DOI: 10.3745/KIPSTB.2008.15-B.2.123

향상된 자동 독순을 위한 새로운 시간영역 필터링 기법

이 종 석† · 박 철 훈††

요 약

자동 독순(automatic lipreading)은 화자의 입술 움직임을 통해 음성을 인식하는 기술이다. 이 기술은 잡음이 존재하는 환경에서 말소리를 이용한 음성인식의 성능 저하를 보완하는 수단으로 최근 주목받고 있다. 자동 독순에서 중요한 문제 중 하나는 기록된 영상으로부터 인식에 적합한 특징을 정의하고 추출하는 것이다. 본 논문에서는 독순 성능의 향상을 위해 새로운 필터링 기법을 이용한 특징추출 기법을 제안한다. 제안하는 기법에서는 입술영역 영상에서 각 픽셀값의 시간 궤적에 대역통과필터를 적용하여 음성 정보와 관련이 없는 성분, 즉 지나치게 높거나 낮은 주파수 성분을 제거한 후 주성분분석으로 특징을 추출한다. 화자독립 인식 실험을 통해 영상에 잡음이 존재하는 환경이나 존재하지 않는 환경에서 모두 향상된 인식 성능을 얻음을 보인다.

키워드 : 자동 독순, 필터링, 특징추출, 잡음에 대한 강인함

A New Temporal Filtering Method for Improved Automatic Lipreading

Jong-Seok Lee† · Cheol Hoon Park††

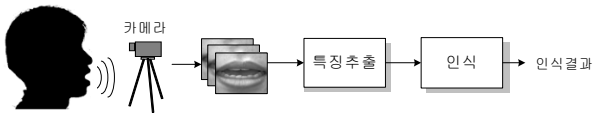
ABSTRACT

Automatic lipreading is to recognize speech by observing the movement of a speaker's lips. It has received attention recently as a method of complementing performance degradation of acoustic speech recognition in acoustically noisy environments. One of the important issues in automatic lipreading is to define and extract salient features from the recorded images. In this paper, we propose a feature extraction method by using a new filtering technique for obtaining improved recognition performance. The proposed method eliminates frequency components which are too slow or too fast compared to the relevant speech information by applying a band-pass filter to the temporal trajectory of each pixel in the images containing the lip region and, then, features are extracted by principal component analysis. We show that the proposed method produces improved performance in both clean and visually noisy conditions via speaker-independent recognition experiments.

Key Words : Automatic Lipreading, Filtering, Feature Extraction, Noise-Robustness

1. 서 론

자동 독순(automatic lipreading)은 (그림 1)에 나타난 것과 같이 화자의 입술 움직임에 대한 관찰을 통해 음성을 인식하는 기술이다. 말소리를 이용한 기존의 음성인식이 잡음이 존재하는 환경에서 성능이 크게 저하되는데 반해 영상정보는 소리 잡음에 영향을 받지 않기 때문에, 자동 독순은 음성인식을 보완하여 강인한 성능을 얻는 기술로써 최근 주목을 받고 있다[1].



(그림 1) 자동 독순의 과정

자동 독순 시스템에서 중요한 단계는 기록된 영상으로부터 인식에 적절한 특징을 추출하는 것이다. 좋은 인식 성능을 위해서는 화자간 외모 차이나 조명 조건 등의 변이에 불변하면서 각 발음 클래스를 구분하는데 중요한 정보를 포함하는 특징을 추출해야 한다. 영상에서 특징을 추출하는 방법은 크게 두 가지 기법으로 나눌 수 있다. 첫째는 영상에서 입술의 윤곽선을 찾아내고 이를 통해 입술의 높이나 너비와 같은 기하학적 특징이나 윤곽선을 모델링하는 모델의 파라미터를 특징으로 사용하는 윤곽선 기반 방식이다[2,3]. 둘째는 입술영역을 포함하는 영상에 이산 코사인 변환

※ 본 연구는 2007년 한국과학기술원 BK21 정보기술사업단에 의하여 지원되었음.
† 정 회 원 : 한국과학기술원 전자전산학부 연수연구원
†† 정 회 원 : 한국과학기술원 전자전산학부 교수
논문접수 : 2007년 2월 21일, 심사완료 : 2007년 11월 4일

(discrete cosine transform)이나 주성분분석(PCA: principal component analysis)과 같은 변환을 적용하고 변환된 영상의 일부를 특징으로 사용하는 픽셀값 기반 방식이다[4,5]. 이 두 기법들 중 더 많이 쓰이는 것은 픽셀값 기반 방식이다. 이는 윤곽선 기반 방식이 입술 윤곽선을 추적하는 과정에서 오차를 낼 수 있고 입 안쪽 정보를 표현하지 못한다는 단점이 있기 때문이다[6].

이처럼 다양한 특징 추출 기법이 연구되었지만, 말소리를 이용한 기존의 음성인식 분야에서 좋은 인식성능을 얻기 위한 특징추출 기법이나 인간의 청각기관을 모델링하는 기법 등이 많이 연구되어 있는 것과 비교할 때 자동 독순 분야에서는 특징 추출에 대한 연구가 아직 많이 부족한 실정이다. 또한 영상에 잡음이 존재하는 환경에 대한 연구는 최근에서야 이루어지고 있으며[7,8] 이러한 환경에서 잡음에 강인한 특징을 추출하는 것에 대한 연구는 전무한 실정이다. 일반적으로 영상의 획득이나 전송 과정에서는 잡음이 포함될 가능성이 있으며[9], 이러한 환경에서 강인한 성능을 얻을 수 있는 기법에 대한 연구가 필요하다.

본 논문에서는 잡음이 존재하거나 존재하지 않는 환경에서 좋은 독순 성능을 내는 영상 특징을 추출하기 위한 필터링 기법을 제안한다. 제안하는 기법은 입술 영역 영상의 각 픽셀의 시간 궤적에 대해서 인식에 중요한 음성의 정보 이외의 주파수 영역을 차단하는 대역통과필터(BPF: band-pass filter)를 적용하는 것이다. 필터의 설계 과정에서는 심리학적 근거, 주파수 영역에서의 분석 및 실험적 근거에 의해 통과 대역을 결정한다. 실험 결과를 통해 필터링 기법에 의한 인식 성능은 필터링을 하지 않을 때에 비해 잡음이 존재하는 환경이나 존재하지 않는 환경 모두 상당한 향상을 얻을 수 있음을 보인다.

이하 논문의 구성은 다음과 같다. 2장에서는 사용된 데이터베이스, 입술 영역 추출 과정, 특징 추출 및 인식기에 대해 간략히 설명한다. 3장에서는 제안하는 필터링 기법과 필터의 설계 과정을 설명한다. 4장에서는 화자독립 인식 실험을 통해 제안하는 기법의 성능을 알아본다. 마지막으로 5장에서 결론을 내리고 논문을 맺는다.

2. 기본 독순 시스템

2.1. 데이터베이스

본 논문에서 사용한 데이터베이스는 우리말 숫자 '일'부터 '구', 그리고 '영'과 '공'을 포함한 숫자 데이터베이스(DIGIT)와 우리나라 16개 주요 도시 이름으로 구성된 도시이름 데이터베이스(CITY)이다[10]. 56명의 화자가 각 발음을 세 번씩 고립단어 형태로 발음한 것을 비디오 카메라를 이용하여 30Hz의 프레임 비율로 기록한 것이다. 일률적인 인공조명 없이 자연스러운 연구실 환경에서 기록되었으며 각 동영상은 화자의 입술 주변 얼굴 부분을 포함한다. 인식 실험은 화자독립 형태로 진행된다. 각 데이터베이스별로 28명 화자의 발음을 학습에, 14명 화자의 발음을 제안하는 필터의 설계에, 그리고 나머지 14명 화자의 발음을 인식 테스트에 사용한다. <표 1>에 데이터베이스의 상세 사항을 정리하였다.

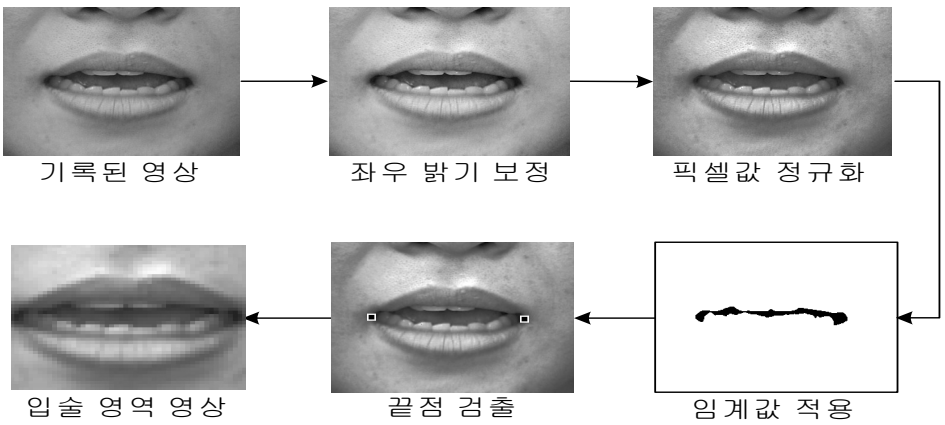
〈표 1〉 데이터베이스의 상세 사항

참여 화자	56명 (남 37명, 여 19명)
기록 대상	발음하는 화자의 입술주변 얼굴 영역
해상도	720 × 480
프레임비율	30 Hz
조명 조건	자연스러운 연구실 조명 (일률적인 인공조명 미설치)
기록 단어	DIGIT: 일, 이, 삼, 사, 오, 육, 칠, 팔, 구, 공, 영 CITY: 서울, 대전, 대구, 부산, 울산, 인천, 광주, 전주, 경주, 제주, 강릉, 춘천, 수원, 충주, 남원, 공주(각 단어를 세 번씩 반복)
인식실험	화자 독립 (학습용 데이터, 설계용 데이터, 테스트용 데이터로 구분)

2.2. 입술영역 추출

특징 추출을 위해서는 먼저 기록된 얼굴 영상에서 입술을 찾는 과정이 필요하다. 이 과정은 입술 영역을 추출하는 것과 동시에 화자별 피부색의 차이나 기록시의 조명조건 차이 등을 보정하기 위한 전처리 과정을 포함한다[11]. 아래에서 그 과정을 간략히 설명한다. ((그림 2) 참조)

먼저, 영상에 포함된 얼굴 좌우의 밝기 차이를 제거한다.



(그림 2) 입술 영역의 추출 과정

이를 위해 좌우 조명의 변화를 선형적으로 모델링하고 로그 영역에서 이러한 변화를 뺌으로써 제거한다[3]. 이 때 좌우 일부 영역의 평균 픽셀값을 계산하고 이 두 값의 선형보간으로 밝기 변화를 모델링한다. 다음으로, 모든 영상의 픽셀값이 같은 확률분포를 따르도록 정규화한다. 전체 학습데이터에 대한 영상의 픽셀값이 정규분포로 근사화될 수 있음을 관찰하였기 때문에 각 영상이 이 분포를 따르도록 히스토그램 명세화(histogram specification)[12]을 통해 영상의 픽셀값을 변환한다. 이를 통해 화자나 조명에 의한 영상간의 픽셀값 분포의 차이를 줄일 수 있다. 다음으로, 임계값을 적용하여 입술의 양 끝점을 찾는다. 입을 다물었을 때는 입술 사이가 그림자에 의해 어둡게 나타나며 입을 벌렸을 때는 입 안쪽이 어둡게 나타나기 때문에 임계값을 적용하여 입술 양 끝점을 찾을 수 있다. 이렇게 찾은 양 끝점을 바탕으로 영상마다 다른 회전이나 크기 변화가 보정된 44×50 픽셀 크기의 입술 영역 영상을 얻는다.

2.3 특징 추출

특징추출의 첫 단계는 2.2절에서 얻은 입술 영역 영상의 각 픽셀에 대해 발음 전체에 대한 평균을 제거하는 것이다. 이렇게 함으로써 각 발음간 밝기의 차이를 제거한다. 다음으로 PCA를 적용하여 특징을 얻는다. PCA를 이용한 특징추출 기법은 다른 변환 기법과 비교했을 때 비슷하거나 더 우수한 성능을 보이는 것으로 알려져 있다[13]. 평균이 제거된 입술 영상의 픽셀값을 2200차원(=44×50) 열벡터로 만든 것을 x, 학습 데이터 전체에 대한 x의 평균을 m이라 할 때, PCA를 적용하여 얻는 특징벡터는

$$y = P^T(x - m) \tag{1}$$

로 구해진다. 여기서 P의 행의 수는 x의 차원과 같으며, P의 각 열은 학습 데이터의 공분산 행렬을 고유치 분해(eigenvalue decomposition)하여 얻는 고유벡터(eigenvector)를 고유치 크기 순서대로 정렬한 것이다. 얻어진 고유벡터 중 일부만을 사용하여, 즉 P의 열의 수를 x의 차원인 2200보다 적게 함으로써 낮은 차원의 특징벡터를 얻을 수 있다. 실험을 통해 12차원의 특징벡터가 인식에 적절함을 관측하였다. 따라서 각 영상 x에 대해 식 (1)과 같이 변환을 적용한 후 최종적으로 12차원의 특징벡터 y를 얻는다.

2.4. 인식기

인식에는 음성인식에서 가장 많이 사용되는 은닉 마르코프 모델(HMM: hidden Markov model) [14]을 사용한다. 관측확률 분포가 가우시안 혼합 모델(Gaussian mixture model)로 주어지는 연속 HMM을 사용하는데, HMM의 상태의 수는 각 단어별로 포함된 음소의 수에 비례하도록 하고 각 상태마다 3개의 가우시안 함수를 사용한다. 이는 실험을 통해 좋은 성능을 얻는 설정을 선택한 것이다. HMM 파라미터의 학습은 기대-최대(expectation-maximization)기법[14]을 사용하여 수행한다. 인식 과정에서는 클래스를 알 수 없는 인식용 데이터가 주어졌을 때 이를 모든 클래스에 대한 HMM에 입력하고 가장 높은 확률값을 내는 HMM을 선택하여 인식결과를 얻는다. 인식 성능은 테스트 데이터의 단어들 중 오인식된 단어의 수의 백분율로 정의되는 오인식율(%)로 나타낸다.

윗 절에서 얻은 특징과 연속 HMM을 이용한 인식 결과 오인식율은 DIGIT 데이터베이스의 462개 단어(=14명×11단어×3번반복)와 CITY 데이터베이스 672개 단어(=14명×16단어×3번반복)에 대해 각각 45.9%와 37.8%로 나타났다. 우리말 숫자인 DIGIT 데이터베이스에는 단음절이며 입술 모양으로 구분이 어려운 단어들이 존재하기 때문에 이들간의 오인식율이 높았다. 상대적으로 발음간 구분이 더 쉬운 CITY 데이터베이스의 경우에는 더 낮은 오인식율을 얻었다.

3. 제안하는 필터링 기법

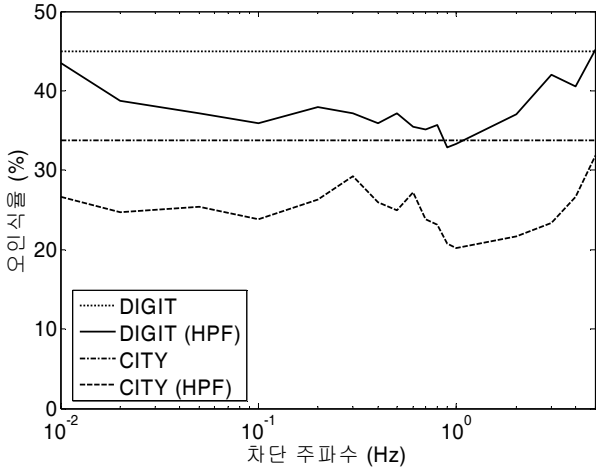
입술 영역의 각 좌표별 픽셀값의 시간에 대한 궤적은 0Hz인 직류(dc) 성분부터 샘플링 주파수 30Hz의 절반인 15Hz의 주파수 성분까지 포함한다. 하지만 인식에 도움이 되는 성분은 그 일부분에 국한되어 있다. 심리학 분야의 연구에서는 음절에 대한 입술의 움직임은 대략 4Hz의 주파수 성분으로 이루어져 있고, 이 주파수를 기본으로 하여 높은 성분의 주파수 성분이 변조 형태로 포함되어 있는 것으로 알려져 있다[15,16].

본 논문에서 제안하는 기법은 입술영역 영상에서 각 픽셀값의 시간 궤적에 대해 BPF를 적용하여 음성 정보와 관련이 없는 주파수 성분들, 즉 너무 빨리 변하거나 너무 느리게 변하는 성분들을 제거하고자 하는 것이다. 이러한 성분들은 화자나 조명과 같은 음성과 무관한 변화들이나 잡음에 의한 효과에 의한 것이기 때문에 이를 제거함으로써 인식 성능의 향상을 얻을 수 있다. 입술 영역 영상의 각 픽셀에 대해 필터링을 한 후 2.3절에서 설명한 것과 같이 평균 제거와 PCA를 적용하여 향상된 성능을 내는 특징을 얻는다.

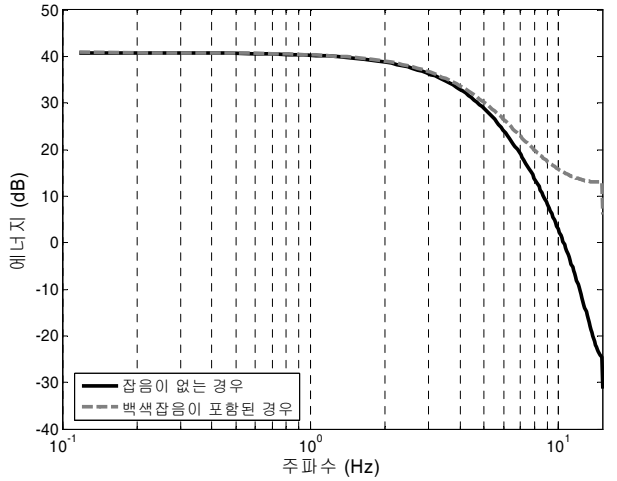
이하에서 BPF의 저역 차단 주파수(lower cut-off frequency)와 고역 차단 주파수(higher cut-off frequency)를 결정하는 방법을 설명한다. 본 논문에서 필터는 Butterworth 필터[17]로 설계한다. Buterworth 필터는 Chebyshev 필터나 elliptic 필터와 같은 필터들과는 달리 통과대역이나 차단대역에 굴곡(ripple)이 존재하지 않아 원하는 주파수 대역을 왜곡없이 전달하고 원치않는 주파수 대역을 완전히 차단할 수 있다. Butterworth BPF는 짝수의 차수를 가지게 되는데, 데이터베이스에 포함된 동영상의 프레임 수가 크지 않기 때문에 필터의 차수를 작은 값으로 설정한다. 따라서 제안하는 필터는 4차 Butterworth BPF로 구현한다.

3.1 저역 차단 주파수의 결정

저역 차단 주파수를 결정하기 위해 5Hz 이하의 주파수를



(그림 3) HPF의 차단 주파수에 따른 오인식율의 변화



(그림 4) 잡음이 존재하지 않는 경우 및 백색잡음이 존재하는 경우에 대한 주파수에 따른 에너지의 분포

차단 주파수로 하는 고역통과필터(HPF: high-pass filter)를 설계하고 설계용 데이터에 대한 실험을 통해 가장 좋은 성능을 내는 경우를 선택한다. 저역 차단 주파수를 0.01Hz에서 5Hz까지 변화시키면서 Butterworth HPF를 설계하고 적용하여 특징을 추출하고 그 인식 성능을 (그림 3)에 나타내었다. 통과대역과 차단 대역 사이에서 최종 BPF와 비슷한 roll-off factor(-40dB/decade)를 얻기 위해 필터의 차수는 2차로 하였다.

HPF를 적용함으로써 향상된 성능을 얻을 수 있음을 그림에서 확인할 수 있다. 대략 0.6~1Hz의 차단 주파수를 가질 때 두 데이터베이스 모두에 대해 좋은 성능을 보이고 있다. 두 데이터베이스에 대해 최적의 차단 주파수는 약간의 차이가 있으나 두 경우 모두 좋은 성능을 나타내는 값인 0.9Hz를 차단 주파수로 한다. 그러한 2차 Butterworth HPF는 다음의 식으로 얻어진다[17].

$$H(z) = \frac{0.8752 - 1.7504z^{-1} + 0.8752z^{-2}}{1 - 1.7347z^{-1} + 0.7660z^{-2}} \quad (2)$$

위의 식으로 주어지는 필터는 0.9Hz에서 최대응답크기 대비 3dB의 감쇠를 보이고 통과대역과 차단대역 사이에서 -40 dB/decade의 roll-off factor를 보인다.

3.2. 고역 차단 주파수의 결정

고대역 주파수 신호를 차단하는 것은 잡음에 의한 오염을 제거하기 위한 것이다. 영상에는 기록이나 전송 과정에서 잡음이 포함될 수 있으며, 채널 잡음, 영상 기록시 포함되는 잡음, CCD 카메라의 잡음 등 여러 잡음들은 백색잡음으로 근사화하여 표현할 수 있다[9].

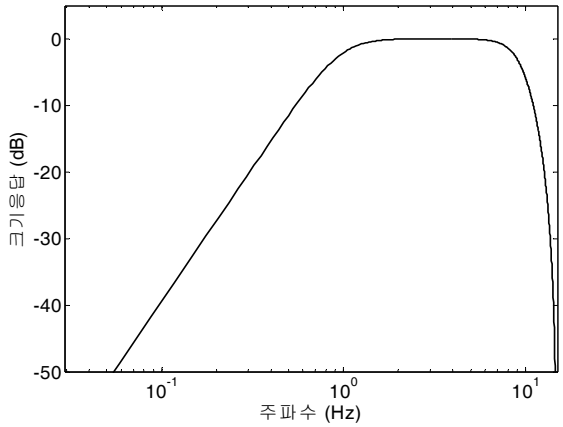
고역 차단 주파수를 결정하기 위해 잡음이 존재하지 않는 경우와 백색잡음이 존재하는 경우의 주파수 스펙트럼 분석을 수행하였다. (그림 4)는 각 경우에 대해 주파수에 따른 에너지의 분포를 나타낸 것이다. 필터 설계용 데이터의 모든 단어에 대해 각 발음의 평균적인 에너지의 분포를 나타낸다. 그림에서 보는 것과 같이 잡음의 효과는 높은 주파수 대역의 에너지를 증가시키는 것으로 나타난다. 이러한 고주파 대역의 에너지 증가는 대략 8~9Hz 이상에서 두드러진다.

기존의 심리학 연구에서는 다양한 프레임 비율의 입술 움직임 동영상에 대한 인간의 독순 인식 성능을 비교한 바 있다[18]. 그 결과, 16.7Hz 정도의 프레임 비율이 인식에 충분한 것으로 나타났으며, 그 이상의 프레임 비율은 인식에 큰 도움이 되지 못하였다. 이 결과는 16.7Hz의 절반인 8.35Hz 이하의 변화율을 가지는 정보가 인식에 충분함을 의미한다.

이상의 스펙트럼 및 심리학적 분석들을 통해 잡음에 의한 오염이 심하면서 인식에 중요한 정보가 많지 않은 고주파 대역을 제거하여 강인한 인식 성능을 얻을 수 있음을 알 수 있다. 따라서, 제안하는 BPF의 고역 차단주파수는 인식에 중요한 정보의 변화율이 포함되는 8.35Hz에 약간의 여유를 둔 9Hz로 설정한다. 이러한 여유를 두는 것은 차단 주파수가 통과대역의 최대응답보다 3dB 감쇠된 응답을 얻는 주파수이므로 8.35Hz에서도 감쇠없이 신호가 통과되도록 하기 위함이다. 앞 절에서 정한 저역 차단 주파수 0.9Hz와 고역 차단 주파수 9Hz를 가지는 4차 Butterworth BPF는 다음과 같이 얻을 수 있다[17].

$$B(z) = \frac{0.3307 - 0.6614z^{-2} + 0.3307z^{-4}}{1 - 1.4261z^{-1} + 0.4618z^{-2} - 0.1566z^{-3} + 0.1754z^{-4}} \quad (3)$$

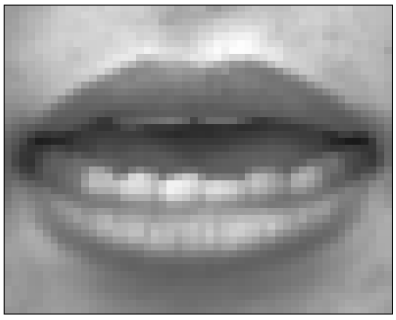
(그림 5)는 제안하는 BPF의 주파수 응답을 나타낸다. Butterworth 필터를 사용하였기 때문에 통과 대역에서 매우 평평한 응답을 보이며 이를 통해 왜곡 없는 신호의 전달이 가능하다. 그리고 설정한 차단 주파수 밖의 저역 및 고역 주파수 대역 신호를 차단할 수 있다.



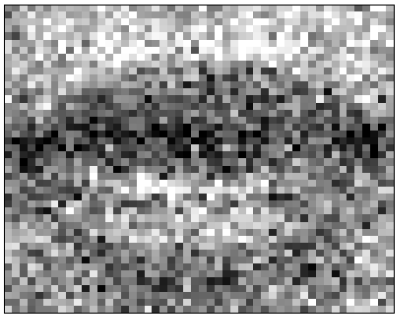
(그림 5) 제안하는 BPF의 주파수 응답

4. 실험 및 결과

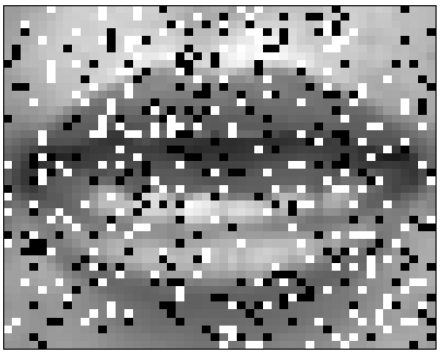
본 장에서는 제안하는 특징추출 기법의 성능을 2.1절에서 서술한 두 데이터베이스에 대한 인식실험을 통해 알아본다. 제안하는 기법에서 특징 추출의 과정은 입술 영역 영상열에 식 (3)으로 주어진 BPF를 적용하고 각 발음에 대한 평균 제거 후 PCA를 통해 각 영상마다 12개의 특징을 얻는다. BPF를 사용하지 않은 경우(2.3절 참조)와 인식 성능을 비교하며, 두 경우 모두 인식기로써 2.4절에서 설명한 연속 HMM을 사용한다. 실험에서 사용한 영상의 예를 (그림 6)에 나타내었다.



(a)



(b)

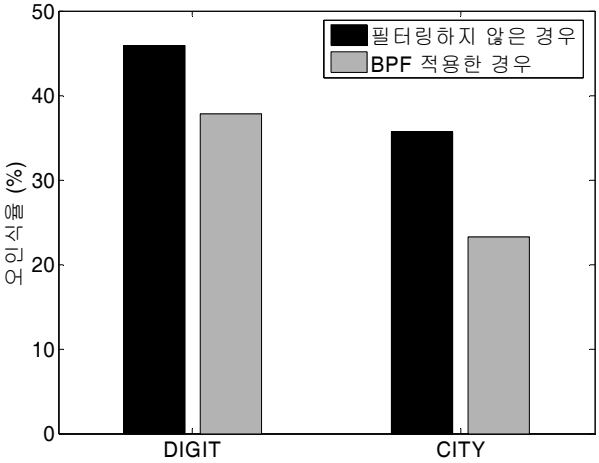


(c)

(그림 6) 실험에 사용된 입술영역 영상의 예 (a) 잡음이 없는 영상 (b) 백색잡음이 포함된 영상 (PSNR=15dB) (c) 충격잡음이 20% 포함된 영상

4.1. 잡음이 없는 환경

먼저, 제안하는 기법을 잡음이 없는 영상에 대해 실험하여 성능을 알아본다. (그림 7)은 BPF를 적용하지 않은 경우와 적용한 경우의 오인식율을 비교한다. 그림에서 보는 것과 같이 제안하는 기법에 의해 인식의 성능이 크게 향상되는 것을 볼 수 있다. 각 데이터베이스에 대한 오인식율은 DIGIT의 경우 45.9%에서 35.7%로, CITY의 경우 37.8%에서 23.2%로 감소하였다. 이는 각각 22.2%와 38.6%의 상대적 오인식율 감소를 얻은 것이다. 이러한 성능 향상은 낮은 주파수 대역에 포함되어 있는 음성정보와 무관한 성분들을 제거함으로써 얻는 이득이다.



(그림 7) 잡음이 없는 경우 제안하는 필터링 기법의 성능 비교

4.2 백색잡음이 존재하는 환경

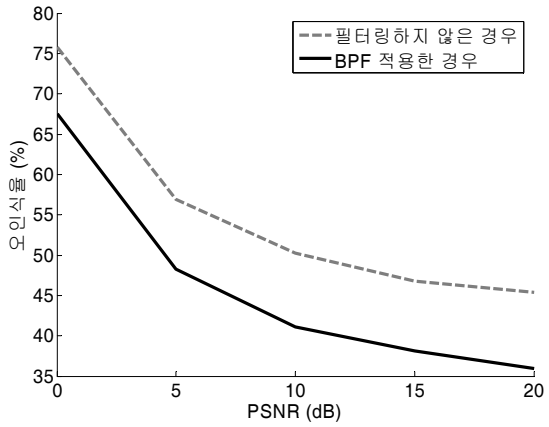
다음으로, 영상에 백색잡음이 포함된 경우에 대해 제안하는 기법의 성능을 알아본다. 백색잡음은 평균이 0인 가산잡음으로써 각 픽셀에 정규분포의 무작위값을 더하는 형태로 영상에 포함된다. 즉, 잡음이 존재하지 않는 영상을 I, 잡음신호를 η라 하면, 잡음섞인 영상 K는

$$K(m,n) = I(m,n) + \eta(m,n) \tag{4}$$

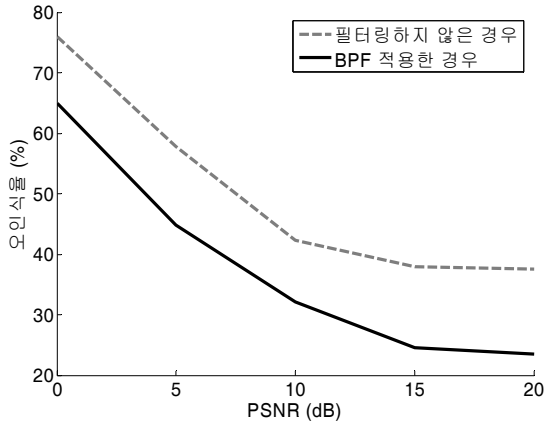
로 얻어진다. 여기서 (m,n) 은 픽셀 좌표이다. 백색잡음의 양은 다음 식으로 정의되는 최대신호 대 잡음비(PSNR: peak-signal-to-noise ratio)로 표현한다.

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{(\text{maximum pixel value})^2 \times M \times N}{\sum_{m=1}^{M}\sum_{n=1}^{N}(I(m,n)-K(m,n))^2}\right) \tag{5}$$

여기서, M 과 N 은 영상의 가로와 세로의 크기를 의미한다. 잡음신호 η 의 크기를 조정하여 PSNR을 20~0dB까지 5dB 간격으로 한 영상을 생성하여 인식 실험을 하였으며 그 결과를 (그림 8)에 나타내었다.



(a)



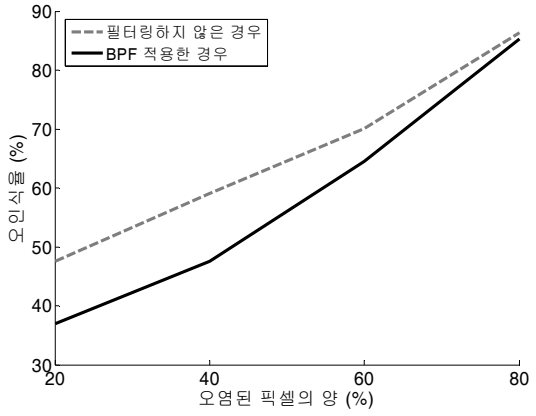
(b)

(그림 8) 백색 잡음이 포함된 경우 제안하는 기법의 성능 (a) DIGIT 데이터베이스 (b) CITY 데이터베이스

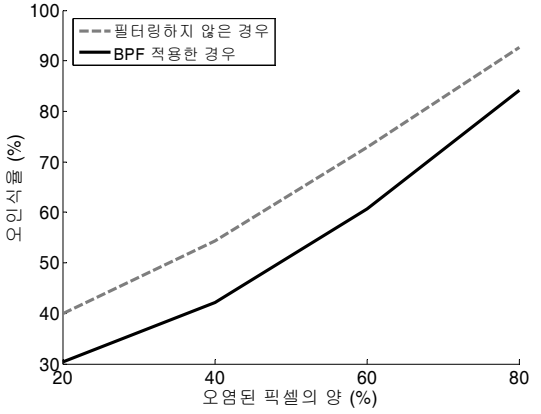
그림의 결과로부터 다양한 잡음 수준에 대해서 제안하는 기법을 사용하여 더 낮은 오인식율을 얻음을 확인할 수 있다. 잡음에 의한 고대역 주파수의 오염 효과를 제거함으로써 강인한 영상 특징을 추출할 수 있음을 알 수 있다. 상대적 오인식율 감소량은 각 데이터베이스마다 여러 잡음 수준에 대해 평균 16.7%와 26.7%로 나타났다.

4.3. 충격잡음이 존재하는 환경

영상에 포함될 수 있는 다른 형태의 잡음으로 충격(impulse) 잡음을 들 수 있다. 충격잡음은 영상 내의 각 픽셀값이 무작위로 0 또는 최대픽셀값으로 바뀌는 형태의 잡음이다. 제안하는 특징 추출 기법의 성능을 충격잡음이 포함되어 있는 영상에 대해 알아본다. 오염된 영상은 각 입술 영역 영상의 픽셀중 20%~80%를 무작위로 선택하여 각 픽셀마다 0 또는 최대픽셀값으로 변화시켜 얻는다. (그림 9)는 오염픽셀의 비율에 따라 제안하는 기법의 오인식율을 비교하여 나타낸 것이다. 충격잡음에 대해서도 제안하는 기법이 다양한 잡음수준에 대해 더 낮은 오인식율을 나타내는 것을 볼 수 있다. 충격잡음 역시 백색잡음과 마찬가지로 고주파의 에너지를 증가시키기 때문에 BPF를 통과시켜 잡음효과를 억제할 수 있다.



(a)



(b)

(그림 9) 충격잡음이 포함된 경우 제안하는 기법의 성능 (a) DIGIT 데이터베이스 (b) CITY 데이터베이스

5. 맺음말

본 논문에서는 자동 독순의 성능을 향상시키기 위한 필터링 기법을 제안하였다. 제안하는 필터는 음성정보와 관련이 적고 잡음에 의해 손상되기 쉬운 저대역 및 고대역의 주파수 성분을 제거하는 BPF로써, 인식 실험, 스펙트럼 분석 및 심리학적 근거에 기반하여 차단 주파수를 결정하였다. 필터링을 거친 영상에 PCA를 적용하여 최종 특징을 얻었다. 실험 결과 잡음이 존재하는 경우와 존재하지 않는 경우 모두 상당한 인식 성능의 향상을 얻을 수 있었다. 추후 과제로써 제안된 기법을 연속음성과 같이 다양한 데이터베이스에 대해 적용하는 연구를 계속할 것이다.

참 고 문 헌

[1] C. C. Chibelushi, F. Deravi, J. S. D. Mason, “A review of speech-based bimodal recognition,” IEEE. Trans. Multimedia, Vol. 4, No. 1, pp. 23-37, 2002.

[2] H. Yao, W. Gao, W. Shan, and M. Xu, "Visual features extracting and selecting for lipreading," in Proc. Int. Conf. Audio- and Video-based Biometric Person Authentication, Guildford, UK, pp. 251-259, Jun. 2003.

[3] 이종석, 심선희, 김소영, 박철훈, “제어되지 않은 조명 조건하에서 입술 움직임의 강인한 특징추출을 이용한 바이모달 음성인식,” Telecommunications Review, 제14권 제1호, pp. 123-134, 2004년 2월.

[4] C. Bregler and Y. Konig, “Eigenlips for robust speech recognition,” in Proc. Int. Conf. Acoustics, Speech and Signal Processing, Vol. 2, Adelaide, Austria, pp. 669-672, 1994.

[5] G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu, “A cascade image transform for speaker independent automatic speechreading,” in Proc. Int. Conf. Multimedia and Expo, Vol. 2, New York, pp. 1097-1100, 2000.

[6] P. Scanlon and R. Reilly, “Feature analysis for automatic speechreading,” in Proc. Int. Conf. Multimedia and Expo, Tokyo, Japan, pp. 625-630, Apr. 2001.

[7] G. Potamianos and C. Neti, “Audio-visual speech recognition in challenging environments,” in Proc. Eurospeech, Geneva, Switzerland, pp. 1293-1296, Sep. 2003.

[8] K. Saenko, T. Darrell, J. Glass, “Articulatory features for robust visual speech recognition,” in Proc. Int. Conf. Multimodal Interfaces, State College, PA, pp. 152-158, Oct. 2004.

[9] A. Amer and E. Dubois, “Fast and reliable structure-oriented video noise estimation,” IEEE. Trans. Circuits and Systems for Video Technology, Vol. 15, No. 1, pp. 113-118, Jan. 2005.

[10] 이종석, 박철훈, “시청각 음성인식을 위한 정보통합: 신뢰도 측정방식의 비교와 신경회로망을 이용한 통합 기법,” Telecommunications Review, 제17권 제3호, pp. 538-550, 2007년 6월.

[11] 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, and Cybernetics, pp. 198-202, Taipei, Taiwan, Oct. 2006.

[12] R. C. Gonzalez and R. E. Woods, ‘Digital Image Processing,’ Prentice-Hall, Upper Saddle River, NJ, 2001.

[13] S. Lucey, “An evaluation of visual speech features for the tasks of speech and speaker recognition,” in Proc. Int. Conf. Audio- and Video-based Biometric Person Authentication, Guildford, UK, pp. 260-267, Jun. 2003.

[14] X. Huang, A. Acero, and H.-W. Hon, “Spoken Language Processing,” Prentice-Hall, Upper Saddle River, NJ, 2001.

[15] J. J. Ohala, “The temporal regulation of speech,” in Auditory Analysis and Perception, eds., G. Fant and M. A. Tatham, Academic Press, London, UK, pp. 431-453, 1975.

[16] K. Munhall and E. Vatikiotis-Bateson, “The moving face during speech communication,” in Hearing by Eye II: Advances in the Psychology of Speechreading and Audio-Visual Speech, eds., R. Campbell, B. Dodd, and D. Burnham, Psychology Press, Hove, UK, pp. 123-142, 1998.

[17] J. G. Proakis and D. G. Manolakis, ‘Digital signal processing,’ Prentice-Hall, Upper Saddle River, NJ, 1996.

[18] M. Vitkovitch and P. Barber, “Visible speech as a function of image quality: effects of display parameters on lipreading ability,” Applied Cognitive Psychology, Vol. 10, pp. 121-140, 1996.

이 종 석

e-mail : jslee@nnmi.kaist.ac.kr

1999년 한국과학기술원 전기및전자공학과 학사

2001년 한국과학기술원 전자전산학과 공학석사

2006년 한국과학기술원 전자전산학과 공학박사

2006년~현재 한국과학기술원 전자전산학부 연수연구원

관심분야 : 시청각 음성인식, 멀티모달 인터페이스

박 철 훈

e-mail : chpark@kaist.ac.kr

1984년 서울대학교 전자공학과 학사

1985년 Caltech 전자공학과 공학석사

1990년 Caltech 전자공학과 공학박사

1991년~현재 한국과학기술원 전자전산학부 교수

관심분야 : 지능시스템, 신경회로망, 최적화, 지능제어