

# 음성인식에서 입술 파라미터 열화에 따른 견인성 연구

김진영<sup>\*</sup>, 신도성<sup>\*</sup>, 최승호<sup>\*\*</sup>  
전남대학교 전자공학과<sup>\*</sup>, 동신대학교 정보통신공학과<sup>\*\*</sup>

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

Jinyoung Kim<sup>\*</sup>, Dosung Shin<sup>\*</sup>, Seungho Choi<sup>\*\*</sup>  
Chonnam Nat'l Univ, Dongshin Univ.<sup>\*\*</sup>  
E-mail : Kimjin@dsp.chonnam.ac.kr

### 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 with lip reading experiments. Also, a measure of lip parameter degradation is proposed. This measure can be used in the determination of weighting values of video information.

### I. 서론

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 매우 저하되기 때문이다.

립리딩(lip-reading)은 음성인식 분야 중 잡음 환경

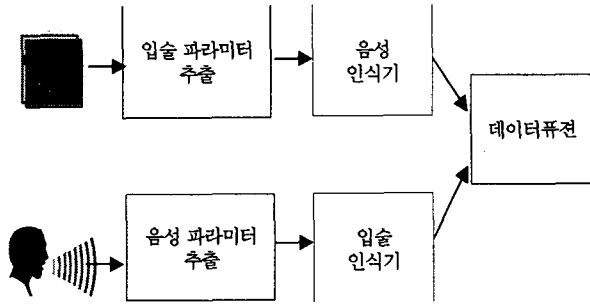
에서 현저하게 떨어지는 인식률을 높이기 위한 보상 방법의 하나로써, 화자의 입술을 포함한 영상 정보는 발성의 조음현상을 반영하고 있기 때문에, 오염된 음성 파라미터를 보완하는 정보로서 이용되고 있다[1~4]. 그런데, 립리딩을 이용하는 시청각음성인식에서 주요한 문제는 입술정보와 음성정보를 어떻게 혼합할 것인가 있다. 그런데, 지금까지 알려진 방법으로는 DI (direct identification model), SI(separate identification model) 그리고 DR(dominant recording model)이 있다.

한편, 시청각 음성인식에서 입술정보를 사용하는 이유는 음성 파라미터의 잡음에 의한 열화인데, 실제 시각 정보인 입술 정보 추출 또한 조명과 같은 다양한 조건하에서 상당한 추정오차를 갖게 된다. 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 추정시에 오차를 가정하여 시청각음성인식 통합방법의 견인성에 대하여 실험하였다.

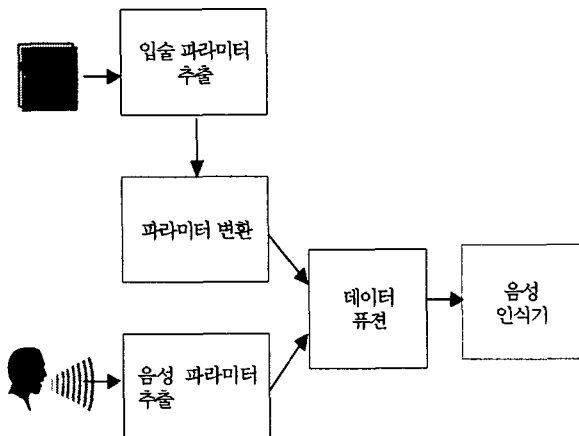
### III. 시청각정보의 통합방법

본 논문에서 검토한 시청각 통합방법은 위의 서론에서

언급한 바와 같이 DI, SI 그리고 DR 방법으로 나뉜다. 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 방법의 통합방법을 보여 준다.



(a) SI 통합방법



(b) DR 통합방법

그림 1. 시청각 음성인식 통합 방법

그림에 보인 바와 같이 SI 방법에서는 각각의 파라미터에 대하여 인식 확률을 구한 후 가중치에 의하여 통합하며, DR 통합에서는 입술 파라미터를 음성 파라미터로 변환 후에 적절한 통합방법에 의하여 두 파라미터를 통합한 후 음성 인식기에 입력으로 사용하게 된다. 그러므로, DR 통합에서는 입술 파라미터를 음성 파라미터로 변환을 시켜야 하는데, 본 연구에서는 이를 위하여 선형회귀(linear regression)을 이용하였다 [6]. 그 방법은 다음의 식 (1)과 같다.

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

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

위 식에서  $\hat{x}_a$  는 입술정보로부터 추정된 음성 특징 파라미터이며,  $x_v$  는 입술 특징 파라미터이다.

### III. 시청각음성 데이터베이스

본 연구는 통합방법에 대한 검증을 위하여 간단하게 한국어 단모음을 대상으로 하였다. 또한, 입술영상으로부터 입술에 관한 정보를 획득하기 쉽도록 하기 위하여 입술과 주요부분에 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 30번씩 발음한 것이다.

마커의 위치는 Famous Tracker 툴을 사용하여 추적하였다. 추적된 마커의 위치정보로부터 입술 특징 파라미터를 결정하였는데, 그림 2에 보인 바와 같이 입술의 폭과 입술의 높이 그리고, 윗입술에서 턱까지의 거리, 코로부터 입술 위까지의 z 방향 거리, 코로부터 턱까지의 z 방향거리 등이다.

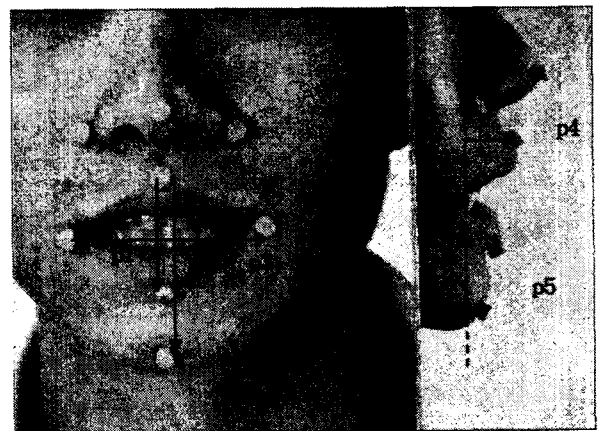


그림 2. 마커를 부착한 얼굴이미지 및 입술 특징 파라미터

#### IV. 립리당의 견인성 평가 실험

본 장에서는 위에서 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 방법으로서 상태 3의 GMM(Gaussian mixture model) 기반의 HMM(hidden Markov model)을 사용하였다.

##### 4.1 입술 파라미터의 추정오차 구현

본 연구에서는 입술 파라미터를 입술 부위에 마커를 부착하고, 부착된 마커로부터 정확하게 입술 파라미터를 추출하였으므로, 추정된 입술 파라미터에는 추정 오차가 없다고 가정하여도 무방하다. 한편, 현재 마커가 붙지 않은 상태에서 입술 파라미터를 추정하는 자동화된 프로그램이 없으므로, 다양한 환경에서, 실제 상황의, 오류가 있는 입술 파라미터를 구할 수가 없다. 따라서, 본 연구에서는 정확하게 구해진 입술 파라미터에 인공의 잡음을 더하여 오차가 있는 입술 파라미터 추출을 구현하였다. 즉 오염된 입술 파라미터는 다음의 식과 같이 표현된다.

$$\bar{x}_v = x_v + n_v \quad (2)$$

위 식에서 잡음은 불규칙 잡음으로서 입술 파라미터와는 전혀 상관도가 없다.

##### 4.2 시청각 통합방법의 입술 파라미터의 열화정도에 따른 견인성

본 절에서는 4.1절에 기반한 오염된 입술 파라미터로부터 립리딩 수행결과를 설명한다. 그런데, DR 통합 방법은 음성 파라미터로 변환하여 인식하므로, 어떠한 음성 파라미터를 사용할 것인가가 중요한 문제이다. 본 연구에서는 참고문헌 [6]의 실험결과를 배경으로 하여, 캡스트럼 계수를 사용하였다. 왜냐하면, 참고문헌 [6]에 보인 바와 같이 캡스트럼 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 대하여 상대인 평균오차의 크기

를 말한다. 즉, 평균오차가 10%라는 것은 참 값으로부터 +/-로 평균적으로 10%의 오차값을 갖는 것을 의미한다.

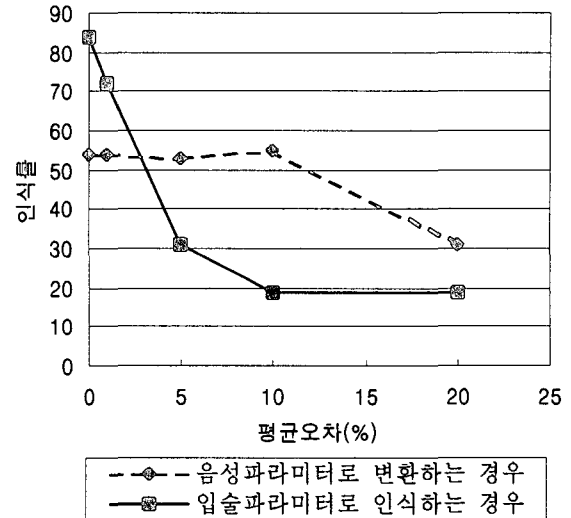


그림 3. 입술파라미터 열화에 따른 립리딩 인식성능

그림 3에 의하면, 입술 파라미터의 추정시에 오차가 없을 때에는 입술 파라미터를 가지고 립리딩을 하는 것이 성능이 우수함을 알 수 있다. 무잡음시에는 입술 파라미터로 인식하는 SI 방법은 립리딩 성능이 약 84% 정도이지만, DR 방법의 경우에는 55% 정도로 인식성능이 크게 저하됨을 알 수 있다. 그러나, 입술 파라미터의 추정시에 오차가 발생할 경우, SI 방법론의 립리딩 성능은 인식성능이 크게 저하되면, DR 통합의 경우에는 잡음이 약 10%정도까지 되더라도, 인식 성능이 저하되지 않음을 살펴볼 수 있다. 결국 잡음량이 10%정도 된다고 할 때, DR 통합방법은 55%의 인식률을 유지하고 있지만, SI 통합에서는 인식률이 19%정도까지 크게 저하된다. 그러므로 그림 3의 실험결과로부터, 입술 파라미터 추정에 잡음이 발생한다고 하면, SI 통합방법보다는 DR 통합방법을 사용하는 것이 타당하다는 결론을 내릴 수 있다.

#### V. 결론

본 논문에서는 최근 잡음환경하 음성인식의 성능을 향상시키기 위해서 연구되고 있는 시각정보를 이용한 bimodal 음성인식의 견인성 문제를 검토하였다. 음성의 잡음에 대한 성능저하는 널리 알려져 있는 문제가

지만, bimodal 음성인식에서 잡음에 의한 입술 파라미터의 오염과, 이에 따른 성능저하는 아직 고려의 대상이 아니었다. 본 논문에서는 입술 파라미터의 추정시 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 dominant recording 통합 방법이 잡음에 대하여 견인함을 발견하였다.

향후, 우리는 시청각 DB를 확충하여 연구결과를 더욱 객관적으로 입증할 것이다. 또한, 음성과 영상의 정보에 잡음이 섞인 경우, 자동적으로 잡음의 양을 추정하고, 음성과 영상 정보의 가중치를 자동적으로 결정할 수 있는 방법에 대하여 연구하고자 한다.

## 참 고 문 헌

- [1] Rajeev Sharma, Vladimir I. Pavlovic, Thomas S, Huang, "Toward Multi-modal Human-Computer Interface", Proceedings of the IEEE Vol. 86. No 5. May 1998.
- [2] Gerasimos Potamianos, Hans Peter Graf, Eric Cosatto, "An Image Transform Approach for HMM based Automatic Lipreading", Processing Of the Int. Conf. On Image Processing. pp. 173-177, 1998.
- [3] C.Bregler and Yochai Konig, "Eigenlips' for Robust Speech Recognition", Proc. IEEE Int. Conf. On Acoustics, Speech and Signal Processing, pp. 669-672, 1994
- [4] T.Chen, H. P. Graf, and K. Wang, "Lip-synchronization using speech-assisted video processing", IEEE Signal Processing Lett., vol 2, pp. 57-59, 1995.
- [5] L. Girin, J. Schwartz, and G Feng, "Audio-visual enhancement of speech in noise," JASA vol109-6, pp.3007-3020, 2001 June.
- [6] 김진영, 최용진, 최승호, "입술정보를 이용한 음성 파라미터 예측 및 음성인식 성능향상," 한국음향학회 추계학술대회 논문집, pp.65-68. 2002년 11월 2일.