음성인식에서 입술 파라미터 열화에 따른 견인성 연구

Robustness of Bimodal Speech Recognition on Degradation of Lip Parameter Estimation Performance

김 진 영* · 민 소 희* · 최 승 호**
Jinyoung Kim · Sohee Min · Seungho Choi

ABSTRACT

Bimodal speech recognition based on lip reading has been studied as a representative method of speech recognition under noisy environments. There are three integration methods of speech and lip modalities as like direct identification, separate identification and dominant recording.

In this paper we evaluate the robustness of lip reading methods under the assumption that lip parameters are estimated with errors. We show that the dominant recording approach is more robust than other methods through lip reading experiments.

Keywords: Bimodal Speech, Lip Reading, Dominant Recording Degradation

1. 서 론

최근 심한 잡음 환경에서 음성인식의 성능을 향상시키기 위한 연구가 활발히 진행되고 있다. 이는 음성인식이 실험실과 같이 잡음을 거의 배제한 환경에서는 뛰어난 인식성능을 보이고 있으나 소음이 많이 발생하는 자동차 내부, 사무실, 거리 같은 실생활에 적용할 때는 성능이 크게 저하되기 때문이다.

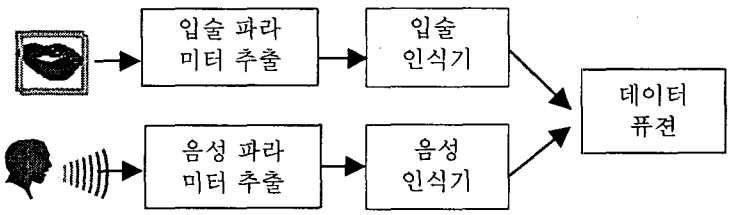
립리딩(lip-reading)은 음성인식 분야 중 잡음 환경에서 현저하게 떨어지는 인식률을 높이기 위한 보상 방법의 하나로서, 화자의 입술을 포함한 영상 정보는 발성의 조음현상을 반영하고 있기 때문에, 오염된 음성 파라미터를 보완하는 정보로서 이용되고 있다[1~4]. 그런데, 립리딩을 이용하는 시청각음성인식에서 입술정보와 음성정보를 어떻게 혼합할 것인가가 주요한 문제인데 지금까지 알려진 시청각음성인식 통합방법으로는 direct identification model (DI), separate identification model (SI) 그리고 dominant recording model (DR)이 있다[5].

* 전남대학교 전자컴퓨터정보통신공학부
** 동신대학교 멀티미디어통신공학과

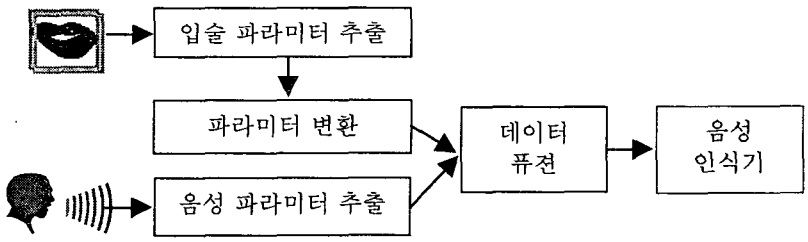
한편, 음성파라미터가 잡음에 의해 쉽게 열화되기 때문에 시청각 음성인식에서 입술정보를 사용하는데, 시각 정보인 입술 정보도 영상에 가장 영향을 미치는 조명과 같은 다양한 조건 하에서 상당한 추정오차를 갖게 되며, 실제 사용환경에서는 이러한 오차를 피할 수 없다. 따라서, 오차를 갖는 입술 파라미터를 립리딩에서 사용하게 되면, 바이모달 음성인식은 장점을 지니지 못하게 된다. 그러므로, 시각정보를 입술인식에 성공적으로 활용하기 위해서는 입술정보를 이용한 인식이 추정오차에 대하여 얼마나 강인한지, 어떠한 통합방법이 강인성을 갖는지에 대한 연구가 필요하다. 본 논문에서는 입술 파라미터의 추정 시에 오차를 가정하여 시청각음성인식 통합방법의 견인성에 대하여 실험하였다.

2. 시청각정보의 통합방법

본 논문에서 검토한 시청각 통합방법은 위의 서론에서 언급한 바와 같이 DI, SI 그리고 DR 방법으로 나뉜다. DI통합방법은 음성과 입술 파라미터를 단순 통합하여 인식한 것이므로 본 논문에서는, SI 및 DR 방법의 견인성을 비교 검토하였다. 물론, 잡음환경 하에서 시청각 음성인식의 인식률의 베이스라인(baseline)은 시각정보만을 사용하는 경우이므로, 순수히 입술 정보만을 사용한 인식률을 검토하면 될 것이다. 다음의 그림 1은 separate identification과 dominant recording 통합방법을 보여주고 있다.



(a) SI 통합방법



(b) DR통합방법

그림 1. 시청각 음성인식 통합 방법

그림에 보인 바와 같이 SI 방법에서는 각각의 파라미터에 대하여 구별된 음성인식과 입술인식기를 대상으로 인식 확률을 구한 후 가중합 (weighted sum) 방법[6]에 의하여 통합을 하며, DR 통합에서는 입술 파라미터를 음성 파라미터로 변환 후에 두 파라미터를 통합한 후 음성 인식기에 입력으로 사용하여 인식확률을 구하게 된다. 그러므로, DR 통합에서는 입술 파라미터를 음성 파라미터로 변환을 시켜야 하는데, 본 연구에서는 이를 위하여 선형회귀(linear regression)을 이용하였다. 선형회귀 방정식은 아래의 (1)식과 같다.

$$\begin{bmatrix} \hat{x}_{a1} \\ \hat{x}_{a2} \\ \vdots \\ \hat{x}_{ap} \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1(q+1)} \\ t_{21} & t_{22} & \cdots & t_{2(q+1)} \\ \vdots & \vdots & \cdots & \vdots \\ t_{p1} & t_{p2} & \cdots & t_{p(q+1)} \end{bmatrix} \begin{bmatrix} 1 \\ x_{v1} \\ \vdots \\ x_{vq} \end{bmatrix} \tag{1}$$

$$\hat{x}_a = T x_v$$

위 식에서 $\hat{x}_a$는 입술정보로부터 추정된 음성 특징 파라미터이며, x_v는 입술 특징 파라미터이다. 즉, 입술 파라미터는 음성파라미터로 선형변환을 사용하여 변환된다. 파라미터의 추정은 최소자승법(least squares method)에 의하여 쉽게 구할 수 있다.

3. 시청각음성 데이터베이스

본 연구는 통합방법에 대한 검증을 위하여 간단하게 한국어 단모음을 대상으로 하였다. 또한, 입술영상으로부터 입술에 관한 정보를 획득하기 쉽도록 하기 위하여 입술과 주요 부분에 마커(marker)를 부착하고 촬영하였다. 또한 입술 파라미터의 깊이방향(z-축방향)의 정보를 얻기 위하여 거울을 이용하여 카메라의 영상에 얼굴의 옆 모습이 투영되도록 하였다. 그림 2는 본 연구에서 촬영된 얼굴영상 중 한 프레임을 그린 것이다. 구축된 시청각 음성 DB는 한국어 단모음 '아', '이', '우', '에', '오'를 두 명의 화자가 30 번씩 발음한 것이다.

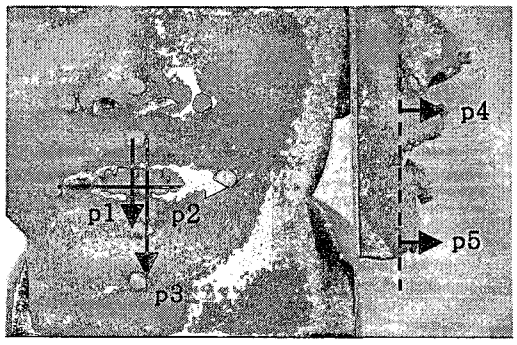


그림 2. 마커를 부착한 얼굴이미지 및 입술 특징 파라미터

마커의 위치는 Famous Tracker 툴을 사용하여 추적하였다. 추적된 마커의 위치정보로부터 입술 특징 파라미터를 결정하였는데, 그림 2에 보인 바와 같이 입술의 폭(p2)과 입술의 높이(p1) 그리고, 윗입술에서 턱까지의 거리(p3), 코로부터 입술 위까지의 z 방향 거리(p4), 코로부터 턱까지의 z 방향 거리(p5) 등이다.

4. 립리딩의 견인성 평가 실험

본 장에서는 위의 2장에서 설명한 DR 방법과 SI 방법의 견인성을 검토한다. 즉, 입술 파라미터를 음성 파라미터로 변환하여 인식하는 립리딩과, 입술 파라미터를 그대로 이용하여 인식하는 립리딩의 견인성에 대하여 실험결과를 설명한다. 본 연구에서는 인식의 방법으로 GMM (Gaussian mixture model) 기반의 HMM (hidden Markov model)을 사용하였다.

4.1 입술 파라미터의 추정오차 구현

본 연구에서는 입술 부위에 마커를 부착하고, 부착된 마커로부터 FamousTracker라는 툴을 사용하여 정확하게 입술 파라미터를 추출하였으므로, 추정된 입술 파라미터에는 추정 오차가 없다고 가정하여도 무방하다. 한편, 현재 마커가 붙지 않은 상태에서 입술 파라미터를 추정하는 자동화된 프로그램이 존재하지 않기 때문에, 다양한 환경에서 오류가 있은 실제 사황의 입술 파라미터를 얻을 수가 없다. 따라서, 본 연구에서는 정확하게 구해진 입술 파라미터에 인위적으로 오차를 더하여, 추정 오차가 첨가된 입술 파라미터를 얻었다. 즉 오염된 입술 파라미터는 다음의 식과 같이 표현된다.

$$\overline{x_v} = x_v + n_v \tag{2}$$

위 식에서 첨가된 추정오차는 불규칙 변수(random variable)로서 입술 파라미터와는 전혀 상관도가 없은 가우시안 불규칙 변수이다. 물론, 입술 라미터의 오염도에 따라, 가우시안 분포의 표준편차는 변화된 게 된다.

4.2 시청각 통합방법의 입술 파라미터의 열화정도에 따른 견인성

본 절에서는 4.1절에 기반한 오염된 입술 파라미터로부터 립리딩 수행결과를 설명한다. 그런데, DR 통합방법은 음성 파라미터로 변환하여 인식하므로, 어떠한 음성 파라미터를 사용할 것인가가 중요한 문제이다. 본 연구에서는 참고문헌 [6]의 실험결과를 배경으로 하여, 켑스트럼 계수를 사용하였다. 왜냐하면, 참고문헌 [6]에 보인 바와 같이 켑스트럼 파라미터를 중간 파라미터로 사용하는 경우 가장 우수한 인식 성능을 보였기 때문이다.

다음의 그림 3은 입술 파라미터의 열화정도에 따라 SI통합에서의 립리딩 인식률과 DR 통합에서의 인식률을 보여주는 그림이다. 그림에서 평균 오차는 입술 파라미터의 참 크기값에 대하

여 상대인 평균치 오차의 크기를 말한다. 즉, 평균오차가 10%라는 것은 참값으로부터 +/-로 평균적으로 10%의 오차값을 갖는 다는 것을 의미한다.

그림 3에 의하면, 입술 파라미터의 추정시에 오차가 없을 때에는 입술 파라미터를 가지고 립리딩을 하는 것이 성능이 우수함을 알 수 있다. 평균오차가 0%일 때는 입술파라미터로 인식하는 SI 방법은 립리딩 성능이 약 84% 정도이지만, DR 방법의 경우에는 55%정도로 인식성능이 SI 방법에 비하여 좋지 않음을 관찰할 수 있다. 그러나, 입술 파라미터의 추정 시에 평균오차가 10%정도 된다고 할 때, DR 통합방법은 55%의 인식률을 유지하고 있어 인식성능이 저하되지 않음을 볼 수 있고 SI 방법에서는 인식률이 19%정도까지 크게 저하된다. 그러므로 그림 3의 실험결과로부터, 입술 파라미터 추정에 오차가 발생한다고 하면, SI 통합방법보다는 DR 통합방법을 사용하는 것이 타당하다는 결론을 내릴 수 있다. 즉, 실제 얻어진 입술정보로부터 입술파라미터를 구하고자 할 때에는 항상 추정오류가 존재하기 때문에 DR 방법을 사용하는 것이 바람직하다고 판단된다.

한편, 입술파라미터의 추정 오류를 고려할 때, 인식성능은 DR 방법의 경우 50~60% 정도의 인식률을 보이고 있다. 이 정도의 인식률은 당연히 입술 파라미터 만을 대상으로 인식하는 것이 의미가 없어 보일 것이다. 그러므로, 입술파라미터는 항상 음성정보와 통합되어 인식을 하게 되는 것이며, 바이모달 음성인식기의 의의는 시청각정보를 통합하여 얻어지는 시너지효과에 있는 것이다.

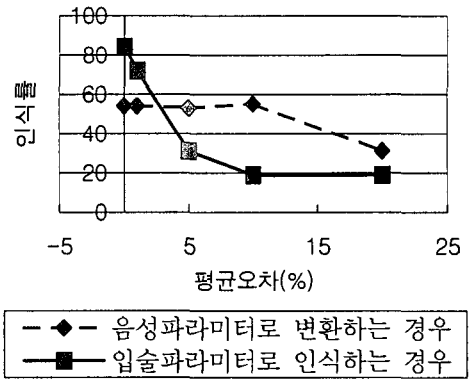


그림 3. 입술파라미터 열화에 따른 립리딩 인식성능

5. 결 론

본 논문에서는 최근 잡음환경 하에서 음성인식의 성능을 향상시키기 위해 연구되고 있는 시각정보를 이용한 바이모달 음성인식의 견인성 문제를 검토하였다. 음성의 잡음에 대한 성능저하는 널리 알려져 있는 문제이지만, 바이모달 음성인식에서 추정오차에 의한 입술 파라미터의 오염과, 이에 따른 성능저하는 아직 고려의 대상이 아니었다. 본 논문에서는 입술 파라미터의 추

정시 오차가 발생할 수밖에 없는 점을 고려하여, 잡음이 섞인 음성 파라미터를 립리딩에 사용하는 경우 립리딩의 성능을 평가하였다. 본 연구의 실험 결과 SI 통합보다는 입술 파라미터를 음성 파라미터로 변환하여 사용하는 DR (dominant recording) 통합 방법이 추정오차에 대하여 견인함을 발견하였다.

향후, 우리는 시청각 DB를 확충하여 연구결과를 더욱 객관적으로 입증할 것이다. 또한, 음성과 영상의 정보에, 잡음 및 조명의 영향으로 인하여, 추정오차가 존재하는 경우, 자동적으로 잡음의 양을 추정하고, 음성과 영상 정보의 가중치를 자동적으로 결정할 수 있는 방법에 대하여 연구하고자 한다.

참 고 문 헌

[1] Sharma, Rajeev, Vladimir I. Pavlovic & Thomas S. Huang. 1998. "Toward multi-modal human-computer interface." *Proceedings of the IEEE*, Vol. 86, No. 5.

[2] Potamianosm, Gerasimos, Hans peter Graf & Eric Cosatto. 1998. "An image transform approach for HMM based automatic lipreading." *Proceedings of the Int. Conf. on Image Processing*, 173-177.

[3] Bregler, C. & Yochai Konig. 1994. "Eigenlips for robust speech recognition." *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 669-672.

[4] Chen, T., H. P. Graf & K. Wang. 1995. "Lip-syncronization using speech-assisted video processing." *IEEE Signal Processing Lett.*, Vol 2, 57-59.

[5] Girin, L., J. Schwartz & G. Feng. 2001. "Audio-visual enhancement of speech in noise." *JASA*, Vol. 109, No. 6, 3007-3020.

[6] 김진영, 최용진, 최승호. 2002. "입술정보를 이용한 음성파라미터 예측 및 음성인식 성능향상." *한국음향학회 추계학술대회 논문집*, 65-68.

접수일자: 2003. 4. 28.
게재결정: 2003. 5. 31.

▲ 김진영
광주광역시 북구 용봉동 300 (우: 500-757)
전남대학교 전자컴퓨터정보통신공학부
Tel: +82-62-530-1757 Fax: +82-62-530-1759
E-mail: beyondi@chonnam.ac.kr

▲ 민소희
광주광역시 북구 용봉동 300 (우: 500-757)
전남대학교 전자컴퓨터정보통신공학부 IA&R LAB
Tel: +82-62-530-1750 Fax: +82-62-530-1759
E-mail: shmin3@hanmail.net

▲ 최승호
전남 나주시 대호동 252 (우: 500-757)
동신대학교 멀티미디어통신공학과
Tel: +82-61-330-3194
E-mail: shchoi@dsu.ac.kr