이미지 변환과 HMM에 기반한 자동 립리딩

김진범, 김진영
전남대 전자공학과, 고품질 전기·전자 부품 및 시스템연구센터
Tel) 062-530-0472, Fax) 062-530-0472

Automatic Lipreading Based on Image Transform and HMM

Jin Bum Kim , Jin Young Kim
Electronic Engineering, Chonnam National University &
Research Center for High-Quality Electric Components and Systems
E-Mail : kimjin@dsp.chonnam.ac.kr

Abstract

This paper concentrates on an experimental results on visual only recognition tasks using an image transform approach and HMM based recognition system. There are two approaches for extracting features of lipreading, a lip contour based approach and an image transform based one. The latter obtains a compressed representation of the image pixel values that contain the speaker's mouth results in superior lipreading performance. In addition, PCA(Principal component analysis) is used for fast algorithm. Finally, HMM recognition tasks are compared with the another.

I. 서론

최근 음성인식 분야에서는 잡음 하에서의 인식률을 높이기 위한 보조 수단으로서 화자의 입술을 포함한 영상 정보를 이용하는 연구가 활발히 진행되고 있다. 음성과 영상 정보를 이용한 립리딩(lipreading)이 추가된 바이모달(bimodal) 형태는 그밖에 화자의 얼굴표정이나 머리의 회전, 손짓, 눈동자의 움직임 등의 종합적인 정보를 갖는 멀티모달(multimodal) 형태와 함께 HCI(human-computer interface) 분야에서 중요한 부분을 차지하고 있다[1].

본 논문에서는 바이모달에서 필요한 립리딩에 관련된 화자의 입술 영상 정보를 처리하는 방식에 있어서, 입술 윤곽선의 특정 정보만을 이용한 방식보다 안정적이고 인식률의 개선을 가져오는 영상 변환 방식으로 접근하여 독립적인 단어 인식에 대해 실험하였다.

일반적으로 입술 윤곽선 기반 방식에 비해 화자의 입술 영상 전체를 변환하여 처리하는 방식은 보다 많은 데이터를 처리하게 되므로 수행 속도를 고려하여 모든 영상은 그레이 레벨로 변환하여 분석, 처리하였다. 또한 영상처리의 관건인 다양한 조명의 변화를 균일화하기 위해 입력영상을 4영역으로 분할하여 각 부분의 명암에 대한 보상 처리를 하는 알고리즘을 적용했으며, 각 화자와 단어마다 카메라와의 거리에 따라 입술크기가 달라질 수 있으므로 이에 대한 정규화도 고려하였다. 또한 영상의 선형 변환을 위한 알고리즘으로는 DCT(Discrete Cosine Transform)와 DWT(Discrete Wavelet Transform)의 2가지 방식을 적용하였다.

마지막으로 주성분 분석(PCA : Principal Component Analysis) 알고리즘을 사용하여 처리할 데이터 양을 줄임으로서 보다 빠른 알고리즘의 구현이 용이하게 하였고, 이 결과를 화자 독립 인식이 가능한 HMM 인식 시스템에 적용하여 인식실험을 수행하였다.

수행하였다. 그림 3은 각각의 결과를 영상으로 보인 것이다[4],[5].

2.3. 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.

본 논문에서는 X_{it} 의 원소들 간의 상관구조를 나타내는 공분산 Σ 에 기반한 PCA를 고려함으로서 일반성을 유지하였다.

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

여기서 E는 p 개의 고유벡터(eigenvector) e_i 들을 열로 하는 크기 (pxp)인 직교행렬이고 Δ 는 Σ 의 고유값(eigenvalue) δ_i 를 대각원소로 하는 크기 (pxp)인 대각행렬이다. 이는 다음 식 (5)로 표현될 수 있다.

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

이때 고유값과 각각의 고유값에 대응되는 고유벡터 e_i 의 짝을 δ_i 의 크기순으로 배열하고 가장 큰 고유값 m 개에 해당하는 고유벡터의 짝 E_m 을 이용하여 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} , $j = 1, 2, \cdots, m$, $(m \leq p)$를 X_{it} 의 주성분으로 추출할 수 있다. 이때 고유값들의 합을 $tr(\Delta) = (\delta_1 + \delta_2 + \cdots + \delta_p)$ 이라 하면 m 개의 주성분에 의해 설명되는 부분은 $(\delta_1 + \delta_2 + \cdots + \delta_m)/tr(\Delta)$ 이 될 것이다.

그림 4에서는 DCT 변환을 수행한 결과에 대해 실험적인 결과를 산출해 낸 것으로, M=N=16 일 때 m 개의 고유값들에 대한 누적 백분율을 보이고 있다. 여기에서 PCA를 거친 후 원래 정보의 80%를 갖는 데이터 벡터 수(m)는 9개 였고, 90%를 갖는 데이터 벡터 수는 23개 였다.

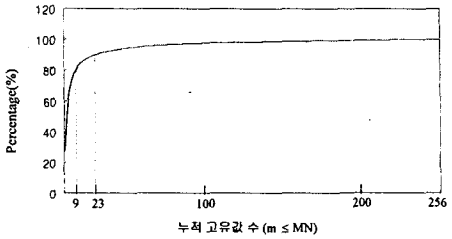

그림 4. 첫 m개의 고유값에 대한 누적 백분율

III. HMM 기반 인식 실험

3.1. HMM 규정

본 논문에서는 임의의 시간 t 에서 화자의 입술 영역을 담고있는 영상에 대해 선형 변환을 거쳐 최종적으로 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)를 모든 o_{jt} 마다 덧붙여 사용하였다.

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

여기서, k_1 , k_2 는 가중치로서 각각 2, 4이고, t 는 프레임 수, j 는 특징 벡터 o_{jt} 의 원소 수, R_j 는 j 번째 특징 벡터이고, D_{jt} 는 임의의 프레임 t 에서의 j 번째 특징 벡터에 해당하는 델타 파라미터의 값이다.

3.2. 실험 및 결과

실험에 사용된 영상 데이터는 20대 남자 52 명이 22개의 단어를 평상시 발음으로 발음한 영상을 디지털 카메라를 사용하여 30 frame/sec의 속도로 저장한 것이다. 22개의 단어는 정보서비스를 제공해 줄 때의 매뉴얼 단어를 선택하였다.

HMM의 학습화 과정에 52명의 영상 데이터가 사용되고 실제 인식률을 시험하고자 하는 데이터는 52명 중 임의의 중복된 18명이 같은 22개 단어에 대해 다시 발음한 영상 데이터를 사용하여 표 2에서와 같이 상태수와 가지 수를 변화시키면서 실험을 수행했다.

수행하였다. 그림 3은 각각의 결과를 영상으로 보인 것이다[4],[5].

2.3. 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.

본 논문에서는 X_{it} 의 원소들 간의 상관구조를 나타내는 공분산 Σ 에 기반한 PCA를 고려함으로서 일반성을 유지하였다.

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

여기서 E는 p 개의 고유벡터(eigenvector) e_i 들을 열로 하는 크기 (pxp)인 직교행렬이고 Δ 는 Σ 의 고유값(eigenvalue) δ_i 를 대각원소로 하는 크기 (pxp)인 대각행렬이다. 이는 다음 식 (5)로 표현될 수 있다.

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

이때 고유값과 각각의 고유값에 대응되는 고유벡터 e_i 의 짝을 δ_i 의 크기순으로 배열하고 가장 큰 고유값 m 개에 해당하는 고유벡터의 짝 E_m 을 이용하여 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}, $j = 1, 2, \cdots, m$, $(m \leq p)$를 X_{it} 의 주성분으로 추출할 수 있다. 이때 고유값들의 합을 $tr(\Delta) = (\delta_1 + \delta_2 + \cdots + \delta_p)$ 이라 하면 m 개의 주성분에 의해 설명되는 부분은 $(\delta_1 + \delta_2 + \cdots + \delta_m)/tr(\Delta)$ 이 될 것이다.

그림 4에서는 DCT 변환을 수행한 결과에 대해 실험적인 결과를 산출해 낸 것으로, M=N=16 일 때 m 개의 고유값들에 대한 누적 백분율을 보이고 있다. 여기에서 PCA를 거친 후 원래 정보의 80%를 갖는 데이터 벡터 수(m)는 9개 였고, 90%를 갖는 데이터 벡터 수는 23개 였다.

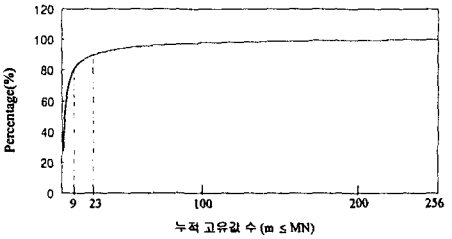

그림 4. 첫 m개의 고유값에 대한 누적 백분율

III. HMM 기반 인식 실험

3.1. HMM 규정

본 논문에서는 임의의 시간 t 에서 화자의 입술 영역을 담고있는 영상에 대해 선형 변환을 거쳐 최종적으로 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)를 모든 o_{jt} 마다 덧붙여 사용하였다.

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

여기서, k_1 , k_2 는 가중치로서 각각 2, 4이고, t 는 프레임 수, j 는 특징 벡터 o_{jt} 의 원소 수, R_j 는 j 번째 특징 벡터이고, D_{jt} 는 임의의 프레임 t 에서의 j 번째 특징 벡터에 해당하는 델타 파라미터의 값이다.

3.2. 실험 및 결과

실험에 사용된 영상 데이터는 20대 남자 52 명이 22개의 단어를 평상시 발음으로 발음한 영상을 디지털 카메라를 사용하여 30 frame/sec의 속도로 저장한 것이다. 22개의 단어는 정보서비스를 제공해 줄 때의 매뉴얼 단어를 선택하였다.

HMM의 학습화 과정에 52명의 영상 데이터가 사용되고 실제 인식률을 시험하고자 하는 데이터는 52명 중 임의의 중복된 18명이 같은 22개 단어에 대해 다시 발음한 영상 데이터를 사용하여 표 2에서와 같이 상태수와 가지 수를 변화시키면서 실험을 수행했다.

표 2. 영상 변환 기반 HMM 인식 결과 (화자 52 명 training/중복된 18 명 testing , PCA 80%) 단위: %

DCT	S 3	S 4	S 5	S 6
M 3	38.64	35.35	40.66	45.96
M 4	41.16	39.90	44.44	42.93
M 5	39.14	41.41	44.44	43.43
M 6	39.39	39.39	42.17	45.20

DWT	S 3	S 4	S 5	S 6
M 3	36.87	39.90	44.70	45.20
M 4	40.40	38.38	45.20	45.20
M 5	37.12	40.66	42.68	43.18
M 6	36.36	40.66	44.95	46.97

이에 대해 입술 윤곽선의 특정 정보만을 이용하는 방식으로 HMM 인식실험을 거친 결과가 표 3에 있다.

영상의 선형 변환 알고리즘으로 사용한 DCT와 DWT의 결과는 HMM 인식률에서 별다른 차이를 보이지 않았다.

표 3. 입술 윤곽선 기반 HMM 인식 결과 (화자 52 명 training / 중복된 18 명 testing) 단위: %

	S 3	S 4	S 5	S 6
M 3	20.45	25.75	26.51	27.02
M 4	20.96	27.77	26.26	28.78
M 5	24.74	27.77	29.29	31.06
M 6	27.52	29.04	31.06	31.56

표 4. 영상 변환 기반 HMM 인식 결과 (화자 52 명 training/중복된 18 명 testing , PCA 90%) 단위: %

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

DWT	S 3	S 4	S 5	S 6
M 3	43.18	44.95	49.24	48.99
M 4	39.90	47.47	49.49	51.52
M 5	44.95	48.74	48.99	50.25
M 6	42.68	44.95	49.49	52.02

VI. 결 론

본 논문에서는 립리딩을 위한 특징 추출 방식에 있어서 화자의 입술 ROI에 대한 영상 선형 변환 알고리즘을 사용하였다. 기존의 입술 윤곽선 기반 방식은 입술 윤곽선 추출 오류가 발생할 경우 인식률에 영향을 주는 반면, 영상 변환 기반 방식은 보다 안정적인 알고리즘으로 평가될 수 있고 HMM 인식률 면에서도 다소 우수한 성능을 보이고 있다.

이 실험에서는 22개의 한정된 단어를 가지고 분석하였는데, 보다 광범위한 단어에 적용하여 볼 필요가 있으며 음성정보와의 실시간 인식처리 시스템 구현을 위해 보다 적은 데이터 양으로 보다 빠른 처리가 가능한 알고리즘에 대한 연구가 진행되어야 할 것으로 본다.

참고문헌

[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] Rafael C. Gonzalez, Richard E. Woods, "Digital Image Processing " , Addison - Wesley Publishing Company.

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

[4] Mulcahy, Colm Ph.D, "Image compression using the Haar wavelet transform" , Spelman Science and Math Journal, 22-31.

[5] Martin Vetterli, Jelena Kovacevic, Wavelets and Subband Coding. Prentice Hall. Englewood Cliffs, NJ 07632.

[6] Gerasimos Potamianos, Hans Peter Graf and Eric Cosatto, "An Image Transform Approach for HMM Based Automatic Lipreading" 1998 IEEE.

[7] 박병구, 김진영, 최승호, "잡음 환경 하에서의 바이모달 음성인식", '98 한국음향학회 학술발표대회 논문집. pp111-114, 1998년 7월.

[8] 박병구, 김진영, 임재열, "입술 파라미터 선정에 따른 바이모달 음성인식 성능 비교 및 검증", 한국음향학회지 제 18 권, 제 3 호, pp68-72, 1999년 4월.

[9] 박병구, 김진영, 최승호, "바이모달 음성인식의 음성정보와 입술정보 결합방법 비교", 한국음향학회지 제 18 권 4호, 0031-37, 1999년 6월.

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