한국어 입술 독해에 적합한 시공간적 특징 추출

오현화, 김인철, 김동수, 진성일
경북대학교 전자전기컴퓨터학부

Experiments on Various Spatial-Temporal Features for Korean Lipreading

Hyun-Hwa Oh, In-Cheol Kim, Dong-Su Kim, Sung-Il Chien
School of Electrical Engineering and Computer Sciences,
Kyungpook National University
ohh@palgong.knu.ac.kr, kiminc@palgong.knu.ac.kr,
dskim@palgong.knu.ac.kr, sichien@ee.knu.ac.kr

Abstract

Visual speech information improves the performance of speech recognition, especially in noisy environment. We have tested the various spatial-temporal features for the Korean lipreading and evaluated the performance by using a hidden Markov model based classifier. The results have shown that the direction as well as the magnitude of the movement of the lip contour over time is useful features for the lipreading.

I. 서론

음성의 음향 정보만을 이용한 일반적인 음성인식 시스템은 잡음과 음성간의 간섭이 많이 존재하는 실제 환경 하에서 성능이 현저히 저하되어 그 적용에 어려움을 나타내고 있다. 최근 들어 이러한 문제를 해결하기 위하여 입술의 모양과 움직임 등의 영상 정보를 이용한 입술 독해 (lipreading) 기술을 음성 인식 기술과 결합하여 인식 성능을 향상시키고자 하는 연구가 진행되고 있다 [1][2][3].

입술 독해에 대한 기존의 연구에는 입술 영상의 명암 정보로부터 필터링(filtering)과 차원 감소(dimension reduction) 등의 전처리 과정을 거쳐 입술 영역의 특징 벡터를 추출하는 방법 [4]과 입술 윤곽선 상의 형태에 기반한 입술 모델을 구성하고 파라메터를 추정하여 이용하는 방법 등이 있다 [5][6]. 그리고 입술 독해를 위하여 time delayed neural network(TDNN) [7], neural network(NN) [2], hidden Markov model(HMM) [6] 등의 인식기를 이용하고 있다. 이와 같이 입술 독해를 위한 연구가 진행되고 있으나 한국어 입술 독해에 대한 연구는 현재 미진한 상태이다. 그러므로 본 논문에서는 한국어에 적합한 입술 독해 시스템의 구현을 위한 기초 연구로서 입술 움직임의 특징 추출과 성능 검증에 관한 연구를 수행한다. 동영상 데이터로부터 입술 윤곽선 상의 특징 점을 검출하고 한국어 입술 독해에 적합한 다양한 시공간적 특징들을 추출한다. 그리고 추출된 특징 벡터들의 성능을 검증하기 위하여 HMM에 기반한 인식기를 이용한다. 실험을 통하여 입술 윤곽선의 시공간적 움직임 정도와 방향이 반영된 특징 벡터가 입술 독해를 위하여 유용함을 확인할 수 있다.

II. 한국어 모음에 대한 입술 독해

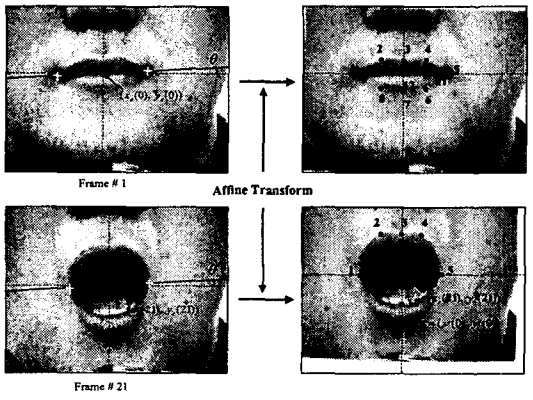
본 논문에서는 한국어 모음에 대한 입술 모양과 움직임의 시공간적 특징들을 연구하기 위하여 동영상 데이터베이스의 입술 영상에서 입술 윤곽선 상의 특징 점들을 수동으로 검출한다. 특징 점들의 수동 검출은 특징 점의 자동 검출을 위한 영상 처리 과정에서 조명 등의 영향으로 인하여 발생할 수 있는 특징 점 좌표의 오 검출을 배제하여 보다 정확하게 추출된 입술 특징을 분석하고 그 성능을 평가하기 위한 것이다.

화자가 발음을 하는 동안 얼굴의 움직임으로 인하여 동영상의 각 프레임마다 입술의 위치(translation)와 회전(rotation) 변화 정도가 다르게 나타난다. 본 논문에서는 이와 같은 위치와 회전 변화를 보정하고 동영상 데이터의 모든 프레임(frame)에 대하여 동일한 방식으

로 특징 벡터를 추출하기 위하여 식 (1)과 같은 유사 변환(affine transform)을 수행한다.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x - x_c \\ y - y_c \end{bmatrix} + \begin{bmatrix} T_x \\ T_y \end{bmatrix} \quad (1)$$

여기서 (x_c, y_c)는 첫 번째 프레임의 입술 중심 좌표이며 (T_x, T_y)는 첫 번째 프레임의 입술 중심에 대한 현재 프레임의 입술 중심의 위치 변화를 나타낸다. 그림 1은 식 (1)에 의하여 위치와 회전 변화가 보정된 영상이다. 이와 같이 위치와 회전이 보정된 영상에 대하여 그림 1(b)와 같이 입술 끝점으로부터 계산된 입술 폭을 등간격으로 분배하여 입술의 내부와 외부 윤곽선(contour) 상의 특징 점들을 수동으로 검출한다.



(a) (b)

그림 1. 입술의 위치와 회전 변화의 보정을 위한 유사 변환과 수동으로 검출된 입술의 특징 점.

입술 윤곽선 상에서 검출된 특징 점들로부터 동영상에서의 시공간적인 입술 모양과 움직임에 대한 특징들을 다음과 같이 추출한다.

- $\mathbf{f}_a$: 입술의 내/외부의 높이(height)와 폭(width)
- $\mathbf{f}_b$: 검출된 입술 특징 점들의 시간에 따른 궤적(trajectory)
- $\mathbf{f}_c$: 검출된 입술 특징 점들의 시간에 따른 움직임의 크기(magnitude)
- $\mathbf{f}_d$: 검출된 입술 특징 점들의 시간에 대한 움직임의 속도(velocity)

$\mathbf{f}_a$는 입술 내외부의 높이와 폭, 외부의 좌우 높이로 구성된 6차원의 벡터들의 집합이며 식 (2)와 같이 계산된다.

$$\mathbf{f}_a = [\ \mathbf{f}_{a1},\ \mathbf{f}_{a2}, \ldots,\ \mathbf{f}_{aN_k}\ ]^T \quad (2)$$

$$\mathbf{f}_{ai} = [h_O(i), w_O(i), h_{OL}(i), h_{OR}(i), h_I(i), w_I(i)],$$

$$h_O(i) = H_O(i)/H_O(1) \qquad w_O(i) = W_O(i)/W_O(1)$$

$$h_{OL}(i) = H_{OL}(i)/H_{OL}(1) \qquad h_{OR}(i) = H_{OR}(i)/H_{OR}(1)$$

$$h_I(i) = H_I(i)/H_I(1) \qquad w_I(i) = W_I(i)/W_I(1)$$

$$i = 1,2,\ldots,N_k$$

여기서 $\mathbf{f}_{ai}$는 i 번째 프레임에서 산출된 성분들로 구성된 벡터이며 N_k는 k번째 동영상을 구성하는 프레임의 총 개수이다. 일반적으로 화자마다 입술의 크기가 다르며 동일한 모음에 대해서도 입술의 벌림 정도에 차이가 있다. 그러므로 식 (2)와 같이 첫 번째 프레임에서 산출된 값으로 정규화하여 i 번째 프레임의 특징 벡터를 구성한다. 본 논문에서는 화자가 입술을 닫은 상태에서 발음을 시작하도록 하였으므로 데이터베이스의 첫 번째 프레임에서 $H_I(1)$과 $W_I(1)$는 정의되지 않는다. 그러므로 $h_I(i)$와 $w_I(i)$는 $H_O(1)$와 폭 $W_O(1)$을 이용하여 정규화한다. 그림 2는 입술이 닫힌 프레임과 열린 프레임에 대하여 추출된 특징 벡터 $\mathbf{f}_{a1}$과 $\mathbf{f}_{a21}$의 성분들을 나타낸다.

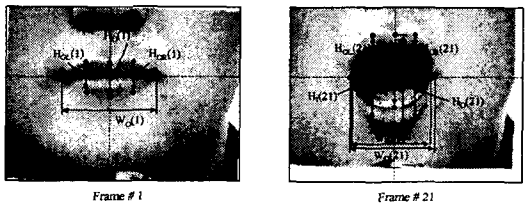


그림 2. 모음 /ㅏ/에 대해서 입술이 닫힌 경우와 열린 경우의 특징 벡터 $\mathbf{f}_{a1}$과 $\mathbf{f}_{a21}$.

화자가 발음을 하는 동안 모음에 따라서 위 입술과 아래 입술의 벌림 정도가 서로 다르다. 그리고 각 모음마다 입술의 내/외부 윤곽선 상에서 검출된 특징 점들의 움직임 방향이 서로 다르게 나타난다. 그러므로 본 논문에서는 아래와 위 입술의 벌림 정도의 차이와 특징 점들의 시공간적인 움직임을 충분히 반영할 수 있는 특징 벡터 $\mathbf{f}_b$를 다음과 같이 추출한다.

$$\mathbf{f}_b = [\ \mathbf{f}_{b1},\ \mathbf{f}_{b2}, \ldots,\ \mathbf{f}_{bN_k}\ ]^T \quad (3)$$

$$\mathbf{f}_{bi} = [u_1(i), v_1(i), \ldots, u_{12}(i), v_{12}(i)],$$

$(u_I(i), v_I(i))$는 첫 번째 프레임 영상에서 검출된 j 번

째 특징 점의 좌표를 기준으로 하여 i 번째 프레임에서 검출된 j 번째 특징 점의 상대좌표를 나타내며 화자의 입술 크기의 차이를 고려하여 다음과 같이 정규화하여 계산된다.

$$u_j(i) = (x_j(i) - x_j(1)) / W_O(1)$$
$$v_j(i) = (y_j(i) - y_j(1)) / W_O(1) \quad (4)$$

여기서 $i = 1,2,\ldots,N_k$ 이다. 그림 3은 첫 번째 프레임에서의 검출된 특징 점들의 좌표를 각각의 원점으로 한 상대 좌표($u_j(i), v_j(i)$), 즉 시간에 따른 특징 점들의 궤적(trajectory)을 그래프로 표현한 것이다.

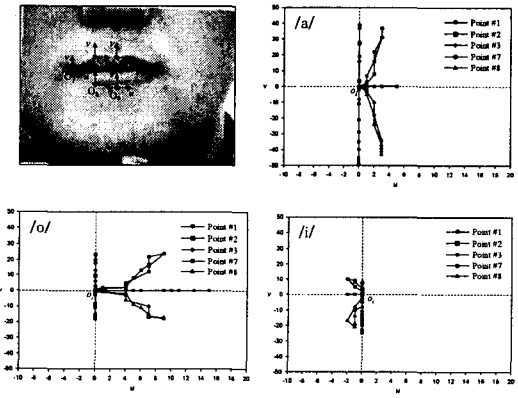


그림 3. 모음 /ㅏ/, /ㅗ/, /ㅣ/에 특징 벡터 $\mathbf{f}_b$.

그림 3에서 보는 바와 같이 특징 벡터 $\mathbf{f}_b$에서는 각 모음에 대하여 입술 움직임의 고유한 동적 특징이 뚜렷이 나타나고 있음을 확인할 수 있다. 즉, 위 입술과 아래 입술의 벌림 정도, 입술 폭의 벌림 정도, 그리고 특징 점들의 움직임 방향 등이 모음마다 다르게 나타나고 있음을 확인할 수 있다. 나머지 모음들에 대해서도 각각 특징 점들의 고유한 궤적을 관찰할 수 있다.

$\mathbf{f}_c$는 특징 점들의 시간에 따른 움직임 방향은 제외하고 크기만을 고려한 특징 벡터로서 다음과 같이 추출하여 구성한다.

$$\mathbf{f}_c = [\ \mathbf{f}_{c1},\ \mathbf{f}_{c2},\ldots,\ \mathbf{f}_{cN_k}\ ]^T$$
$$\mathbf{f}_{ci} = [M_1(i), M_2(i), \ldots, M_{12}(i)] \quad (5)$$

여기서 i 번째 프레임의 j 번째 특징 점의 움직임 크기 $M_j(i)$는 $\mathbf{f}_b$에서 산출된 $u_j(i)$과 $v_j(i)$로부터 다음과 같이 산출된다.

$$M_j(i) = \sqrt{u_j(i)^2 + v_j(i)^2} \quad (6)$$

그리고 입술 독해에 있어서 입술의 움직임 속도가 기여하는 정도를 평가하기 위하여 각 특징 점의 움직임 속도를 산출하여 특징 벡터 $\mathbf{f}_d$ 를 구성한다.

Ⅲ. 실험 결과 및 고찰

본 논문에서는 추출된 다양한 시공간적 특징 벡터 $\mathbf{f}_a$, $\mathbf{f}_b$, $\mathbf{f}_c$, $\mathbf{f}_d$ 의 성능을 검증하기 위하여 이산 HMM 기반의 입술 독해 인식기를 구현하였다. 그림 4는 본 논문에서 사용된 표준 left-to-right HMM을 나타낸다. HMM의 파라메터 추정을 위하여 Baum-Welch 알고리듬을 사용하였으며, 학습된 HMM을 이용한 인식 수행을 위하여 Viterbi 알고리듬을 이용하였다.

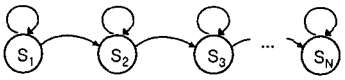


그림 4. N개의 상태를 가지는 left-to-right HMM.

본 논문에서는 15명의 화자로부터 /ㅏ/, /ㅐ/, /ㅓ/, /ㅗ/, /ㅜ/, /ㅡ/, /ㅣ/의 7개 모음을 10번씩 발음하여 총 875개의 동영상 데이터를 획득하였다. 실험 데이터의 525개는 학습용 데이터로 사용하였으며 나머지 350개는 테스트용 데이터로 사용하였다. 표 1 특징 벡터 $\mathbf{f}_a$, $\mathbf{f}_b$, $\mathbf{f}_c$, $\mathbf{f}_d$ 를 이용하여 인식 실험을 수행한 결과로서 인식 결과가 top 1인 경우와 top 2내에 포함되는 경우의 인식률을 나타낸다.

표 1. 특징 벡터를 이용한 인식 실험 결과

특징 벡터		$\mathbf{f}_a$	$\mathbf{f}_b$	$\mathbf{f}_c$	$\mathbf{f}_d$
인식률 (%)	top1	47.43	62.86	56.86	36.86
	top1, top2	75.71	84.29	80.86	58.00

표 1의 인식률은 다수의 화자에 대하여 입술 독해 실험을 수행한 결과이다. 기존의 입술 독해 연구에서는 대부분 한 화자에 대한 입술 독해 실험을 수행하였으며, 다수의 화자에 대한 입술 독해 연구에서는 인식률이 40~60% 정도로 나타나고 있다 [2][4]. 입술 독해에 대한 연구가 음성 인식 등의 분야와 같이 활발히 진행되지 않은 현재, 입술 독해 연구를 위한 공인된 데이

터베이스가 전무하므로 입술 독해 인식기의 성능을 비교하는 데에는 어려움이 있다. 그럼에도 불구하고 표 1의 결과는 기존의 연구에서 나타난 인식률과 비교하여 우수하다고 판단된다.

표 1의 결과에서 인식 결과가 top 1인 경우 특징 벡터 $\mathbf{f}_b$에 대한 인식률이 $\mathbf{f}_a$, $\mathbf{f}_c$, $\mathbf{f}_d$와 비교하여 높게 나타남을 확인할 수 있다. 특징 벡터 $\mathbf{f}_a$는 모음을 발음하는 동안 입술의 전반적인 벌림 정도를 나타내므로 {/ㅏ/, /ㅐ/, /ㅓ/}와 {/ㅡ/, /ㅣ/}와 같이 입술의 높이와 폭이 서로 비슷한 모음들에 의한 혼동이 인식률에 크게 영향을 미친 것으로 사료된다. 특징 점의 움직임 크기만이 반영된 특징 벡터 $\mathbf{f}_c$의 경우는 56.86%로 $\mathbf{f}_b$의 경우보다 다소 낮은 인식률을 보인다. 이 결과로부터 특징 점이 시간에 따라 이동한 크기뿐만 아니라 방향이 입술 독해에 있어서 중요한 요소임을 알 수 있다. 특징 벡터 $\mathbf{f}_d$는 특징 점이 시간에 따라 이동한 속도, 즉 발음 속도를 나타낸다. 발음 속도는 화자마다 차이가 있으며 동일한 화자의 경우에 있어서도 화자의 심리 상태에 따라서 차이가 있다. 그러므로 실험 결과로부터 발음 속도는 입술 독해에 있어서 유용한 특징이 아님을 확인할 수 있다. 본 논문에서 가장 우수한 성능을 보인 특징 벡터 $\mathbf{f}_b$의 경우 62.86%의 입술 독해 인식률을 나타낸다. 이러한 결과로부터 시간에 대한 입술 특징 점의 상대적 좌표인 $\mathbf{f}_b$에 입술 움직임의 고유한 동적 특징이 잘 반영되었음을 알 수 있으며 입술 독해를 위하여 매우 유용한 특징임을 확인할 수 있다. 또한 인식 결과가 top 2 이상이 되는 조건을 만족하는 경우의 인식률은 84.29%로 매우 높게 나타나고 있다. 이러한 결과로부터 HMM인식기의 top 1과 top 2 출력 값으로부터 신뢰도(reliability)를 검증하여 보다 우수한 성능의 입술 독해 인식기를 구현할 수 있을 것으로 사료된다.

Ⅳ. 결론

본 논문에서는 한국어에 적합한 입술 독해를 위한 연구의 기초 단계로서 입술 모양에 대한 다양한 시공간적 특징 벡터들을 추출하고 그 성능을 검증하였다. 다수의 화자로부터 획득된 동영상 데이터를 이용하여 HMM에 기반한 인식 실험을 수행한 결과에서 기존의 연구 결과와 비교하여 우수한 성능의 인식률을 획득하였다. 특히 시간에 대한 입술 특징 점의 상대적 좌표를 구성성분으로 하는 특징 벡터는 한국어 모음에 대한 입술 움직임의 고유한 동적 특징을 잘 반영하므로 우수한 성능을 나타내었다.

향후 연구에서는 본 논문에서 연구된 특징 벡터를 자동으로 추출하여 자동 입술 독해 시스템을 구현하고자 한다. 자동 입술 독해 시스템은 음성인식 시스템과 결합되어 잡음이 많이 존재하는 실제 환경 하에서 음성 정보만을 이용한 기존의 음성인식 시스템의 성능 저하를 개선할 수 있을 것으로 기대된다. 또한 화상회의, 은행 ATM 시스템, 역무 자동화 시스템, 그리고 자동 다이얼링 시스템 등의 응용분야에서 음성 인식 시스템과 결합되어 유용하게 이용될 것으로 기대된다.

※ 본 연구는 한국과학재단 목적기초연구(1999-2-303-001-3) 지원으로 수행되었음.

Reference

[1] E. D. Petajan, "Automatic Lipreading to Enhance Speech Recognition," *ICPR-85*, pp. 40-47, 1985.

[2] G. J. Wolff, K. V. Prasad, D. G. Stork, and M. Hennecke, "Lipreading by Neural Networks: Visual Preprocessing, Learning and Sensory Integration," *NIPS-94*, pp. 1027-1034, 1994.

[3] P. L. Silsbee and A. C. Bovik, "Computer Lipreading for Improved Accuracy in Automatic Speech Recognition," *IEEE trans. on Speech and Audio Processing*, vol. 4, no. 5, pp. 337-351, Sep. 1996.

[4] G. Potamianos, H.P. Graf, and E. Cosatto, "An Image Transform Approach for HMM Based Automatic Lipreading," *ICIP-98*, vol. 3, pp. 173-177, Chicago, 1998.

[5] M.E. Hennecke, K.V. Prasad, and D.G. Stork, "Using Deformable Template to Infer Visual Speech Dynamics," *Proc. of 28th Asilomar Conf. on Signals, Systems and Computers*, vol. 1, pp. 578-582, 1994.

[6] J. Luettin, N.A, Thacker, and S.W. Beet, "Visual Speech Recognition Using Active Shape Models and Hidden Markov Models," *ICASSP-96*, vol. 2, pp. 817-820, 1996.

[7] P. Duchnowski, M. Hunke, D. Busching, U. Meier, and A. Waibel, "Toward Movement-invariant Automatic Lip-reading and Speech Recognition," *ICASSP-95*, vol. 1, pp. 109-112, 1995.