한국음향학회지 제22권 제5호 pp. 375~382 (2003)

시간영역 필터를 이용한 립리딩 성능향상에 관한 연구

A Study on Lip-reading Enhancement Using Time-domain Filter

신 도 성*, 김 진 영**, 최 승 호***
(Do Sung Shin*, Jin Young Kim**, Seung Ho Choi***)

*전남대학교 전자공학과, **전남대학교 정보통신공학부 & RRC HECS, ***동신대학교 정보통신공학과
(접수일자: 2002년 12월 4일; 수정일자: 2003년 5월 20일; 채택일자: 2003년 6월 9일)

현재 음성인식 분야에서는 잡음이 심한 환경에서 음성 인식률을 향상시킬 수 있는 바이모달의 한 형태인 립리딩 기술에 관한 연구가 활발히 진행되고 있다. 립리딩 연구에 있어서 가장 중요한 것은 정확한 입술 이미지를 찾아내는 것이다. 그러나 조명변화, 화자의 발음습관, 입술 모양의 다양성, 입술의 회전과 크기 변화 등의 환경 변화 요인 때문에 안정적인 성능을 예측하기가 힘든 실정이다. 본 논문에서는 보다 안정적 성능을 얻기 위해 시간영역에서 이미지를 임펄스 응답 필터링을 수행을 통해 향상된 인식성능을 보였다. 또한 본 연구에서는 입술 전체 영상을 대상으로 처리하는 립리딩 기법의 사용으로 인해 발생하는 데이터 용량 증가를 고려해 영상의 정보는 손실하지 않고 그 특징만을 추출하여 데이터의 양을 줄일 수 있는 주성분 분석을 전처리 과정으로 사용하였다. 본 연구에서는 영상정보만을 사용하여 음성인식 성능 관찰을 위해 자동차 내에서 서비스가 가능한 22단어를 선정하여 인식실험을 하였다. 이 단어들의 인식 성능을 비교하기 위하여 음성 인식 알고리듬으로 잘 알려진 HMM (Hidden Markov Model)을 이용하였다. 실험결과 PCA (Principal Component Analysis)만 하였던 경우 립리딩이 64%의 인식률을 보인 반면, 시간영역필터를 립리딩에 적용시 72.7%로 인식률의 향상을 보였다.
핵심용어: 립리딩, 주성분 분석, 시간영역, 필터, HMM
투고분야: 음성처리 분야 (2,5)

Lip-reading technique based on bimodal is to enhance speech recognition rate in noisy environment. It is most important to detect the correct lip-image. But it is hard to estimate stable performance in dynamic environment, because of many factors to deteriorate Lip-reading's performance. There are illumination change, speaker's pronunciation habit, versatility of lips shape and rotation or size change of lips etc. In this paper, we propose the IIR filtering in time-domain for the stable performance. It is very proper to remove the noise of speech, to enhance performance of recognition by digital filtering in time domain. While the lip-reading technique in whole lip image makes data massive, the Principal Component Analysis of pre-process allows to reduce the data quantity by detection of feature without loss of image information. For the observation performance of speech recognition using only image information, we made an experiment on recognition after choosing 22 words in available car service. We used Hidden Markov Model by speech recognition algorithm to compare this words' recognition performance. As a result, while the recognition rate of lip-reading using PCA is 64%, Time-domain filter applied to lip-reading enhances recognition rate of 72.4%.
Keywords: *Lip-reading, PCA, Time-domain, Filter, HMM*
ASK subject classification: *Speech signal processing (2,5)*

I. 서 론

최근 음성인식 분야에서는 심한 잡음 환경에서 인식률을 높이기 위한 연구가 활발히 진행되고 있다. 현재 인식 기술 수준은 실험실과 같이 잡음을 거의 배제한 환경에서는 뛰어난 인식률을 보이고 있으나 소음이 많이 발생하는 자동차 내부, 사무실, 거리 같은 실생활에 적용할 때는 인식률이 매우 저하된다.

립리딩은 음성인식 분야 중 잡음 환경에서 현저하게 떨어지는 인식율을 높이기 위한 보상 방법으로 화자의 입술을 포함한 영상 정보를 이용하는 목적으로 연구되었다[1-4].

립리딩은 아직 실용화 단계에 있지 못하지만, 많은 연

책임저자: 신도성 (bombool@hanmail.net)
500-757 광주광역시 북구 용봉동 전남대학교
전자공학과 DSP Lab
(전화: 062-530-0472; 팩스: 062-530-0472)

구가 진행 중이다. 그 방법으로는 모델기반과 이미지 기반 방법이 있으며, 본 논문에서는 이미지를 기반으로 하여 영상정보를 음성정보에 이용하는 립리딩 기술을 바탕으로 연구하였다. 실험에 사용된 이미지 기반 방법은 입술 전체 영상을 처리하므로 잘못된 파라미터로 인해 인식율이 저하하는 다른 방법보다는 안정된 인식률을 보이는 장점이 있는 반면, 입술 전체영상을 특징 파라미터로 사용하기 때문에 데이터 용량의 증가에 따른 인식 속도 저하가 발생한다.

이같은 문제해결을 위해 본 논문에서는 입력된 전체 입술 영상을 그레이 변환 후 입술 관심영역 (ROI: Region Of Interest) 추출을 통해 입술만을 따로 분리하고 다운샘플링을 과정을 통해 입술의 정보손실을 최소화하는 범위 내에서 데이터 처리량을 줄이기 위한 작업을 수행하였다. 그리고 입술형태가 좌우 대칭인 점에 착안하여 입술 ROI 영상의 절반만을 이용해서 영상을 처리하였다. 접어진 영상은 입력 이미지를 시간영역에서 필터링하는 과정을 통해 입술 이미지에서 불필요한 정보를 제거하였다. 이처럼 처리된 이미지에 대해 PCA를 수행하여 중요한 몇 개의 특징 파라미터만을 추출해 처리하므로써 인식 속도를 현저히 단축하고 인식률을 증가시킬 수 있었다.

본 논문에서는 안정적인 성능을 보이는 립리딩 구현을 위해 립리딩 성능을 저하시키는 원인을 분석하고 그 보상 방법으로 몇 단계 전처리 과정 중 PCA를 통해 추출된 파라미터를 사용한 성능 향상 방법과 시간 영역 필터를 이용한 방법에 대해 잡음이 배제된 실내에서 인식 실험을 수행하고 그 결과를 비교 분석하여 인식률을 살펴보았다.

II. 기준 립리딩 시스템의 구현

실험을 위해 이용된 데이터는 일반인 남성 화자 70명을 대상으로 운전자들이 요구하는 정보에 대해 서비스가 가능한 22단어를 발음하여 영상 데이터 및 음성 데이터를 구축하였다.

영상 데이터의 수집은 30 Frames/sec 속도로 컬러 이미지로 코에서부터 턱까지만 320 x 240 크기로 고정하여 수집하였고 음성데이터는 일반 마이크를 사용하여 수집하였다.

실험에 사용된 단어들은 일반적으로 자동차 주행 중에 사용자가 요구할 수 있는 정보들에 대하여 서비스가 가능한 단어들로 선정하여 데이터를 구축하였다.

본 논문에서 기준 립리딩시스템은 이미지 기반 립리딩

표 1. 단어인식 실험을 위한 선정단어
Table 1. The selected words for recognition experimen.

number	word	number	word
1	메뉴명	12	문화정보
2	뉴스	13	증권정보
3	메인메뉴	14	종합지수
4	정치	15	등락종목
5	경제	16	종목시세
6	사회	17	투자정보
7	스포츠	18	교통정보
8	방송정보	19	교통하나
9	표준FM	20	교통둘
10	음악FM	21	교통셋
11	연예정보	22	교통넷

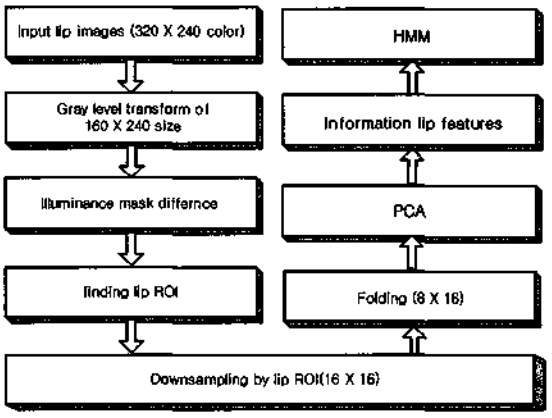

그림 1. 기준 립리딩 알고리즘 블록도
Fig. 1. Diagram of baseline lipreading algorithm.

방법을 적용하여 모델링하였으며 그림 1에 본 연구를 위해 구현한 기준 립리딩 알고리듬의 블록도를 도시하였다.

음성과 영상의 동기화를 위해 화자가 단어를 발음하는 구간 동안에 음성 분석 알고리듬을 이용하여 음성의 시작점과 끝점을 찾고 이를 토대로 영상 정보에서 발음의 시작점과 끝점을 맞추는 동기화 방법을 이용하였다.

그림 2에 발음 구간 동안의 음성과 영상 동기화 과정을 나타내었다.

영상에서 화자가 발음하는 동안에 정확한 음성구간의 결정은 피치를 이용한 방법과 신호의 ZCR (Zero Cross Rate)을 이용한 음성 구간의 시작점 끝점 검출 알고리듬을 이용하였다[5]. 동기화에 의해 얻어진 프레임들은 먼저 각 프레임에 대해 8 bits 그레이 영상으로 변환하여 이 영상에 대해 명암 마스크 차감을 시켜 일차적인 명암의 영향을 제거한 후 ROI를 찾는다. 찾아진 ROI는 다시 16 × 16 사이즈 이미지로 다운 샘플링하여 파라미터 양을 줄인다. 입술 ROI의 특징 파라미터는 매 프레임마다 추출

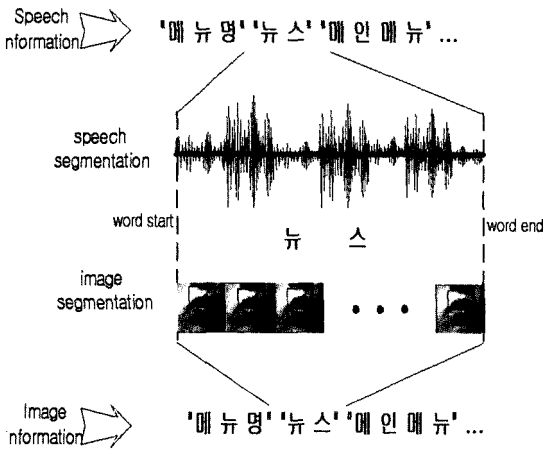

그림 2. 발음 구간 동안의 음성과 영상 동기화
Fig. 2. Synchronization of speech and image signals.

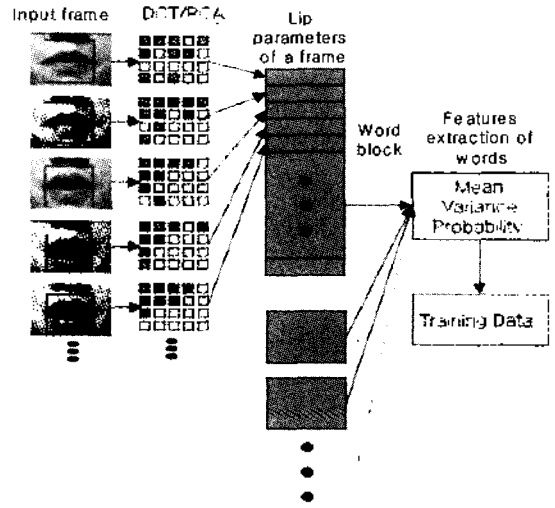

그림 3. 립리딩에서 영상 특징 추출
Fig. 3. Overview of the visual features extraction for lip-reading.

하게 되고, 그 결과 각각의 단어를 발음하는 구간 전체 프레임에 대해서 특징 파라미터를 가지게 된다. 이 파라미터들은 화자가 발음하는 음성 구간 동안만 추출한다. 입술 특징 파라미터는 프레임별로 추출하게 되고, 발음하는 구간 전체 프레임에 대해서 입술파라미터를 가지게 된다. 그림 3에 영상에서 입술 파리미터 처리과정과 PCA 적용 과정에 대하여 도식하였다.

입술 ROI 안의 영상은 입술과 입술 주변의 피부가 대상이 되며 발음하는 동안 화자의 입술 움직임은 계속적으로 변화가 된다. 이미지 크기와 비례하는 파라미터 수를 가지고 있으므로 PCA 과정을 통해 데이터의 양을 축소하였다. PCA 과정을 거치게 되면 입력된 입술 영상이 갖는 특징 정보를 거의 포함 소수 m 차원 파리미터로 축약시킬 수 있고 여기서 추출된 주성분들이 최종적으로 HMM 인식 알고리듬에 사용될 입술 특징 파라미터들로 확정된다.

III. 시간영역 필터링에 의한 립리딩 알고리듬

3.1. 립리딩 성능 저하 원인

립리딩의 인식율 저하에 영향을 주는 요인은 화자의 변화, 화자의 움직임, 그리고 조명의 변화로 크게 세 가지로 볼 수 있다.

먼저 화자 변화에 대한 요인을 살펴보면 화자 독립 시스템에서 화자들은 다양한 발음습관을 보이며 같은 단어를 발음할 경우에도 입술이 움직임과 혀와 치아의 보이는 정도가 달라 인식 성능의 저하에 원인이 된다. 다음으로 화자의 움직임에 대한 요인을 살펴보면, 실제 환경에서 인간은 말을 할 때 고정 자세로 하지 않는다. 설령 똑바로 서있다고 하더라도 행동 습관에 의해서 발생하는 머리의 움직임 또는 발음 습관에 의한 입술의 움직임은 일정하게 카메라에 포착되지 않는다. 이처럼 다양하게 포착되는 입술 때문에 인식율은 저하되게 된다. 마지막으로 주위 환경 변화에 따른 조명의 변화를 들 수 있다. 조명은 실생활에서 보면 아침, 정오, 저녁에 각각 조사 강도와 방향이 다르며, 그에 따른 영상의 왜곡 또한 판이하게 다르다. 조명은 카메라로부터 입력되는 영상에서 실제 색 정보를 왜곡시켜 영상 분석을 통해 파라미터를 추출하여 인식하는 방식을 사용하는 경우 매우 큰 성능 저하를 보이게 된다[6].

그림 4에서 보여주는 결과는 조명 변화가 없는 실내 환경과 시간대별 빛의 조사 방향을 다르게 하여 실험한 동적 환경에서의 립리딩 성능을 비교한 결과이며, 동적 환경에서 인식율이 크게 저하하고 있음을 보여 주고 있다. 실내 환경을 100% 기준으로 잡았을 경우 55% 정도의 왜곡을 발생시킨다. 이러한 왜곡 현상은 실내 환경의 데

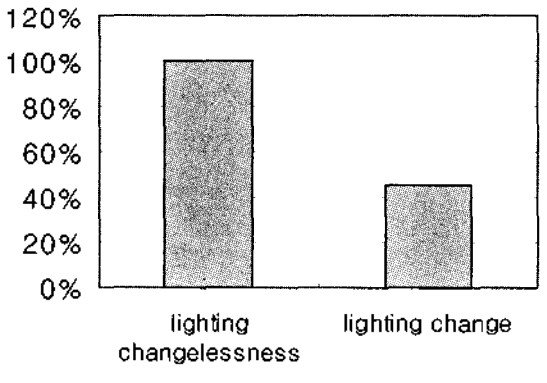

그림 4. 조명 무변화시와 조명변화시 인식율 비교
Fig. 4. Comparison of recognition rate when lighting change and changelessness.

이터로 학습화를 시켜서 테스트한 결과이기 때문이기도 하겠지만, 그만큼 이미지 기반 방법이 명암이나 조도 변화에 민감하게 반응을 하기 때문이다[10-13].

본 논문에서는 이러한 성능저하를 보상하기 위해서 명암이나 조도 변화에 강인한 시간 영역 필터링기법을 사용하였으며, 이 방법을 통해 실제 명암의 변화나 잡음 등이 줄어들어 인식성능이 향상되었음을 확인하였다.

3.2. 시간 영역 필터링을 이용한 특징 파라미터 추출

인식기술이 사용되는 실제 환경은 단순한 변환들의 집합 특히, 주위의 임펄스 응답의 컨벌루션이나 주위 환경 잡음의 합에 의해서 모델화를 할 수 있다. 이런 주위 환경들의 시간적 특성은 음성의 시간적 특성과 매우 다른 경향을 보인다. 이와 같이 음성과 잡음이 다른 점을 이용하여 시간 영역에서 음성 인식 향상에 강인하게 작용하는 필터를 연구하였다. 이 필터는 잡음 환경하에서 견인하게 잡음을 제거할 수 있어 음성 인식 성능을 효과적으로 향상시킨다[10].

본 논문에는 이 시간 영역 필터의 장점을 기본 립리딩 시스템에 적용하여 새로운 시스템을 구성하였다. 실제 실험에서는 고역통과 필터링과 대역통과 필터링을 입술 ROI 이미지에 적용하여 인식률을 실험해 보았으며, 각각의 필터식은 다음과 같다.

고역통과 필터식

$$Y_t[n,m] = 0.9858 \times (X_t[n,m] - X_{t-1}[n,m]) + 0.9716 \times Y_{t-1}[n,m]$$

저역통과 필터식

$$Y_t[n,m] = 0.8638 \times (X_t[n,m] + X_{t-1}[n,m]) - 0.7257 \times Y_{t-1}[n,m]$$

여기서 $y_t[n,m]$은 시간 t에서 (n, m) 픽셀 좌표의 필터링된 이미지 출력 값이다. $X_t[n,m]$은 입력 이미지의 픽셀 값, $X_{t-1}[n,m]$은 시간 t의 과거 값이 현재 입력에 영향을 주는 무한 임펄스 응답 필터이다. 위의 저역 통과 필터식은 대역 통과 필터링 수행을 위해 고역 통과 필터링의 출력 값을 입력으로 하여 실행되며, 실제 출력 값은 대역 통과 필터링을 수행한 결과 값과 동일하다.

그림 5는 이 무한 임펄스 응답 필터의 임펄스 응답 특성을 나타내고 그림 6은 주파수 응답 특성을 나타내고 있다.

하나의 영상 프레임을 주파수 영역으로 변환하면 변하지 않는 부분은 저주파 영역, 급변하는 부분은 고주파 영역으로 나타나게 된다. 이와 마찬가지로 입술이 변하는 동영상을 시간의 영역이 아닌 주파수 영역으로 살펴본다면, 변하지 않는 부분은 저주파 영역에 변화가 심한 부분은 고주파 영역에 나타날 것이다. 또한 시간의 흐름에 따라서 픽셀 값이 이전의 픽셀 값과 많은 차이가 나면 고주파 영역에 도시되고 그렇지 않으면 저주파 영역에 도시된다. 이같은 특성을 이용해 적절한 필터를 사용하여 중요한 정보만을 추출하는 것이 필터링의 목적이라 할 수 있다.

립리딩처럼 입술 영상만을 통하여 단어를 인식하기 위해서는 입술의 움직임이 매우 중요하다. 30 Hz (frames/sec)의 속도로 입술 영역이 찾아진 데이터들은 시간의 흐름에 따라 입술 ROI는 계속적으로 변하는 부분과 변하지 않는 부분으로 나누어진다. 즉 단어를 발음하는 동안 입술 영역은 계속하여 변하고 상대적으로 입술 주변 영역들은 변화가 적다. 이때 발음을 하면서 계속적으로 변하는 부분은 고주파 영역에서 나타나고, 변화가 적은 부분은 저주파 영역에 나타나게 된다. 이를 바탕으로 고역 통과 필터와 대역 통과 필터를 각각 적용하여 실험하고 결과를 비교·분석하였다.

그림 7에 각각의 실험 방식을 도식하였다. 시간영역 필

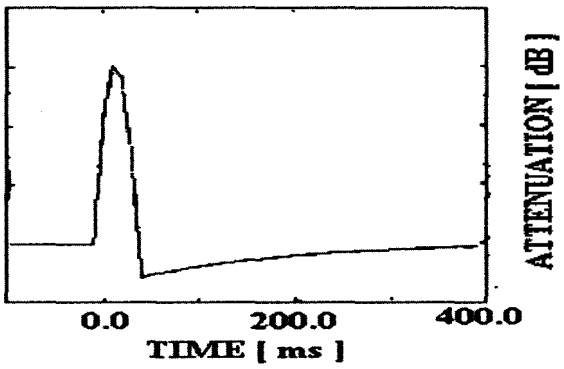

그림 5. 시간 영역 필터의 임펄스 응답
Fig. 5. Impulse response of time-domain filter.

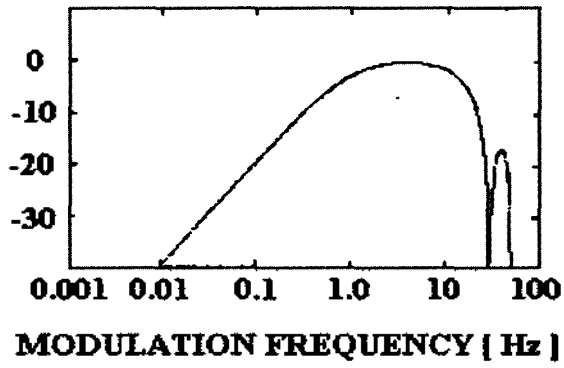

그림 6. 시간영역 필터의 주파수 응답
Fig. 6. Frequency response of time-domain filter.

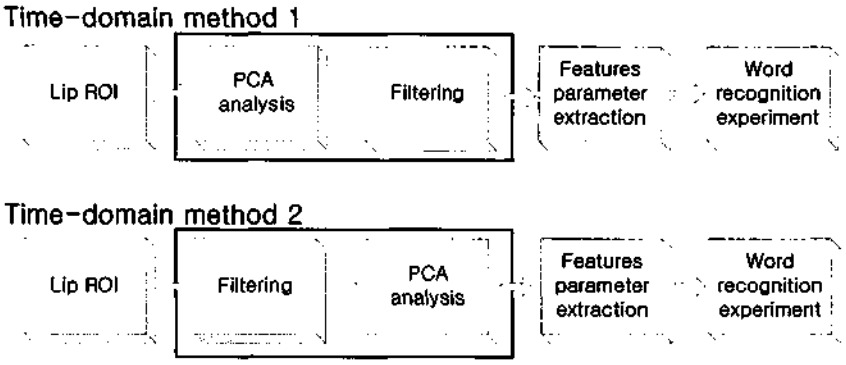

그림 7. 실험 방법 비교
Fig. 7. Comparison of experiment methods.

터링 방법 1은 PCA를 수행한 후 구해진 파라미터에 필터링을 적용한 방법이고 시간영역 필터링 방법 2는 본 논문에서 제안한 방법이다.

IV. 실험 결과

4.1. 필터링 수행유무에 따른 주성분 개수변화

실험은 크게 2가지 조건으로 나누어 수행하였다. 첫번째 조건은 필터링을 적용하지 않고 PCA만 수행하여 얻어진 주성분 개수에 대한 변화 유무에 대해서 실험했으며, 두번째로 필터링을 기준 립리딩 시스템에 적용하여 주성분 개수의 변화를 살펴 보았다. 이 경우 먼저 PCA 수행 후 얻어진 결과 파라미터를 대상으로 필터링을 한 시간영역 필터링 방법 1과 먼저 필터링을 수행 후 얻어진 파라미터를 대상으로 PCA를 한 시간영역 필터링 방법 2로 분류하여 총 3가지 경우에 대해 주성분 개수변화 실험을 수행하였다.

이 실험에서 사용되는 주성분의 개수는 반으로 접은 이미지를 아무런 필터링을 거치지 않고 PCA만을 수행하여 얻어진 파라미터의 수를 필터에 모두 사용하게 되므로 필터링에 의해서는 파라미터의 개수가 줄어들지 않는다. 단지 추출된 파라미터 성분에서 고주파나 저주파 영역의 정보만을 제거하는 것이다. 실험 결과에서도 PCA 90%를 적용하였을 때는 24개, PCA 95%를 적용할 경우는 44개의 주성분 개수가 그대로 사용되어 인식속도 향상을 기대할 수 없었다. 그래서 본 논문에서는 필터에 의해 정보를 손실하지 않고 잡음성분만을 제거하고 파라미터 수를 줄이 인식속도를 향상시키기 위해 필터링을 먼저 수행한 후 PCA를 통해 주성분을 추출하는 방법을 제안하였다. 비례의 80%, 90%와 95%는 주성분 백분율을 나타내며 PCA한 결과값이 각각 80%, 90%와 95%의 원래 신호 정보를 가지고 있음을 뜻한다.

제시한 필터링 수행 유무에 따라 PCA를 통해 추출된 주성분의 개수는 그림 8에서 보는 바와 같이 많은 차이를 보인다. 그림 8는 입술 ROI를 16 × 16으로 다운 샘플링을 한 후 반으로 접어 마주보는 픽셀들의 평균값을 8 × 16 이미지로 만들어 필터링한 결과를 누적 백분율에 의해 구해진 PCA 개수를 비교한 것이다.

그림 8에서 보는 바와 같이 입술 ROI 파라미터에 대해 필터링을 수행하면 PCA의 개수가 필터링을 전혀 하지 않은 것에 비해 절반 이상 줄어드는 것을 볼 수 있다. 결과적으로 시간영역 필터링 방법 2가 더 적은 파라미터를 추출한다는 것을 알 수 있으며, 따라서 파라미터 수를 줄여 인식 속도를 향상시킬 수 있는 장점이 있음을 알 수 있다.

본 논문에서는 기준 립리딩 시스템에 시간영역 필터링을 적용한 방식을 두가지로 분류하여 인식 실험을 하였다. 그 결과를 분석하면 PCA를 먼저 수행한 후 필터링을 한 시간영역 필터링 방법 1은 먼저 수행한 후 PCA를 먼저 수행한 후 PCA 를 수행한 시간 영역 필터링 방법 2에 비해 좋지 못한 인식 결과를 보였다. 필터링을 하지 않고 PCA

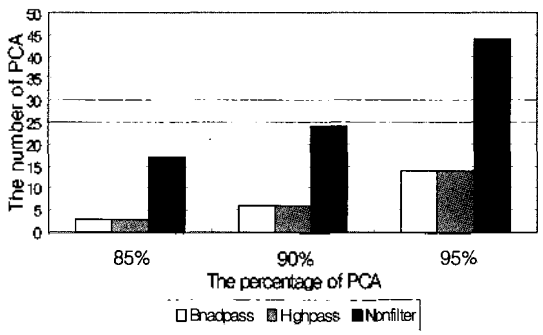

그림 8. 필터에 따른 주성분의 개수
Fig. 8. The number of PCA parameter by filter sort.

를 먼저 적용하는 시간영역 필터링 방법 1은 PCA를 수행하기 전에 먼저 필터링을 하여 본래 이미지에서 불필요한 정보를 제거한 후 주성분을 추출하는 제안한 방식과 비교해서 좋지 못한 인식 결과를 보였다. 이는 음성의 정보특성과 영상의 정보특성이 서로 다르기 때문이며, 시간영역 필터링 방법 1을 수행할 경우 인식률이 떨어지는 것은 영상 인식에 중요한 파라미터들의 손실에 의한 결과라 할 수 있다.

인식률 비교를 위한 실험은 22단어를 대상으로 실험실 내에서 수행하였다. 영상 녹화는 남성 화자 70명을 대상으로 코에서부터 턱까지 촬영하였다. 이 중 학습화를 위해 52명의 영상을, 나머지 18명의 영상은 테스트에 사용하였으며, 실험은 다음과 같은 방법으로 이루어졌다.

먼저 시간영역 필터링 방법 1로 PCA를 한 후 필터링을 통한 단어 인식 성능을 분석하였다. 다음으로 본 논문에서 제안한 방법으로 본래 이미지를 먼저 필터링하고 그 결과에 대한 특징 파라미터를 추출하여 인식 성능을 비교 분석하였다.

그림 9는 입력 이미지를 전처리 과정을 수행한 영상이미지로서 폴딩되기 이전의 다운 샘플링한 16 × 16 원 영상 이미지를 보여 주고 있다.

그림 10과 그림 11은 그림 9에 본영상의 입술 모양이 좌우가 기하학적 대칭성을 이루는 것에 착안해서 8 × 16로 접은 이미지로 시간영역 필터링 방법 1을 사용해서 PCA를 수행한 후 고역통과 필터링과 대역통과 필터링을 하여 얻어진 영상 이미지이다.

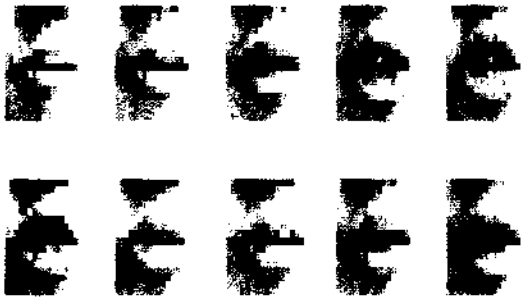

그림 9. 16×16으로 다운샘플링한 원 영상
Fig. 9. Downsampled 16×16 original image.

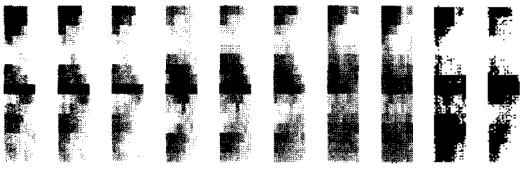

그림 10. 시간영역 필터링 방법 1: 고역통과 필터링 (8×16)
Fig. 10. Time-domain method 1: highpass filtering (8×16).

그림 12와 그림 13은 제안한 시간영역 필터링 방법 2를 사용하여 각각 고역 통과 필터와 대역 통과 필터를 사용하여 필터링한 후 PCA 수행한 결과 이미지를 보여 주고 있다.

두 실험 방법에 의해 나온 영상들은 모두 본래 영상에 비하여 약간은 흐려진 느낌이 든다. 이 현상은 픽셀마다 필터에 의해서 제거된 정보 때문이다. 두 가지 방법의 결과 영상만을 비교해 보면 시간영역 필터링 방법 1의 결과가 시간영역 필터링 방법 2의 결과에 비하여 더 많이 흐려진 것을 확인할 수 있다. 이는 시간영역 필터링 방법 1에서 중요한 영상 정보가 필터에 의해 더 많이 제거되었음을 보여 준다.

이 실험 결과에서 보면 PCA에 의하여 추출된 정보가 필터링에 의해 이미지 정보가 필터 영역에 따라 제거됨을 알 수 있으며, 따라서 시간영역 필터링 방법 1을 이용하는 것 보다 필터링을 하지 않고 PCA만 하여 파라미터를 추출한 것이 더 좋은 인식율을 나타낸다는 것을 예측할 수 있다. 이는 원 영상으로부터 추출된 파라미터를 필터링하는 것이므로 필터에 의해서 파라미터들의 정보가 손실되어 인식 성능의 저하를 가져온다는 것을 알 수 있는 결과이다.

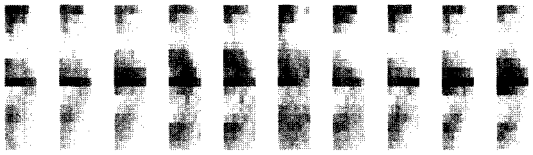

그림 11. 시간영역 필터링 방법 2: 대역통과 필터링 (8×16)
Fig. 11. Time-domain method 1: bandpass filtering (8×16).

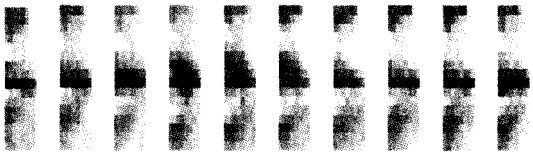

그림 12. 시간영역 필터링 방법 2: 고역통과 필터링 (8×16)
Fig. 12. Time-domain method 2: highpass filtering (8×16).

그림 13. 시간영역 필터링 방법 2: 대역통과 필터링 (8×16)
Fig. 13. Time-domain method 2: bandpass filtering (8×16).

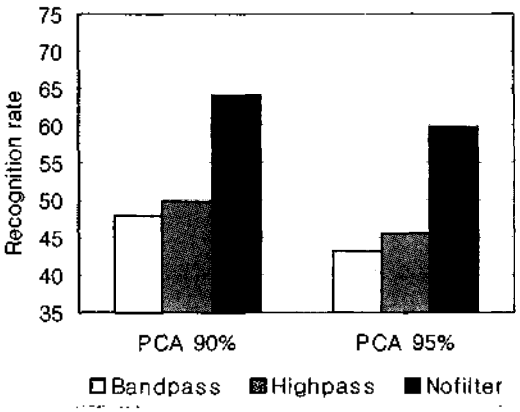

그림 14. 시간영역 필터링 방법 1의 인식율
Fig. 14. Recognition rate of Time-domain method 1.

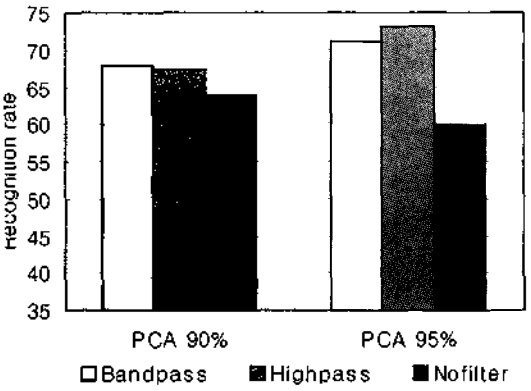

그림 15. 시간영역 필터링 방법 2 인식율
Fig. 15. Recognition rate of Time-domain method 2.

그림 14는 시간영역 필터링 방법 1을 사용해 인식 실험한 결과이다. 실험결과를 분석해 보면 필터링을 수행하지 않는 쪽의 인식률이 더 좋은 결과를 나타내고 있음을 확인할 수 있다. 이는 필터에 의해 추출된 특징 파라미터가 기존의 특징 파라미터와 비교해서 더 적은 정보를 갖고 있음을 알 수 있다. 그림 15는 본 논문에서 제안한 시간영역 필터링 방법 2를 사용하여 인식 실험한 결과이다.

위의 두 결과에서 알 수 있듯이 시간영역 필터링 방법 1에서는 필터링을 하지 않는 것이 더 좋은 인식률을 보인다. 반면 시간영역 필터링 방법 2에서는 필터링을 한 경우가 인식률이 더 좋은 것을 알 수 있다. 필터링을 수행한 데이터의 특징 파라미터도 그림 8에서 보이는 바와 같이 필터링을 하지 않는 것보다 훨씬 적음을 알 수 있다. 이렇게 많은 파라미터 수의 차이는 HMM 알고리듬에 의하여 단어를 인식할 때 속도를 절감시킬 수 있어 실시간 인식에 가능성을 보여준다.

그림 16은 PCA 90%에서 필터링을 먼저 한 후 PCA 처리한 방법과 PCA를 먼저 수행한 후 필터링을 한 두 방법의

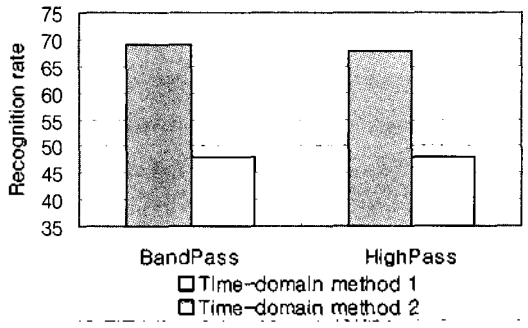

그림 16. PCA 90%에서 시간영역 필터링 방법 1과 2 비교
Fig. 16. Comparison of Time-domain method 1 and 2 in PCA 90%.

인식율을 비교한 것이다. 그림에 도식된 결과를 살펴보면 필터링을 먼저 수행한 후 PCA를 수행한 것이 더 좋은 인식 성능을 보임을 알 수 있다.

전처리로 필터링을 하면 입술 ROI의 데이터에서 변하지 않는 파라미터나 잡음이 제거되며, 이렇게 제거된 데이터에서의 주성분을 추출하므로 파라미터 량도 훨씬 축소할 수 있지만, 아무런 필터를 거치지 않고 주성분을 먼저 검출하게 되면 필터에 의해서 제거되었던 불필요한 정보들이 파라미터로 검출될 수도 있기 때문이다.

결과적으로 제안된 시간영역 필터링 방법 2가 립리딩의 성능 향상에는 효과적임을 알 수 있다. 또한 대역 통과 필터와 고역 통과 필터의 인식율만을 비교하여 보면 거의 비슷한 인식성능을 보임을 알 수 있다. 같은 개수의 주성분을 갖으면서도 인식율이 비슷하다는 것은 영상은 음성과는 달리 시간영역에서 볼 때 고주파 영역에 존재하는 급변하는 정보는 많지 않다는 것을 알 수 있다.

V. 결 론

본 논문에서는 영상 정보만으로 단어를 인식하는 방법으로 효과적인 입술 파라미터를 추출하는 립리딩에 대하여 살펴보았다.

영상이미지 기반 립리딩 방법에 시간영역 필터링의 적용은 조명의 영향에 의한 성분을 제거하여 파라미터 수를 줄여줄 뿐 아니라 인식성능을 개선시켜 주었다. 또한 필터링를 한 결과 출력 파라미터에 대해 PCA을 적용하므로써 다시 몇 개의 특징 파라미터들로 주성분 개수를 줄일 수 있다.

본 연구의 결과에서도 필터를 립리딩에 적용한 후 인식률 변화를 살펴보면 필터링을 하지 않고 PCA만 수행했을

경우 64%의 인식율을 보인 반면, 필터링을 수행하였을 경우는 72.7%으로 인식율 향상을 보였다. 그리고 파라미터 개수의 감소로 인해서 인식처리 속도 또한 현저히 단축되었다.

립리딩은 영상 정보만을 통해 단어를 인식하는 것이므로 입술 영역을 견고하게 잘 찾는 것이 매우 중요하다. 본 실험에서는 실내라는 안정적인 환경에서 입술 ROI를 추출하여 립리딩 실험을 하여 인식결과를 얻었다. 그러나 실제 립리딩의 응용 범위는 실생활에 적용되어야 하므로 추후 연구에서는 립리딩의 응용 영역은 실제 현실영역을 대상이므로 동적환경에서 보다 효율적 향상된 립리딩 방식에 대한 연구를 하고자 한다.

참 고 문 헌

1. R. Sharma, V. I. Pavlovic, and T. S. Huang, "Toward multimodal human-computer interface," *Proceedings of the IEEE*, 86 (5), May 1998.
2. G. Potamianosm, H. P. Graf, and Eric Cosatto, "An image transform approach for HMM based automatic lipreading," *Processing Of the Int. Conf. On Image Processing*, 173-177, 1998.
3. C. Bregler and Y. Konig, "Eigenlips' for robust speech recognition," *Proc. IEEE Int. Conf. On Acoustics, Speech and Signal Processing*, 669-672, 1994.
4. T. Chen, H. P. Graf and K. Wang, "Lip-syncronization using speech-assisted video processing," *IEEE Signal Processing Lett.*, 2, 57-59, 1995.
5. J. Luettin and N. A. Tracker, "Speechreading using probabilistic models," *Computer vision and Image Undestanding*, 65 (2), 163-178, Feb. 1997.
6. G. Engel, D. Greve and E. Schwartz, "Space-variant active vision and visually guided robotics," *ICPR*, 487-490, 1994.
7. 박병구, 김진영, 임재열, "입술 파라미터 선정에 따른 바이모달 음성인식 성능 비교 및 검증," 한국음향학회지, 18 (3), 68-72, 1999.
8. 박병구, 김진영, 최승호, "바이모달 음성인식의음성정보와 입술정보 결합방법비교," 한국음향학회지, 18 (4), 31-37, 1999.
9. 신도성, 김진영, 이주헌, "동적 환경에서의 립리딩 인식성능저하 요인분석에 대한 연구," 한국음향학회지, 21 (5), 471-477, 2002.
10. H. Hermansky, N. Morgan "RASTA Processing of Speech," *IEEE Transaction on Speech and Audio Processing*, 2, 587-589, October 1994.

저자 약력

● 신 도 성 (Do Sung Shin)

1993년 2월: 동신대학교 정보통신 공학과 졸업
1998년 2월: 전남대학교 대학원 전자공학과 (석사)
1999년 3월~현재: 동대학원 (박사수료)

● 김 진 영 (Jin Young Kim)

1986년 2월: 서울대학교 전자공학과 졸업
1988년 2월: 서울대학교 전자공학과 석사
1994년 2월: 서울대학교 전자공학과 박사
1994~1995년: 한국통신 소프트웨어 연구소
1995~현재: 전남대학교 전자공학과 부교수

● 최 승 호 (Seung Ho Chio)

1981년 2월: 전북대학교 물리학과 (이학사)
1984년 8월: 명지대학교 전자공학과 (공학사)
1984년 2월: 명지대학교 전자공학과 (공학석사)
1984년 3월: 명지대학교 전자공학과 (공학박사)
1992년 3월~현재: 동신대학교 정보통신공학과 교수