RASTA 필터를 이용한 립리딩 성능향상에 관한 연구

신도성*, 김진영*, 최승호**, 김상훈***
*전남대학교 전자공학과, *전남대학교 정보통신공학과 & RRC HECS
**동신대학교 정보통신공학과, ***ETRI

A Study on Lip-reading enhancement using RATSTA fileter

Dosung Shin*, Jinyoung Kin*, Seungho Choi*, Sanghun Kim***
*Dept. of Electronic Engineering Chonnam University,
*Dept. of Information and Communication & RRC HECS, Chonnam University,
**Dept. of Information and Communication DongShin University, ***ETRI
E-mail : jesus33@dsp.chonnam.ac.kr

Abstract

Lip-reading technology that is studied them is used to compensate speech recognition degradation in noise environment in bi-modal's form. The most important thing is that search for correct lips area in this lip-reading. But, it is hard to forecast stable performance in dynamic environment. Used RASTA filter that show good performance to remove noise in the speech to compensate. This filter shows that improve performance of using time domain of digital filter. To this experiment observes performance of speech recognition only using image information, service chooses possible 22 words and did recognition experiment in car. We used hidden Markov model by speech recognition algorithm to compare this words' recognition performance.

I. 서론

립리딩은 음성인식 분야 중 잡음 환경에서 현저하게 떨어지는 인식율을 높이기 위한 보상방법의 하나로 화자의 입술을 포함한 영상 정보를 이용하는 목적으로 연구되었다[1~4]. 그 방법으로는 모델기반과 이미지 기반방법이 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 인식률이 저하하는 다른 방법보다는 안정된 인식률을 보이는 장점이 있는 반면, 입술 전체영상을 특징 파라미터로 사용하기 때문에 데이터 용량의 증가에 따른 인식 속도 저하가 발생한다. 이와 같은 문제해결을 위해 본 연구에서는 입력된 전체 입술 영상에 대해서 전처리 과정을 통하여 데이터 처리량을 줄이기 위한 작업을 수행하였으며, 안정적인 성능을 보이는 립리딩 구현을 위해 립리딩 성능을 저하시키는 원인을 분석하였다[5]. 그리고 그 보상 방법으로 몇 단계 전처리 과정 중 PCA를 통한 주성분 분석으로 추출된 파라미터를 사용한 성능향상 방법과 시간 영역에서 RASTA필터를 이용한 방법에 대해 잡음이 배제된 실내에서 인식 실험을 수행하고 그 결과를 비교·분석하여 인식률을 살펴보았다.

II. 구현된 립리딩 시스템의 구조

실험을 위해 구현한 립리딩 시스템의 전체적인 구조를 그림 1에 도식하였다. 입력된 입술영상은 이진 영상처리를 위해 명암영상 형태로 변환된다. 변환된 이미지를 명암변화나 조명방향의 변화에 따른 영향을 보상하기 위해 선형 명암마스크를 이용하여 일차적인 명암의 영향을 제거한 후 입술 영역만을 추출하는 작업을 한다. ROI 추출 블록에서는 2진 영상 변환을 통해 찾아낸 입술 안쪽 영역을 기준으로 입술의 중심 높이와 좌우 폭을 찾아 입술만을 따로 분리해 낸다.

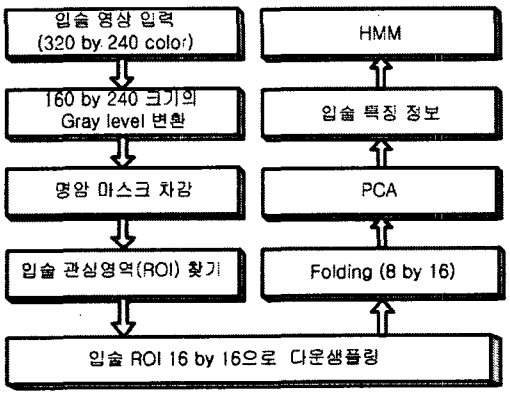

그림 1. 구현된 립리딩시스템 구조

분리된 입술 ROI는 다운샘플링(Downsampling) 과정을 거쳐 입술의 정보손실을 최소화하는 범위 내에서 데이터 처리량 을 줄이기 위한 작업을 수행한다.

Folding 블록에서는 다운샘플링된 영상 이 화자의 머리 회전이나 상하좌우 기울어짐이 없다는 전제 하에 인간의 입술형태가 좌우 대칭인 점에 착안하여, 입술 ROI 영상의 절반만을 이용해서 영상을 처리하게 된다. 입술을 반으로 접는 방법을 적용하면 데이터 처리량도 감소할 뿐만 아니라 이후 **HMM** 인식 실험을 위해 추출되는 중요 파라미터 수도 현저히 감소하게 된다. PCA 블록에서는 통계적 알고리즘인 주성분 분석이 갖는 원래 정보를 적절히 선형 변환시켜 정보를 가능한 많이 보존하는 새로운 차원으로 전체 체계의 특성을 요약하는 특징을 이용한다.

매 프레임마다 15개의 파라미터를 추출하고 단어에 대한 음성구간 동안의 파라미터를 생성하였으며 단어에 대한 파라미터는 HMM의 학습화 과정을 거쳐 각 단어에 대한 평균과 분산과 가중치를 계산하였다. 이렇게 하여 22단어에 대한 각 단어의 특징을 가지고 테스트 데이터의 입력과 비교를 하여 최적의 단어를 검출한다.

Ⅲ. RASTA 필터링을 적용한 립리딩 시스템

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 음성 인식 향상에 강인하게 작용하는 필터를 연구한 것이 **RASTA** 필터이다. 이 필터는 잡음 환경 하에서 견인하게 잡음을 제거할 수 있어 음성 인식 성능을 효과적으로 향상시킨다[6].

본 논문에는 **RASTA**필터의 장점을 립리딩에 적용하여 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 $Y_t[n,m]$는 시간 t에서 $[n,m]$픽셀 좌표의 필터링된 이미지 출력 값이다. $X_t[n,m]$는 입력 이미지의 픽셀 값, $X_{t-1}[n,m]$는 시간 t의 과거 값이 현재 입력에 영향을 주는 **IIR** 필터이다. 위의 저역통과 필터식은 대역 통과 필터링 수행을 위해 고역통과 필터링의 출력 값을 입력으로 하여 실행되며, 실제 출력 값은 대역 통과 필터링을 수행한 결과 값과 동일하다.

립리딩에서 입술 영상만으로 단어를 인식하기 위해서는 입술의 움직임이 매우 중요하다. 30Hz (**frames /sec**)의 속도로 입술 영역이 찾아진 데이터들은 시간의 흐름에 따라 입술 **ROI**는 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. 이 특성을 이용해 적절한 필터를 사용해 중요정보만을 추출하는 것이 필터링의 목적이라 할 수 있으며, 이를 바탕으로 고역 통과 필터와 대역 통과 필터를 각각 적용하여 실험하고 결과를 비교·분석하였다.

그림 3에 각각의 실험 방식을 도식하였다. **RASTA** 방법 1은 일반적인 **RASTA** 필터링 방법이고 **RASTA** 방법 2는 본 논문에서 제안한 방법이다.

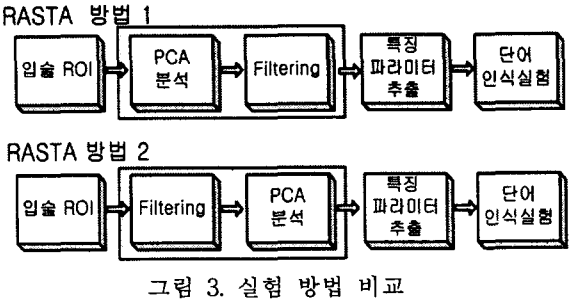

그림 3. 실험 방법 비교

Ⅳ. 실험결과

4.1 RASTA 필터링 적용 방식에 따른 인식률 비교

인식률 비교를 위한 실험은 22단어를 대상으로 실험실 내에서 수행하였다. 영상 녹화는 남성 화자 70명을 대상으로 코에서부터 턱까지 촬영하였다. 이중 학습화를 위해 52명의 영상을, 나머지 18명의 영상은 테스트에 사용하였으며, 실험은 다음과 같은 방법으로 이루어졌다. 먼저 일반적인 **RASTA** 방법으로 **PCA**를 한 후 필터링을 통한 단어 인식 성능을 분석한 후 본 논문에서 제안한 방법으로 본래 이미지를 먼저 필터링하고 그 결과에 대한 특징 파라미터를 추출하여 인식 성능을 비교·분석하였다.

그림 3은 입력 이미지를 전처리 과정을 수행한 영상 이미지로서 폴딩되기 전의 다운샘플링한 16×16 원 영상 이미지를 보여 주고 있다.

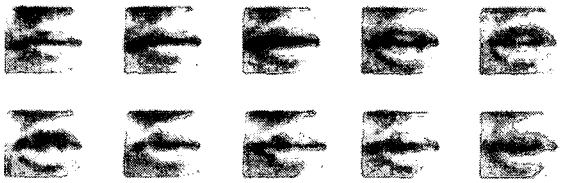

그림 3. 16×16으로 다운샘플링한 원 영상

그림 4와 그림 5는 그림 4에 본영상의 입술 모양이 좌우가 기하학적 대칭성을 이루는 것에 착안해서 8×16로 접은 이미지로 제안한 **RASTA** 방법을 사용하여 각각 고역 통과 필터와 대역통과필터를 사용해 필터링한 후 **PCA** 수행한 결과이미지를 보여 주고 있다.

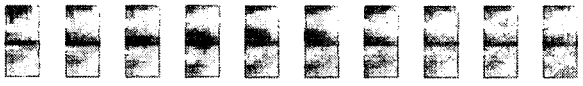

그림 4. RASTA 방법 2 : 고역통과 필터링(8×16)

그림 5. RASTA 방법 2 : 대역통과 필터링(8×16)

결과 영상들은 모두 본래 영상에 비하여 약간은 블러된다. 이 현상은 픽셀마다 필터에 의해서 제거된 정보 때문이며, 원인은 주성분 분석에 의해 추출된 정보가 필터링 수행의 결과로 이미지 정보가 필터 영역에 따라 제거되었기 때문이다. 이는 원 영상으로부터 추출된 파라미터를 필터링하는 것이므로, 파라미터의 정보가 손실되어 인식성능 저하를 가져온다는 것을 알 수 있다.

이 실험에서 사용되는 주성분 개수는 반으로 접은 이미지를 PCA만을 수행해 얻은 파라미터 수를 필터에 모두 사용하게 되므로 필터링에 의해서는 파라미터의 개수가 줄어들지 않는다. 단지 추출된 파라미터 성분에서 고주파나 저주파 영역의 정보만을 제거하는 것이다. 실제로 실험결과를 보더라도 PCA 90%를 적용하였을 때는 24개 PCA 95%를 적용할 경우는 44개의 주성분 개수가 그대로 사용되어 인식 속도향상을 기대할 수 없었다.

본 논문에서는 필터에 의해 정보를 손실하지 않으면서 잡음성분 만을 제거하고 파라미터 수를 줄이기 위해 필터링을 먼저 수행한 후 PCA를 통해 주성분을 추출하는 방법을 제안하였다. 그리고 추출된 특징 파라미터는 HMM 알고리듬에 적용해 인식실험을 하였다.

그림 6은 **RASTA** 방법 1을 사용해 인식 실험한 결과이다. 실험결과를 분석해 보면 필터링을 하지 않는 것의 인식률이 더 좋은 결과를 나타내고 있다. 이는 필터에 의해 추출된 특징 파라미터가 기존의 특징 파라미터보다 더 적은 정보를 갖고 있음을 알 수 있다.

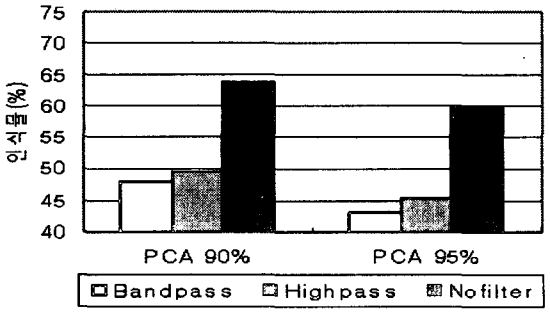

그림 6. RASTA 방법 1의 인식율

그림 7은 **RASTA** 방법 2를 사용하여 인식 실험한 결과이다.

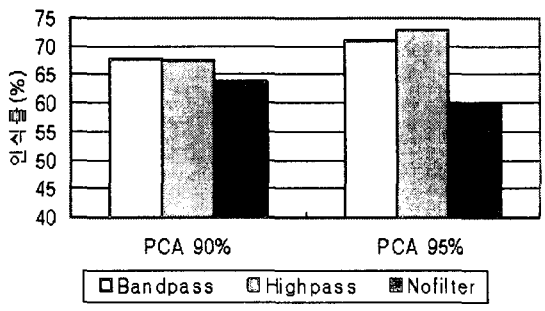

그림 7. RASTA 방법 2 인식율

실험결과로 알 수 있듯이 RASTA 방법 1에서는 필터링을 하지 않는 것이 더 좋은 인식률을 보고 있다. 반면 RASTA 방법 2에서는 필터링을 한 경우가 인식률이 더 좋은 것을 알 수 있다.

4.2 필터링 수행유무에 따른 주성분 개수 변화

그림 8은 입술 **ROI**를 16×16으로 다운샘플링을 한 후 폴딩하여 픽셀들의 평균값을 8×16 이미지로 만들어 필터링한 결과를 누적 백분율로 구한 **PCA** 개수를 비교한 것이다. 필터링 수행 유무에 따라 PCA를 통해 추출된 주성분 개수는 보는 바와 같이 많은 차이를 보인다. 입술 **ROI** 파라미터에 대해 필터링을 수행하면 **PCA**의 개수가 필터링을 전혀 하지 않은 것에 비해 절반이상 줄어드는 것을 볼 수 있다. 따라서 파라미터 수를 줄일 수 있고, 인식 속도 또한 단축시킬 수 있는 장점이 있음을 알 수 있다. 이렇게 많은 파라미터 수의 차이는 HMM 알고리듬에 의하여 단어를 인식할 때 속도를 절감시킬 수 있어 실시간 인식에 가능성을 보여준다.

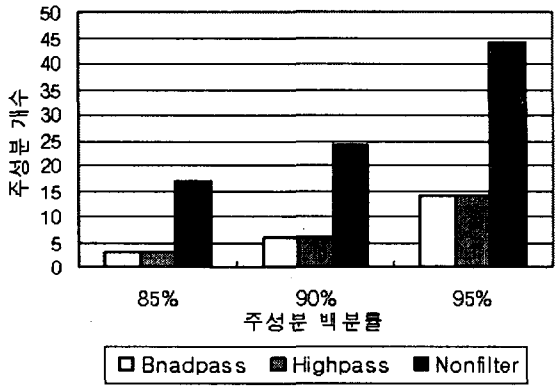

그림 8. 필터에 따른 주성분의 개수

위의 두 실험의 결과를 보면 필터링을 먼저 수행하고 **PCA**를 한 것이 더 좋은 인식 성능을 보임을 알 수 있다. 이는 전처리로 필터링을 하면 입술 **ROI**의 데이터에서 변하지 않는 파라미터나 잡음이 제거되며, 이렇게 제거된 데이터에서의 주성분을 추출하므로 파라미터 량도 훨씬 축소할 수 있지만, 필터를 거치지 않고 주성분을 먼저 검출하게 되면 필터에 의해서 제거되었던 불필요한 정보들이 파라미터로 검출될 수도 있기 때문이다.

결과적으로 제안된 **RASTA** 방법 2가 립리딩의 성능 향상에는 효과적임을 알 수 있다. 또한 대역 통과 필터와 고역 통과 필터의 인식율 만을 비교하여 보면 거의 비슷한 인식성능을 보임을 알 수 있다. 그리고 작은 수의 파라미터만을 가지고도 단어를 인식하기 위한 입술 정보를 효율적으로 나타내는 것이 가능하고 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. 결론

본 논문에서는 영상 정보만으로 단어를 인식하는 방법으로 효과적인 입술 파라미터를 추출하는 립리딩에 대하여 살펴보았다. 영상이미지 기반의 립리딩에서 RASTA 필터링은 조명의 영향에 의한 성분을 제거하여 파라미터 수를 줄여줄 뿐 아니라 인식성능을 개선시켜준다. 본 연구에서도 **RASTA**필터를 립리딩에 적용한 후 인식률 변화를 살펴 본 결과 **PCA**만 수행했을 경우 64%의 인식율을 보인 반면, **RASTA**를 적용하였을 경우는 72.7%로 인식률 향상을 보였다.

참 고 문 헌

[1] Rajeev Sharma, Vladimir I. Pavlovic, Thomas S, Huang, "Toward Multimodal Human- Computer Interface" ,Proceedings of the IEEE Vol. 86. No 5. May 1998

[2] Gerasimos Potamianosm, Hans peter Graf, Eric Cosatto, "An Image Transform Approach for HMM based Automatic Lipreading", Processing Of the Int. Conf. On Image Processing. pp. 173-177, 1998.

[3] C.Bregler and Yochai Konig, "Eigenlips' for Robust Speech Recognition", Proc. IEEE Int. Conf. On Acoustics, Speech and Signal Processing, pp. 669-672, 1994

[4] Potamianos, G.; Graf, H.P; Cosatto, E., "An image transform approach for HMM based automatic lipreadingn", Image Processing, 1998. ICIP98. Proceedings. 1998, International Conference on, 1998, pp173-1777, vol.3

[5] 신도성, 김진영, 동적 환경에서의 립리딩 인식성능 저하 요인분석에 대한 연구, 한국음향학회지 제 21권, 제 5호, pp.471-477

[6] Hynek Hermansky ,Nelson Morgan, "RASTA processing of speech", IEEE Transaction on Speech and audio processing Vol.2, NO4, October 1994.