시간영역 이미지 필터링에 의한 립리딩 성능 향상

Time domain Filtering of Image for Lip-reading Enhancement

이 지 은*, 김 진 영*, 이 주 헌**

(Jeeeun Lee *; Jinyoung Kim *; Joohun Lee **)

전남대학교 전자공학과, 동아방송대학 인터넷 방송과 **

Multimedia DSP Lab., Dept. of Electronic Engineering, Chonnam Natl Univ. *

Dept. of Internet Broadcasting, Dong-Ah Broadcasting Collge **

(jini@dsp.chonnam.ac.kr *; kimjin@dsp.chonnam.ac.kr *)

요 약

립리딩은 잡음 환경 하에서 음성 인식 성능을 향상을 위해 영상정보를 이용한 바이모달(bimodal)음성인식으로 연구되었다[1][2]. 그 일환으로 이미 영상정보를 이용한 립리딩은 구현되었다. 그러나 현재까지의 시스템들은 환경의 변화에 강인하지 못하다. 본 논문에서는 이미지 기반 립리딩 방법을 적용하여 입술 영역을 보다 안정적으로 찾아 성능을 향상 시켰다. 그러나 이 방법은 많은 데이터량을 처리해야 하므로 전처리 과정이 필요하다. 전처리로 입력영상을 그레이 레벨로 변환하는 방법과, 입술을 반으로 접는 방법, 그리고 주성분 분석(PCA: Principal Component Analysis)을 사용하였다. 또한 인식 성능 향상을 위해 음성에서 잡음 제거나 분석·합성에 효과적인 성능을 보이는 RASTA(Relative Spectral)필터를 적용하여 시간 영역에서의 변화가 적은 성분이나 급변하는 성분, 그 밖의 잡음 등을 제거하였다. 그 결과 72.7%의 높은 인식 성능을 보였다.

1. 서 론

현재 음성 인식기는 무 잡음 환경에서 우수한 인식률을 을 보이나 잡음 하에서는 안정적인 성능을 기대 하기란 어렵다. 이에 영상정보를 음성 인식에 사용하는 립리딩 기술이 연구되었다. 립리딩의 방법에는 이미지 기반, 모델 기반, 모션 기반등이 있다. 본 논문에서는 가장 안정적으로 입술영역을 찾는 이미지 기반 방법을 이용하였다[3]. 이 방법은 입술 영역을 파라미터로 사용하므로 많은 데이터량을 가지고 있다. 이를 보상하기 위하여 몇 단계의 전처리를 사용하였다. 먼저 입력된 영상을 8bits 그레이 레벨로 변환시켜 입술을 찾았다. 이렇게 찾아진 입술은 반으로 접고, PCA를 통하여 중요 파라미터만을 추출하였다. 또한 인식 성능 향상을 위하여 음성에 사용되는 RASTA 필터를 적용하여 보았다[4]. 그 결과 필터링 하지 않은 것에 비하여 인식률이 떨어졌다. 이것은 영상과 음성이 다른 특성을 가지고 있다는 것이므로, RASTA필터 방법을 응용하여 필터링을 한 후 PCA를 하는 방법을 적용하여 보았다.

2. 입술 모델링을 위한 전처리

본 실험에서 입력된 영상은 320×240 의 컬러 이미지로 캠코더에 의해서 코에서부터 턱까지만 촬영하였다. 이 크기의 영상을 그대로 처리하기란 방대한 데이터 량을 계산해야 하므로 몇 가지 전처리 단계를 거쳐 중요 파라미터만을 추출하여 사용하였다.

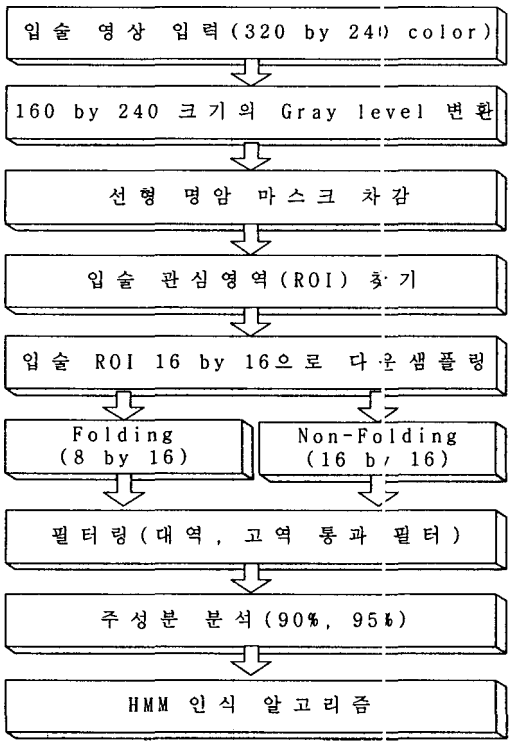

그림 1. 립리딩 알고리즘

먼저 입력 된 컬러 영상을 정보의 왜곡 없이 160×120 크기로 다운 샘플링하고, 8bits 그레이 레벨로 변환시킨다. 이렇게 변환된 이미지를 그대로 사용한다면 화자의 얼굴의 색이나 명암 변화에 따라 평균 명암의 밝기가 변하므로 이를 보상하기 위하여 선형 명암 마스크 차감을 하였다. 명암 차감된 영상은 X축 프로젝션에 의해 입술의 좌우 폭과 중심점을 찾고, Y축 프로젝션에 의해 입술의 중심 높이를 찾는다. 이런 방법을 통하여 입력 영상에서 입술 영역만을 정확히 파악할 수 있다. 입술만을 포함한 ROI(Region Of Interest) 영역의 크기는 '입술 폭×1.2' 를 한 변으로 하는 정사각형으로 추출해 낸다. 이때 화자마다 입술의 크기가 달라 추출되는 ROI의 크기도 각기 다르나 일정한 개수의 특징 파라미터를 추출하려면 입술 ROI의 정규화 과정이 필요하다. 따라서 추출한 입술 ROI는 일정하게 16×16 크기가 되도록 다운 샘플링을 한다. 이렇게 하여 얻어진 입술 ROI는 립리딩 인식실험을 하기 위한 특징 파라미터를 구할 수 있게 된다[5].

전처리에 거쳐 구해진 입술 ROI는 15×16 픽셀을 가지고 있다. 그레이 레벨의 영상이라 하지만 한 프레임당 256개의 픽셀을 처리하기란 많은 데이터 량이다. 이것을 보상하기 위해 입술 ROI 영상을 입술의 좌우 대칭인 것에 착안하여 ROI를 절반으로 접이 파라미터 수를 줄이는 방법을 사용하였다[6]. 반으로 접은(Folding) 이미지는 접지않은(Non-Folding) 이미지에 비해 데이터 량이 절반으로 감소된다. 이때 접어진 영상은 Y축을 기준으로 좌우 대칭 값의 평균으로 구해지므로 영상의 잡음요소 및 좌우 측면의 조명의 불균형에 대하여 강인함을 갖게 된다.

3. 시간 영역의 이미지 필터링

본 실험에서 적용한 필터는 음성처리의 한 분야인 자동 음성인식 시스템 구현에 좋은 성능을 보이는 RASTA 필터이다.

RASTA 필터는 스펙트럼 파라미터 신호 중 변하지 않거나 천천히 변하는 성분(환경적 요인, 화자의 발음 특성)뿐만 아니라 급변하는 잡음 제거에도 좋은 특성을 보이고 있다. 대역 통과 필터는 일반적으로 1~10Hz의 대역폭을 가지고 있으며 제로 주파수에서 샤프한 제로를 특성을 나타낸다. 이런 특성은 저주파 영역의 변하지 않는 잡음이나 천천히 변하는 잡음을 제거 할 수 있을 뿐만 아니라 고주파 영역의 급변하는 잡음도 충분히 제거할 수 있다. 또한 고역 통과 필터에 의해서 채널의 잡음같이 느리게 변하는 성분을 제거할 수 있다.

본 실험에서는 위에서 살펴본 RASTA의 장점을 립리딩에 적용하여 고역 통과 필터와 저역 통과 필터를 각각 이미지에 적용하여 실험하였다.

고역 통과 필터 식

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

저역 통과 필터 식

$$Y_t[n,m] = 0.8638 \times (X_t[n,m] + X_{t-1}[n,m]) - 0.7275 \times Y_{t-1}[n,m]$$

저역 통과 필터링은 고역 통과 필터를 거친 이미지를 입력으로 받아 처리하므로 결과적으로 출력은 대역 통과 필터링 한 것과 동일하다.

입력 영상은 전처리에 의하여 입술 ROI를 찾는다. 찾아진 ROI는 데이터 량을 줄이기 위하여 PCA를 통하여 중요한 파라미터만 추출된다. 이 데이터들은 발음하는 구간동안 픽셀 값들이 변한다. 즉 시간이 흐름에 따라 입술 영역은 계속하여 변할 것이고, 입술 주변 영역들은 상대적으로 변화가 적을 것이다. 이때 계속적으로 변하는 입술 부분은 저주파가 아닌 영역에 나타나고, 상대적으로 변화가 적은 부분은 저주파 영역에 나타날 것이다. 이를 근거로 고역 통과 필터와 대역 통과 필터를 사용하였다. 고역 통과 필터는 저주파 영역의 변화가 적은 데이터를 제거하고, 대역 통과 필터는 고주파 영역의 급변하는 데이터를 제거한다. 그러나 실제 RASTA를 적용한

방법의 인식률이 PCA만 한 것보다 좋지 않은 결과를 보였다.

본 논문에서는 이 결과를 기반으로 RASTA를 응용하여 먼저 필터링을 하고 PCA를 하는 실험을 해보았다. 그 결과 제시된 방법은 작은 개수의 파라미터로 향상된 인식률을 보였다. 이것은 필터링에 의해 잡음이 제거되므로 보다 정보를이 좋은 주성분들이 추출되었음을 보여준다.

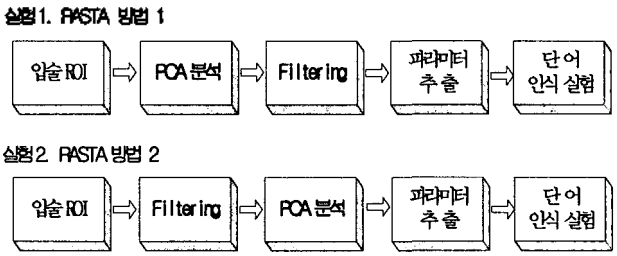

그림 2. 파라미터 추출 방법 비교

실험 1은 주성분 분석에 의하여 추출된 정보를 고역 통과 필터에 의하여 저주파 영역을 제거하거나, 대역 통과 필터에 의하여 고주파 영역과 저주파 영역을 제거한다. 실험 2 는 1의 실험을 역으로 한 것으로 필터링에 의해 제거한 정보를 주성분 분석을 거쳐 특징 파라미터를 추출한다.

아래 그림3은 실험 2방법으로 구해진 원이미지 이고 그림 4, 5, 6, 7 은 이것을 대역 통과 필터링 또는 고역 통과 필터링을 한 입술 ROI를 나타낸다.

그림 3. 16 by 16으로 다운샘플링 된 이미지

그림 4. 16 by 16 이미지의 고역 통과 필터링

그림 5. 8 by 16 이미지의 고역 통과 필터링

그림 6. 16 by 16 이미지의 대역 통과 필터링

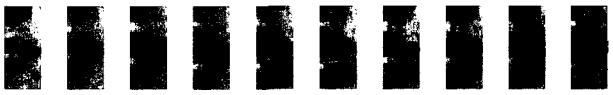

그림 7. 8 by 16 이미지의 대역 통과 필터링

4. 인식 실험 결과

실험에 사용된 입력 영상은 컬러 이미지로 코에서부터 턱까지 30Hz(frames/sec)로 촬영하여, 남성 화자 70명의 데이터가 구축되었다. 이중 52명은 학습데이터에 사용하고, 18명은 실험데이터로 사용하였다.

립리딩 인식 실험은 음성 인식 알고리즘으로 잘 알려진 HMM(Hidden Markov Model)을 이용하였다[7][8].

표 1은 주성분 분석만 한 데이터의 인식률이 실험 1의 실제 RASTA 보다 좋은 인식 결과를 보임을 나타낸다. 이는 영상의 정보와 음성의 정보의 형태가 다르기 때문일 것이다. 이 결과를 바탕으로 필터링을 먼저 한 후 PCA를 하는 방법 실험 2를 적용하여 보았다.

표 1. 실제 RASTA필터 적용결과 (실험 1)

	PCA 90% / 인식률	PCA 95% / 인식률
Bandpass	24 / 48.0%	44 / 43.2%
Highpass	24 / 49.7%	44 / 45.5%
No filter	24 / 63.9%	44 / 59.8%

이 실험에서 사용되는 주성분의 개수는 Folding하여 PCA만 한 것의 파라미터 수와 같고 이를 대역 통과 필터와 고역 통과 필터에 모두 사용하게 된다. 즉 필터링에 의해서는 파라미터의 개수가 줄어들지 않으며, 단지 추출된 파라미터 성분에서 고주파나 저주파 영역의 정보만을 제거하는 것이다.

표 2. 필터링 후 PCA 분석 결과(실험 2)

		PCA85%	PCA90%	PCA95%
Folding (8 by 16)	Bandpass	3	6	14
	Highpass	3	6	14
	No filter	17	24	44
Non-folding (16 by 16)	Bandpass	5	12	35
	Highpass	6	13	38
	No filter	24	41	84

표 2를 보면 알 수 있듯이 PCA만 통해 추출된 주성분의 개수와 필터링을 거쳐 PCA를 한 주성분의 개수는 많은 차이가 있다. 결과적으로 입술 ROI는 필터링에 의해 잡음이 제거되며, 이 파라미터에 PCA 적용하면 적은 수의 파라미터로 원이미지 정보를 대부분을 포함한다. 또한 줄어든 파라미터 수로 인하여 인식 속도를 줄일 수 있는 장점이 있다.

표 3. 실험 2의 인식 결과

		PCA90%/인식률	PCA95%/인식률
Folding (8 by 16)	Bandpass	6 / 68.9%	14 / 71.0%
	Highpass	6 / 67.4%	14 / 72.7%
	No filter	24 / 63.9%	44 / 59.8%
Non-folding (16 by 16)	Bandpass	12 / 67.4%	35 / 69.9%
	Highpass	13 / 67.9%	38 / 68.4%
	No filter	41 / 57.3%	84 / 51.5%

표3은 실험 2의 PCA 개수와 인식률을 나타낸 것인데, 필터를 사용한 방법이 PCA만 적용한 방법에 비하여 적은 파라미터 수로 좋은 인식결과를 나타낸다. 결과를 보면, 고역 통과 필터를 사용하였을 때 72.7%의 최고 인식률을 보인다. 이는 영상은 음성과는 달리 시간영역에서 볼 때 고주파 영역에도 단어를 인식할 수 있는 중요한 정보가 들어있다는 것을 보여준다.

표 3의 Folding과 Non-Folding을 비교해보면 추출된 파라미터 수는 많아 졌지만 인식률은 오히려 떨어진 것을 볼 수 있다. 이것은 파라미터 수가 많아질수록 그 파라미터 안에 들어가는 정보들의 중요도가 떨어져 오히려 인식률을 저하를 초래한다. 또한 많은 데이터양은 학습화 시간과 인식 실험 시간이 길어 그 효율 면에서도 떨어진다. 그러므로 필터링과 Folding은 립리딩에 있어서 불필요한 정보를 제거하고 이미지 기반이라는 단점을 보안하는데 매우 효과적이라 할 수 있다.

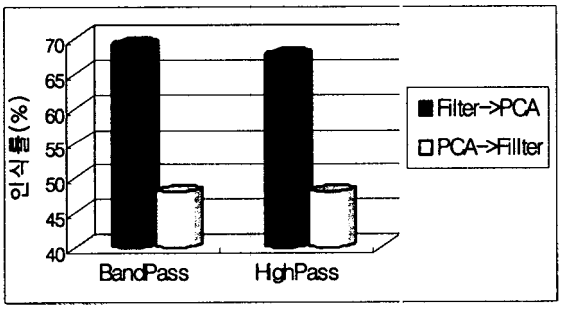

그림 8. PCA90%에서 실험 방법에 따른 비교

그림 8은 실험 1과 실험 2의 인식률을 비교 한 것이다. 필터링을 하고 PCA를 한 실험 2가 더 좋은 성능을 보임을 알 수 있다.

5. 결론 및 향후 계획

본 논문에서는 영상 정보만으로 단어를 인식하는 방법으로 효과적인 파라미터를 추출하는 립리딩에 대하여 살펴보았다. 립리딩에 있어서 영상 한 장의 정보도 중요하나 단어를 인식하기 위해서는 그 영상의 움직임 또한 매우 중요한 정보이다. 따라서 입술 움직임 정보를 추출하기 위해 시간영역에서 RASTA 필터를 적용하였다. 그 결과 주성분 분석만 하였던 63.9%인식률에 비하여 72.7%로 성능 향상을 보였다.

립리딩처럼 입술 정보만으로 단어를 인식하기란 음성 정보에 비하여 정보량이 매우 작다. 또한 화자마다 입술의 모양과 발음 습관, 환경의 변화 등의 제약 사항이 많기 때문에 안정적인 인식률을 기대하기란 어렵다. 이를 보상하기 위해 환경에 견인한 방법이 요구된다. 또한 본 논문의 실험결과에 따라 시간 영역 필터링이 좋은 성능을 보였으므로, 다양한 필터링에 대한 실험을 통하여 인식 성능을 조사해 볼 필요성이 있다.

참 고 논 문

[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]민덕수, 김진영, "음성인식기 성능 향상을 위한 립리딩 알고리즘", 호남·충청지역 학술대회 1999년 10월

[3]Jinyoung Kim, Joohun Lee, Katsuhiko Shirai," A Study on Various Factors Concerned with Lip-reading Performance at Dynamic Environment", Proceedings of ICSP 2001 August.

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

[5]G. Engel ,D. Greve and E. Schwartz, "Space-variant active vision and visually guided robotics" , ICPR pp. 487-490., 1994.

[6]DukSoo Min, JinYoung Kim, "Robust Lip Extraction and Tracking of the Mouth Region",ITC-CSCC, vol 2, pp. 927-931. 2000

[7]민덕수, 김진영, "Lipreading에 기반을 둔 HMM을 이용한 단어 인식", 신호처리 합동학술대회, 한국음향학회 발표, 1999년 10월

[8]김진범, 김진영, "이미지변환과 HMM에 기반한 자동 립리딩", 대한전자공학회 추계합동학술대회, 대한전자공학회 발표, 1999년 11월.