한국지능시스템학회 논문지 2010, Vol. 20, No. 3, pp. 348-355

시각 음성인식을 위한 영상 기반 접근방법에 기반한 강인한 시각 특징 파라미터의 추출 방법

Robust Feature Extraction Based on Image-based Approach for Visual Speech Recognition

송민규* · Thanh Trung Pham* · 민소희* · 김진영* · 나승유* · 황성택**

Song Min Gyu*, Thanh Trung Pham*, So Hee Min*, Jing Young Kim*, Seung You Na*, Hwang Sung Taek**

* 전남대학교 전자공학과
** 삼성전자 정보통신총괄 통신연구소

요 약

음성 인식 기술의 발전에도 불구하고 잡음 환경하의 음성 인식은 여전히 어려운 분야이다. 이를 해결하기 위한 방안으로 음성 정보 이외에 시각 정보를 이용한 시각 음성인식에 대한 연구가 진행되고 있다. 하지만 시각 정보 또한 음성과 마찬가지로 주위 조명 환경이나 기타, 다른 요인에 따른 영상잡음이 존재하며, 이런 영상잡음은 시각 음성 인식의 성능 저하를 야기한다. 따라서 인식 성능 향상을 위해 시각 특징 파라미터를 어떻게 추출하느냐는 하나의 관심분야이다. 본 논문에서는 HMM기반 시각 음성인식의 인식 성능 향상을 위한 영상 기반 접근방법에 따른 시각 특징 파라미터의 추출 방법에 대하여 논하고 그에 따른 인식성능을 비교하였다. 실험을 위해 105명에 화자에 대한 62단어의 데이터베이스를 구축하고, 이를 이용하여 히스토그램 매칭, 입술 접기, 프레임 간 필터링 기법, 선형마스크, DCT, PCA 등을 적용하여 시각 특징 파라미터를 추출하였다. 실험결과, 제안된 방법에 의해 추출된 특징 파라미터를 인식기에 적용하였을 때의 인식 성능은 기본 파라미터에 비해 약21%의 성능 향상이 됨을 알 수 있다.

키워드 : 시각 음성인식, 히스토그램 매칭, 입술 접기 기법, RASTA 필터링

Abstract

In spite of development in speech recognition technology, speech recognition under noisy environment is still a difficult task. To solve this problem, Researchers has been proposed different methods where they have been used visual information except audio information for visual speech recognition. However, visual information also has visual noises as well as the noises of audio information, and this visual noises cause degradation in visual speech recognition. Therefore, it is one the field of interest how to extract visual features parameter for enhancing visual speech recognition performance. In this paper, we propose a method for visual feature parameter extraction based on image-base approach for enhancing recognition performance of the HMM based visual speech recognizer. For experiments, we have constructed Audio-visual database which is consisted with 105 speackers and each speaker has uttered 62 words. We have applied histogram matching, lip folding, RASTA filtering, Liner Mask, DCT and PCA. The experimental results show that the recognition performance of our proposed method enhanced at about 21% than the baseline method.

Key Words : Visual speech recognition, Histogram matching, Lip folding, Rasta filter, PCA

1. 서 론

오늘날 음성 인식 기술은 비약적인 발전을 하였고, 저 잡음 환경 하에서의 인식 성능을 충분히 신뢰할만한 수준에 도달하였다. 하지만 음성 인식 기술의 발전에도 불구하고, 여전히 심한 잡음 환경에서의 인식 성능은 신뢰감을 주지 못한다[1]. 이런 문제점을 해결하기 위해 심한 잡음 환경에서의 인식 성능을 향상시키기 위한 방법으로 잡음 제거 알고리듬 연구, 잡음에 강한 음성 파라미터 연구 등이 진행되고 있지만 만족할만한 성능을 보이지 못하고 단일 모달리티에 의한 음성인식은 한계점을 보였다. 그러던 중 시각정보가 음성정보의 인지에 영향을 미친다는 McGurk 효과(1976년)[2]가 밝혀져 널리 알려진 후, 1990년대 중반 이래로 흥미로운 연구의 주제가 되어 왔으며, 여러 연구를 통하여 시청각 음성인식(audio-visual speech recognition)이 잡음 환경 하 음성인식의 성능을 향상시킬 수 있는 대안으로서

접수일자 : 2009년 2월 3일
완료일자 : 2010년 2월 12일
본 연구는 삼성전자 프로젝트와 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음 (NIPA-2010-C1090-1011-0008)

인정받게 되었다[3-6].

시청각 음성인식이 기존 음성인식과 다른 점은 음성특징 뿐만 아니라 시각정보에서 시각 특징도 추출해야 한다는 점이다. 하지만 시각 음성인식에 필요한 시각 정보 또한 음성과 마찬가지로 주위 조명환경이나 기타, 다른 요인에 따른 영상잡음이 존재하며, 이런 영상잡음은 시각 음성 인식의 성능 저하를 야기한다. 따라서 인식 성능 향상을 위한 시각 특징 파라미터를 어떻게 추출하느냐는 하나의 관심분야이다.

본 논문에서는 HMM 기반 시각 음성인식의 인식 성능 향상을 위한 영상 기반 접근 방법에 따른 강인한 시각 특징 파라미터를 추출하는 방법에 대하여 논하고 그에 따른 인식 성능을 비교하였다. 본문의 구성은 제 2장에서는 시각 음성인식을 위한 선행 연구와 기본 시스템에 관한 내용을, 제 3장에서는 강인한 특징 파라미터를 획득하기 위한 여러 방법들에 대하여 논한다. 제4장에서는 실험 및 결과에 대해서 논하고 제 5장에서는 결론을 기술한다.

2. 기준 시스템

본 논문에서는 인식성능의 향상을 위해 강인한 시각 특징 파라미터를 구하는 방법들을 제시하고 이를 미리 구축한 HMM기반 시각 음성 인식기에 적용하여 각 방법에 따른 인식성능을 비교하는데 중점을 둔다. 시각 특징을 얻기 위한 접근 방법으로는 다양한 방법들이 존재하는데[7-8], 그 중 영상 기반 접근 방법은 영상 자체 또는 영상의 코딩값이나 영상 처리 후의 영상 자체를 시각 특징으로 사용하는 방법이다. 이와 같은 영상 기반 접근 방법은 다른 접근 방법에 비해 정보량이 많아 데이터 처리 면에서 손해를 보지만 정보량을 줄이기 위한 방법들을 적용하면 극복 가능하며, 인식 성능 면에서 우수한 결과를 보인다[9]. 하지만 부가적으로 입술의 위치를 정확히 찾는 과정이 필요하다. 입술의 위치를 정확히 찾지 못할 경우 추출된 시각 특징 파라미터들의 편차가 커져 인식 성능의 저하를 야기하기 때문이다. 하지만 입술 윤곽선을 정확히 찾는 것에 비해 입술의 위치를 정확히 찾는 것은 훨씬 쉬운 문제이다. 따라서 본 논문에서는 영상 기반 접근 방법이 시각 음성인식의 성능 면에서 더 적합하다고 판단하고 이를 기반으로 시각 특징 파라미터를 구하는데 있어서 데이터 처리량을 줄이고 인식 성능을 향상시키기 위한 방법에 대해 연구하였다.

우선 영상 기반 접근 방법을 위한 선행 조건으로는 정확한 입술의 위치 검출이 필요하다. 입술의 검출 방법에는 다양한 방법이 존재하나, 본 연구에서는 눈 정위 기반 입술의 검출 방법을 사용하였다. 그림 1은 입술의 검출 방법을 보여준다. 눈 탐색은 적응 임계값을 이용하여 영상을 이진화한 후 눈 후보영역으로 분할하고, GMM 기반으로 검증한다[10]. 그 후 탐지된 눈 정위를 기반으로 양 눈과 입의 기하학적 구조를 이용하여 입술 후보영역을 구한 후 K-means 집단화를 통해 입술의 영역을 분할하고 분할된 영역들에 필터를 적용하여 입술을 검출하였다[11-12]. 이 방법은 본 연구 이전에 진행된 연구로 자세한 내용은 참고문헌을 참조하길 바란다[13-14].

인식 성능의 비교실험을 위해선 자체 제작한 데이터베이스 영상에서 얻어진 입술 ROI데이터에 다운샘플링과 PCA를 적용한 특징 파라미터를 기본 파라미터로 정하고 인식실험을 하였다. 다음 그림 2는 기본 파라미터의 인식 성능

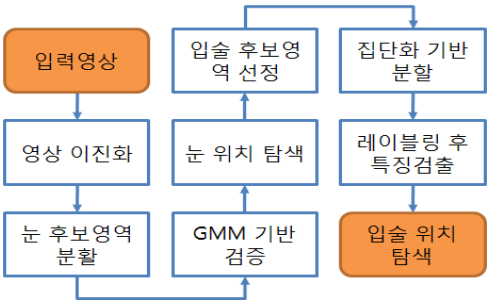


그림 1. 입술 검출 과정.
Fig. 1. Lip Detection Flowchart.

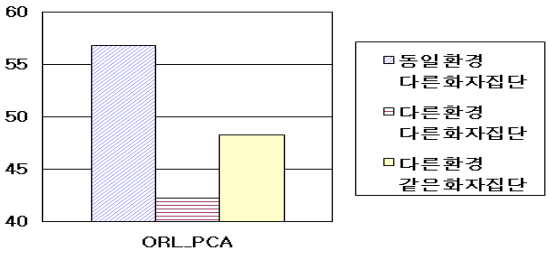


그림 2. 기본 방법의 인식실험결과.
Fig. 2. Result of recognition using baseline method.

을 나타내며 실험조건은 4장을 참고하길 바란다. 기본 실험결과 서론에서 언급한 바와 같이 서로 다른 환경에 의한 조명의 변화가 인식성능 저하현상을 일으킴을 알 수 있다. 다음 3장에서는 인식성능 향상을 위한 방법으로 설명한다.

3. 강인한 특징 파라미터 획득 방법

영상 기반 접근 방법에 따른 시각 특징 파라미터를 얻기 위해 앞장에서 언급한 입술 정위 방법에 의해 얻어진 입술 ROI(region of interest)사용한다. 입술 ROI(region of interest)에 대해 데이터 처리량을 줄이기 위해 32x32 크기로 다운샘플링(down-sampling)을 거치고 여기에서 입술의 기하학적 대칭성에 의거하여 입술영상을 절반인 32x16 크기로 접는 과정을 거친다. 입술 ROI를 절반으로 접게 되면 이후 처리과정에 소요되는 데이터 처리량을 줄일 수 있을 뿐만 아니라 최종적으로 HMM 인식 파라미터로 사용될 특징 파라미터의 개수 또한 줄어듦으로써 여러 가지 측면에서 부담을 낮출 수 있는 장점이 있다. 다음 RASTA 필터링의 프레임간 필터링을 수행한다. 그리고 선택적으로 DCT를 수행하고, 최종적으로 특징파라미터 수를 감축하기 위하여 PCA 에 기반 한 변환을 수행하여 특징 파라미터를 얻는다.

한편, 입술 영상의 색상정보는 조명에 따라서 그 분포가 쉽게 변한다. 본 논문에서는 이러한 문제를 해결하기 위한 방안으로써 히스토그램 매칭(histogram matching)을 전처리로서 수행하였다. 결국 입술 특징 파라미터를 얻기 위한 과정은 그림 3과 같이 정리된다. 그림에서 괄호 안에 표시되어 있는 블록은 선택적 사항이다. 즉 '선형마스크 차감'과 'DCT' 블록은 부가적으로 적용되는 블록이다. 뒤에서 설명되겠지만 선형마스크 차감법은 성능 개선에 도움이 되지 못하였고, DCT는 성능 개선량이 작아 기본 블록에서는 제외되어 있다. 다음 각 절에서는 강인한 특징 파라미터를 얻기 위한 방법에 대하여 자세하게 설명한다.

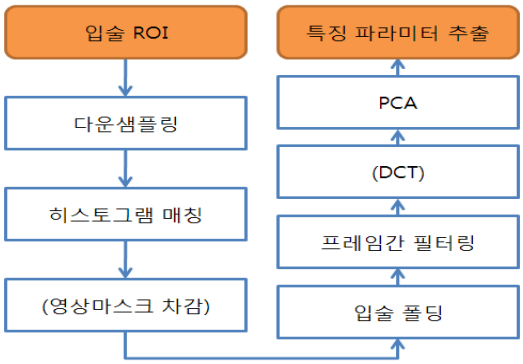


그림 3. 입술 특징 파라미터 추출과정.
Fig. 3. Block diagram for Lip features extraction.

3.1 히스토그램 매칭

인식을 위해 사용된 학습데이터들의 녹화 환경과 실제 환경들에는 많은 차이가 있지만 그중 조명의 차이가 존재한다. 이 조명의 변화는 실제 인식성능에 큰 영향을 끼치는 요인 중 하나이다. 이러한 조명의 변화에 따른 인식성능의 저하를 막기 위해 히스토그램 매칭 기법을 이용하여 조명의 변화를 보상하였다. 히스토그램 매칭은 서로 다른 두 영상을 비교하여 상대적으로 비슷한 히스토그램을 갖게 하는 방법이다.

그림 4는 학습데이터 환경에서의 히스토그램과 실제 테스트 환경에서의 히스토그램을 비교한 것이다. 학습데이터 환경에서의 히스토그램의 분포 범위는 50-220 정도의 사이에 걸쳐 분포하지만 실제 테스트 환경에서의 히스토그램의 분포 범위는 30-170 정도의 사이에 걸쳐 분포하고 있음을 알 수 있다. 이러한 서로 다른 히스토그램의 분포는 인식률 결과의 저하를 야기하는 원인이 되기도 한다. 따라서 본 연구에서는 조명 보상방법으로써 하나의 매핑 함수를 만들어 실제 테스트 환경의 히스토그램을 학습데이터 환경의 히스토그램과 비슷한 분포를 가지도록 매핑하여 조명의 변화를 보상하였다. 그림 5는 매핑 함수를 보여준다.

본 논문에는 다음 2개의 방법으로 히스토그램 매칭에 의한 조명의 변화를 보상하였다.

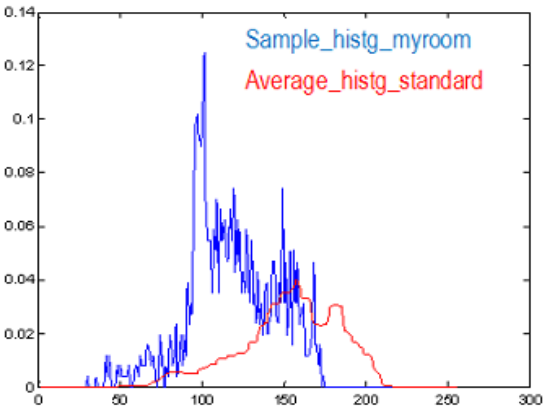


그림 4. 학습데이터와 실제 환경의 히스토그램 차이.
Fig. 4. Histogram for train and test data.

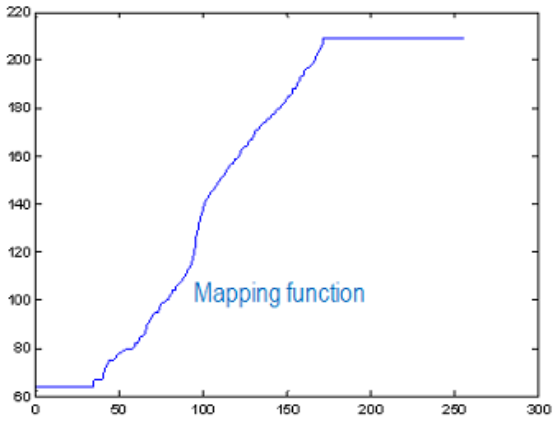


그림 5. 학습데이터와 실제 환경 사이의 히스토그램 매핑함수의 예.
Fig. 5. Schematic view of the Histogram mapping function.

- 방법1 – 모든 프레임에 대하여 별도의 히스토그램 매칭 함수를 구하여 조명을 보상
- 방법2 – 초기 프레임을 대상으로 히스토그램 매칭 함수를 구하고 나머지 프레임에 대해서는 첫 프레임에서 결정한 매칭 함수를 사용하여 조명을 보상

처음 방법1과 같이 히스토그램 매칭을 적용한 결과 오히려 인식결과가 떨어지는 현상이 발생하였다. 두 번째 방법으로 실제 테스트 영상의 맨 처음 프레임과 학습데이터의 처음 프레임사이의 히스토그램을 이용하여 히스토그램 매핑 함수를 만든 후 그 매핑 함수를 실제 테스트 영상의 모든 프레임에 적용하여 조명을 보상 하였고 그 결과 히스토그램 매칭을 사용하지 않았을 때 보다 사용했을 때가 더 높은 인식 성능을 보였다. 인식 실험 결과는 뒤에서 설명한다.

3.2 입술영상 접기[15]

영상 선형 변환 접근법은 입술 윤곽선 기반 접근법에 비해 많은 장점을 가지는 방법이긴 하지만 데이터 처리량이 훨씬 증가하게 되는 단점이 있다. 입술 윤곽선 기반 접근법의 경우에는 보통 입술의 폭, 안쪽 입술의 높이, 바깥쪽 입술의 높이 정도를 입술 특징 파라미터로 사용하므로 파라미터 개수가 보통 3개 정도 밖에 되지 않지만 영상 변환 접근법의 경우는 입술 영상 전체를 대상으로 입술 특징 파라미터를 추출해야 하므로 윤곽선 기반 접근법보다 보통 훨씬 많은 파라미터들이 생성되게 된다. 이러한 이유 때문에 앞서 입력 입술 ROI의 크기를 32x32로 다운샘플링 하였으나, 이 크기에서도 여전히 입술 윤곽선 기반이나 모델 기반접근법보다는 훨씬 많은 개수의 파라미터가 생성된다. 따라서 본 연구에서는 2D-DCT 또는 PCA를 적용하여 영상 선형 변환을 거치기 전에 앞서 처리할 픽셀 데이터를 최대한 줄임으로써 데이터 처리량을 감소시킬 수 있는 입술 폴딩 기법을 사용하였다.

우선 입술 ROI 영상은 화자의 머리 회전이나 상하좌우 기울어짐이 거의 없다는 전제 하에 입술의 기하학적 대칭성에 의거하여 입술 ROI영상을 절반으로 접을 수 있을 것이다. 입술의 모양이 좌우가 같은 대칭이라면 입술을 수직 축

을 기준으로 해서 절반으로 접은 영상도 원래의 입술 영상이 갖는 주요한 정보를 대부분 포함할 수 있을 것이므로 정보의 왜곡도 거의 없을 것이다. 이와 같은 근거에서 입술 ROI 영상을 절반으로 접게 되면 원래 32x32 크기의 ROI가 32x16 크기의 영상으로 줄어들게 된다. 접어진 입술영상의 각 픽셀들은 좌우 점대칭 되는 두 픽셀 값의 평균값이 된다. 이것은 32x32 크기의 영상을 그대로 영상 선형 변환할 때보다 데이터 처리량이 절반으로 감소하게 됨을 의미한다. 이러한 방법을 적용하면 데이터 처리량도 감소할 뿐만 아니라 이후 영상 변환을 거치고 HMM 인식 파라미터로 사용될 입술 특징 파라미터 개수에 있어서도 많은 차이를 보이게 된다. 즉 이러한 방법을 통해 원래의 ROI 영상 크기를 절반으로 줄임으로써 픽셀 데이터 처리량 및 이후 PCA를 통해 생성되는 특징 파라미터들의 수를 감소시킬 수 있다는 것을 의미한다. 뿐만 아니라, 접어진 32x16 ROI의 픽셀 값들은 대칭되는 픽셀들의 평균값들이므로 영상잡음 요소 및 좌우 측면 조명의 불균형에 대한 강인함을 갖게 된다고 볼 수 있다. 그림 6은 대칭성에 근거한 입술영상의 접는 방법을 보여준다.

그림 6. 대칭성에 근거한 입술영상 접기.
Fig. 6. Lip folding based on symmetricalness.

3.3 프레임간 필터링[16]

실제 환경에서 인간은 말을 할 때 고정된 자세에서 말을 하지 않는다. 설령 똑바로 서있다고 하더라도 행동 습관에 의해서 발생하는 머리의 움직임 또는 발음 습관에 의한 입술의 움직임 등은 일정하지 않다. 이처럼 다양한 입술의 움직임은 인식성능을 저하시킨다. 또한 주위 환경 변화에 따른 조명의 변화도 인식률 저하의 원인이 된다. 조명은 실생활에서 보면 오전, 오후, 저녁의 시간대에 따라 각각 조사강도와 방향이 다르며, 그에 따른 영상의 왜곡 또한 다르다. 조명은 카메라로부터 입력되는 영상에서 실제 색 정보를 왜곡시키며, 영상 분석을 통해 파라미터를 추출하여 인식하는 방식을 사용하는 경우 인식률 성능에 큰 영향을 미친다. 따라서 본 연구에서는 이러한 성능저하를 보상하기 위해서 명암이나 조도 변화에 강인한 RASTA 필터를 사용하였다.

RASTA 필터는 원래 음성과 잡음이 서로 다른 점을 이용하여 인식 향상에 강인하게 작용하는 필터로 연구되었다. 이 필터는 잡음 환경 하에서 견인하게 잡음을 제거할 수 있어 음성 인식 성능을 효과적으로 향상 시킨다. 본 연구에서는 RASTA필터의 장점을 시각 음성인식에 적용하여 시스템을 구성하였다. 다음 그림 7은 고역통과 필터와 저역통과 필터의 주파수 응답이고, 각각의 필터 식은 다음식과 같다.

고역통과 필터식 :

$$Y_t[n,m] = 0.9859 \times (X_t[n,m] - X_{t-1}[n,m]) + 0.9716 \times Y_{t-1}[n,m] \quad (1)$$

저역통과 필터식 :

$$Y_t[n,m] = 0.8638 \times (X_t[n,m] + X_{t-1}[n,m]) - 0.7257 \times Y_{t-1}[n,m] \quad (2)$$

여기서 $Y_t[n,m]$는 시간 t에서 (n,m) 픽셀 좌표의 필터링된 이미지 출력값이다. $X_t[n,m]$는 입력 이미지의 픽셀 값, $X_{t-1}[n,m]$는 시간 t의 과거 값이 현재 입력에 영향을 주는 IIR필터이다. 위의 저역 통과 필터 식은 대역 통과 필터링 수행을 위해 고역 통과 필터링의 출력 값을 입력으로 하여 실행되며, 실제 출력 값은 대역 통과 필터링을 수행한 결과 값과 동일하다.

하나의 영상 프레임을 주파수 영역으로 변환하면 변하지 않는 부분은 저주파 영역, 급변하는 부분은 고주파영역으로 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 이전 픽셀 값과 차이가 나면 고주파 영역에 도시되고 그렇지 않으면 저주파 영역에 도시된다. 이를 이용해 적절한 필터를 사용하여 중요한 정보만을 추출하는 것이 필터링의 목적이다.

시각 음성인식처럼 입술 영상만을 통하여 단어를 인식하기 위해서는 입술의 움직임이 매우 중요하다. 연속된 프레임에서 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 저역 통과 필터를 적용하면 보다 강인한 파라미터들을 구할 수 있다.

3.4 선형 마스크 차감법

선형 마스크 차감법은 입력 영상에서 조명을 제거하기 위하여 사용되는 방법 중 하나인데, 얼굴 인식 등에서 성공적으로 사용된 바 있다. 선형 마스크는 주어진 영상을 선형식으로 모델링한 후 이를 차감하는 방식으로써 다음 식과 같이 정의된다.

$$I_M(x,y) = \alpha x + \beta y + \gamma \quad (3)$$

$$\tilde{I}(x,y) = I(x,y) - I_M(x,y) \quad (4)$$

위식에서 파라미터 값 α, β 그리고 γ는 $\tilde{I}$의 에너지 값을 최소화 시키도록 최적화된다. 그러나 위와 같은 선형 마스크 차감법을 사용하여도 100% 조명의 영향을 제거할 수 없다. 조명변화는 더하기(+) 조작이면서도 곱하기(*) 조작이기도 하기 때문이다.

3.5 주파수 변환 - DCT

영상처리에서는 주파수 분석의 방법으로 DCT를 사용하는 것이 매우 일반적이다. 본 연구에서도 영상 접기를 거친 영상 즉 32×16의 영상에 대하여 DCT 변환을 시도하였다. DCT 변환의 크기는 원 입력 영상의 크기인 32×16이며 그 중 고주파 부분을 제외한 저주파 부분만을 사용하기 위해 주파수 도메인에서 16×8부분을 사용하였다. 다음 그림8은 DCT 적용과정을 보여준다.

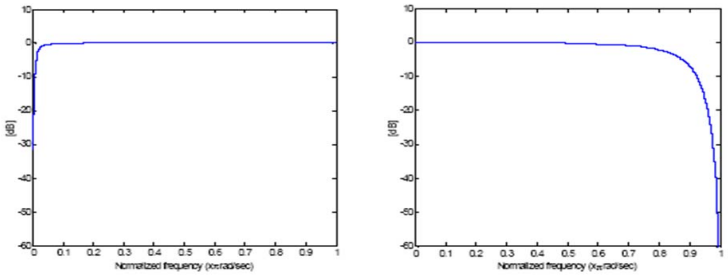


그림 7. 고역통과 필터(좌)와 저역통과 필터(우)의 주파수 응답.
Fig. 7. High pass fileter(left) and Low pass fileter(right) Frequency Response.

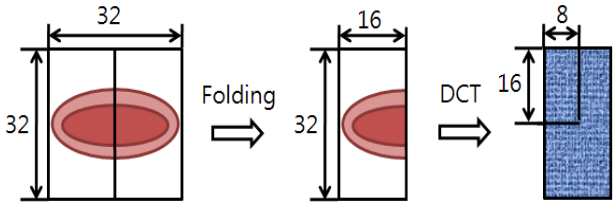


그림 8. DCT를 적용한 예.
Fig. 8. Example of DCT Window.

3.6 주성분 분석 - PCA

통계적 알고리듬인 주성분 분석(PCA) 알고리듬은 임의 시간 t에서 벡터 $X=[X_{1t}, X_{2t}, \ldots, X_{p(=MN)t}]$를 적절히 선형변환 시켜 그것이 가지는 정보를 가능한 많이 보존하는 소수 m개의 새로운 인공변수를 창조함으로써, $p-$차원 변이를 $m-$차원으로 축소하여 전체 체계의 특성을 요약할 수 있다.

여기에서 다음 식에 보인바와 같인 X_{it}의 원소들간의 상관구조를 나타내는 공분산 $\sum$에 기반한 PCA를 고려하여 그 일반성을 유지하였다.

$$\Sigma = E\triangle E' \tag{5}$$

여기서 E는 p개의 고유벡터(eigenvector) $e_i = \{e_{1i}, e_{2i}, \ldots, e_{p(=MN)i}\}$들을 열로 하는 크기 $(p\times p)$인 직교행렬이고 $\triangle$는 $\sum$의 고유값(eigenvalue) δ_i를 대각원소로 하는 크기 $(p\times p)$인 대각행렬이다. 이는 다음 식으로 표현될 수 있다.

$$E=(e_1, e_2, \ldots, e_p 0, \triangle = diag(\delta_1, \delta_2, \ldots, \delta_p) \tag{6}$$

이때 고유값과 각각의 고유값에 대응되는 고유벡터 e_i의 짝을 δ_i의 크기순서($(\delta_1 \geq \delta_2 \geq \ldots \geq \delta_p)$로 배열하고 맨 첫 번째부터 가장 큰 고유값 m개에 해당하는 고유벡터의 짝 E_m을 이용하여 데이터 벡터 X_{it}에 대한 다음과 같은 직교변환 $o_{jt} = E_m' X_{it}$를 고려할 때 이 변환에 의해 새로이 창조되는 m개의 특징벡터 $o_{jt}, j=1,2,\ldots,m(m \leq p)$를 X_{it}의 주성분으로 추출할 수 있다. 여기서 주성분 개수 m을 구하기 위해 고유값 δ_i의 합을 $tr(\triangle)=(\delta_1+\delta_2+\ldots+\delta_p)$이라 하면, m개가 가지는 원본 데이터에 대한 정보율은 $(\delta_1+\delta_2+\ldots+\delta_m)/tr(\triangle)$이 될 것이다.

표 1. DB 단어목록
Table1. DB Word List

그룹	소그룹
G0	단축다이얼, 전화번호그룹정보, 전화번호찾기, 전화번호추가, 이름으로걸기, 번호로걸기, 4자리번호검색
G1	메시지, 발신메시지, 수신메시지, 보낸메시지보기, 문자메시지보기, 예약메시지, 임시보관메시지, 문자메시지수신함, 소 리샘연결, 음성메시지, 연락처등록, 연락처보기, 첨부파일보관함, 이모티콘보내기
G2	모닝콜, 알람, 기념일, 시간표, 현재시각, 작업
G3	단어장보기, 음성메모, 음성으로읽기, 음성메모녹음, 최근검색단어, 단어검색
G4	벨/진동선택, 수신벨선택
G5	암호설정, 비밀번호/잠금설정
G6	전원설정, 조명설정, 배터리량, 종료
G7	촬영, 동영상, 동영상편집기, 사진보기, 애니콜앨범, 그림 및 비디오, 포토스튜디오
G8	황성택, 오상욱, 김상호, 민소희, 박아론, 김율희
G9	하나, 둘, 셋, 넷, 다섯, 여섯, 일곱, 여덟, 아홉
G10	전화, 멀티메일, 일정관리, 메모장, 소리및알림설정, 암호설정, 전원설정, 카메라, 이름검색

4. 실험 및 분석

4.1. 실험 데이터베이스

4.1.1 화자구성 및 발성량

실험에 사용한 DB의 화자의 수는 총105명의 남여화자이다. 화자 중 남자는 54명, 여자는 51명으로 약 50%의 성비로 구성되었다. 연령별 구성은 20대가 74명(남:39 여:35), 30대는 31명(남:15 여:16)이며 연령별 비율은 70대 30으로 20대 화자를 더 많이 포함하였다. 총 발성단어는 62개 단어로 숫자와 이름 그리고 핸드폰에 사용되는 용어들로 10개의 그룹으로 표1과 같다.

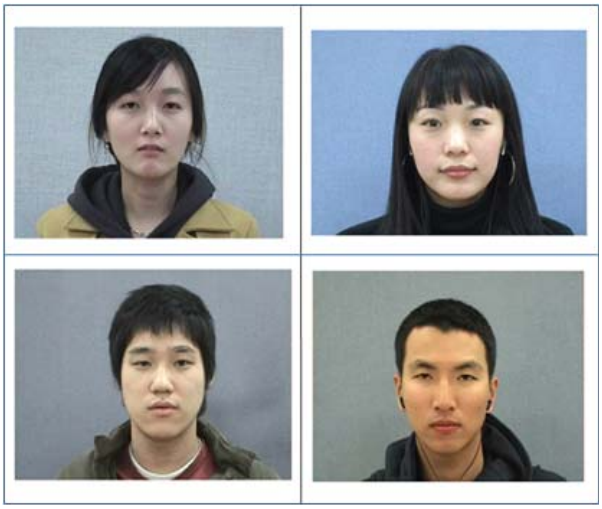


그림 9. 표준 DB의 예
Fig. 9. Example of Standard DB.

그림 10. 실내 DB의 예
Fig. 10. Example of Indoor DB.

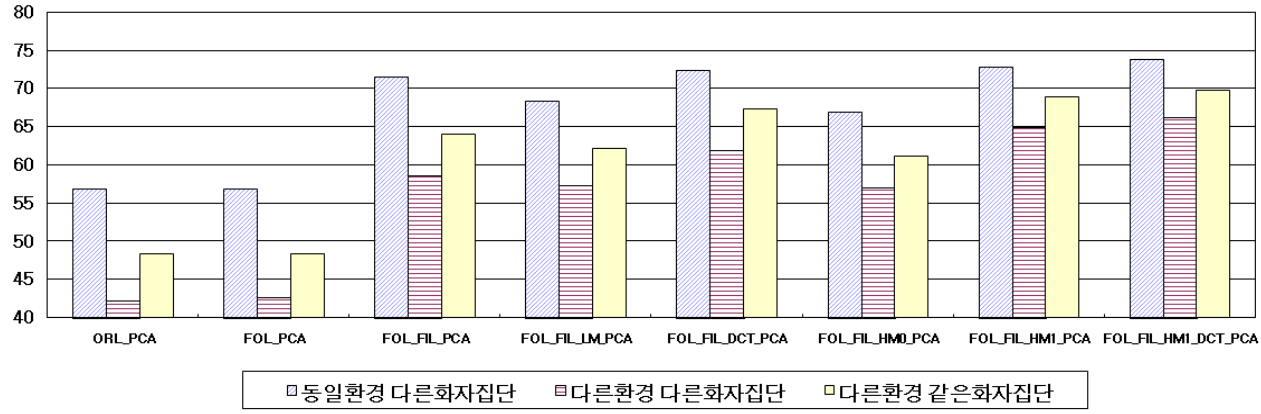


그림 11. 다양한 방법에 따른 특징 파라미터별 인식 성능비교
Fig. 11. Comparison of recognition rate(%) using different features extraction methods.

표 2. 'FIL+HM1+DCT+PCA'의 그룹별 인식률과 평균 인식률
Table 2. Group and average recognition rate for 'FOL+FIL+HM1+DCT+PCA'

	G0	G1	G2	G3	G4	G5	G6	G7	G8	G9	Avg
단어개수	7	13	6	6	2	2	4	7	6	9	
SInter 인식률	77	69	83	83	92	94	69	82	83	48	73.9
IInter 인식률	69	58	79	76	89	82	65	75	74	40	66.1
IIntr 인식률	73	65	79	69	92	87	72	78	80	46	69.8

4.1.2 녹화 조명 환경

녹화조명 환경은 표준/실내로 구성하였다. 표준환경은 조명이 균일한 조용한 실험실 환경이며, 실내환경은 자유로운 배경과 잡음이 존재하는 환경이다. 표준 환경에서 촬영을 할 때, 뒤 배경을 적절히 처리하는 것이 중요하다. 표준환경에서는 푸른색 파티션을 사용하여 얼굴의 선명도가 강조되도록 촬영하였다. 실내환경은 일반적인 실험실 환경으로 표준환경에 비해 다양한 배경과 조명의 변화환경에서 촬영하였다. 그림 9, 10는 표준/실내 DB의 장면들을 보여준다.

4.2 실험 결과

인식실험은 앞 절에서 언급한 시청각 음성 DB 중 표준 및 실내 환경에서 녹화된 DB를 대상으로 HMM기반 시각 음성인식기를 이용하여 수행 하였다. 인식실험은 학습데이터와 테스트 데이터를 구분하여 수행하였는데 정리하면 다음과 같다.

- 학습데이터 : 표준 환경 DB 1-80번 화자
- 테스트데이터 :
  - 1. 표준 환경 DB 81-105번 화자 : SInter
  - 2. 실내 환경 DB 81-105번 화자 : IInter
  - 3. 실내 환경 DB 1-80번 화자 : IIntr

위 테스트데이터 중 1,2번째 데이터는 학습시에 사용되지 않은 화자그룹이며 3번째 데이터는 학습시와 동일
한 화자그룹들이다. 그리고 1번 데이터는 학습시와 동일한 환경이며, 2,3번 데이터는 학습시와 다른 환경이다.

한편 그림 11에서 실험 시 특징 파라미터의 종류를 설명하는 약자들을 사용하였는데 이를 정리하면 다음과 같다.

- ORL : 기본 32x32 영상
- FOL : 접힌 영상 32x16
- LM : 선형 매스크(linear mask) 차감법
- FIL : bandpass-filtered 영상 (참조: RASTA)
- DCT : 이산 코사인 변환
- PCA : 주축분석
- HM0 : 히스토그램 매칭(histogram matching)
  이 경우 모든 프레임에 대하여 별도의 매칭 함수를 구한다.
- HM1 : 히스토그램 매칭
  이 경우 초기 프레임을 대상으로 매칭 함수를 구하고 나머지 프레임에 대해서는 첫 프레임에서 결정한 매칭 함수를 사용한다.

그림 11은 립리딩 성능(인식률)을 보여주고 있다. 인식 성능은 그룹별 인식 실험을 통하여 얻었는데, 그룹별 인식률을 얻은 후, 그룹에 속한 단어수를 가중값으로 사용하여 평균 인식률을 구하였다. 그림에서 알 수 있듯이 원영상에 다운샘플링과 PCA를 적용한 기본 파라미터를 사용한 경우 평균적으로 약 49.1%의 인식률을 보였다. 가장 우수한 성능은 'FOL+FIL+HM1+DCT+PCA'의 경우 69.9%를 보였는데, 이는 입술접기, 공간필터링, 히스토그램 매칭1, DCT 그리고 PCA를 반영했을 때의 결과이다. 한편 3가지 테스트의 결과를 살펴보면, 녹화환경이 바뀌거나 또는 학습에 참여하지 않는 화자에 대해서 인식성능이 저하됨을 알 수 있다. 또한 DCT를 적용한 경우 인식률이 그렇지 않은 경우에 비하여 약간 상승하였다. 한편 선형매스크를 사용하는 경우, 인식률이 향상되지 않음을 알 수 있다. 표 2는 가장 좋은 성능을 보인 'FOL+FIL+HM1+DCT+PCA'에 대하여 그룹별 인식률과 평균 인식률을 보여주고 있다.

5. 결 론

본 논문은 시각 음성인식에 있어 인식 성능 향상을 위한 영상 기반 접근방법에 따른 시각 특징 파라미터를 추출하는 방법을 제안하였다. 시각 특징 파라미터를 추출하는 방법은 히스토그램 매칭, 선형 마스크 차감법, 입술 접기 기법, 프레임 간 필터링, DCT, PCA 등을 적용하였고 그 중 선형 마스크 차감법과 DCT 방법은 인식 성능이 향상되지 않거나 향상 정도가 미비하였다. 실험 결과 선형 마스크 차감법과 DCT를 제외한 방법을 적용하여 추출한 파라미터를 인식기에 적용한 결과 기본 영상에 PCA만을 적용한 파라미터 보다 평균적으로 약 21%정도 인식 성능이 향상되었다. 특히 학습데이터와 테스트데이터의 환경과 화자가 모두 다른 경우 더 큰 성능 향상을 보였다.

본 논문에서는 시각 음성인식 성능의 향상을 위한 시각 정보를 이용한 파라미터 추출방법에 대해 연구를 진행하였으나, 향후 시각 정보와 음성 정보를 같이 이용한 통합 파라미터를 추출하는 방법이나 혹은 각각의 파라미터를 이용하여 나온 인식 스코어값을 통합하는 방법 등을 적용하여 시청각 기반으로 음성인식의 성능 향상을 위한 연구를 진행할 것이다.

참 고 문 헌

[1] Pedro J. Moreno, "Speech Recognition in Noisy Environment," *Ph.D. Thesis, ECE Department*, CMU, May 1996.

[2] McGurk, Harry and MacDonald, John, "Hearing lips and seeing voices," *Nature*, Vol. 264(5588), pp. 746 - 748, 1976.

[3] S. Dupont and J. Luettin, "Audio-Visual Speech Modelling for Continuous Speech Recognition," *Proceedings of IEEE Transactions on Multimedia*, pp.141-151, 2000.

[4] J. N. Gowdy, A. Subramanya,. C. Bartels, J. Bilmes, "DBN-based muti-stream models for audio-visual speech recognition." *proc. IEEE Int. conf. Acoustics, Speech, and Signal Processing*, pp.993-996, 2004.

[5] Jeff A. Bilmes and Chris Bartels, "Graphical Model Architectures for Speech Recognition," *IEEE Signal Processing Magazine*, vol.22, pp.89-100, 2005.

[6] Jean-Luc Schwartz, , Frédéric Berthommier and Christophe Savariaux, "Seeing to Hear Better: Evidence for Early Audio-Visual Interactions in Speech Identification," *ERIC Journal Articles : Reports-Research, Cognition*, vol.93, no.2, pp. 69-pp.78, Sep, 2004.

[7] G. Potamanos, H.P. Graf, E. Cosatto, "An image transform approach for HMM based automatic lipreading", *Proceedings of the International Conference on Image Processing*, vol.3, pp. 173-177, Chicago, U.S.A., July 1998.

[8] C. C. Chibelushi, F. Deravi, and J. S. Moson, "A review of speech-based bimodal recognition," *IEEE Trans. Multimedia*, vol.4, no.1, pp23-37, Mar. 2002.

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

[10] T. T. pham, J. Y. Kim, S. Y. Na, S. T. Hwang, "Robust Eye Localization for Lip Reading in Mobile Environment," *Proceddings of SCIS&ISIS in Japan*, pp.385-388, 2008.

[11] MacQueen, J. B. "Some Methods for Classification and Analysis of Multivariate

Observations," *In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281-297. 1967.
[12] Andrew W. Moore, "K-means and Hierarchical Cl-ustering", *Tutorial Slides in School of Computer Science Carnegie Mellon University*, http://www.cs.cmu.edu/~awm, http://www.autonlab.org/tutorials/kmeans11.pdf
[13] T. T. Pham, M. G. Song, J. Y. KIm, S. Y. Na, S. T. Hwang, "A Robust Lip Center Detection in Cell Phone Environment," *Proceedings of IEEE Symposium on Signal Processing and Information Technology*, pp.390-395, Sarajevo, December, 2008.
[14] 송민규, 김진영, T. T. Pham, 황성택, "모바일환경에서의 시각 음성인식을 위한 눈 정위 기반 입술의 검출에 대한 연구", *한국퍼지 및 지능시스템학회 논문지*, 제 19권 제 4호, pp. 478-484.
[15] 김진범, 김진영, "입술의 대칭성에 기반한 효율적인 립리딩 방법," *전자공학회논문지*, 제 37권, 제 5호, pp.105-114, 2000.
[16] 신도성, 김진영, 최승호, "시간영역 필터를 이용한 립리딩 성능향상에 관한 연구," *한국음향학회지*, 제 22권, 제 5호, pp.375-382, 2003

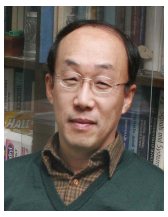
저 자 소 개

송민규 (Song Min Gyu)
제19권 4호(2009년 8월호) 참조

Thanh Trung Pham
제19권 4호(2009년 8월호) 참조

김진영 (Jin Young Kim)
제19권 4호(2009년 8월호) 참조

황성택 (Hwang Sung Taek)
제19권 4호(2009년 8월호) 참조



나승유 (Seung You Na)
1977년 : 서울대학교 졸업.
1984년 : Univ. of Iowa, Dept. of ECE Master
1986년 : Univ. of Iowa, Dept of ECE Ph. D
2005년 ~ 현재 : 전남대학교 교수

관심분야 : 지능제어 및 계측, 신호처리
Phone : 062-530-1757
Fax : 062-530-1759
E-mail : syna@jnu.ac.kr

민소희 (So Hee Min)
1993년 : 전남대학교 졸업
2009년 : 동 대학원 석・박사 통합과정 졸업
2009년 ~ 현재 : 전남대학교 전자정보통신 공학부 위촉연구원

관심분야 : 시청각 신호처리, 화자인식
Phone : 062-530-0370
Fax : 062-530-1759
E-mail : shmin3@jnu.ac.kr