNN은 4개 층으로 이루어진 다층퍼셉트론의 구조를 가지며 각 층은 입력 특징의 추상화 기능을 수행한다. BMNN는 제 3층인 융합층에서 잡음에 의한 음성 정보의 손실을 보상하기 위하여 음성과 영상 정보를 통합하는 기능을 수행한다. 또한, 잡음 환경에서 음성 인식률을 향상시키기 위해 사용자가 사용하는 명령어들의 순차 패턴을 나타내는 문맥정보를 이용한 후처리 방법을 제안한다. 잡음환경에서 BMNN은 단순히 음성만을 사용한 것 보다 높은 성능을 보임으로써 그 타당성을 확인할 수 있을 뿐 아니라, 특히 문맥을 이용한 후처리를 하였을 경우 잡음환경에서 90%이상의 인식률을 달성하였다.

따라서 본 논문에서는 제안한 문맥정보와 같은 사용자 행동패턴이 잡음환경에서 강인한 음성인식을 위해 고려할 수 있는 새로운 정보로 이용될 수 있다는 가능성을 제시한다.

향후 연구로는 제안한 문맥정보를 이용한 후처리 방법을 화자 독립에서도 적용될 수 있도록 문맥정보 표현 방법에 대한 연구를 진행할 것이다. 또한 문맥정보 인식기에서 인식할 고립단어 개수가 많아지면 은닉층의 노드 수가 증가할 필요가 있기 때문에 모델 생성 시 계산량이 많아지는 문제점을 가지게 된다. 따라서 클러스터링을 통해 유사한 고립단어들로 군집하여 인식할 대표 단어들을 생성함으로써 보다 효율적인 모델 생성 방법에 대해 연구할 것이다. 마지막으로 BMNN과 문맥정보 인식기 결과를 결합하는데 있어, 본 논문에서는 순차 결합 방법을 제안하여 BMNN 결과를 문맥정보 인식기 결과보다 항상 더 중요하게 반영함으로써 상황에 따른 적응성이 떨어질 수 있다. 따라서 상황에 따라 두 결과에 대한 판단 가중치를 고려할 수 있도록 신경망, HMM 혹은 퍼지이론을 적용한 보다 일반적인 결합방법에 대해 연구할 것이다.

참 고 문 헌

[1] Claude C. Chibelushi, Farzin Deravi, " A Review of Speech-Based Bimodal Recognition", *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 23-37, March, 2002.

[2] 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 Int. Conf. on* , vol. 1, pp. 181-186, 2001.

[3] Gemello, R.; Albesano, D.; Mana, F.; Moisa, L.; "Multi-source neural networks for speech recog nition: a review of recent results" , *Neural Net works, 2000. IJCNN 2000, Proc. of the IEEE -INNS-ENNS Int. Joint Conf. on* , vol. 5, pp. 265-270, 2000.

[4] Xiaozheng Zhang; Merserratt, R.M.; Clements, M.; , " Bimodal fusion in audio-visual speech recognition " , *Image Processing. 2002. Proc.. 2002 Int. Conf. on* ,vol.1, pp. 964-967, 2002..

[5] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. J. Lang, "Phoneme Recognition Using Time -Delay Neural Networks", *IEEE Trans. on Acou stics, Speech and Signal Processing.* vol.37,

no.3, pp. 328-339, March 1989.
[6] Haffiner,P., and Waibel, A. "Multi-State Time Delay Neural Networks for Continuous Speech Recognition". *In Advances in Neural Information Processing Systems* 4, Morgan Kaufmann Publishers, 1992.
[7] Joe Tebelskis,"Speech Recognition using Neural Networks", CMU-CS-95-142, May 1995.
[8] 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.
[9] 이상원, 박인정, "잡음환경에서 음성-영상 정보의 통합 처리를 사용한 숫자음 인식에 관한 연구", 전자공학회논문지, 제38권 CI편, 제3호, pp.61-67, 2001년 5월
[10] 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 Au dio Processing*, vol.7, no.1, pp. 55-69, January 1999.
[11] L.Reveret, C. Benoit, "Lip Parameters Extraction Based on Projection of Raw Images onto Refe rence Shapes", *Proc. of IEEE First Workshop on Multimedia Signal*, pp.1~6, June, 1997
[12] 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.
[13] 남상엽, 전은희, 박인정, "실시간 임베디드 음성 인식 시스템", 전자공학회논문지, 제40권 CI편, 제1호, pp.74-81, 2003년 1월

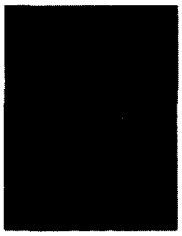
별지.1

● 실험에 사용한 고립단어들

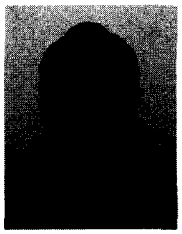
NO	고립단어	실험에 사용유무		NO	고립단어	실험에 사용유무	
		화자종속	화자독립			화자종속	화자독립
1	도움말		0	33	뒤로	0	0
2	닫기		0	34	앞으로	0	0
3	열기		0	35	브라우저 끝	0	0
4	확인		0	36	브라우저 시작	0	0
5	취소	0	0	37	재생기 시작		0
6	아니오		0	38	점프	0	0
7	예		0	39	앞점프		0
8	서버로 전송		0	40	종료		0
9	문서인식		0	41	정지	0	0
10	문자추출		0	42	재생	0	0
11	사진저장		0	43	일번리스트	0	0
12	사진캡춰		0	44	이번리스트	0	0
13	동영상저장		0	45	삼번리스트	0	0
14	활영끝		0	46	사번리스트	0	0
15	활영시작		0	47	오번리스트	0	0
16	선택	0	0	48	육번리스트		0
17	이전	0	0	49	칠번리스트		0
18	다음	0	0	50	팔번리스트		0
19	왼쪽	0	0	51	구번리스트		0
20	오른쪽	0	0	52	십번리스트		0
21	위로	0	0	53	일번항목	0	0
22	아래로	0	0	54	이번항목	0	0
23	가운데	0	0	55	삼번항목	0	0
24	우측		0	56	사번항목	0	0
25	좌측		0	57	오번항목	0	0
26	하단	0	0	58	육번항목		0
27	상단	0	0	59	칠번항목		0
28	목차	0	0	60	팔번항목		0
29	중지	0	0	61	구번항목		0
30	새로고침	0	0	62	십번항목		0
31	즐겨찾기	0	0	63	히스토리	0	
32	홈으로	0	0	64	일시정지	0	

저 자 소 개

김 명 원(정회원)
1972년 서울대학교 응용수학과 졸업
1981년 Univ. of Messachusetts Computer Science 석사
1986년 Univ. of Texas Computer Science 박사
1975년~1978년 한국과학기술연구소 연구원
1985년~1987년 AT&T Bell Labs. 연구원
1987년~1994년 한국전자통신연구소 책임연구원
2000년~2001년 미국 IBM T.J Watson 연구소 방문과학자
1994년~현재 숭실대학교 컴퓨터학부 교수
<주관심분야: 신경망, 퍼지, 유전자알고리즘,패턴인식,자동추론,기계학습,데이터마이닝, creativity engineering 등>



이 순 신(정회원)
1999년 숭실대학교 컴퓨터 학부 졸업
2004년 숭실대학교 컴퓨터학과 석사
2004년~현재 LG CNS 솔루션 센타 DW/BI팀
<주관심분야: 인공지능,신경망,데이터마이닝,데이터 웨어하우징, CRM>



한 문 성(정회원)
1977년 서울대학교 수학과 졸업
1977년~1979년 전국경제인연합회 경제기술 조사부
1980년~1981년 한국IBM
1981년~1988년 Indiana Univ. Computer Secience 박사과정
1989년~현재 한국 전자통신연구원 책임연구원
<주관심분야: 음성인식, 패턴인식, 신경망, 기계학습>

류 정 우(정회원)
1998년 숭실대학교 인공지능학과 졸업
2000년 숭실대학교 컴퓨터학과 석사
2000년~현재 숭실대학교 컴퓨터학과 박사 과정
<주관심분야: 인공지능, 기계학습, 퍼지, 유전자알고리즘>