입술정보를 이용한 음성 특징 파라미터 추정 및 음성인식 성능향상*

민소희(전남대), 김진영(전남대), 최승호(동신대)

<차 례>

<Abstract>

Estimation of speech feature vectors and enhancement of speech recognition performance using lip information

So-Hee Min, Jin-Young Kim, Seung-Ho Choi

Speech recognition performance is severly degraded under noisy envrionments. One approach to cope with this problem is audio-visual speech recognition.

In this paper, we discuss the experiment results of bimodal speech recongition based on enhanced speech feature vectors using lip information. We try various kinds of speech features as like linear predicion coefficient, cepstrum, log area ratio and etc for transforming lip information into speech parameters. The experimental results show that the cepstrum parameter is the best feature in the point of reconition rate. Also, we present the desirable weighting values of audio and visual informations depending on signal-to-noiso ratio.

* 주제어: 입술인식, 바이모달, DRI통합, 선형변환

---

* 본 연구는 2002년도 한국전자통신연구원 네트워크기술연구소 음성정보연구센터의 연구비 지원으로 수행되었습니다.

1. 서 론

최근 음성인식 분야에서는 심한 잡음 환경에서 인식률을 높이기 위한 연구가 활발히 진행되고 있다. 현재 인식기술 수준은 실험실과 같이 잡음을 거의 배제한 환경에서는 뛰어난 인식성능을 보이고 있으나 소음이 많이 발생하는 자동차 내부, 사무실, 길거리와 같은 실생활에 적용할 때는 인식성능이 매우 저하된다.

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 음성파라미터를 보완하는 정보로서 이용되고 있다[1-4]. 그런데 립리딩을 이용하는 시청각 음성인식에서 주요한 문제는 입술정보와 음성정보를 어떻게 혼합할 것인가에 있다.

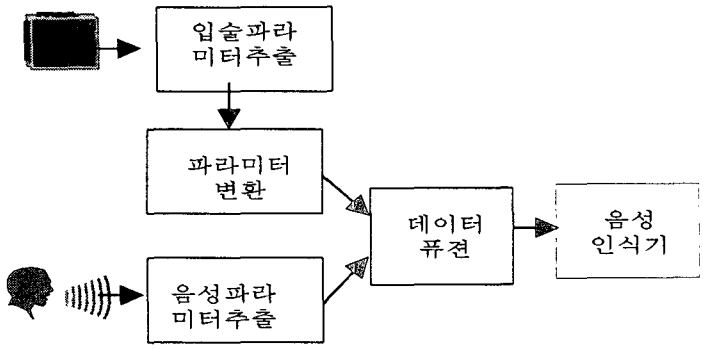
지금까지 널리 사용된 방법은 크게 초기통합(early integration)과 후기통합(late integration)으로 나눌 수 있는데 이 두 방법은 인간의 뇌가 시각정보를 반영하는 과정(process)을 모두 표현할 수 있다. 최근에 발표된 논문[5]에 의하면, 인간의 뇌에서는 시각정보를 이용하여 음성정보를 개선하고, 개선된 정보를 통하여 음성을 인식하는 메커니즘(dominant recording model, DRM)이 있다고 하여 이를 시청각음성인식의 방법으로 검토한 바가 있다. (물론, 이 방법도 초기통합의 한 방법으로 생각할 수 있다.)

본 논문에서는 최근에 제시된 DRM 모델을 기반으로 한 시청각정보의 통합 및 음성인식 성능향상에 대하여 실험결과를 논의하고자 한다. 본 논문에서는 DRM 방법을 이용하는 경우 여러 가지의 음성특징 파라미터 중 어느 파라미터가 가장 유효한지에 대하여 실험하였으며, 잡음음성과의 통합 시에 필요한 가중값이 어떻게 SNR에 따라 결정될 수 있는지에 대하여 실험하였다.

2. 시청각 정보의 통합방법

시청각 바이모달 정보를 통합하는 방법은 크게 직접식별 (direct identification, DI), 구별식별 (separate identification, SI), 그리고 우세레코딩식별 (dominant recording identification, DRI)로 나누어 볼 수 있다. 여기서 DI와 DRI는 초기통합에 해당하며, SI는 후기통합에 해당한다. 직접식별은, 음성 파라미터와 입술 파라미터를 확장된 벡터로 통합하여 학습 및 인식을 수행하는 것이며, 구별식별은 음성인식기와 입술 인식기의 결과를 가중치에 의하여 통합하는 것이다. 그리고, DRI는 입술 파라미터를 이용하여 입술 파라미터를 개선하고, 이를 사용하여 음

성인식을 하는 방법이다. 그림 1은 DRI 방법을 표현한 것이다.

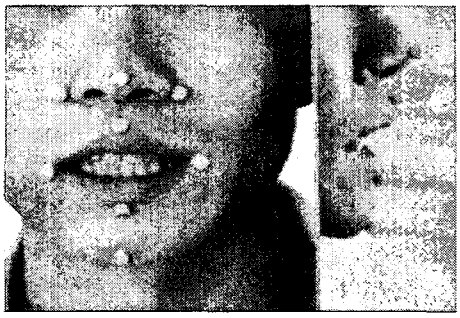


<그림 1> DRI 통합 방법

그림에 보인 바와 같이 DRI 방법에서는 입술 파라미터를 음성 파라미터로 변환 후에 적절한 통합방법에 의하여 두 파라미터를 통합한 후 음성 인식기에 입력으로 사용하게 된다.

3. 시청각음성 데이터베이스

본 연구에서는 DRI 방법의 시청각정보 통합방법을 실험하기 위하여 시청각 음성 DB를 구축하였다. 본 연구는 DRI 방법을 적용하기 위한 기초연구이므로, 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 옆모습이 투영되도록 하였다. 그림 2는 본 연구에서 촬영된 얼굴영상 중 한 프레임을 그린 것이다.



<그림 2> 마커를 부착한 얼굴이미지
및 입술 특징 파라미터

다음의 표 1은 구축된 음성 시청각 DB의 스펙을 나타낸다.

<표 1> 구축된 시청각음성 DB의 스펙

녹음단어명	아, 이, 우, 에, 오의 한국어 단모음 5개
녹음화자 수 및 회수	20대 1인, 각 30회발성
A/D 변환	음성 : 8kHz/16bit 영상 : 30frames/sec

한편, 본 논문에서 입술 파라미터는 마커의 위치를 추적한 후에 구하였는데, 마커의 위치는 Famous Tracker 툴을 사용하여 추적하였다. 추적된 마커의 위치 정보로부터 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.

4. 음성인식 실험 결과

본 장에서는 위에서 설명한 DRI 방법을 구축된 시청각음성 DB에 적용하여, 도출된 시청각음성인식 실험결과에 대하여 설명한다. 본 연구에서는 음성인식 방법으로서, 가우시안 혼합(Gaussian mixture) HMM (Hidden Markov Model)을 사용하였으며, 인식대상의 단어가 모음인 음소이기 때문에 상태개수가 3개인 단어 단위 인식기를 사용하였다. 세 종류의 인식실험이 이루어졌는데 다음과 같다.

첫째, 어떤 음성 특징 파라미터가 견인한가.
둘째, 입술 파라미터를 음성 파라미터로 변환 후 어떤 특징 파라미터가 유효한가.
셋째, 잡음의 정도에 따라, 시청각정보의 혼합비율은 어느 값이 최적인가에 관한 실험이다. 다음 절에서 이 세 가지에 대한 실험결과를 설명한다.

4.1. 음성특징 파라미터의 견인성

음성 특징 파라미터의 견인성에 관한 문제는 널리 알려져 있으며, 일반적으로 켑스트럼 파라미터가 널리 사용되고 있다. 본 논문에서는, 입술파라미터의 음성 파라미터 변환 후 각 특징 파라미터의 유효성을 검증하기 전에 여러 파라미터의 견인성을 다시 한번 검토하였다. 다음의 표는 각종 깨끗한 환경의 음성과 잡음음성에 대한 여러 파라미터에 대한 성능을 보인 것이다.

<표 2> 음성 특징 파라미터의 견인성

파라미터 종류	깨끗한 음성	잡음환경 음성 (SNR=10dB)
선형예측계수	99	24
반사계수	100	24
면적비계수	100	50
로그면적비계수	99	33
라인스펙트럼계수	100	20
켑스트럼	100	60

우리는 표 2의 실험결과로부터 켑스트럼 파라미터가 잡음환경 하에서 견인한 인식 파라미터임을 다시 확인할 수 있다.

4.2. 입술파라미터의 변환과 인식성능

우리는 윗 절에서 음성특징 파라미터의 견인성을 고찰하였다. 본 절에서는 입술 파라미터를 위의 여섯 가지의 파라미터로 변환하였을 때, 인식의 성능을 검토하고자 한다. $\hat{x}_a$를 입술정보로부터 추정된 음성특징 파라미터, x_v를 입술특징 파라미터라고 할 때, 변환 식은 다음과 같은 함수로 표현할 수 있다.

(1) $$\hat{x}_a = f(x_v)$$

추정함수 f로는 선형회귀에 의한 선형함수, 다층퍼셉트론과 같은 비선형함수를 사용할 수 있으나, 본 연구에서는 DB의 개수가 많지 않은 관계로 선형회귀에 의한 선형함수를 추정함수로서 사용하였다. 즉,

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

이다. 식 (2)에서 p는 음성 특징 파라미터의 차수이며, q는 입술 파라미터의 차수이다. 물론, q=5이고 p는 켑스트럼 파라미터의 경우는 13이고, 나머지 파라미터는 8이다. 위의 식에서 선형회귀 행렬 T는 최소자승법 또는 최소평균자승오차법을 사용하여 쉽게 구할 수 있는데, 다음과 같다.

(3) $$T = A_{x_v x_v}^{-1} C_{x_v x_a}$$

여기서, 행렬 A는 자기상관행렬이고, C는 상관행렬이다. 다음의 표 3은 입술 특징 파라미터로부터 음성 특징 파라미터를 추정하고 이를, 원래의 깨끗한 음성으로 학습된, 음성인식기에 입력하여 얻은 인식률을 보여준다.

<표 3> 입술 파라미터를 음성 특징 파라미터로 변환 후 인식률
(단, M은 가우시안 혼합계수)

파라미터 종류	M=1	M=2
선형예측계수	63	63
반사계수	62	54
면적비계수	69	46
로그면적비계수	62	54
라인스펙트럼계수	32	48
켑스트럼	84	68

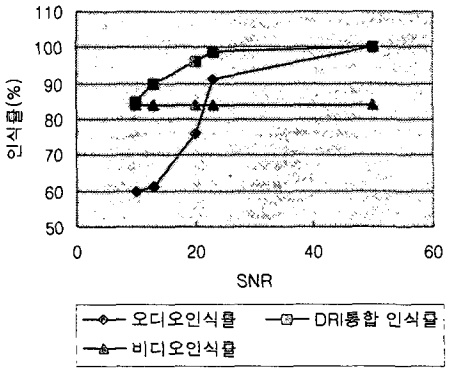
우리는 위의 표 3에서, 음성 특징 파라미터 중 켑스트럼 계수가 입술 파라미터와 가장 상관도가 높다는 사실을 관찰할 수 있으며, 만약, 입술 특징 파라미터로부터 추정된 켑스트럼 파라미터를 사용한다면, 최소 84%(M=1인 경우)의 인식률을 얻을 수 있다는 사실을 알 수 있다. 즉, DRI 방법에 의한 시청각 음성인식은 잡음환경 하에서 인식률 향상을 도모하는 데 매우 유용할 것이다.

4.3. 신호대잡음비에 따른 DRI 통합

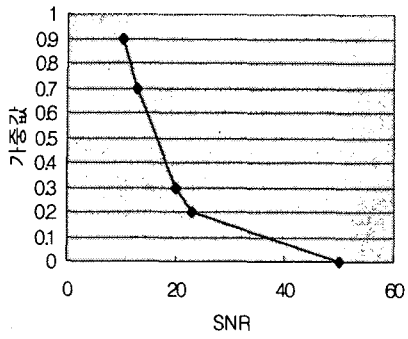
우리는 4.2절에서 입술 파라미터를 음성 특징 파라미터로 변환하여 사용하면, 잡음환경 하에서 음성인식률의 향상을 꾀할 수 있음을 확인하였다. 그렇다면, 문제는 원음성 파라미터와 추정음성 파라미터를 어떻게 혼합할 것인가이다. 본 연구에서는 가중 값을 두어 혼합하는 방법을 사용하였다. 식 (4)는 이를 표현한 것이다.

$$x_a^{mixed} = \lambda \hat{x}_z + (1-\lambda) x_a \tag{4}$$

다음의 그림 3은 다양한 신호대잡음비에서 가중 값 λ를 변환시키면서 구한 최적의 인식률을 그림으로 보인 것이다. 그림에 보인 바와 같이 SNR이 20dB의 경우 오디오정보만을 사용한 경우는 인식률이 76% 정도이나, 최적 가중치에 의해 비디오 정보와 혼합된 경우 인식률은 약 97%를 보여, 실제 인식률이 향상되고 있음을 확인할 수 있었다. 한편 그림 4는 SNR에 따른 최적 가중 값을 보여주는 그림으로서, SNR이 작을수록 입술정보의 가중치가 점차 증가하고 있음을 알 수 있다.



<그림 3> DRI 통합에 의한 인식실험결과



<그림 4> SNR에 따른 최적 가중값

5. 결 론

본 논문에서는 시각정보를 이용하여, 잡음환경 하에서 음성인식의 성능을 향상시키기 위한 기초연구로서, 최근 새롭게 제시된 통합방법 DRI에 대하여 실험결과를 보였다. 본 논문에서 새롭게 밝혀진 사실은 여러 음성 파라미터 중 켑스트럼 파라미터가 입술 파라미터와 가장 상관도가 높아, DRI 통합 시에는 켑스트럼을 사용하는 것이 바람직하다는 사실이다.

향후 연구로서 우리는 시청각 음성 DB를 확장하여 더욱 객관적인 결과를 확보하는 것, 그리고 입술 특징 파라미터에 잡음이 섞인 경우 어떻게 이를 예측하고 가중 값을 결정할 것인가에 대한 연구를 진행할 계획이다.

참 고 문 헌

[1] Sharma, R., I. Vladimir, Pavlovic, S. and Thomas, Huang (1998), Toward Multi-modal Human-Computer Interface, *Proceedings of the IEEE 86(5).*

[2] Potamianosm, G., H. P. Graf, and E. Cosatto (1998), An Image Transform Approach for HMM based Automatic Lipreading, *Processing Of the Int. Conf. On Image Processing,* pp.173~177.

[3] Bregler, C., & Y. Konig (1994), Eigenlips for Robust Speech Recognition, *Proc. IEEE Int. Conf. On Acoustics, Speech and Signal Processing*, pp.669~672.

[4] Chen, T., H. P. Graf, and K. Wang (1995), Lip-syncronization using speech-assisted video processing, *IEEE Signal Processing Lett. 2*, pp.57~59.

[5] Girin, L., J. Schwartz, and G. Feng (2001), Audio-visual enhancement of speech in noise,

JASA 109(6), pp.3007~3020

접수일자: 2002년 11월 13일
게재결정: 2002년 12월 12일

▶ 민소희(So-Hee Min)
주소: 500-757 광주광역시 북구 용봉동 300
소속: 전남대학교 공과대학 전자공학과 대학원 신호처리실험실
전화: 062) 530-0472
Fax: 062) 530-0472
E-mail: shmin3@hanmail.net

▶ 김진영(Jin-Young Kim)
주소: 500-757 광주광역시 북구 용봉동 300
소속: 전남대학교 공과대학 전자공학과
전화: 062) 530-1757
Fax: 062) 530-0472
E-mail: kimjin@dsp.chonnam.ac.kr

▶ 최승호(Seung-Ho Choi)
주소: 520-714 전남 나주시 대호동 252번지
소속: 동신대학교 정보과학대학 정보통신공학부 멀티미디어통신공학전공
전화: 061) 330-3194
Fax: 061) 330-2209
E-mail: shchoi@white.dsu.ac.kr