청각 및 시각 정보를 이용한 강인한 음성 인식 시스템의 구현

Constructing a Noise-Robust Speech Recognition System using Acoustic and Visual Information

이 종 석*, 박 철 훈
(Jong-Seok Lee and Cheol Hoon Park)

Abstract : In this paper, we present an audio-visual speech recognition system for noise-robust human-computer interaction. Unlike usual speech recognition systems, our system utilizes the visual signal containing speakers' lip movements along with the acoustic signal to obtain robust speech recognition performance against environmental noise. The procedures of acoustic speech processing, visual speech processing, and audio-visual integration are described in detail. Experimental results demonstrate the constructed system significantly enhances the recognition performance in noisy circumstances compared to acoustic-only recognition by using the complementary nature of the two signals.

Keywords : audio-visual speech recognition, noise-robustness, integration

I. 서론

음성은 인간과 컴퓨터의 인터페이스를 위한 가장 자연스러운 의사소통 방법 중 하나로써, 컴퓨터를 통한 음성의 인식은 지난 수십년간 많은 발전을 거쳐 실세계에서 응용되는 수준에 이르렀다. 그러나 아직 해결해야 할 몇 가지 문제가 남아있는데 그 중 하나는 실제 작동 환경에서 피할 수 없는 잡음에 의한 성능의 저하이다. 잡음이 존재하지 않는 경우 현재 음성인식 기술은 90% 이상의 높은 인식율을 보이지만, 잡음이 존재하는 환경에서는 성능이 크게 저하된다. 잡음은 음성인식이 적용되는 현장에서 다양한 형태로 존재하며, 다양한 종류와 수준의 잡음에 대해 강인한 성능을 얻기 위한 노력이 절실하다.

잡음에 대한 강인함을 위한 여러 방법들 중, 말소리 정보 이외에 시각정보, 특히 화자의 입술 움직임 정보를 함께 사용하는 시청각 음성인식(audio-visual speech recognition) 기법이 최근 주목을 받고 있다. 시청각 음성(audio-visual speech)은 시각적(visual) 측면과 청각적(acoustic) 측면을 모두 포함하는 음성을 의미한다. 인간은 귀로 말소리를 듣는 외에 입술의 움직임을 관찰함으로써 시끄러운 환경에서도 높은 인식 능력을 보인다는 것이 알려져 있다[1]. 시청각 음성인식은 이러한 인간의 특성을 공학적으로 모방하는 것으로써, 시각정보를 이용하는 것은 청각정보(말소리)에 비해 성능이 다소 낮지만 소리잡음에 영향을 받지 않기 때문에 강인한 음성인식을 위한 도구로 사용될 수 있다[2].

시청각 음성인식 시스템을 구성하는 중요한 세 부분은 청각특징(acoustic feature) 추출, 시각특징(visual feature) 추출, 그리고 시각정보와 청각정보의 통합이다. 이 중 청각특징을 추출하는 것은 기존의 음성인식 분야에서 많이 연구되어 있다[3]. 따라서 본 논문에서는 기록된 동영상으로부터 시각특징을 추출하는 기법과 두 정보를 효과적으로 통합하는 기법을 중점적으로 다루어 다양한 잡음 환경에서 강인한 성능을 보이는 시청각 음성인식 시스템의 구현 과정을 보인다.

시각특징의 추출은 기록된 영상에서 인식에 중요한 입술 움직임 정보를 효과적으로 표현하는 낮은 차원의 특징벡터를 얻는 것이다. 좋은 인식 성능을 위해서 특징벡터는 다음의 두 요건을 만족해야 한다. 첫째, 각 발음 클래스를 구분할 수 있는 중요한 음성 정보를 포함해야 한다. 둘째, 화자의 외모나 피부색, 조명의 변화, 얼굴의 움직임이나 카메라와의 거리 등과 같은 음성 외의 변화 요인에 불변해야 한다. 시각특징을 추출하기 위한 기존의 기법은 크게 윤곽선 기반 방식과 픽셀값 기반 방식으로 나뉜다[2]. 전자의 경우, 동영상에서 입술의 윤곽선을 추적하고 입술의 높이나 너비와 같은 기하학적 정보를 특징으로 사용하거나 윤곽선의 모델을 정의하고 그 모델의 파라미터를 특징으로 사용하는 기법이다. 후자는 기록된 영상에서 입술영역의 영상을 얻고 영상변환을 적용해 특징을 얻는 기법이다. 윤곽선 기반 방식에서는 입 안쪽의 혀와 이빨의 변화나 입술의 돌출과 같이 인식에 중요한 정보를 잃게 되고, 윤곽선의 추적 과정에서 생길 수 있는 오차가 인식율을 저하시킬 수 있기 때문에 픽셀값 기반 방식이 더 좋은 인식율을 보인다[4].

본 논문에서는 제어되지 않은 조명 하에서 기록된 영상으로부터 조명이나 화자별 차이와 같은 인식에 불필요한 변화를 제거하고 중요한 정보를 획득하기 위한 픽셀값 기반 특징 추출 과정을 보인다.

시각정보와 청각정보를 통합하는 문제는 최종 인식 결과를 얻기 위해 두 정보를 언제, 어떠한 방식으로 합치는가에 관한 것이다. 두 정보의 통합 방식은 크게 두 정보의 특징벡터를 하나로 합쳐 인식기에 입력하는 초기통합 기법과 각각을 따로 인식한 후 그 결과를 통합하는 후기통합 기법으로 나눌 수 있다[2]. 일반적으로 강인한 인식 시스템을 위해서는 후기통합 기법을 선호하는 경향이 있는데 그 이유는 다음 세 가지를 들 수 있다. 첫째, 후기통합에서는 각 정보가 독립적으로 인식된 후 합쳐지기 때문에 음성에 존재하는 잡음의 수

* 책임저자(Corresponding Author)
논문접수 : 2007. 5. 15., 채택확정 : 2007. 6. 25.
이종석, 박철훈 : 한국과학기술원 전자전산학부 전기및전자공학전공
(jslee@nnmi.kaist.ac.kr/chpark@kaist.ac.kr)
※ 본 연구는 2007년도 한국과학기술원 BK21 정보기술사업단에 의하여 지원되었음.

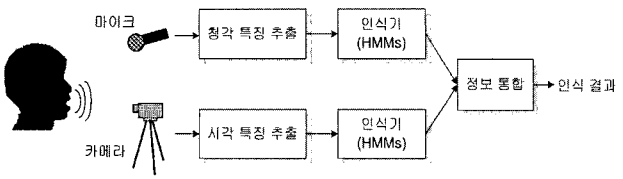


그림 1. 시청각 음성인식 시스템의 전체 구조.

Fig. 1. Overall structure of the audio-visual speech recognition system.

준에 따라 최종 결과에 각 정보가 기여하는 정도를 조절하는 기법을 사용하기가 상대적으로 쉽다. 둘째, 초기통합은 두 신호의 완벽한 동기화를 가정하지만 후기통합은 유연한 동기화를 제공할 수 있다. 실제로 시각음성과 청각음성은 정확히 동기되어 있지 않으며 때로는 말소리보다 입술이 수십~수백 밀리세컨드 정도 먼저 움직이기도 한다[5]. 셋째, 후기통합기법에 기반한 시스템은 사전에 구축되어있는 단일정보기반 인식 시스템을 활용하여 구성할 수 있지만 초기통합 시스템의 구현은 완전히 새로운 시스템을 만들어야 한다.

본 논문에서 구현하는 시스템에서는 특별히 후기통합에서 청각모듈과 시각모듈에 대한 최적의 가중치를 결정하는 문제를 다룬다. 가중치는 최종 인식 결과를 얻을 때 각 모듈에 의존하는 정도를 결정하는 변수이다. 최적 가중치를 결정하는 기법은 인식 대상 음성에 포함된 잡음의 수준에 따라 각 모듈의 상대적 가중치를 자동으로 조절하는 것으로써, 시스템을 다양한 잡음 환경에서 사용하기 위해서는 반드시 필요하다. 주어진 시청각 음성 데이터에 대해 통합 가중치를 적절히 선택하지 못하는 경우 두 정보의 시너지 효과를 얻을 수 없을 뿐만 아니라 심지어는 어느 한 정보만을 이용한 인식 결과보다도 못한 성능을 얻을 수도 있다.

본 논문은 지금까지 논의한 사항들을 고려하여 잡음에 강인한 성능을 보이는 후기통합 기법 기반 시청각 음성인식 시스템을 구현한다. 특히 시각정보의 특징 추출과 각 정보를 통합하는 과정을 자세히 설명한다. 시스템의 설계 목표는 다양한 잡음의 종류와 수준에 대해 잡음에 대한 사전 지식이 없는 상황에서 항상 강인한 인식 성능을 얻는 것이며, 실험에 의한 성능 평가를 통해 이를 확인한다.

이하 논문의 구성은 다음과 같다. II장에서는 전체 시스템의 구조 및 사용된 음성 데이터베이스를 설명한다. III장과 IV장에서는 각각 청각특징 추출과 시각특징 추출에 대해 설명한다. V장에서는 두 모듈을 통합하는 기법에 대해 설명한다. VI장에서는 시스템에 대한 성능을 실험을 통해 평가하고 마지막으로 VII장에서 결론으로 논문을 맺는다.

II. 시청각 음성인식 시스템 구조

1. 시스템 구조

본 논문에서 구현하는 시청각 음성인식 시스템의 전체 구조는 그림 1과 같다. 마이크와 비디오 카메라로 화자의 말소리와 입술의 움직임을 각각 기록하고, 기록된 각 신호에서 인식에 적절한 특징을 추출한다. 추출된 특징은 각각의 인식기에 입력되고 인식기의 출력을 통합하여 최종 인식 결과를 얻는다.

인식기는 음성인식에서 가장 많이 사용되는 은닉 마르코프 모델(HMM: Hidden Markov Model)로 구성된다[3]. 하나의 HMM은 학습을 통해 하나의 클래스에 속하는 여러 화자의 발음을 모델링하는데, 학습은 기대-최대(expectation-maximization) 알고리즘으로 할 수 있다. 인식 단계에서는 소속 클래스를 알지 못하는 데이터가 입력되면 이를 모든 HMM에 입력하여 가장 높은 확률값을 보이는 HMM을 선택하여 인식 결과를 얻는다.

2. 시청각 음성 데이터베이스

개발하는 시청각 음성인식 시스템의 성능 평가를 위해 두 가지 우리말 고립단어 데이터베이스를 이용한다. 첫째는 우리말 "일"부터 "구"까지, 그리고 "공"과 "영"을 포함한 11개의 숫자로 이루어진 데이터베이스이고, 둘째는 우리나라 16개 주요 도시 이름으로 이루어진 데이터베이스이다[6]. 데이터베이스의 작성에는 총 56명(남자 38명, 여자 18명)이 참가하였고 화자마다 각 단어를 세 번씩 발음하였다. 조용한 연구실 환경에서 마이크를 통해 말소리를 16kHz의 샘플링 주파수로 저장하고 비디오 카메라를 통해 화자의 입술 주변 얼굴 부분을 30Hz의 프레임비율로 동영상으로 저장하였다. 영상 기록시 천장의 형광등 외에 별도의 인위적 조명은 설정하지 않아 영상마다 밝기 및 빛의 각도가 일정하지 않게 기록되어 있다.

인식 실험은 화자 독립 방식으로 이루어지며, 실험 결과의 신뢰도를 높이기 위해 56명의 화자를 네 모둠으로 나누고 세 모둠(42명)을 학습에, 나머지 한 모둠(14명)을 인식테스트에 사용하는 과정을 돌아가면서 네 번 반복하였다.

III. 청각특징 추출

청각신호에서 특징을 추출하는 것은 기존의 여러 연구에서 다루고 있다. 개발된 기법 중 가장 많이 쓰이는 것은 멜-주파수 켑스트럼 계수(MFCC: Mel-Frequency Cepstral Coefficient)이다[3]. MFCC는 사람의 청각모델을 바탕으로 하고 있으며 잡음이 없는 경우 및 있는 경우 모두 우수한 인식 성능을 내는 것으로 알려져 있다. 입력된 신호를 한번에 10ms씩 움직이며 25ms 크기의 프레임으로 분할하고 각 프레임별로 푸리에 변환, 필터뱅크 에너지 계산, 로그 변환, 이산 코사인 변환을 거쳐 12차 MFCC와 정규화된 프레임 에너지, 그리고 그것들의 시간 미분인 동적 특징(dynamic feature)을 청각특징으로 사용한다.

IV. 시각특징 추출

서론에서 서술한 바와 같이 시각특징은 더 좋은 성능을 보이는 것으로 알려진 픽셀값 기반 방식으로 추출된다. 기록된 영상에 전처리 과정을 수행한 후 입술 영역을 추출하고, 입술영역 영상에 변환 기법을 적용하여 특징을 얻는다. 그 과정을 그림 2에 요약하였다. 여기서 서술하는 시각특징 추출 기법은 II-2절에서 설명한 데이터베이스의 영상(화자의 입 주변만 포함하는 영상)에 대한 것으로써, 얼굴 전체가 보이거나 배경이 포함되는 경우에는 입술 주변 영역을 검출하는 과정이 먼저 이루어져야 한다.

먼저, 기록된 영상에서 좌우의 밝기 차이를 보정한다. 이후 단계에서 좌우 입술 끝점을 정확하게 추출하기 위해 영상의

좌우간 밝기 차이를 제거하고자 하는 것이다. 좌우 일부 영역의 평균 픽셀값을 계산한 후 좌우 밝기의 변화를 그 두 값의 선형 보간(linear interpolation)으로 모델링한다. 로그 영역에서 이 밝기 변화를 뺌으로써 좌우 밝기 차이가 제거된 영상을 얻는다[7].

다음으로, 영상들 사이의 밝기 차이를 보정하기 위해 픽셀값 분포를 정규화한다. 이는 화자간 얼굴 색의 차이와 기록 시마다 다른 조명 차이에 의해 생기는 영상간 픽셀값 분포 차이를 제거하기 위한 과정이다. 학습 데이터의 전체 영상에 대해 픽셀값의 분포가 대략적으로 가우시안 분포를 가짐을 확인하였다. 따라서 각 입력 영상에 대해 히스토그램 명세화(histogram specification)[8] 과정을 통해 영상 내 픽셀값이 이 가우시안 분포를 따르도록 정규화한다.

위의 두 과정에 의해 전처리된 영상에 임계값을 적용하여 입술의 양 끝점을 추출한다. 입을 다물었을 때는 양 입술의 사이가 어둡게 나타나고, 입을 벌렸을 때는 입 안쪽이 어둡게 나타나기 때문에, 임계값을 적용한 영상에서 검은 부분의 양쪽 끝 부분이 입의 양 끝점이 된다. 찾은 양 끝점을 바탕으로 회전과 크기 변화에 불변하는 44×50픽셀 크기의 입술 영역 영상을 얻는다.

입술 영역 영상을 얻은 후 각 픽셀 좌표별로 하나의 발음 전체에 대한 평균을 제거한다. 길이가 총 T 프레임인 발음에서 t번째 프레임에 해당하는 입술영역 영상의 좌표 (m,n)에서의 픽셀값을 $I(m,n,t)$ 이라 하면, 평균이 제거된 영상의 픽셀값 $J(m,n,t)$ 은 다음 식으로 주어진다.

$$J(m,n,t) = I(m,n,t) - \frac{1}{T}\sum_{k=1}^{T} I(m,n,k),\ 1 \le m \le M, 1 \le n \le N \quad (1)$$

여기서 M 과 N 은 영상의 가로 및 세로길이이다. 이 과정은 각 발음별로 다른 조명 효과나 화자의 피부색에 의한 직류(DC) 성분을 제거하는 과정이다.

평균이 제거된 영상에 주성분분석(PCA: Principal Component Analysis)을 적용하여 정적 특징(static feature)을 얻는다. 평균이 제거된 영상의 픽셀값을 n_0 $(= M \times N)$차원 열벡터로 만든 것을 $\mathbf{x}$ 라 하면, $\mathbf{x}$ 에 대한 n 차원 정적특징벡터 $\mathbf{s}$ 는 다음과 같다.

$$\mathbf{s} = P^T(\mathbf{x} - \overline{\mathbf{x}}) \quad (2)$$

여기서 $\overline{\mathbf{x}}$ 는 모든 학습 데이터(42명의 화자에 대한 데이터)에 대한 $\mathbf{x}$의 평균이고, $P\,(\in \mathbb{R}^{n_0 \times n})$의 각 열은 학습 데이터의 모든 $\mathbf{x}$ 에 대한 공분산 행렬에서 얻은 고유벡터(eigenvector) 중 고유치(eigenvalue)가 큰 $n(\ll n_0)$개에 대한 고유벡터이다. 논문에서 구현하는 시스템에서는 n=12로 하여 12개의 PCA 계수를 특징으로 얻는다. 그림 3은 학습 데이터에 대해 입술영역 영상의 평균 영상과 상위 4개의 주성분(principal component)에 대한 변화를 나타낸 것이다. 각 주성분마다 서로 다른 입술의 변화를 나타내는 것을 볼 수 있다. 첫번째 주성분은 주로 입이 열리고 닫히는 변화를 나타낸다. 두번째 주성분은 아랫입술의 돌출과 이빨의 유무를 나타낸다. 세번째 주성분은 윗입술의 돌출과 아랫입술 아래의 그림자의 변화를, 네번째 주성분은 주로 입이 벌어질 때 이빨의 유무를 나타냄을 관찰할 수 있다.

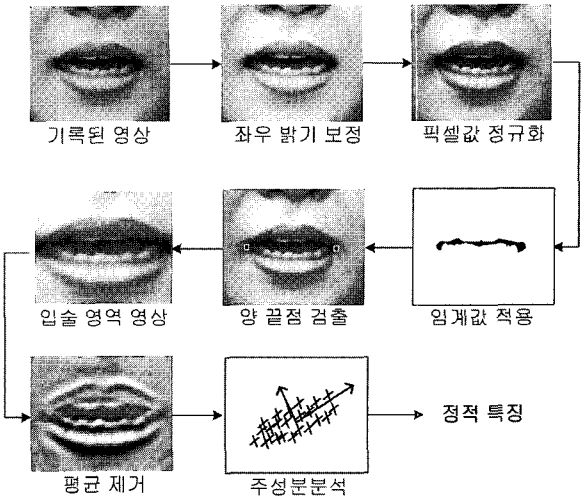


그림 2. 기록된 영상에서 특징을 추출하는 과정.

Fig. 2. Procedure of extracting visual features from a recorded image.

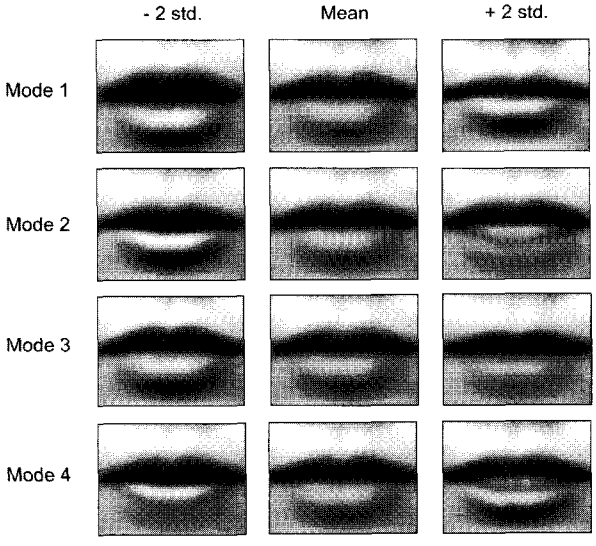


그림 3. 주성분분석에 의한 입술영상의 주요 변화분석.

Fig. 3. First four principal modes of variation for the lip region images.

청각특징 추출 과정에서와 같이 인식율의 향상을 위해 정적 특징의 시간 미분인 동적 특징을 함께 사용한다. 결과적으로 각 프레임별로 총 24차원의 특징벡터를 얻는다.

V. 시청각 통합 인식

시각 모듈과 청각 모듈 각각의 인식기(HMMs)가 인식을 수행한 후 그들의 출력으로부터 통합된 인식 결과를 얻는다. 주어진 데이터 O 에 대해 최종 인식결과 클래스 C^* 는 다음과 같이 주어진다.

$$C^* = \arg\max_i \{\gamma \log P(O \mid \lambda_A^i) + (1-\gamma)\log P(O \mid \lambda_V^i)\} \quad (3)$$

여기서 λ_A^i 와 λ_V^i 는 각각 청각 및 시각 모듈에서 i 번째 클

래스에 대한 HMM이며, $\log P(O|\lambda_A^i)$ 와 $\log P(O|\lambda_V^i)$ 는 그들의 출력인 로그우도(log-likelihood)이다. 통합가중치 γ 는 0과 1 사이의 값을 가지는데, 각 정보에 대한 최종 인식 결과의 의존도를 나타낸다. 잡음이 적어서 청각모듈의 성능이 시각모듈에 비해 좋을 경우 최종 결정이 청각모듈에 더 의존하도록 가중치는 상대적으로 큰 값을 가진다. 반대로 잡음이 많은 경우 가중치는 작은 값을 가져야 한다.

기존의 연구에서는 가중치 결정의 가장 간단한 형태로써 여러 잡음 조건에 대해 항상 같은 값으로 설정하는 방법[9]이나 사용자가 결정하는 방법[10]이 고려되기도 했다. 또한 포함된 잡음의 수준, 즉 신호대잡음비(SNR: Signal-to-Noise Ratio)를 안다고 가정하고 가중치를 SNR의 함수로 정하는 방법도 있으나[11], 이러한 정보가 항상 정확하게 주어지는 것은 아니다. 시스템의 작동 환경에서 얻는 적응 데이터를 이용해 가중치를 결정하는 방법도 기존의 연구에서 고려된 바 있다[12].

본 논문에서는 잡음의 종류나 수준에 대한 사전정보나 추가의 적응 데이터 없이 자동적으로 가중치 γ 를 결정하기 위해 신경회로망을 이용한 기법을 제안한다. 제안하는 방법에서는 각 모듈의 인식기인 HMM들의 출력으로부터 모듈의 신뢰도를 추정하고 이를 바탕으로 가중치를 결정한다. 신뢰도는 HMM들의 출력의 분포로부터 얻어진다. 즉, 잡음이 많지 않을 때는 청각 HMM의 출력간에 큰 차이가 있지만 잡음이 많아지면 이 차이는 점차 줄어드는 경향을 이용하여 각 인식기의 신뢰도를 다음과 같이 정의한다[13].

$$S_k = \sum_i \{\max_j \log P(O|\lambda_k^j) - \log P(O|\lambda_k^i)\},\ k \in \{A, V\} \quad (4)$$

즉, 주어진 데이터에 대한 전체 HMM들의 로그우도를 얻은 후 최대값에서 나머지 값을 각각 뺀 값의 합을 신뢰도로 정한다. 신경회로망은 각 모듈의 신뢰도와 최적 가중치 사이의 입출력 사상(mapping)을 모델링한다(그림 4). 이론적으로 신경회로망은 은닉뉴런의 수가 충분하다면 임의의 함수를 임의의 오차범위 이내로 근사화할 수 있다[14]. 주어진 데이터를 이용한 학습을 통해 데이터에 내재된 입출력 관계의 연속적 사상을 만들고 학습되지 않은 데이터에 대해 일반화할 수 있기 때문에, 신뢰도와 가중치간의 사상을 근사화함으로써 다양한 잡음 환경에서 적절한 통합 가중치를 추정하기 위한 도구로써 적합하다.

신경회로망의 학습은 다음의 과정으로 이루어진다. 먼저

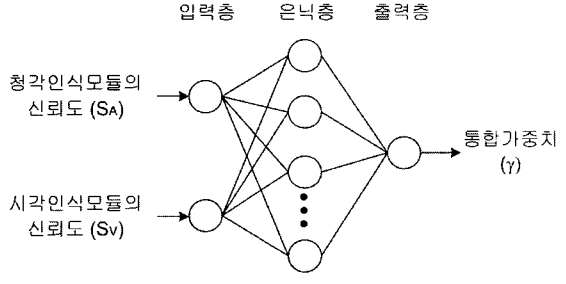


그림 4. 통합 가중치 추정을 위한 신경회로망.

Fig. 4. Neural network for estimating integration weights.

여러 SNR의 학습 데이터에 대해 청각 및 시각 HMM의 출력의 신뢰도를 (4)에 의해 계산한다. 다음, 각 데이터에 대해 γ 값을 0부터 1까지 0.01씩 증가시키면서 (3)에 의해 옳은 인식 결과를 내는 γ 값을 저장한다. 마지막으로 신뢰도와 최적 γ 값을 입출력 학습 데이터로 하여 신경회로망을 학습한다. 학습에는 잡음이 없는 데이터와 20dB, 10dB, 0dB의 백색잡음 섞인 데이터를 사용하고, 학습되지 않은 잡음조건에 대해서는 신경회로망의 일반화 성능을 통해 적절한 가중치를 얻는다.

신경회로망을 사용하는 과정에서 기존의 일반적인 문제들과의 다른 점은 학습데이터에 대한 출력의 목표값이 하나의 값이 아닌 범위로 주어진다는 것이다(그림 5). 따라서, 학습 과정에서 사용되는 신경회로망의 오차함수를

$$e(y) = y_d - y \quad (5)$$

에서

$$e(y) = \begin{cases} \gamma_l - y & \text{for } y < \gamma_l \\ 0 & \text{for } \gamma_l \le y \le \gamma_u \\ y - \gamma_u & \text{for } \gamma_u < y \end{cases} \quad (6)$$

로 수정한다. 여기서 y 는 신경회로망의 출력, y_d 는 기존의 학습규칙에서의 학습목표값, γ_l 과 γ_u 는 각각 옳은 인식결과를 내는 가중치 범위의 하한 및 상한값이다. 하한값과 상한값 사이의 값(그림 5에서 회색영역)은 모두 옳은 인식 결과를 낸다.

VI. 성능 평가

1. 잡음 데이터베이스

잡음이 존재하는 인식 환경을 모의실험하기 위해 많이 사용되는 가산잡음 데이터베이스인 NOISEX-92[15]을 이용하였다. 네 가지 잡음, 즉 백색잡음, F-16 조종석 잡음, 공장잡음 그리고 조종실잡음을 음성의 청각신호에 더하여 0dB부터 25 dB의 SNR을 가지는 잡음섞인 음성 데이터를 생성하여 실험에 사용하였다.

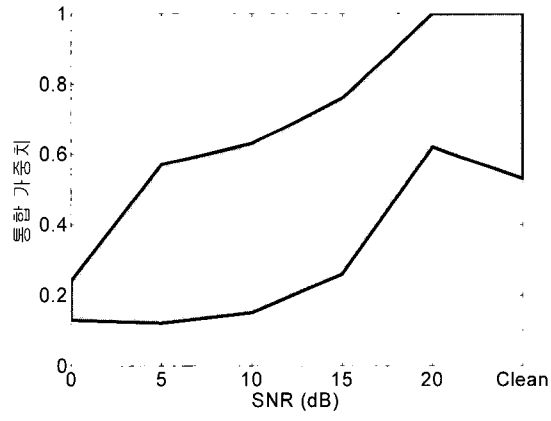
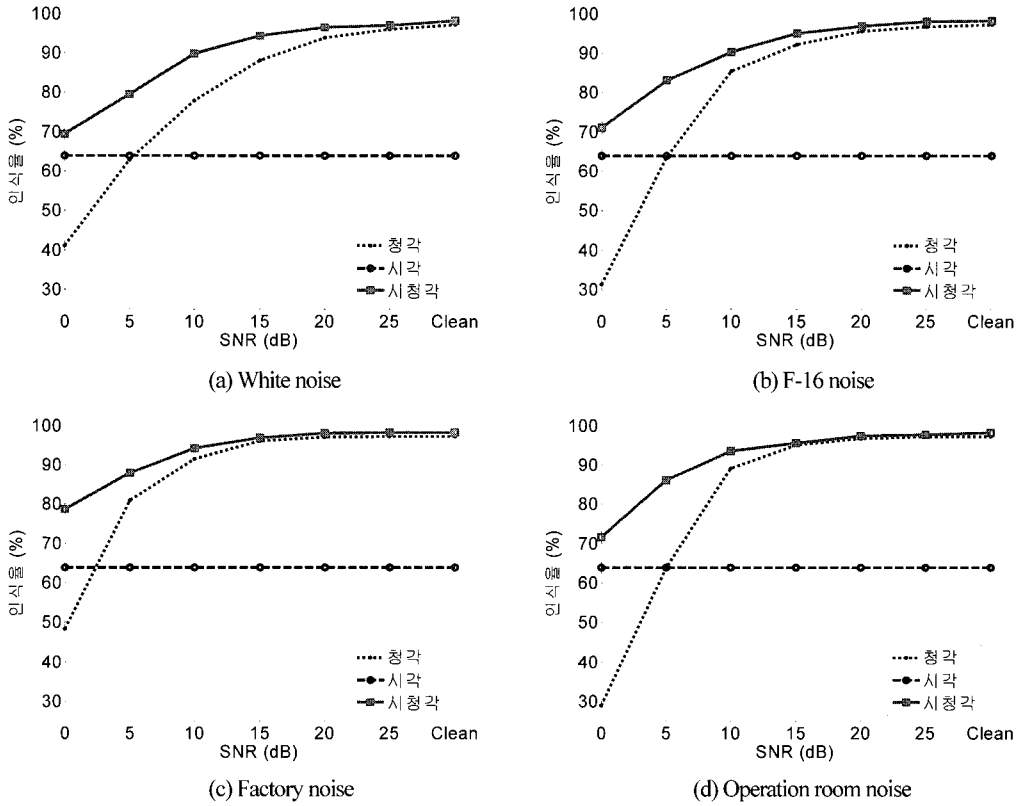


그림 5. 백색잡음 섞인 데이터에 대해 SNR에 따른 최적 통합 가중치 변화의 예.

Fig. 5. Example of the optimal weight as a function of the SNR for a datum corrupted by white noise.

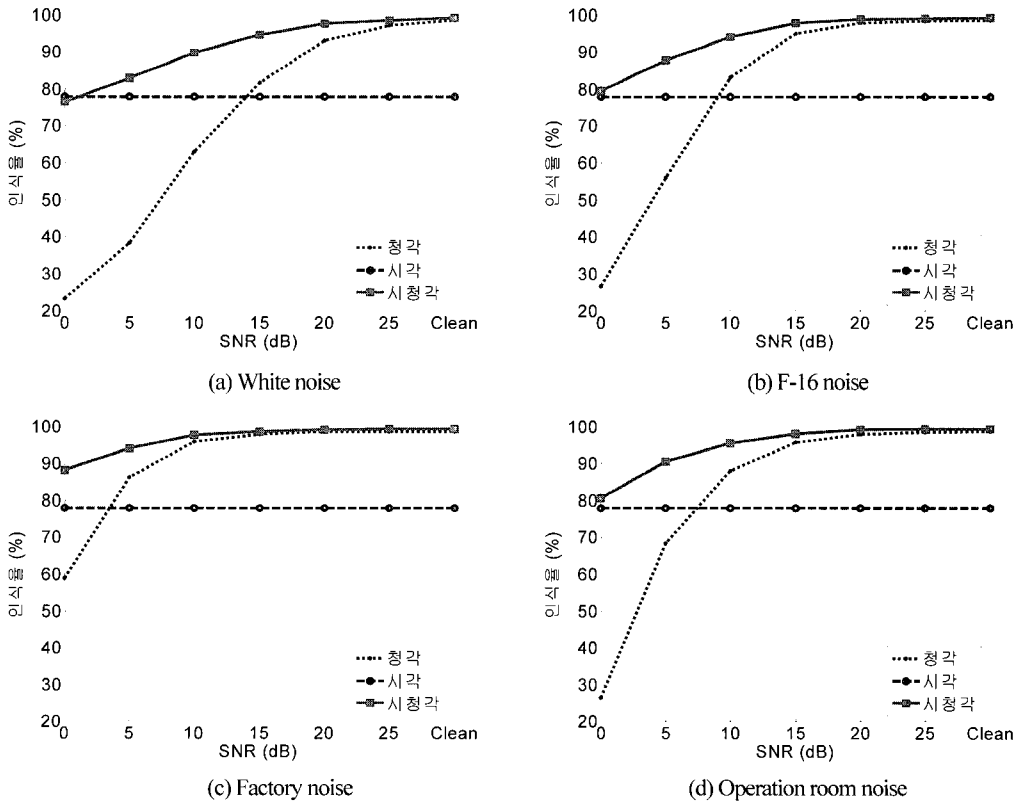


(a) White noise (b) F-16 noise

(c) Factory noise (d) Operation room noise

그림 6. 숫자 데이터베이스에 대한 인식 결과.

Fig. 6. Recognition results for the digit database.



(a) White noise (b) F-16 noise

(c) Factory noise (d) Operation room noise

그림 7. 도시이름 데이터베이스에 대한 인식 결과.

Fig. 7. Recognition results for the city name database.

2. 실험 조건

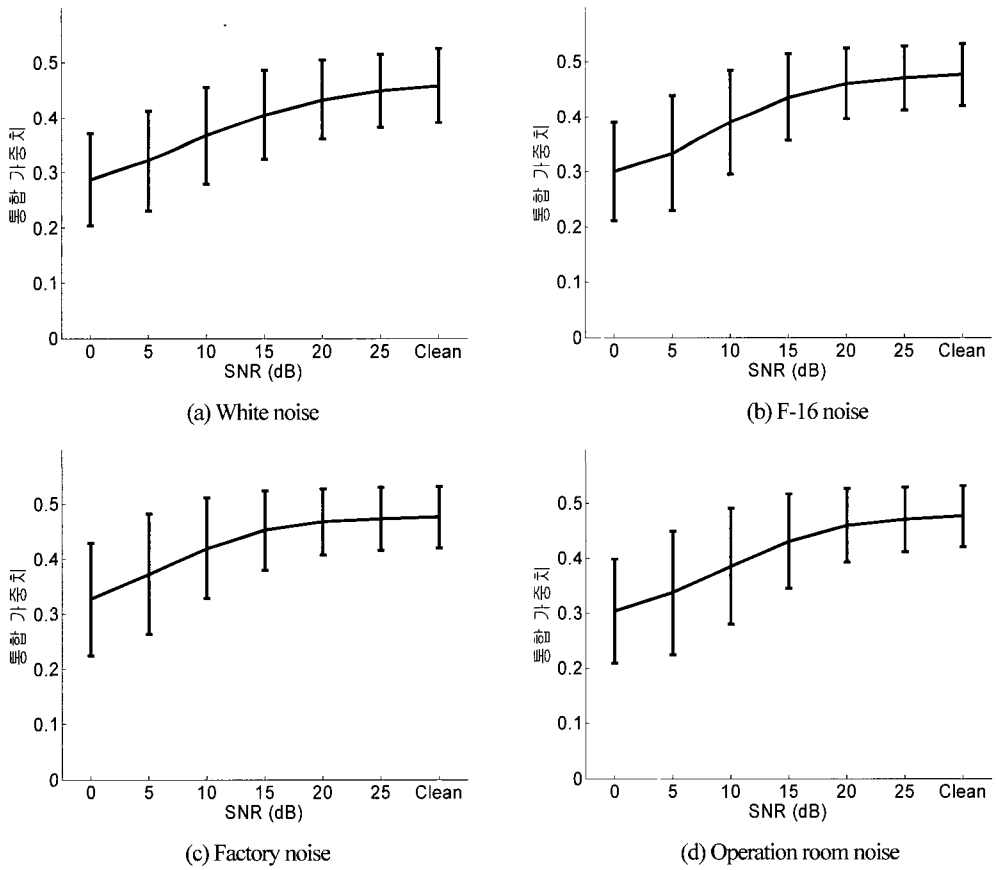
II-2절에서 설명한 숫자 및 도시이름 데이터베이스를 대상으로 청각 정보만을 이용한 경우, 시각 정보만을 이용한 경우 및 시청각 정보를 모두 이용한 경우에 대해 화자독립 인식실험을 수행한다.

청각 및 시각모듈의 인식기에서 각 HMM의 상태의 수는 해당 단어의 음소의 수에 비례하게 설정하였다. 그리고 각 HMM의 상태의 가우시안 함수는 3개를 사용하였다. 통합을 위한 신경회로망은 시그모이드 특성을 가지는 5개의 은닉 뉴런으로 구성된 은닉층을 가지는데, 더 이상의 뉴런은 성능 향상에 도움이 되지 않았다. 신경회로망의 학습에는 여러 알고리즘들 중 가장 빠르고 좋은 성능을 보이는 것으로 알려져 있는 Levenberg-Marquardt 알고리즘[14]을 사용하였다.

3. 인식 결과

그림 6과 7은 각 데이터베이스에 대해 시각 또는 청각 모듈 단독의 성능과 통합 인식 시스템의 성능을 나타낸 것이다. 결과에서 다음과 같은 사실들을 확인할 수 있다. 첫째, 청각 정보만을 이용한 경우의 인식율은 잡음이 존재하지 않는 경우 두 데이터베이스 모두 거의 100%에 가깝지만 포함된 잡음이 많아질수록 인식율은 급격히 저하되며 0dB에서의 인식율은 잡음의 종류에 따라 30%에도 못미치는 경우가 있다. 둘째, 시각정보만을 이용한 인식은 잡음의 수준에 상관없이 항상 일정한 인식율을 유지하는데, 각 데이터베이스에 대해 각각 63.9%와 78.0%로 나타났다. 이 값은 잡음없는 음성에 대해 청각신호만을 이용한 경우보다는 낮지만 청각신호에 잡음이 많아질 때에 비해서는 높은 값이다. 셋째, 두 정보를 모두 이용하여 인식하는 경우 하나의 정보만을 이용한 경우보다 최소한 비슷하거나 더 나은 인식율을 보인다. 특히 5~15dB의 구간에서는 두 정보를 이용한 상승효과가 두드러지는 것을 관측할 수 있다. 청각신호만을 이용했을 때에 비해 두 정보를 모두 사용한 경우 상대적 오인식율의 감소량은 각 데이터베이스에 대해 39.4%와 60.4%로 나타났다. 특히 잡음이 심한 0~10dB의 구간에 대한 상대적 오인식율 감소는 각각 48.4%와 66.9%로 나타나, 잡음환경에서 강인한 인식성능을 얻고자 하는 시스템의 목표를 달성하였음을 알 수 있다. 넷째, 제안하는 통합 가중치 결정 방식이 학습에 사용되지 않은 환경에 대해 성공적으로 동작함을 볼 수 있다. 가중치를 결정하는 신경회로망의 학습에는 깨끗한 신호와 백색잡음이 섞인 0dB, 10dB, 20dB의 신호만을 사용하였는데 그 외의 잡음수준과 다른 세 잡음 종류에 대해서도 강인한 인식성능을 위한 가중치를 얻는데 성공하였다.

그림 8은 숫자 데이터베이스에 대해 신경회로망에 의해 결정된 가중치 값의 SNR에 따른 변화를 나타낸 것이다. 데이터 전체에 대한 평균과 표준편차를 나타내었다. 결정된 가중치가 SNR에 비례하는 것을 볼 수 있는데, 이를 통해 제안된 통합 기법이 SNR에 따라 가중치를 자동적으로 결정하는



(a) White noise (b) F-16 noise

(c) Factory noise (d) Operation room noise

그림 8. 숫자 데이터베이스에 대한 SNR에 따른 통합 가중치의 변화.

Fig. 8. Integration weights with respect to the SNR for the digit database.

것을 확인할 수 있다.

VII. 결론

본 논문에서는 다양한 잡음 환경에서 잡음에 대한 사전지식 없이 강인한 성능을 보이는 음성인식을 위한 시청각 음성인식 시스템을 구현하였다. 말소리정보 이외의 부가적인 음성 정보로써 입술영역의 영상으로부터 시각특징을 추출하는 과정과 신경회로망을 이용하여 청각정보와 시각정보를 효과적으로 통합하는 기법을 보였다. 잡음에 의해 기존의 청각정보를 이용한 인식의 성능이 크게 떨어지는데 반해 시각정보를 함께 사용함으로써 다양한 잡음 환경에서 성능을 크게 향상시키는 것을 확인하였다. 추후과제로써 구현된 시스템을 연결단어 또는 연속음성의 인식으로 확장하는 연구가 진행중이다.

참고문헌

[1] L. A. Ross, D. Saint-Amour, V. M. Leavitt, D. C. Javitt, and J. J. Foxe, "Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments," *Cerebral Cortex*, vol. 17, no. 5, pp. 1147-1153, 2007.

[2] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A review of speech-based bimodal recognition," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 23-37, Mar. 2002.

[3] X.-D. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall, 2001.

[4] P. Scanlon and R. Reilly, "Feature analysis for automatic speechreading," in *Proc. Int. Conf. Multimedia and Expo*, pp. 625-630, 2001.

[5] C. Benoît, "The intrinsic bimodality of speech communication and the synthesis of talking faces," *The Structure of Multimodal Dialogue II*, M. M. Taylor, F. Nel, and D. Bouwhuis, Eds. Amsterdam, The Netherlands: John Benjamins, pp. 485-202, 2000.

[6] J.-S. Lee and C. H. Park, "Training hidden Markov models by hybrid simulated annealing for visual speech recognition," in *Proc. Int. Conf. Systems, Man, Cybernetics*, pp. 198-202, Oct. 2006.

[7] 이종석, 심선희, 김소영, 박철훈, "제어되지 않은 조명 조건하에서 입술움직임의 강인한 특징추출을 이용한 바이모달 음성인식," *Telecommunications Review*, 제 14 권, 제 1 호, pp. 123-134, 2. 2004.

[8] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Addison-Wesley Publishing Company, 2001.

[9] T. J. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," *IEEE Trans. Audio, Speech, Language Processing*, vol. 14, no. 3, pp. 1082-1089, May 2006.

[10] A. Verma, T. Faruquie, C. Neti, and S. Basu, "Late integration in audio-visual continuous speech recognition," in *Proc. Workshop on Automatic Speech Recognition and Understanding*, pp. 71-74, Dec. 1999.

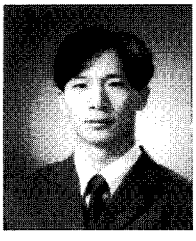
[11] G. F. Meyer, J. B. Mulligan, and S. M. Wuerger, "Continuous audio-visual digit recognition using N-best decision fusion," *Information Fusion*, vol. 5, no. 2, pp. 91-101, June 2004.

[12] S. Tamura, K. Iwano, and S. Furui, "A stream-weight optimization method for multi-stream HMMs based on likelihood value normalization," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, pp. 469-472, Mar. 2005.

[13] T. W. Lewis and D. M. W. Powers, "Sensor fusion weighting measures in audio-visual speech recognition," in *Proc. Conf. Australasian Computer Science*, pp. 305-314, 2004.

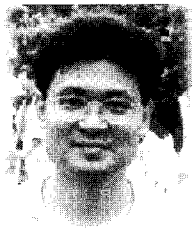
[14] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, UK, 1995.

[15] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247-251, 1993.



이 종 석

1999년 한국과학기술원 전기및전자공학과 학사. 2001년 한국과학기술원 전자전산학과 석사. 2006년 한국과학기술원 전자전산학과 박사. 2006년~현재 한국과학기술원 전자전산학부 연수연구원. 관심분야는 시청각 음성인식, 멀티모달 인터페이스, 패턴인식.



박 철 훈

1984년 서울대학교 전자공학과 학사. 1985년 Caltech 전자공학과 석사. 1990년 Caltech 전자공학과 박사. 1991년~현재 한국과학기술원 전자전산학부 교수. 관심분야는 지능시스템, 신경회로망, 최적화, 지능제어.