1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 2(s) 호

입술정보추출 및 파라미터 선정 방법에 따른 바이모달 음성인식 성능 비교

박병구* 김진영* 최승호**
전남대학교 전사공학과*
동신대학교 정보통신공학과**

Effects of Extraction Method and Choice of Lip Parameters on the Bi-modal Speech Recognition

Byung-Ku Park*, Jin-Young Kim*, Seung-Ho Choi**
Dept. of Electronics Engineering, Chonnam National Univ.*
(E-mail : park@dsp.chonnam.ac.kr, kimjin@dsp.chonnam.ac.kr)
Dept. of Information and Communication Engineering, DongShin Univ.**

요약

음성신호와 영상신호를 함께 이용하는 바이모달(Bi-modal)음성인식에서 어떤 입술 파라미터를 사용하는가에 따라 인식시스템의 성능이 달라진다. 그래서 본 논문에서는 이미지에 관계한 입술파라미터를 견인하게 추출하기 위한 방법으로 x 프로파일(profile)을 이용한 방법을 사용하였다. 파라미터를 선정을 달리하여 실험한 결과 15dB이상에서는 안쪽입술의 2개의 파라미터를 이용한 경우가, 10dB 이하에서는 4개의 입술파라미터를 이용한 경우가 더 좋은 인식률을 보였다. 안쪽 입술 파라미터를 이용한 경우가 바깥쪽 입술 파라미터를 이용한 경우보다 더 좋은 인식률을 보였다.

Ⅰ. 서론

기존 음성인식시스템이 가지고 있는 단점인 잡음환경에서 인식률의 한계를 극복하기 위해 음성정보와 입술정보를 이용한 바이모달 음성인식시스템에 대한 연구가 최근에 매우 활발하다[1]. 바이모달 음성인식 시스템의 성능에 영향을 미치는 요소는 여러 가지가 존재할 수 있다. 그 중 파라미터 추출방법에 따라서 그리고 어떤 파라미터를 선택하느냐에 따라서 인식률이 영향을 받는다[2]. 그래서 본 논문에서는 입술 파라미터 선정에 따라서 바이모달 음성인식 성능이 어떻게 변화하는 가를 살펴보고자 한다. 2장에서 바이모달 음성인식 시스템에 대해서 설명을 하고 음성파라미터 추출과정을 설명하였으며 3장에서는 입술모양을 파라미터 화하는 방법을 보였다. 4장과 5장에서는 추출된 입술파라미터와 음성파라미터를 이용하여 두 형태의 정보를 결합하는 방법과 인식 실험한 결과를 보이고 성능비교를 하였다.

Ⅱ. 시스템구성 및 음성파라미터 추출

그림1과 같이 음성신호뿐만이 아니라 입술영상도 함께 저장하기 위해서 두 대의 컴퓨터를 이용하여 입술영상은 TMC-7 CCTV카메라를 이용하여 Oculus-Tcx 이미지보드에 영상을 1초에 18프레임을 저장하고 영상과 동기를 맞추기 위해서 FX케이블을 이용하였고 음성신호는 마이크를 통하여 ASPI(Atlanta Signal Processors, Inc.) 회사의 TMS320C31 DSP칩을 내장한 ELF보드를 이용해서 음성을 저장하도록 구성하였다.

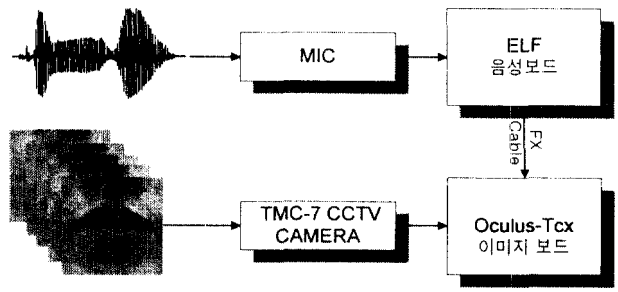

그림 1. 바이모달 음성인식 시스템 구성

바이모달(Bi-modal)음성인식은 음성신호에서 파라미터를 추출하는 부분과 입술 이미지에서 입술파라미터를 추출하는 부분으로 구성되어있다. 그중 음성인식에 사용하기 위해 음성파라미터로 12차 LPC(Linear Prediction Coding) 셉스트럼(Cepstrum)계수를 이용하였다. 8kHz로 저장한 음성신호를 1프레임당 256샘플로 나누어서 100샘플씩 이동시키면서 해밍윈도우를 사용하였다[3].

Ⅲ. 입술정보 추출

입술이미지는 1초에 18프레임을 저장시키서 한번 저장할 때마다 50개의 연속프레임을 100 × 100크기의 TIFF(Tagged Image File Format)형식의 컬러 이미지파일로 저장시켰다. 한번 연속프레임을 저장할 때마다 발음을 한 번 발음하여 저장시켜 실험에 이용하였다. 입술파라미터로는 가장 입 모양을 대표적으로 나타낼 수 있는 바깥입술의 높이(H1)와 폭(W1) 안쪽입술의 높이(H2)와 폭(W2)을 입술인식에 이용하였으며 추출과정은 그림2에서 볼 수 있고 각 블록별 과정은 다음과 같다.

1. Oculus-Fex 이미지보드를 통해 획득한 칼라이미지를 흑백영상으로 전환한다.
2. 메디안필터(Median Filter)를 이용하여 윤곽선을 보호하면서 잡음을 제거한다.
3. 잡음제거 후 그림3.B는 세로축 평균 색깔분포인 y 프로파일(profile)의 값을 나타내는데 여기 입술 이미지에서 이 분포를 살펴보면 입 모양을 나타내는 영역에서 비슷한 모양을 나타내는 것을 알 수 있다. 이러한 y 프로파일 값을 이용해서 입술의 위와 아래폭을 추출하여 바깥 입술의 높이를 계산할 수 있다.
4. 그림4.A와 같이 Sobel 윤곽 추출 자를 이용하여 입술의 윤곽선을 추출한 후 잡음제거 과정을 거치면 그림4.B와 그림을 얻을 수 있다[4].
5. 이 윤곽선이 추출된 이미지에서 바깥 입술의 폭을 계산할 수 있다.
6. 안쪽 입술의 높이는 전 과정에서 구해진 안쪽 입술의 폭으로부터 입술 중앙값을 계산할 수 있고 이 중앙값으로부터 중앙부위의 부분적인 y 프로파일을 따로 계산해내서 그것의 미분 값을 이용해서 그림3.D와 같이 안쪽 입술의 높이를 추출할 수 있다.
7. 그림6.C에서 보듯이 입술영역의 x 프로파일을 이용하여 입술 폭을 추출할 수 있다.

7번 과정에서 윤곽선을 이용한 입술파라미터 추출방법을 이용할 경우 윤곽선이 정확하게 추출되지 않으면 안쪽 입술 폭을 계산할 때 오류가 발생할 경우가 생긴다. 그림5.C에서 보듯이 안쪽 입술이 왼쪽을 밖에 잘못 잡았으니 오른쪽 부분은 밖이 오른쪽에서 미치므로 약간 왼쪽으로 치우쳐 있다. 이러한 에러는 안쪽에 입술이나 이빨의 영상이 있을 경우 윤곽선을 가지고는 안쪽 입술의 폭을 추출하는데 문제가 발생한다. 그래서 이러한 점을 보완하기 위해서 x 프로파일(x profile : x축 평균 이미지 분포도)을 이용하였다.

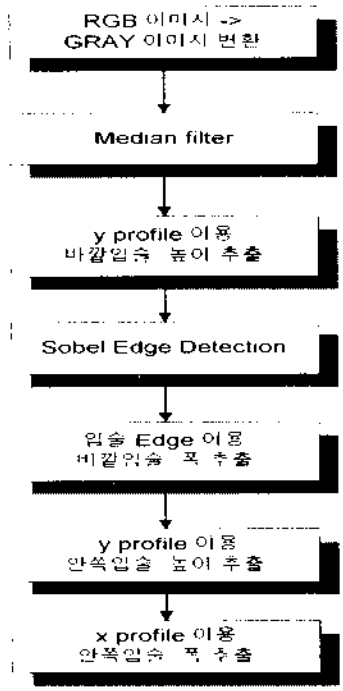

그림 2. 입술파라미터 추출 알고리즘

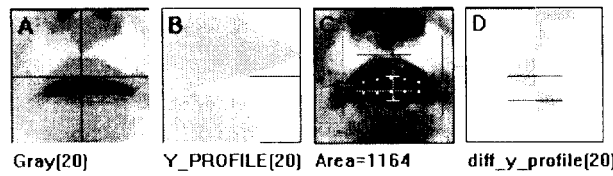

그림 3. 입술 파라미터가 추출된 모습

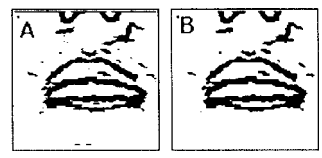

그림 4. Sobel 윤곽추출 후 잡음제거

입술이미지의 x 프로파일을 살펴보면 그림6.C와 같이 입술의 영역에서 일정한 분포를 갖는 것을 알 수 있다. 전 단계에서 구한 바깥 입술의 좌측, 우측 값으로부터 입술의 중앙값을 계산할 수 있고 그림6.C에서처럼 계산된 입술의 중앙값으로부터 좌 우측으로 가면서 x

프로파일의 부분적인 최소 값 지점을 찾아내서 이것을 안쪽 입술의 좌우로 추출하였다. 이 과정에서 입술의 좌우가 대칭이라라고 가정하고 좌측, 우측 값을 재 추정하는 과정을 거치면 그림6.D와 같이 좌 우측이 대칭인 안쪽입술의 폭을 추출할 수 있다.

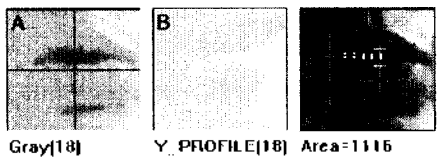

그림 5. 윤곽선 이용한 정우 에러맵생 프레임

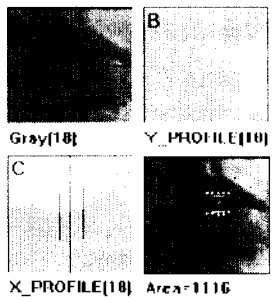

그림 6 x profile이용 안쪽입술 폭 추출

Ⅳ. 음성정보와 입술정보 결합

앞의 음성정보 추출과정과 입술정보 추출과정을 거친 후 추출된 음성과 입술 파라미터를 이용해서 인식실험을 하기 위해서는 음성과 입술 파라미터의 발음구간인 시작점과 끝점을 구해야 한다. 먼저 음성의 시작점과 끝점을 에너지(Energy)를 이용하여 구한 다음 입술 프레임의 시작 프레임과 끝 프레임은 음성 파라미터와 동기를 맞추기 위해서 이미 추출된 음성구간의 시작점과 끝점 정보를 이용해서 입술 이미지의 시작 프레임과 끝 프레임을 추출하였다. 입술프레임은 50프레임이고 음성은 44032byte이므로 한 프레임 당 22016샘플/50프레임=440.32샘플에 해당한다.

위와 과정을 통해서 4개의 입술 특징 파라미터와 12차 LPC 켑스트럼 계수를 이용하여 음성정보와 입술정보를 독립적으로 결합하는 방법(late integration)을 이용하여 통합하였다[5]. 음성정보와 입술정보를 독립적으로 결합하는 방법은 그림7과 같이 음성인식기와 음성인식기를 통하여 나온 각각의 인식스코어를 식1과 같이 각각 가중치(α)를 부과하면서 결합하는 방법이다. 그림 7.(a)는 입술파라미터를 DTW(Dynamic Time Warping)방법을 이용하여 패턴을 비교한 결과이고 그림7.(b)는 음성파라미터를 같은 방식으로 비교한 결과를 나타내고 있다.

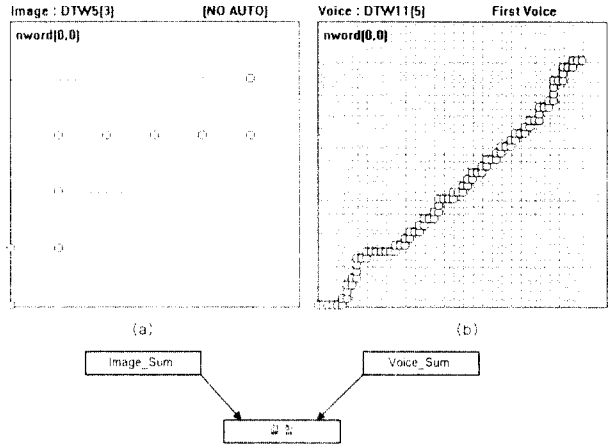

그림 7. 영상정보와 음성정보를 독립적으로 결합

$$S_{av} = \alpha * S_{v} + (1-\alpha) * S_{a} \quad (1)$$

S_{av} : 인식스코어

S_{v} : 시각스코어, S_{a} : 음성스코어

α : 시각가중치, $1-\alpha$: 음성가중치

Ⅴ. 실험 및 결과

바이모달 음성인식 실험을 위해서 160개의 시외전화 번호 지역 명을 실험데이터로 사용하였다. 한 지역 명에 대해서 4개의 데이터를 녹음하여 1개는 기준패턴으로 사용하고 나머지 3개는 테스트패턴으로 사용하였다. 화자종속 고립단어 인식실험을 수행하였고 음성파라미터와 입술파라미터는 각각 DTW(Dynamic Time Warping)방법을 이용해서 패턴끼리 비교를 하였다. 영상은 1초에 18개의 입술영상을 획득했고 음성은 8kHz로 샘플링하여 녹음하였다. 그림8의 (a)-(d)는 각각 시각가중치(α)를 변화시키면서 깨끗한 음성에서부터, 20dB, 15dB, 10dB, 5dB, 0dB의 백색 가우시안(Gaussian) 분포를 갖는 잡음을 섞은 신호에서 바이모달 음성인식 실험 결과를 나타내고 있다. 바깥쪽 입술 2개를 사용한 경우, 안쪽 입술 2개를 사용한 경우, 4개의 파라미터를 모두 사용한 경우에 대해서 각각 실험을 하였다. 안쪽 입술의 폭과 높이를 사용한 경우가 바깥 입술의 폭과 높이를 사용한 경우보다 전체적인 바이모달 음성인식시스템의 더 많은 성능향상을 보이는 것을 알 수 있다. 입술 파라미터 4개를 이용한 경우는 전체적으로 좋은 인식률을 보이고 있음을 알 수 있다. SNR(Signal to Noise Ratio)이 15dB이상에서는 안쪽입술2개를 사용한 경우가 더 좋은 인식률을 보이고 10dB이하에서는 입술파라미터 4개를 사용한 경우가 더 좋은 인식률을 보였다.

VI. 결론

바이모달 음성인식에서 이미지의 색상에 관계한 입술모양을 파라미터 화하여 인식실험에 사용함에 있어서 윤곽선을 이용하여 안쪽입술의 폭을 측정 시 발생할 수 있는 에러를 x profile을 이용하여 견인하게 추출하였다. 입술파라미터로 바깥입술의 높이와 폭 안쪽입술의 높이와 폭, 입술파라미터 4개 모두를 사용한 경우에 대하여 실험을 하였다. 파라미터를 신정을 달리하여 실험한 결과 15dB이상에서는 안쪽입술의 2개의 파라미터를 이용한 경우가, 10dB 이하에서는 4개의 입술파라미터를 이용한 경우가 더 좋은 인식률을 보였다. 전체적으로 바깥쪽 입술보다 안쪽 입술의 변화가 인식률이 좋게 나왔다. 안쪽 입술 파라미터가 바깥쪽 입술보다 인식과정에서 더 많은 정보를 제공한다는 것을 알 수 있었다.

※ 이 논문은 한국과학재단의 '98 핵심전문연구 지원에 의해 이루어진 연구결과물 중 하나입니다.

참고문헌

[1] Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, Victor Zue, "Survey of the State of the Art in Human Language Technology", Center for Spoken Language Understanding, Oregon Graduate Institute, p. 329-362, 1995.

[2] Peter L. Silsbee and Alan C. Bovik "Computer Lipreading for Improved Accuracy in Automatic Speech Recognition" IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING VOL.4, NO. 5, pp 337-351, SETEMBER 1996.

[3] Lawrence Rabiner, Biing-Hwang Juang "Fundamentals of Speech Recognition", PTR Prentice-Hall, 1993.

[4] Earl Gose, Richard Johnsonbaugh, Steve Jost "PATTERN recognition and IMAGE analysis", Prentice Hall, 1996.

[5] Silsbee, P. L., "Sensory Integration in Audiovisual Automatic Speech Recognition", 28th Annual Asilomar Conference on Signals, Systems, and Computers, 1994.

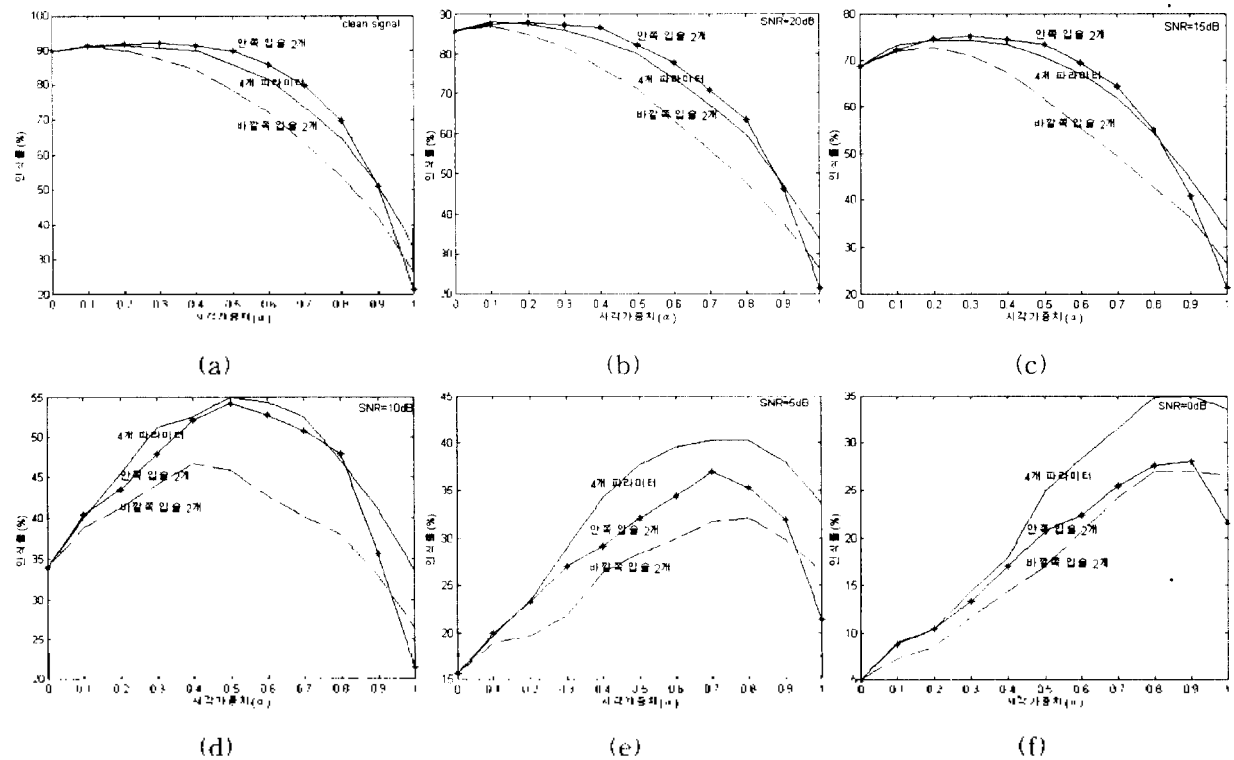

그림 8. 파라미터 신정에 따른 바이모달 음성인식의 인식률