論文2000-37SP-5-11

입술의 대칭성에 기반한 효율적인 립리딩 방법

(An Efficient Lipreading Method Based on Lip's Symmetry)

金 鎭 範 *, 金 秦 永 **

(Jin-Bum Kim and Jin-Young Kim)

요 약

본 논문에서는 영상 변환 기반 자동 립리딩 알고리즘에서 처리하는 데이터 수를 효과적으로 감소시키는데 중점을 두었다. 화자의 입술에 대한 압축된 정보를 갖는 영상 변환 방식이 입술 윤곽선 기반 방식보다 우수한 립리딩 성능을 보이지만 이 방식은 입술 특징 파라미터를 다수 갖게 되므로 데이터 처리량이 많아지고 인식시간이 길어지게 된다. 계산되는 데이터를 줄이기 위해 우리는 입술의 대칭성에 기반하여 입술영상을 수직으로 접는 간단한 방법을 제안한다. 추가적으로 주성분 분석(PCA) 알고리즘을 사용하여 빠른 알고리즘을 고려하였고, HMM을 이용한 단어 인식실험 결과를 보인다. 제안된 방법에서 접어진 입술영상을 이용한 결과, 일반적으로 16×16 입술영상을 사용하는 방법에 비해 특징파라미터 수가 22~47% 감소하였고, HMM(hidden Markov model) 인식 알고리즘을 이용한 단어 인식률에서도 2~3% 개선된 결과를 얻었다.

Abstract

In this paper, we concentrate on an efficient method to decrease a lot of pixel data to be processed with an image transform based automatic lipreading. It is reported that the image transform based approach, which obtains a compressed representation of the speaker's mouth, results in superior lipreading performance than the lip contour based approach. But this approach produces so many feature parameters of the lip that has much data and requires much computation time for recognition. To reduce the data to be computed, we propose a simple method folding at the vertical center of the lip-image based on the symmetry of the lip. In addition, the principal component analysis(PCA) is used for fast algorithm and HMM word recognition results are reported. The proposed method reduces the number of the feature parameters at 22~47% and improves hidden Markov model(HMM) word recognition rates at 2~3%, using the folded lip-image compared with the normal method using 16×16 lip-image.

* 正會員, 全南大學校 電子工學科

(Dept. of Electronic Engineering, Chonnam National Univ.)

** 正會員, 全南大學校 電子工學科

(Dept. of Electronic Engineering, Chonnam National Univ.)

※ 본 논문은 한국과학재단의 '98 핵심전문연구 자원에 의해 이루어진 연구결과물 중 하나입니다.

接受日字:1999年12月30日, 수정완료일:2000年7月15日

Ⅰ. 서 론

근래에 음성인식 분야에서 화자의 연속적인 입술 영상 정보를 이용한 자동 립리딩에 대한 연구가 많은 관심의 대상으로 부각되고 있다. 일반적으로 화자의 음성만을 이용한 인식 시스템의 경우, 현재 깨끗한 음성 환경에서는 인식률 100%에 근접하는 우수한 인식률을 보이고 있으나 주위 잡음이 증가하면 인식률의 저하가 심화되는 취약성을 보이고 있다. 이에 비해, 영상 정보

는 다소 낮은 인식률을 나타내지만 소리 잡음에 거의 영향을 받지 않고 일정한 인식률을 보이기 때문에 잡음환경 하에서 음성인식률 저하를 보상할 수 있는 유력한 자원이 될 수 있다.

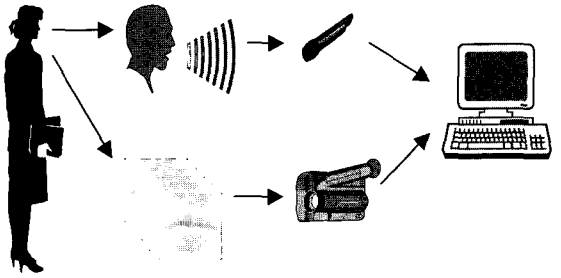


그림 1. 바이모달 인식 시스템 구성도
Fig. 1. *Bimodal Recognition System.*

이러한 음성과 영상정보의 결합을 통한 '바이모달(bimodal)' 음성 인식 시스템은 사람과 컴퓨터간의 상호 의사전달을 목적으로 하는 HCI(human computer interaction) 시스템의 여러 가지 정보전달 자원 중 중요한 일부분이 된다[1].

본 연구에서는 음성과 영상정보의 적절한 결합으로 잡음 환경에 강인한, 바이모달 음성 인식 시스템 구현을 목적으로 우선 오프라인(Off-line)으로 영상 정보만을 이용한 단어 인식 시스템의 성능 개선에 중점을 두었다.

본 실험에서는 먼저 일반적인 영상변환 기반 자동 립리딩 방법을 자체적으로 구현하여 그에 대한 화자 독립적 HMM 인식 알고리즘을 이용한 단어 인식 결과를 보였다. 영상변환 기반 접근법은 그 인식률이나 입술 특징 파라미터 추출에 있어서 입술 윤곽선 기반 방식보다 안정적이고 우수하다고 알려져 있다[2]. 그러나 입술 윤곽선 기반 방식은 소수의 입술 특징 파라미터를 처리하므로 파라미터 추출에서 인식하는데 까지 비교적 적은 데이터 처리량을 갖는데 비해[3], 영상변환 접근법은 화자의 입술 영상 전체의 픽셀들을 변환, 처리하므로 데이터 처리량이 훨씬 많아지게 된다. 영상변환 방식이 실시간 인식 시스템에서 사용되기 위해서는 알고리즘 개선을 통해 데이터 처리량을 줄이는 방법이 요구된다고 보겠다.

이러한 점을 고려하여 본 논문에서는 일반적인 영상변환 방법의 많은 데이터 처리량을 감소시킬 수 있는 간단하고도 효율적인 방법을 제시하고자한다. 본 논문에서 제시하는 알고리즘에서는 입술의 기하학적 대칭성에 착안하여 입술 영상을 입술의 중심점을 기준으로 대칭으로 접어 영상을 절반으로 줄임으로써 픽셀 데이터 처리량을 감소시킴과 동시에 추출되는 특징벡터 수도 줄일 수 있었다. 뿐만 아니라 데이터들의 유사성에 대한 통계적 확률을 계산하는 HMM 패턴인식 알고리즘을 사용하여 단어단위로 인식한 결과에 있어서도 본 실험에서 자체적으로 구현한 일반적인 영상변환 방법에 비해 동등하거나 다소 우수한 결과를 얻을 수 있었다. 부가적으로 데이터 처리량 감소에 따른 HMM 인식 시간 단축효과를 얻을 수 있었다.

본 논문의 구성은 다음과 같다. II장에서는 입술 특징 파라미터 추출을 위한 영상 전처리 과정을 보인다. 여기에서 통계적 알고리즘인 다중선형회귀분석을 이용하여 본 실험실에서 자체적으로 고안한 2진(binary) 영상변환 방법으로 입술 영역을 검색하는 과정을 보인다. III장에서는 2D-DCT (discrete cosine transform)를 적용하여 본 연구에서 구현한 일반적인 영상변환 과정을 보이는데, 여기서 특히 입술영상 전체에 대해 영상변환을 처리하는 방법에 비해 데이터 처리량을 절반으로 감소시킬 수 있는 입술의 대칭성에 기반한 간단하고도 효율적인 방법에 대해 기술하였다. 추가적으로 IV장에서는 빠른 알고리즘을 위해 PCA(principal component analysis)를 사용하여 중요한 입술 특징 정보만을 갖는 소수의 주성분 특징 벡터들을 추출하는 과정을 보인다[2]. V장에서는 본 논문에서 구현한 일반적인 영상변환 방법과 새롭게 제시되는 입술의 대칭성에 기반한 방법에 의해 각각 PCA를 거쳐 추출된 입술 특징 벡터들을 가지고 HMM 인식 시스템을 사용하여 단어 인식 실험을 수행한 결과에 대해 비교하여 살펴본다. 마지막으로 IV장에서는 결론과 향후 연구 계획을 밝힌다.

II. 입술 영상 전처리

1. 영상 전처리

화자의 입술영상은 SONY Digital Camcorder를 사용하여 320×240 픽셀 단위로 30frame/sec의 속도로 입력된 컬러 영상이다. 모든 실험은 52명의 각기 다른 화자로부터 얻어진 입술영상 DB를 가지고 오프라인으로 구현되었다. 빠른 알고리즘 구현을 위해 모든 입술영상은 먼저 160×120 크기로 다운샘플링(downsampling)을 거

친 후 명암영상으로 변환된다. 명암 영상에서의 다양한 조명변화는 이후 인식률에 적지 않은 영향을 미치게 되므로 이에 대해 일차적으로 전반적인 명암 균일화 (equalization) 과정을 거치게 된다[4][5]

균일화 과정을 거친 명암 영상에서 화자의 입술만이 포함된 ROI(region of interest)를 추출하기 위해서 명암 영상은 다시 2진 영상으로 변환된다.

먼저 그림 2의 (a) 명암 영상을 4 부분으로 분할한 뒤 그림 (d)에 보인 각각의 명암 히스토그램을 참조로 나타내었다. 2진 영상변환으로 입술 안쪽 영역만을 찾으려면 4개의 분할영역마다 각각 적절한 변환 임계값을 사용하여 2진 영상 변환을 거치고 결과적으로 변환된 4부분을 다시 결합하면 그림 (b)처럼 입술 안쪽 영역만을 뚜렷이 찾게 된다. 또한 그림 (b)에서 입술 안쪽 영역의 위치를 알기 위해 바로 옆 그림 (c)에는 Y축 프로젝션을 보이고 (b) 그림 바로 아래에 Y축 위치를 참조로 부분적 X축 프로젝션을 그림 (e)에 나타내었다.

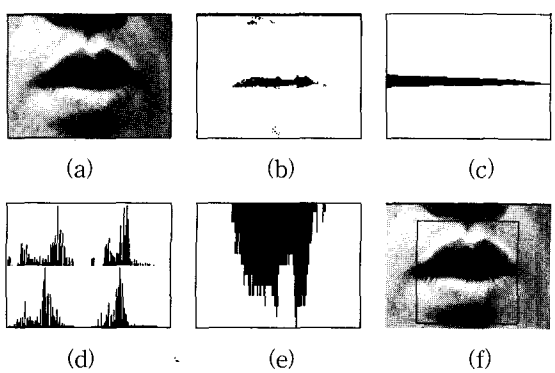


그림 2. 입술 전처리 (a) 입력 명암 영상 (b) 2진 변환 영상 (c) Y축 프로젝션 (d) 4분할 히스토그램 (e) 부분 X축 프로젝션 (f) 입술 ROI

Fig. 2. Lip image preprocessing (a) Gray image (b) Binary transform image (c) Y-projection (d) Histogram on 4 squares (e) Partial X-projection (f) Lip ROI.

2진 영상 변환은 입술의 안쪽 어두운 부분만을 나타나게 할 목적으로 사용되었는데, 영상의 상하좌우 조명의 불균형으로 인해 단순히 전체영상 변환으로는 구하기가 어려웠다. 여기서 본 연구에서는 입술 안쪽 영역만을 보다 명확히 찾기 위해 통계적 분석법인 다중선형회귀분석(MLRA multiple linear regression analysis)을 이용하여 자체적으로 고안한 2진 영상변환방법을 사용하였다.

다중선형회귀분석에서는 여러 개의 독립변수들로부터 하나의 종속변수를 예측한다. 2진 영상변환 임계값 T_i은 예측되는 하나의 종속변수이며 각 분할영역의 평균 명암값들($\overline{I_1}$, $\overline{I_2}$, $\overline{I_3}$, $\overline{I_4}$)을 독립변수들로 하면 다중회귀 모형은 일반적으로 다음 식 (1)과 같은 형태로 표현될 수 있겠다[3].

$$T_i = \beta_0 + \beta_{1i}\overline{I_1} + \beta_{2i}\overline{I_2} + \beta_{3i}\overline{I_3} + \beta_{4i}\overline{I_4} + \varepsilon \qquad (1)$$

여기서 각 독립변수 $\overline{I_1}$, $\overline{I_2}$, $\overline{I_3}$, $\overline{I_4}$ 들의 계수 β_0, β_1, β_2, β_3, β_4 는 추정할 수 있는 알려지지 않은 모집단 값이며, 오차 ε는 0을 중심으로 하는 정규분포 $N(0, \sigma^2)$을 따른다. 이 오차는 실제 임계값에 대한 예측값이 가지게 되는 추정오차인데 완벽하게 정규분포를 따르지는 않으며 무시할 수 있는 값이다.

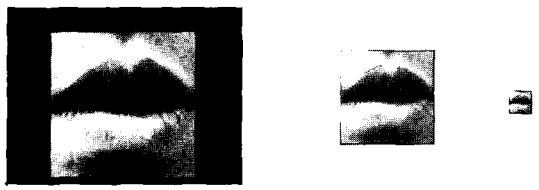
표 1. 2진 영상 변환 시 각 분할 영역에 대한 임계값의 계수들

Table 1. β-coefficients of threshold for binary image transform on each square.

	β_0	β_1	β_2	β_3	β_4
T_1	-100.228	0.691	0.441	0.110	0.094
T_2	-59.469	0.286	0.686	0.014	0.127
T_3	-85.124	0.223	0.233	0.528	0.280
T_4	-50.496	0.213	0.226	0.094	0.545

여기서 β계수들에 대한 추정치를 계산한 후에, 식(1)의 T_i에 대한 예측된 값을 얻기 위하여 추정된 β계수들을 대입하게 된다. 본 실험에서는 β계수들에 대한 추정치를 구하기 위해 52명의 입술영상 DB에서 무작위로 선별된 200개의 입술영상 DB에 대해 4분할된 각 영역의 평균 명암값을 구한 뒤 각각 2진 영상변환 시 입술 안쪽영역만이 시각적으로 잘 드러나는 변환 임계값을 수작업으로 모두 구하였다. 이렇게 구해진 4분할 영역에 대한 평균 명암값과 2진 변환 임계값들의 데이터를 가지고 회귀분석을 적용하면 β계수들에 대한 추정치를 구할 수 있게 된다. 추정된 β계수들은 임계값 T_i의 관찰값과 회귀방정식으로부터 구해진 임계값 T_i의 예측값 사이의 차이를 최소화 시켜주는 값들이다. 실제 T_i값과 예측된 T_i값간의 차이를 잔차(residuals)라고 하는데, 선형회귀분석은 계수들에 대한 최소제곱

추정치를 구하는 것으로서 구체적으로 잔차들의 제곱합을 최소화하는 추정치를 구하게 된다. 표 1에서는 200개의 입술영상을 기반으로 추정된 β계수들을 보이고 있다.



(a) (b) (c)

그림 3. 입술ROI 다운샘플링 (a) 획득된 입술 ROI (b) 64×64 크기로 다운샘플링된 영상 (c) 16×16 크기로 다운샘플링

Fig. 3. Lip ROI downsampling (a) Obtained lip ROI (b) 64×64 downsampled image (c) 16×16 downsampled image.

이렇게 추정된 β계수들과 4분할 영역에 대한 각 평균 명암값을 식 (1)에 대입하면 각 4분할 영역의 2진 영상변환 임계값의 예측치를 얻게 된다. 이 예측 임계값을 적용하여 2진 영상 변환을 수행하면 그림 2. (b)에서처럼 일반적으로 낮은 명암 값을 갖는 입술의 안쪽 영역이 나타나게 된다. 변환된 2진 영상에 대해 그림 2의 (c)에 보이고 있는 Y축 프로젝션(projection)의 최대치를 갖는 부분이 입술의 수평 중심위치가 되고, 그 중심선상으로부터 그림 2의 (e)에 보인바와 같이 부분적 X축 프로젝션을 통해 입술의 폭을 추출할 수 있다.

추출된 입술의 폭으로부터 입술의 수직 중심위치를 구할 수 있고 이를 기준으로 폭과 높이가 1:1의 비를 갖는 ROI로 분리하였다 [그림 2의 (f)]. ROI의 한 변이 갖는 크기는 기본적으로 '화자의 입술 폭×1.1 배' 로 설정했으나, 여기서 입술의 폭은 화자에 따라, 화자와 카메라 사이의 거리변화에 따라 다를 수 있으므로 매 화자의 매 단어에 대해 각각의 첫 프레임에서 구해진 입술 폭을 한 변으로 하는 ROI를 각각의 연속되는 나머지 프레임에 일률적으로 적용하였다.

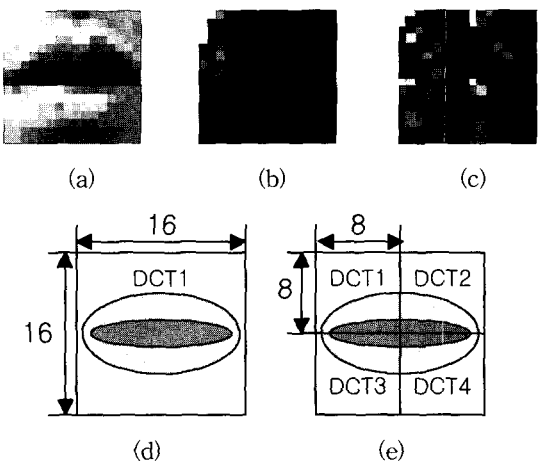
여기에서 매 화자와 단어에 따라 추출된 ROI들은 그 입술의 크기와 폭에 따라 보통 50 - 120 픽셀크기 까지 다양한 크기의 변을 갖는 정사각형의 픽셀 윈도우들이다. 각각 다른 크기로 추출된 ROI를 일정한 크기로 정규화 시키기 위해 2D- interpolation을 단계적으로 적용하여 먼저 64×64 픽셀 윈도우 크기로 일정하게 다운샘플링을 했고, 이를 다시 본 연구에서 적용하고자 하는 영상 선형변환 알고리즘인 2D-DCT에 적용하기 위해 최종적으로 16×16 픽셀 윈도우 크기로 축소했다 [그림 3].

Ⅲ. 영상 선형변환 알고리즘

1. 영상 변환 기반 접근법

앞에 보인바와 같이 전처리를 거쳐 구해진 입술 ROI의 연속적 영상 프레임을 'g' 라 하면 전체 영상 데이터에서 임의의 시간 t의 영상 프레임에 해당하는 g_t는 다음 식 (2)로 표현될 수 있겠다.

$$g_t(x, y, t) : 1 \le x \le M,\ 1 \le y \le N,\ 1 \le t \le T \qquad (2)$$



(a) (b) (c) (d) (e)

그림 4. 2D-DCT를 적용한 일반적 영상변환 (a) 16×16 입술 ROI (b) 16 x 16 2D-DCT 적용결과 (c) 8×8 2D-DCT 적용결과 (d) 16×16 픽셀 윈도우 2D-DCT (e) 8×8 픽셀 윈도우 2D-DCT

Fig. 4. Normal image transform using 2D- DCT (a) 16×16 Lip ROI (b) 16×16 2D- DCT result (c) 8×8 2D-DCT result (d) 16×16 pixel-window 2D-DCT (e) 8×8 pixel-window 2D-DCT.

여기서 M=N=16 이고 T는 연속되는 전체의 영상 프레임 수이다. 프레임 데이터 g_t를 임의의 선형 변환 행렬 P와의 내적을 거치면 식 (3)에서 보듯이 립리딩에 관련된 정보를 갖는 특징 벡터들의 집합 X_t 로 선형 변환시킬 수 있다.

$$X_t = Pg_t \quad (3)$$

P는 임의의 선형변환으로서 본 논문에서는 2D-DCT를 고려하였다. 우선 본 연구에서는 일반적인 영상변환 접근법에서처럼 16×16 ROI 전체에 대해서 16×16 및 8×8 의 4 영역으로 분할한 윈도우 크기로 2D-DCT를 각각 수행하였다 [그림 4. (b),(c)].

그러나 서론에서 밝힌바와 같이 16×16 ROI 전체에 대해 2D-DCT를 적용하면 데이터 처리량도 많아지게 되고 IV장에서 보이듯이 PCA를 이용하여 추출한 입술 특징 파라미터 수에 있어서도 입술윤곽선 기반 방식보다 훨씬 많아지게 된다. 그래서 본 논문에서는 데이터 처리량을 줄임으로써 추출되는 입술특징 파라미터의 수도 감소시키고 그로 인한 전체적인 인식 알고리즘의 처리시간의 단축효과까지 얻을 수 있는 효과적이고도 간단한 방법을 제시하고자 한다.

2. 효율적인 데이터 처리량 축소 알고리즘

립리딩 방식에 있어서 입술 윤곽선 기반 접근법과 입술 영상 선형변환 접근법을 비교하자면, 영상 변환 접근법이 HMM 인식률 면에서 우수할 뿐만 아니라 입술 특징 벡터 추출에 있어서도 보다 안정적이라 할 수 있는데, 그 이유는 입술 윤곽선 기반일 경우 입술영역을 정확히 찾았다해도 명암 변화 및 그 외 다른 요소들에 의해 입술의 특징 파라미터들이 잘못 추출되는 경우가 있을 수 있고 그 때문에 이후 인식률에서도 저하를 초래할 수 있기 때문이다[2].

반면, 데이터 처리량 면에서 보면 영상 변환 접근법이 훨씬 많은 양을 처리하고 있다. 입술 윤곽선 기반에서는 주로 입술 폭, 안쪽 입술 높이, 바깥쪽 입술 높이 정도를 특징 파라미터로 사용하지만 영상 변환 접근법에서는 ROI 픽셀들을 전체 선형변환 한 후 주성분으로 분류되는 입술특징 벡터로 압축시켜도 입술 윤곽선 기반 접근법보다 훨씬 많아지게 된다.

이러한 입술특징 파라미터의 차이는 이후 HMM 인식 알고리즘 적용 시에도 그 데이터 처리량에 있어서 많은 차이가 있고 인식 처리시간에도 많은 영향을 미치게 되므로 보다 빠른 알고리즘 구현을 위해서는 데이터 처리량을 줄이는 방안이 요구된다고 보겠다.

본 논문에서 제안하는 방법은 다음과 같다. 화자의 입술이 입술 중심점을 기준으로 좌우 대칭적 모델임을 볼 때, 화자의 머리방향이 정면을 향하고 있고 입술이 수평선상 상하 기울어짐이 없다는 전제하에 영상을 입술의 중심점을 기준으로 반으로 접을 수 있을 것이다 [그림 5. (b),(c)].

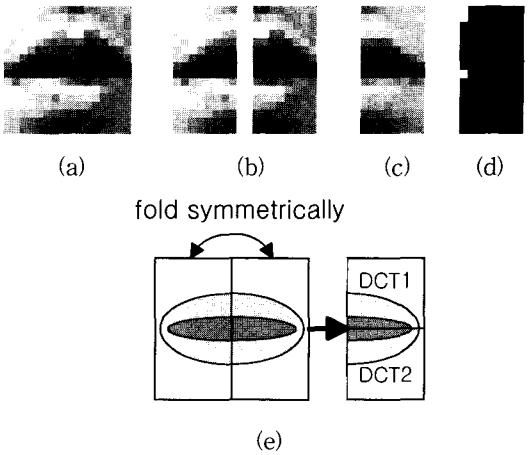


그림 5. 입술대칭성 기반 영상변환 (a) 16×16 입술 ROI (b) 수직으로 분할된 ROI (c) 분할된 ROI를 대칭으로 접어서 합친 그림 (d) 접어진 8×16 ROI에 대한 8×8 2D-DCT 적용결과 (e) 입술의 대칭성 기반 ROI 접기와 2D-DCT 적용

Fig. 5. The image transform based on the symmetry of the lip (a) 16×16 Lip ROI (b) The half-divided ROI vertically (c) Symmetrically folding and adding image with the divided ROI (d) 8×8 2D-DCT implementation on 8×16 folded ROI (e) Folding ROI based on the symmetry of the lip and 2D-DCT implementation.

접어진 입술영상의 각 픽셀들은 좌우 접대칭 되는 두 픽셀 값의 평균값이 된다. 따라서 그림 5의 (e)에서처럼 접어진 8×16 ROI에 대한 2D-DCT는 8×8 픽셀 윈도우 크기로 2번만 수행하면 된다. 앞의 그림 4의 (e) 그림에서 보는바와 같이 16×16 ROI에 대해 8×8 픽셀 윈도우 2D-DCT를 4번 수행한 것에 비해 겨우 절반정도의 계산양인 것이다.

이러한 방법은 원래의 ROI 영상 크기를 절반으로 줄임으로써 픽셀 데이터 처리량 및 이후 IV장에서 PCA를 통해 생성되는 특징 파라미터들의 수를 감소시킬 수 있다. 뿐만 아니라, 접어진 8×16 ROI의 픽셀 값들은 대칭되는 픽셀들의 평균값들이므로 영상잡음 요소 및 좌우 측면 조명의 불균형에 대한 강인함을 갖게 된다고 볼 수 있다.

Ⅳ. 주성분 분석

PCA 알고리즘은 임의 시간 t 에서 원래의 벡터 $X_{it} = [X_{1t}, X_{2t}, \cdots, X_{p(=MN)t}]$를 적절히 선형변환 시켜 그것이 가지는 정보를 가능한 많이 보존하는 소수 m 개의 새로운 인공변수를 창조함으로서, p-차원 변이를 m-차원으로 축소하여 전체 체계의 특성을 요약할 수 있다.

여기에서 다음 식(4)에 보인바와 같이 X_{it}의 원소들 간의 상관구조를 나타내는 공분산 Σ 에 기반한 PCA를 고려하여 그 일반성을 유지하였다.

$$\Sigma = E \Delta E' \qquad (4)$$

여기서 E는 p개의 고유벡터(eigenvector) $e_i = \{e_{1i}, e_{2i}, \cdots, e_{pi}$, (p=MN)}들을 열로 하는 크기 (pxp)인 직교행렬이고 Δ 는 Σ 의 고유값(eigenvalue) δ_i를 대각원소로 하는 크기 (pxp)인 대각행렬이다. 이는 다음 식 (5)로 표현될 수 있겠다.

$$E = (e_1, e_2, \cdots, e_p), \quad \Delta = diag(\delta_1, \delta_2, \cdots, \delta_p) \qquad (5)$$

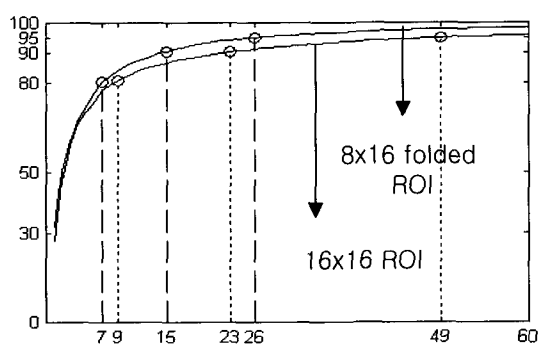


그림 6. 주성분 개수 'm' 에 따른 누적 백분율(PCA 80, 90, 95% 적용 시)

Fig. 6. Accumulate percentage of the principle components 'm' (PCA 80, 90, 95% adapted)

이때 고유값과 각각의 고유값에 대응되는 고유벡터 e_i 의 짝을 δ_i의 크기순서($\delta_1 \geq \delta_2 \geq \cdots \geq \delta_p$)로 배열하고 맨 첫 번째부터 가장 큰 고유값 'm'개에 해당하는 고유벡터의 짝 E_m 을 이용하여 데이터 벡터 X_{it} 에 대한 다음과 같은 직교변환 $o_{jt} = E_m' X_{it}$ 를 고려할 때 이 변환에 의해 새로이 창조되는 'm' 개의 특징벡터 o_{jt} , j = 1, 2, ···, m , (m<=p) 를 X_{it} 의 주성분으로 추출할 수 있겠다.

표 2. 일반적인 16 x 16 ROI와 제안된 8×16 ROI의 주성분 개수 비교

Table 2. Compare the proposed 8×16 ROI with the normal 16×16 ROI.

PCA(%)	제안된 8×16 ROI	16×16 ROI	감소율(%)
80	7	9	22.22
90	15	23	34.79
95	26	49	46.94

여기서 주성분 개수 'm'을 구하기 위해 고유값 δ_i의 합을 $tr(\Delta) = (\delta_1 + \delta_2 + \cdots + \delta_p)$이라 하면, 주성분 'm' 개가 가지는 원본 데이터에 대한 정보율은 $(\delta_1 + \delta_2 + \cdots + \delta_m) / tr(\Delta)$이 될 것이다.

그림 6에서는 본 논문에서 실험한 2D-DCT 영상선형 변환을 수행한 결과에 대해 'm' 개의 고유값들에 대한 누적 백분율을 산출해 낸 것이다. 표2에서 보는 바와 같이 PCA를 거친 후 원래 정보의 80%를 갖는 주성분 벡터 수(m)는 일반적인 영상변환 접근법에 따라 본 연구에서 실험한 16×16 ROI의 경우 9개였던 것이 새롭게 제시된 데이터 처리량 축소 알고리즘에 따라 8×16 접어진 ROI인 경우 7개로 줄었고, 90%를 갖는 주성분 벡터 수는 23개에서 15개로 34.8% 정도 줄어들었으며, 95% PCA의 경우 49개에서 26개로 47% 감소하였음을 알 수 있다. 본 논문에서는 90% PCA 적용 시 추출되는 주성분들을 기준으로 HMM 인식 실험의 파라미터로 사용하였다.

Ⅴ. HMM 기반 인식 실험

1. 영상 특징 정보 전처리

본 논문에서 인식 알고리즘으로 사용한 HMM 인식 알고리즘은 통계적 패턴매칭 방법을 사용한다[6]. 음성인식 알고리즘으로 주로 사용되는 HMM 알고리즘의 관찰(observation) 확률은 대각 공분산 행렬을 가지는 다차원 가우시안 믹스쳐(Gaussian mixture)들로 확률적 모델을 생성하게 된다.

본 연구에서는 임의의 시간 t 에서 화자의 입술 영역

을 담고있는 영상에 대해 선형 변환을 거쳐 최종적으로 PCA를 통해 추출된 소수 m 개의 원소만을 갖는 o_{jt}를 통계적 관찰 특징 벡터로 가정하였다. 각 단어에 대해 o_{jt}의 시간적인 변화를 관찰하여 학습화(training) 과정을 거쳤고 결과적으로 확률적인 모델(model)을 만들었다.

추가적으로 인식률 증대를 목적으로 다음과 같은 델타 파라미터(delta parameter) D_{jt}를 모든 o_{jt} 마다 덧붙여 사용하였다.

$$D_{jt} = k_1(o_{j(t+1)} - o_{j(t-1)}) + k_2(o_{j(t+2)} - o_{j(t-2)})) \quad (j = 1, 2, \cdots, m) \tag{6}$$

여기서, k_1 , k_2 는 가중치로서 각각 2와 4이고, 1 ≤t≤T 는 연속되는 전체 영상 프레임 수, j 는 특징 벡터 o_{jt}의 원소 수, o_j는 j 번째 특징 벡터이고, D_{jt}는 임의의 프레임 t 에서의 j 번째 특징 벡터에 해당하는 델타 파라미터의 값이다.

본 실험에서 구현한 일반적인 16×16 ROI에 대한 영상변환의 경우와 입술대칭성 기반의 8×16 접어진 ROI의 경우, PCA 90% 적용 후 각각 델타파라미터를 추가하면 HMM 인식실험에 사용될 파라미터 개수는 16×16 ROI 일 때 46개, 8 x 16 ROI 일 때 30개가 된다.

표 3. HMM 단어 인식 실험에 사용된 22개의 정보서비스의 매뉴얼 단어들

Table 3. 22 Words in the manual of the information service using by HMM words recognition tasks.

번호	단어명	번호	단어명
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

2. 실험 및 결과

실험에 사용된 영상 데이터는 20대 남/자 52 명이 22개의 단어를 평상시 발음으로 발음한 영상을 SONY 디지털 캠코더를 사용하여 30 frame/sec의 속도로 저장한 것이다. 22개의 단어는 정보서비스를 제공해 줄 때의 매뉴얼 단어로 다음 표 3에 보인바와 같다.

HMM의 학습화 과정에 52명의 영상 데이터가 사용되고 실제 인식률을 시험(testing)하고자 하는 데이터는 52명 중 임의의 중복된 18명이 같은 22개 단어에 대해 다시 발음한 영상 데이터를 사용하여 표 4, 5, 6에서와 같이 HMM 알고리즘에서의 상태(state) 수와 가지(mixture) 수를 변화시키면서 실험을 수행해 보았다.

각 표에서 상태 수는 'S'로 가지 수는 'M'으로 표기한다. 모든 실험은 PCA 90% 적용, 52명 training / 18명 testing을 기준으로 하였다.

HMM에서 상태수라는 것은 임의의 통계적 데이터가 어떤 일정한 확률적 모델링이 될 때 모델링 될 수 있는 상태 수가되며, 가지 수는 그 상태 중 어느 하나에서 나타날 수 있는 경우들의 수를 가리키는 것이다[6].

본 연구에서는 영상정보만으로 인식실험을 하였으므로 어떠한 단어의 발음 구간동안 연속된 영상 프레임들에서 추출된 입술 특징 벡터들을 일정한 상태들로

표 4. 16×16 ROI일 경우, 16×16 2D-DCT 기반 HMM 인식률 (단위 : %)

Table 4. 16×16 ROI, HMM recognition rate based on 16×16 2D-DCT.

DCT	S 3	S 4	S 5	S 6
M 3	42.93	47.98	49.24	51.26
M 4	42.93	47.47	47.72	51.01
M 5	47.47	46.72	45.71	51.52
M 6	44.70	47.22	46.97	51.52

표 5. 16×16 ROI일 경우, 8×8 2D-DCT 기반 HMM 인식률 (단위 : %)

Table 5. 16×16 ROI, HMM recognition rate based on 8×8 2D-DCT.

DCT	S 3	S 4	S 5	S 6
M 3	45.20	47.22	46.97	51.26
M 4	38.38	45.96	48.74	52.02
M 5	43.18	45.45	48.23	47.47
M 6	48.99	47.22	48.48	51.52

나누고 또 그 단어에 대해 여러 사람이 발음한 영상정보들을 각각 비슷한 확률분포를 갖는 집합들로 가우시안 믹스쳐 모델링을 하게 된다. 이렇게 학습을 통해 생성된 확률적 모델을 바탕으로 입력 정보들의 패턴에 대해 확률적 비교를 하는 알고리즘을 사용하여 인식실험을 하였다.

여기서 각각의 상태와 가지 수를 나타내는 'S'와 'M'에 따라 HMM의 인식률 변화가 나타나지만, 단어마다의 어절 수나 연속된 영상의 프레임 수 등의 제한이 있으므로 어느 점에 가면 더 이상 인식률은 올라가지 않게 된다. 보통 가지 수 'M'은 그 증가에 따라 인식률이 비례하여 증가하지는 않으므로, 상태 수 'S'에 따른 인식률의 변화 추이를 주목해야 할 것이다.

표 4, 5에 보인 HMM 단어 인식결과들은 본 연구의 초기에 일반적인 영상변환 접근법에 따라 구현한 16 x 16 입술ROI 전체를 영상 변환 처리하여 PCA를 통해 추출한 입술특징 파라미터들을 사용하여 인식실험을 한 것이고, 아래 표 6은 본 연구에서 새롭게 제시하는 입술의 대칭성을 이용한 알고리즘을 적용하여 8 x 16으로 접어진 입술ROI에 대해 영상 변환하여 역시 PCA를 거쳐 추출된 입술특징 파라미터를 가지고 HMM 인식실험을 수행한 결과에 대한 것이다.

표 6. 입술의 대칭성에 기반하여 8×16 접어진 ROI에 대한 8×8 2D-DCT 적용 HMM 인식률 (단위 : %)

Table 6. HMM recognition rate on 8×16 folded ROI based on the symmetry of the lip.

DCT	S 3	S 4	S 5	S 6
M 3	39.90	49.24	49.24	53.28
M 4	44.70	46.21	50.00	54.04
M 5	46.72	52.02	49.49	52.78
M 6	43.94	47.47	48.74	53.54

표 4, 5에 보는 바와 같이 16 x 16 ROI에 대해 16×16 및 8×8 2D-DCT를 각각 적용한 결과는 별다른 차이를 보이지 않는다는 사실을 알 수 있다. 이 두 가지의 경우 영상변환 후 추출된 특징 파라미터 수는 Ⅳ장에서 이미 보인바와 같이 PCA 90% 적용 시 23개로 같았다.

반면, 표 6에서는 본 논문에서 제안하는 방법에 따라 8×16 픽셀 윈도우 크기로 접어진 ROI에 대해 8×8 2D-DCT를 적용한 결과를 보이고 있는데, 앞에 보인 표 4, 5의 두 결과보다 전반적으로 다소 향상된 인식률을 보이고 있다. 그리고 가지수 4에서의 상태수 변화에 따른 HMM 인식률의 변화 추이를 각각 비교하여 그림 7에 보였다.

본 논문에서 제안하고 있는 입술의 대칭성을 이용한 알고리즘을 적용한 결과가 일반적인 영상변환 방법에 따른 앞의 다른 두 결과보다 전반적으로 2~3% 우수하게 나타나고 있다. 뿐만 아니라 Ⅳ장에서 이미 보인 것과 같이 8×16으로 접어진 입술ROI에 대해 영상변환 후 추출된 특징 파라미터의 수도 PCA 90% 적용 시 15개로써, 일반적인 영상변환 방법에 따라 본 실험에서 구현한 16×16 ROI의 경우 23개였던 것에 비해 약 34.8% 감소하였음을 알 수 있다.

본 논문에서 사용한 입술의 대칭성을 이용한 방법이 영상변환 처리 시 데이터 처리량에 있어서나 이후 HMM 인식 파라미터 수에 있어서 훨씬 적음에도 불구하고 이러한 인식률 향상 결과를 보인 것은 입술 영상을 대칭으로 접으면서 점대칭 되는 각 픽셀들을 합하여 평균값을 취하는 과정이 전체적으로 SNR을 향상시키는 효과를 가져온 것이라 볼 수 있다. 뿐만 아니라, 원래 입술 영상이 가질 수 있는 좌우측면 조명의 불균형에 대해서도 강인함을 가지게 된 원인도 있다고 본다.

추가적으로 변환영상을 절반으로 줄임으로써 데이터 처리량이 감소하게되면 그에 따른 입술특징 파라미터 수가 감소되므로 HMM 인식시간의 단축효과를 기대할 수 있다. 델타 파라미터를 추가한 HMM 인식 파라미터 개수는 16×16 과 8×16 ROI 각각 46개와 30개가 되므로 파라미터 개수의 차이에 따라 HMM 인식 알고리즘 상에서 처리하는 확률 계산량은 줄어들게 되며 따라서 전체적인 인식시간도 단축시킬 수 있게 된다. HMM 인식에 적용되는 파라미터 수에 따라 확률 계산량이 차이가 나며 또한 같은 파라미터 개수에서도 State와 Mixture가 증가할수록 계산량이 훨씬 더 증가하게 된다. HMM 인식률을 높이려면 파라미터 수를 늘려야 하지만 인식시간의 단축을 위해서는 파라미터 수를 줄여 계산량을 감소시켜야 한다. 그러므로 립리딩에 있어서 최종 인식에 적용될 입술특징 파라미터의 수를 줄이면서도 보다 높은 인식률을 얻을 수 있는 알고리즘이 개발될 필요가 있다.

현재 실험에 사용된 단어들은 개수가 적고 발음시간

이 짧아 대폭적인 인식시간의 차이를 보이지 않으나, 향후 수행될 대용량 단어 인식실험에서는 데이터 처리량 및 HMM 인식에 적용될 입술특징 파라미터 수의 감소에 따른 인식시간의 단축효과가 더욱 증대될 것으로 본다.

Ⅵ. 결 론

본 논문에서는 자동 립리딩을 위한 영상 선형변환 접근법에서 입술의 대칭성을 이용한 보다 효율적인 방법을 제안하였다. 단, 이 방법은 화자의 입술 중심점이 추출된 ROI 영상의 중심점과 일치하고, 입술이 수평선상에서 상하 기울어짐이나 좌우 머리회전이 없다는 가정을 전제로 한다. 그러한 전제하에 실험한 결과 본 논문에서 제안하는 방법이 입술의 대칭성을 이용하여 입술ROI를 절반으로 접기 때문에 데이터 처리량도 절반으로 줄이면서 입술 특징 파라미터 수의 감소에 따른 계산량 감소와 우수한 립리딩 결과를 얻을 수 있다는 사실을 확인하였다.

이 실험에서는 22개의 한정된 단어를 가지고 분석하였는데, 보다 광범위한 단어에 적용하여 볼 필요가 있으며, 추가적으로 본 연구에서는 단어 단위로 인식 실험을 수행하였으나 향후 서브 워드(sub-word) 단위의 대용량 단어인식에 대한 연구가 고려될 것으로 본다. 또한 화자의 머리회전 및 입술의 상하 기울어짐 뿐 아니라 카메라와의 거리변화에도 강인한 알고리즘을 위한 연구도 고려해 봐야 할 것이다.

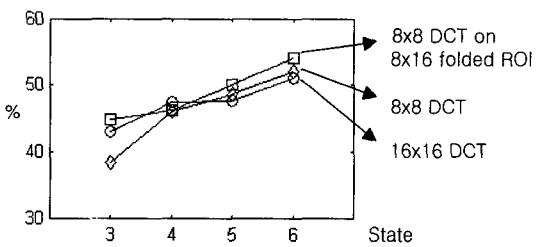


그림 7. 영상 변환 기반 립리딩 HMM 단어 인식 결과 비교 (Mixture 4)

Fig. 7. HMM words recognition results of the image transform based automatic lipreading (at Mixture 4)

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. pp.853-869

[2] Potamianos, G.; Graf, H.P.; Cosatto, E., "An image transform approach for HMM based automatic lipreading", Image Processing, 1998. ICIP98. Proceedings. 1998, International Conference on, 1998, pp.173 -177, vol.3

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

[4] Liëvin M. and Luthon F. "Lip features automatic extraction", Proc. Of the 5th IEEE Int. Conf. On Image Processing. Chicago. Illinois, 1998.

[5] Uwe Meier, Rainer Stiefelhagen, Jie Yang, "Preprocessing of visual speech under real world conditions", Interactive Systems Lab. European Tutorial & Research Workshop on Audio-Visual Speech Processing: Computational & Cognitive Science Approaches (AVSP 97).

[6] Lawrence Rabiner, Biing-Hwang Juang, "Fundamentals of Speech Recognition", Published by PTR Prentice-Hall, Inc. pp.321 -389. 1993.

[7] 박병구, 김진영, 최승호, "잡음 환경 하에서의 바이모달 음성인식", '98 한국음향학회 학술발표대회 논문집. pp.111-114, 1998년 7월

[8] 박병구, 김진영, 임재열, "입술 파라미터 선정에 따른 바이모달 음성인식 성능 비교 및 검증", 한국음향학회지 제 18 권, 제 3 호, pp. 68-72, 1999년 4월

[9] 박병구, 김진영, 최승호, "바이모달 음성인식의 음성정보와 입술정보 결합방법 비교", 한국음향학회지 제 18 권, 제 4 호, pp.31-37, 1999년 6월

저 자 소 개

金 鎭 範(正會員)

1972년 8월 15일생. 1997년 2월 조선대학교 전자공학과 졸업(공학사). 1998년 9월~현재 전남대학교 대학원 전자공학과 석사과정. ※ 주관심분야는 멀티모달 MMI

金 秦 永(正會員)

1962년 4월 26일생. 1986년 2월 서울대학교 전자공학과(공학사). 1988년 2월 서울대학교 전자공학과(공학석사). 1994년 8월 서울대학교 전자공학과(공학박사). 1993년 3월~1994년 12월 한국통신 소프트웨어연구소 전임연구원. 1995년~현재 전남대학교 공과대학교 전자공학과 조교수. ※ 주관심분야는 음성인식 및 음성합성, 멀티모달 MMI