입술 파라미터 선정에 따른 바이모달 음성인식 성능 비교 및 검증

Performance Comparison and Verification of Lip Parameter Selection Methods in the Bimodal Speech Recognition System

박 병 구*, 김 진 영*, 임 재 열**
(Byung Ku Park*, Jin Young Kim*, Jae Yeol Rheem**)

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

요 약

바이모달 음성인식 시스템에서 어떤 입술파라미터를 선정하느냐 그리고 얼마나 견인하게 추출하는 가에 따라서 인식률에 큰 영향을 미친다. 그래서 본 논문에서는 자동 추출 알고리듬을 이용하여 입술파라미터를 추출하고 안쪽 입술 파라미터가 바깥 입술 파라미터보다 바이모달 음성인식 시스템에 더 많은 영향을 미친다는 것을 보였다. 그리고 손으로 추출한 추출알고리듬과 비교하여 자동 추출알고리듬의 신뢰성을 비교하였다.

ABSTRACT

The choice of parameters from various lip information and the robustness of extracting lip parameters play important roles in the performance of bimodal speech recognition system. In this paper, lip parameters are extracted by using an automatic extraction algorithm and inner lip parameters effect on the recognition rate more than outer lip parameters. Compared with a manual extraction algorithm, the automatic extraction method is evaluated about its robustness.

I. 서 론

음성정보와 입술정보를 함께 이용하는 바이모달(Bimodal) 음성인식에서 파라미터 선정은 인식률에 영향을 크게 미치게 된다[1]. 그래서 본 논문에서는 어떠한 파라미터를 선정하느냐에 따라서 바이모달 음성인식 시스템에 영향을 얼마나 미치게 되는지를 자동 알고리듬을 이용하여 입술 파라미터를 추출하고 이 추출된 파라미터의 신뢰성을 검증해 보기 위해 수작업을 통하여 입술파라미터를 추출하여 인식실험을 통하여 비교하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서 입술정보 추출방법에 대해서 설명을 하고, 3장에서는 자동 알고리듬을 이용하여 추출한 입술파라미터와 수작업으로 표시하여 추출한 입술파라미터를 비교한다. 그리고 4장에서는 입술 파라미터 선정에 따른 인식률을 비교해보자 한다.

* 전남대학교 전자공학과
** 한국기술교육대학교 전자공학과

접수일자 : 1999년 1월 5일

II. 입술정보 추출방법

입술정보 추출방법은 여러 가지가 있으나 크게 이미지에 근거한 추출방법(pixel-based method), 모델에 근거한 추출방법(model-based method), 빛의 밝기를 벡터 화하여 추출하는 방법(optical-flow method)등이 있다[2][3]. 본 논문에서는 이중 이미지에 근거한 입술파라미터 추출방법을 이용하였다. 이미지에 근거하여 입술의 모양, 즉 바깥 입술의 높이와 폭, 안쪽 입술의 높이와 폭을 파라미터로 사용하였다. 입술파라미터를 추출하기 위해서 자동 추출 알고리듬을 이용하여 4개의 입술파라미터를 추출하는 방법과 손으로 입술이미지에 표시를 하여 입술 파라미터를 추출하는 방법을 이용하였다.

2.1 자동 입술 파라미터 추출 알고리듬

입술 이미지는 1초에 18프레임을 저장, 한번 저장할 때마다 50개의 연속프레임을 100×100크기의 TIFF-(Tagged Image File Format)형식의 컬러 이미지파일로 저장시켰다. 실제로 자동 입술 파라미터 추출 알고리듬을 이

용하여 추출된 파라미터의 모습은 그림3.D와 같고 추출과정은 다음과 같다.

1. 컬러 입술 이미지를 흑백 입술 이미지로 전환(그림 1.A)
2. 메디안필터(Median Filter)이용 잡음 제거
3. 세로축 평균 색깔분포인 y 프로파일(profile)의 값을 이용해서 입술의 위와 아래쪽을 추출하여 바깥 입술의 높이를 계산(그림 1.B)
4. Sobel 윤곽 추출 자를 이용하여 입술의 윤곽선을 추출
5. 윤곽선이 추출된 이미지에서 바깥 입술의 폭을 계산
6. 전 과정에서 구해진 안쪽 입술의 폭으로부터 입술 중앙값을 계산할 수 있고 이 중앙값으로부터 중앙부근의 부분적인 y 프로파일을 따로 계산해내서 그것의 미분 값을 이용해서 그림 1.D와 같이 안쪽 입술의 높이를 추출
7. 그림 1.C에서 보듯이 입술영역의 x 프로파일을 이용하여 입술 폭을 추출

기존의 방법이 단순한 입술의 x 프로파일과, y 프로파일을 이용한 반면 본 논문에서는 입술의 영역을 우선 구한 다음 그 영역에서의 부분적인 x 프로파일과 y 프로파일을 계산하여 안쪽입술의 높이와 바깥입술의 폭을 추출하였다[4].

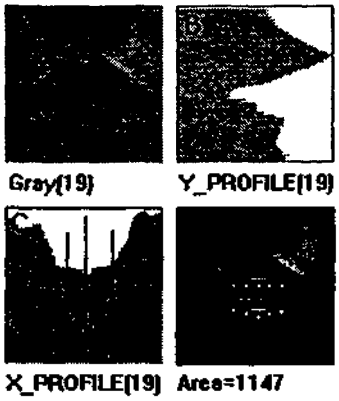

그림 1. 파라미터자동추출
Fig. 1. Automatic parameter extraction.

2.2 수작업을 이용한 입술파라미터 추출

자동으로 추출한 입술 파라미터의 신뢰성을 알아보기 위해 손으로 입술 이미지에 표시를 하여 4개의 입술파라미터를 추출하도록 하였다. 추출방법은 컬러 입술이미지를 보고 바깥 입술의 좌, 우측에 표시를 하면 자동으로 바깥 입술의 폭을 계산하여 파라미터를 저장하도록 프로그램을 구성하였다. 같은 방식으로 바깥 입술의 위쪽, 아래쪽, 안쪽 입술의 좌, 우측, 안쪽입술의 위쪽, 아래쪽에 표시를 하면 자동으로 바깥입술의 높이와 안쪽입술의 폭과 높이를 계산하도록 하였다. 실제 수작업을 통하여 추출하는 모습을 그림 2에 나타내었다.

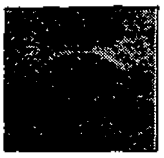

그림 2. 수작업을 이용한 추출
Fig. 2. Manual Parameter Extraction.

III. 추출된 입술파라미터의 비교

그림 3은 지역 이름인 '고성'을 발음한 것에 대해 자동 알고리듬으로 추출한 입술파라미터와 수작업을 통하여 추출된 입술 파라미터를 음성 파형과 함께 보인 것이다. 그림 3.A와 그림 3.B는 자동 알고리듬으로 추출한 입술파라미터이고 그림 3.C와 그림 3.D는 수작업을 통하여 추출된 입술 파라미터이다. 그림 3.E는 음성 파형을 나타낸 것이다. 각각의 단어마다 특유의 입술파라미터 파형을 가지므로 인식이 가능하다. 자동 알고리듬을 이용하여 추출한 경우(그림 3.A, 그림 3.B)는 파라미터 파형이 수작업을 이용하여 추출한 경우(그림 3.C, 그림 3.D)보다 더 안정적이며 불규칙한 성분이 없다는 것을 알 수 있다. 수작업을 이용하여 추출한 방법은 사람의 눈을 통하여 매 순간마다 입술의 경계가 달리 보일 수 있으므로 항상 같은 조건하에서 추출할 수 없는 단점이 있다. 이러한 요소가 그림 3.C와 그림 3.D에서 불규칙한 성분을 만드는 요인이 되고 있다.

파라미터 선정에 따른 바이모달 음성인식 시스템의 성능을 비교하기 위해서 추출된 입술파라미터(바깥 입술의 폭과 높이, 안쪽 입술의 폭과 높이) 중에서 바깥입술의 폭과 높이만 사용한 경우, 안쪽 입술의 폭과 높이만 사용한 경우, 4개의 파라미터를 모두 사용한 경우, 그리고 입술모양 2개를 사용한 네가지 경우에 대해서 비교하였다.

여기서 바깥입술의 폭과 높이, 안쪽입술의 폭과 높이는

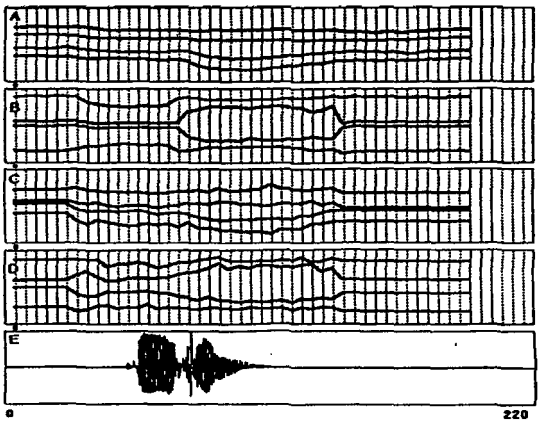

그림 3. 입술파라미터 비교
Fig. 3. Compararisons of lip parameters.

식 1에 의해서, 입술모양 2개의 파라미터는 식 2에 의해서 정규화 과정을 통해 계산된다.

$$OW=OW/W \quad (1)$$
$$OH=OH/W$$
$$IW=IW/W$$
$$IH=IH/W$$

OW : 바깥입술의 폭, OH : 바깥입술의 폭
IW : 안쪽입술의 폭, IH : 안쪽입술의 폭
W : 첫 프레임의 바깥입술의 폭

$$OL=OH/OW \quad (2)$$
$$IL=IH/IW$$

OL : 바깥입술모양 IL : 안쪽입술모양

바깥입술의 폭과 높이, 안쪽입술의 높이와 폭은 저장된 이미지 프레임의 첫 프레임의 바깥입술의 폭으로 나누어서 정규화 과정을 수행했고, 바깥 입술모양과 안쪽 입술모양은 입술모양의 타원형 정도를 수치로 나타냈다. 단순한 바깥입술의 폭과 높이 또는 안쪽입술의 폭과 높이를 이용하는 경우는 입 모양의 정규화를 위해서 인식에 사용되어지는 파라미터들과 함께, 항상 첫 프레임의 바깥입술의 폭을 계산해야 한다는 단점이 있는 반면 입술모양을 이용한 경우는 인식에 사용된 파라미터만을 이용하여 정규화 과정을 수행할 수 있다는 장점이 있다.

IV. 실험 및 결과

실험에 사용한 데이터는 시외전화번호 160개 지역 이름을 대상으로 하였고, 각각의 지명에 대해서 4번 발음한 데이터를 저장하여 임의의 160개를 기준 패턴으로 하고, 나머지 480개를 테스트 패턴으로 사용하였다. 음성 샘플링 주파수는 8kHz로 저장하였고 입술영상은 18frame/sec의 속도로 저장하였다. 음성 특징파라미터로 12차 LPC(Linear Predictive Coding) 켑스트럼(Cepstrum)계수를 이용하였다[5]. 추출된 음성파라미터와 입술 파라미터는 가중치를 사용한 인식 후 결합(late integration)방식을 이용 독립적으로 결합하였다[6].

먼저, 입술 정보만을 이용한 경우의 인식실험 결과를 표 1에 나타냈다. 자동 알고리듬을 이용하여 입술파라미터를 추출한 경우는 4개를 사용한 경우와 바깥 2개를 사용한 경우가 수작업으로 추출한 파라미터를 이용하여 인식실험한 것 보다 인식률이 더 좋게 나타났고 안쪽 2개를 사용한 경우는 수작업으로 추출한 방법이 더 좋은 인식률을 나타냈다. 이러한 결과는 자동 추출 알고리듬은 바깥쪽 입술은 잘 추출하는 편이나 안쪽입술 추출과정은 에러가 발생하는 부분이 있다는 것을 알 수 있다. 다시 말하면, 수작업으로 추출하는 과정이 추출하는 사람의 개인적인 주관에 의해 각각의 파라미터를 추출하는 순간마다 일정하지 않으므로 이러한 결과가 나왔다고 할 수 있다.

각각의 파라미터가 잡음이 섞인 음성신호에서 얼마만큼의 인식률을 나타내는지를 알아보기 위해 음성신호에 각각 0dB, 5dB, 10dB, 15dB, 20dB의 백색 가우시안(Gaussian) 잡음을 섞은 음성에서 실험을 하였다. 표 2는 입술파라미터를 자동 알고리듬을 이용하여 추출한 경우의 인식결과이고, 표 3은 손으로 입술 파라미터를 추출한 경우의 인식결과이다. 그림 4와 그림 5는 표 2와 표 3을 그림으로 나타낸 것이다. 입술파라미터를 알고리듬을 이용하여 추출한 경우와 손으로 입술파라미터를 추출한 경우의 공통된 결과는 음성파라미터만 이용한 경우보다 입술정보를 이용한 실험이 모두 좋은 인식결과를 보였다는 것이다. 그리고 안쪽입술 파라미터 2개를 이용한 경우가 바깥 입술 파라미터 2개를 사용한 경우보다 더 좋은 바이모달 인식결과를 보였다. 또한 파라미터 4개를 사용한 경우가 전체적으로 좋은 인식결과를 보이고 있다. 입술모양 2개를 이용한 경우는 정규화과정이 단순한 반면 잡음이 많이 섞인 음성에서는 인식률이 저조했지만 깨끗한 음성에서는 파라미터 4개를 사용한 경우와 같이 좋은 인식률을 나타냈다.

표 1. 입술정보만 이용한 경우인식률 (단위 : %)
Table 1. Recognition rate using information.

	4개	바깥2	안쪽2	입술모양2
자 동	33.54	26.46	21.46	23.96
수작업	26.46	12.08	21.88	11.04

표 2. 자동 알고리듬을 이용한 경우 인식률 단위(%)
Table 2. Recognition rate using automatic parameter extraction.

	0dB	5dB	10dB	15dB	20dB	clean
4개	[illegible]	[illegible]	[illegible]	74.38	87.92	91.67
바깥2	[illegible]	[illegible]	[illegible]	[illegible]	86.88	91.25
안쪽2	[illegible]	36.88	54.18	75.21	87.71	92.29
입술모양2	[illegible]	[illegible]	[illegible]	[illegible]	87.71	91.88
음성	5.00	15.63	33.96	68.75	85.63	90.00

표 3. 수작업으로 추출한 경우 인식률 단위(%)
Table 3. Recognition rate using manual parameter extraction.

	0dB	5dB	10dB	15dB	20dB	clean
4개	31.67	37.71	52.71	[illegible]	[illegible]	[illegible]
바깥2	14.58	22.29	39.58	72.29	86.88	[illegible]
안쪽2	26.67	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
입술모양2	13.13	22.71	45.21	73.75	[illegible]	[illegible]
음성	5.00	15.63	33.96	68.75	85.63	90.00

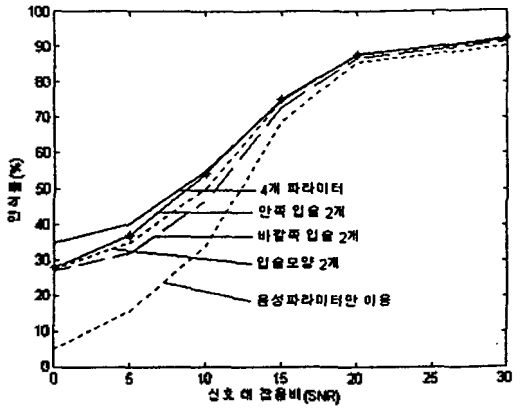

그림 4. 자동 알고리듬을 이용한 경우 SNR에 따른 인식률

Fig. 4. Recognition rate with SNR using automatic parameter extraction.

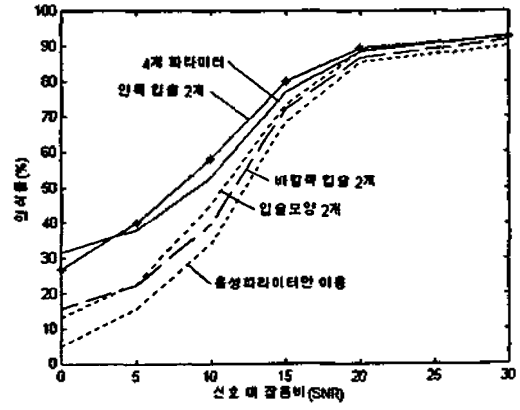

그림 5. 따른 수작업을 이용한 경우 SNR에 따른 인식률

Fig. 5. Recognition rate with SNR using manual parameter extraction

그리고 표 2와 표 3을 비교하면 잡음이 많이 섞인 음성에서는 자동알고리듬을 이용한 경우가 인식률이 높고 잡음이 적을수록 수작업을 이용한 파라미터가 더 좋은 인식률을 보였다.

V. 결 론

음성정보와 입술정보를 이용하는 바이모달 음성인식에서 어떤 파라미터를 인식과정에서 선정하느냐에 따라서 인식률이 달라짐을 보였다. 자동 입술 파라미터 추출 알고리듬과 수작업을 통하여 입술 파라미터를 추출하여 실험한 결과 자동 입술파라미터 추출 알고리듬은 0.83-7.51(%), 수작업을 이용한 입술 파라미터 추출 방법은 0.83-18.34 (%) 정도의 인식률을 나타냈고, 안쪽입술이 바깥입술보다 더 좋은 바이모달 음성인식 성능 향상을 보였다. 안쪽입술 2개를 이용하거나 안쪽과 바깥쪽 입술파라미터 4개를 이용한 경우가 입술모양 2개를 이용한 경우나 바깥 입술파라미터 2개를 이용한 경우보다 더 좋은 인식결과를 보였고 입술파라미터의 정규화과정을 간단하게 하기 위해서 입술모양 2개를 사용할 수도 있다는 것을 보였다. 그러나 자동 입술 파라미터 추출 알고리듬은 안쪽 입술 파라미터를 더 정확하게 추출하도록 알고리듬이 수정되어야 하며, 그로 인하여 더 향상된 바이모달 음성인식 시스템성능을 기대할 수 있다.

참 고 문 헌

1. 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.
2. Marcus E. Hennecke, K. Venkatesh Prasad, David G. Stork, "Using Deformable Templates to Infer Visual Speech Dynamics", CRC-TR-9430, 1994.
3. Goldshen, A.J., Garcia, O.N. & Petajan E., "Continuous Optical Automatic Speech Recognition by Lipreadig", 28th Annual Asilomar Conference on Signals, Systems, and Computers, 1994.
4. Rao, R.R. & Mersereau, "Lip Modeling for Visual Speech Recognition", 28th Annual Asilomar Conference on Signals, Systems, and Computers, 1994.
5. Lawrence Rabiner, Biing-Hwang Juang "Fundamentals of Speech Recognition", PTR Prentice-Hall, 1993.
6. Silsbee, P. L., "Sensory Integration in Audiovisual Automatic Speech Recognition", 28th Annual Asilomar Conference on Signals, Systems, and Computers, 1994.

▲박 병 구(Byung Ku Park) 1971년 5월 19일생

1997년 2월 : 전남대학교 전자공학과 (공학사)

1999년 2월 : 전남대학교 전자공학과 (공학석사)

※주관심분야 : 음성인식 및 신호처리

▲김 진 영(Jin Young Kim)

1986년 2월 : 서울대학교 전자공학과 (공학사)
1988년 2월 : 서울대학교 전자공학과 (공학석사)
1994년 8월 : 서울대학교 전자공학과 (공학박사)
1993년 3월～1994년 12월 : 한국통신 소프트웨어연구소 전임 연구원
1995년～현재 : 전남대학교 고과대학교 전자공학과 조교수
※주관심분야 : 음성인식 및 음성합성, 멀티모달 MMI

▲임 재 열(Jae Yeol Rheem)

1986년 2월 : 서울대학교 전자공학과 (공학사)
1988년 2월 : 서울대학교 전자공학과 (공학석사)
1995년 2월 : 서울대학교 전자공학과 (공학박사)
1995년 9월～현재 : 한국기술교육대학교 전자공학과(조교수)
※주관심분야 : 음성신호처리, DSP, 통신신호처리