1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 1(s) 호

잡음환경에서의 바이모달 음성인식

박병구* 김진영* 최승호**
전남대학교 전자공학과*
동신대학교 정보통신공학과**

Bi-modal speech recognition in noisy environments

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

기존의 음성인식시스템의 잡음환경에서 인식률의 한계를 극복하기 위해 음성신호뿐만이 아니라 입술정보를 결합하여 음성인식에 이용하는 바이모달(Bi-modal) 음성인식이 근래에 제안되어지고 있다. 그래서 바이모달 음성인식 시스템을 실제로 구현해보고 인식 실험을 수행해 보았다. 입술영상은 이미지에 근거한 입술모양을 파라메터화하여 인식실험에 사용하였으며 음성과 입술영상을 각각 인식한 후 인식스코어(Score)에 가중치를 적용하여 통합하는 방법을 사용하였다. 마지막으로 바이모달 음성인식의 잡음환경에서의 성능을 알아보기 위해 음성신호에 여러 레벨의 잡음을 섞어서 실험을 하고 잡음환경에서 인식률의 한계를 입술정보를 이용하여 극복할 수 있다는 것을 보이고자 한다.

Ⅰ. 서론

잡음환경에서 인식률의 한계를 극복하기 위해 여러 가지 음성인식 알고리즘들이 많이 제안되어 왔으나, 인식률 향상에는 한계를 가지고 있다. 그래서 음성신호만을 이용한 음성인식의 한계를 극복하고 인간의 인지구조를 흉내낸 멀티모달(Multi-modal) 음성인식이 근래에 제안되어 왔다[1]. 멀티모달 음성인식이란 음성뿐만 아니라 화자의 시선, 턱의 움직임, 몸의 움직임 그리고 입술모양 등 다양한 정보들을 이용해서 인간의 인지구조를 흉내낸 음성인식방법이다. 이중 입술모양은 가장 음성과 관련이 많고 특징적인 요소가 많기 때문에 입술정보를 이용한 바이모달 음성인식방법을 선택하였다[2][3]. 본 논문에서는 이미지에 근거한 입술 특징 파라메터를 추출하는 방법과 추출된 입술정보와 음성정보의 통합방법, 그리고 인식실험을 통한 입술정보의 음성정보와의 관계를 다루고자 한다.

Ⅱ. 바이모달(Bi-modal) 음성인식시스템

1. 시스템구성

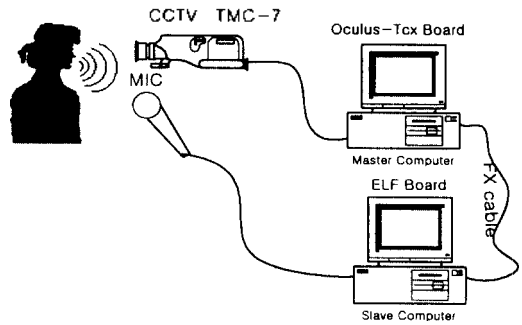

그림 1. 바이모달 음성인식 시스템 구성도

입술영상과 음성을 동시에 저장하기 위해 두 대의 컴퓨터를 사용하여 그림 1과 같이 구성하였다. 마스터 컴퓨터(Master Computer)에서는 입술이미지를 0.11초/1프

래임마다 캡쳐하고 슬레이브컴퓨터(Slave Computer)에서는 마스터컴퓨터에서 FX케이블을 통해서 보내온 동기신호를 이용해서 동시에 음성신호를 저장하도록 구성하였다. 이미지보드는 Oculus-Tcx 보드를 이용해서 CCTV 컬러카메라인 TMC-7으로부터 이미지를 받아서 저장하고 음성은 ASPI(Atlanta Signal Processors, Inc.)회사의 TMS320C31 DSP칩을 내장한 ELF보드를 이용해서 음성을 저장하도록 구성되었다.

2. 음성처리

TMS320C31 DSP칩을 내장한 ELF보드(Board)를 이용하여 8kHz로 샘플링하여 시외진화번호 160개의 지명을 입술영상에 동기를 맞추어서 저장시켰다. 음성파라메터로 12차 LPC(Linear Prediction Coding) 셉스트럼(Cepstrum)계수를 이용해서 고립단어 인식실험에 사용하였다. LPC 셉스트럼 계수를 구하는 방법은 그림 2와 같이 해밍윈도우(Hamming Window)와 Durbin 알고리즘을 이용해서 LPC계수를 먼저 구하고 이 계수를 LPC 셉스트럼 계수로 변환하였다[4].

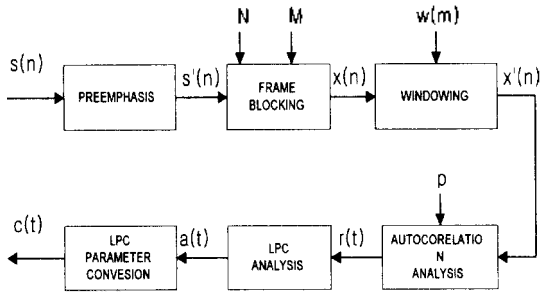

그림 2. 음성파라메터 추출 블럭도

3. 입술영상처리

일정한 실험실 조명아래에서 입술을 화면의 중앙에 위치시키면서 입술영상을 저장하였다. 입술이미지는 1초에 9프레임을 저장시켜서 한번 저장할 때마다 50개의 연속프레임을 100×100크기의 TIFF(Tagged Image File Format)형식의 컬러 이미지파일로 저장시켰다. 한번 연속프레임을 저장할 때마다 단어를 세 번 또는 네 번 발음하여 저장시켜 실험에 이용하였다. 그림 3과 같이 입술 파라메터로는 가장 입모양을 대표적으로 나타낼 수 있는 바깥입술의 높이(H1)와 폭(W1) 안쪽입술의 높이(H2)와 폭(W2)을 입술인식에 이용하였다. 특히 안쪽입술모양이 인식률에 상당히 영향을 미치므로 안쪽 입술모양을 파라메터화 하기 위해서 안쪽입술의 폭을 구한 다음 안쪽입술을 5부분으로 나누어서 각각의 위치에서 입술의 폭(h1, h2, h3, h4, h5)을 구하였다.

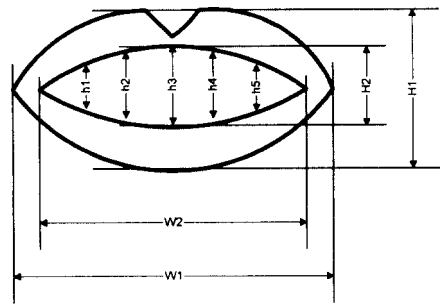

그림 3. 입술 특징 파라메터

먼저 Oculus-Tcx 이미지보드를 통해 획득한 컬러이미지를 흑백영상으로 전환한 다음, 메디안필터(Median Filter)를 이용하여 잡음을 제거하였다. 잡음제거 후 그림 6에서 입술이미지 우측부분은 세로축 색깔분포인 y 프로파일(y profile)의 값을 나타내는데 이 값을 이용해서 입술의 위와 아래쪽을 추출하여 바깥입술의 높이를 계산할 수 있다. 다음 과정으로 그림 5.와 같이 Sobel 윤곽추출자를 이용하여 입술의 윤곽선을 추출한 후 잡음제거 과정을 거치면 그림 5의 오른쪽그림과 같은 이미지를 얻을 수 있다[5]. 이 윤곽선이 추출된 이미지에서 바깥입술의 폭을 계산할 수 있다. 안쪽입술의 높이는 전 과정에서 구해진 안쪽입술의 폭으로부터 입술 중앙값을 계산할 수 있고 이 중앙값으로부터 중앙부근의 부분적인 y 프로파일을 따로 계산해내서 그것의 미분 값을 이용해서 안쪽입술의 높이를 추출할 수 있다. 안쪽입술의 폭은 윤곽선 추출된 이미지를 이용 구할 수 있다.

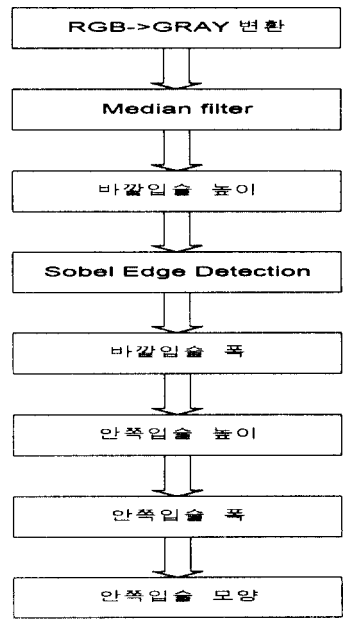

그림 4. 입술파라메터 추출 순서도

그리고 특징적인 안쪽입술모양을 파라메터화 하기 위해서 앞 과정에서 구해진 안쪽입술의 폭을 이용해서 안쪽 입술모양을 5부분으로 나누어서 각각의 안쪽 입술 높이를 측정 파라메터로 만들었다. 이러한 과정을 거쳐서 구해진 파라메터의 모습은 그림 6의 3번째 그림에서 볼 수 있듯이 바깥쪽 입술의 폭과 높이 안쪽 입술의 폭과 높이 그리고 5개의 안쪽 입술모양 파라메터를 이용 모두 9개의 입술특징 파라메터를 인식실험에 이용하였다.

그림 5. Sobel 윤곽추출 후 잡음제거

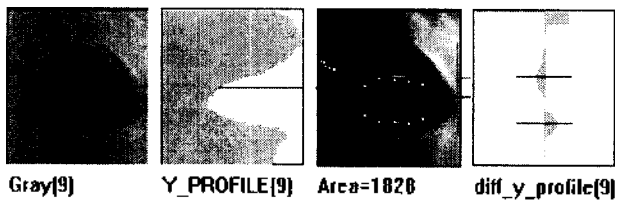

그림 6. 입술파라메터가 추출된 모습

4. 음성정보와 입술정보 결합

위의 과정을 통해서 추출된 음성과 입술 파라메터를 이용해서 인식실험을 하기 위해서는 음성과 입술 파라메터의 발음구간인 시작점과 끝점을 구해야 한다. 먼저 음성의 시작점과 끝점은 에너지(Energy)와 영교차율(ZCR Zero Crossing Rate)를 이용하여 구하였다. 입술파라메터의 시작 프레임과 끝 프레임은 음성 파라메터와 동기를 맞추기 위해서 이미 추출된 음성 파라메터의 음성구간의 시작점과 끝점의 정보를 이용해서 시작프레임과 끝 프레임을 추출하였다. 입술프레임은 50프레임이고 음성은 88064byte이므로 한 프레임 당 44032샘플/50프레임=880.64샘플에 해당한다. 그림 7에서 첫 번째 윈도우는 바깥쪽과 안쪽 입술의 높이변화를 두 번째 윈도우는 바깥쪽과 안쪽 입술 폭의 변화를 보기 쉽게 도시한 그림이다. 추출된 파라메터를 살펴보면 음성이 시작하고 끝날 때의 구간에서 입술 파라메터의 높이와 폭의 변화가 많이 있다는 것을 알 수 있다. 이러한 입술파라메터의 모양들이 각 발음마다 특성이 있으므로 영상 파라메터만의 인식이 가능하다. 입술프레임의 시작 프레임과 끝 프레임, 음성의 시작과 끝점은 두 번째 윈도우와 세 번째 윈도우의 아래쪽과 위쪽에 굵은 선으로 음성구간을 표시하였다.

그림 7. 영상신호의 끝점 검출

위의 과정을 통해서 9개의 입술 특징 파라메터와 12차 LPC 셉스트럼 계수를 이용하여 통합하는 과정은 크게 2가지로 나누어 볼 수 있다. 먼저 파라메터를 합친 후 인식기를 이용해서 인식하는 방법과 영상과 음성부분을 독립적으로 인식기를 이용해서 인식 실험한 후 여기서 나오는 스코어에 가중치를 적용하여 합성하는 방법(그림 8)이 있다. 즉 두 시스템간의 차이는 입술 영상 파라메터와 음성 파라메터를 결합하는 과정이 인식과정보다 전에 있느냐 후에 있느냐에 있다[1].

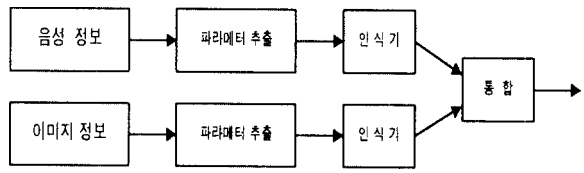

그림 8. 음성과 입술 파라메터를 각각 다른 인식기에 사용한 경우

첫 번째 방법은 인식 전에 입술 영상 파라메터와 음성 파라메터를 형성하는 방법으로 모든 정보이용이 가능하고 성능이 좋으나 입력 데이타가 크므로 인식기가 더 많은 파라메터 데이타를 필요로 하고 관련 시간정보를 알기 위해 음성과 시간의 정확한 동기화가 필요하다. 두 번째 방법으로 음성과 입술 이미지를 각기 다른 파라메터를 형성하여 각각의 인식기에 인식 후 통합하는 방법은 동기화가 잘 이루어지지 않아도 되고 인식속도가 빠르므로 이 논문에서는 두 번째 방법을 이용하였다. 음성과 입술영상의 각각의 인식기를 거쳐 나온 음성정보와 시각정보의 스코어(Score)를 이용해서 전체적인 스코어를 얻는 과정은 식1을 이용하여 계산했다.

$$S_w = k_a S_{wa} + k_v S_{wv} \qquad (k_a + k_v = 1) \qquad (1)$$

S_w : 인식스코어

S_{wa} : 음성스코어, S_{wv} : 시각스코어

k_a : 음성가중치, k_v : 시각가중치

Ⅲ. 실험 및 결과

바이모달 음성인식 실험을 위해서 160개의 시외전화 번호 지역 명을 실험데이터로 사용하였다. 한 지역 명에 대해서 3-4번 발음하여 저장하고 처음 발음한 것은 기준패턴으로 사용하고 나머지패턴들을 테스트패턴으로 사용하였다. 화자종속 고립단어 인식실험을 수행하였고 DTW(Dynamic Time Warping)방법을 이용해서 패턴끼리 비교를 하였다. 영상은 0.1초/1프레임의 속도로 영상을 획득했고 음성은 8kHz로 샘플링하여 저장하였다. 표 1은 백색잡음을 각각 달리 섞어가면서 음성과 시각 가중치를 각각 조금씩 변화시켜 6개의 실험을 한 후 인식률을 살펴본 결과를 나타낸다. 깨끗한 음성에서는 96.23(%)의 인식률을 보이다가 잡음이 많이 섞일수록 음성 파라메터만을 이용한 인식률이 54.78, 26.38, 14.49, 10.14, 7.25(%)로 감소하는 것을 볼 수 있다. 그러나 입술 특징 파라메터만을 이용한 인식실험의 인식률은 46.96으로 일정하게 나타나는 것을 볼 수 있다. 각 실험결과를 시각가중치(α)에 대해서 살펴보면 깨끗한 음성에서는 인식률이 가장 좋은 때는 α=0.3정도이고 SNR이 20dB, 15dB, 10dB, 5dB, 0dB인 경우는 k_c=0.4, 0.7, 0.7, 0.8, 0.9로 값이 변화하였다. 즉 음성에 더 많이 잡음이 섞여질수록 시각가중치를 더 많이 주면 인식률이 높다는 것을 알 수 있다.

또한, 시각가중치의 특정 구간(예 : SNR=10dB에서 $0.5 \le k_c \le 0.9$)에서는 음성만을 이용한 인식률과 입술영상만을 이용한 인식률보다 더 좋은 인식률을 나타내고 있다. 이것은 음성과 영상이 상호 보완관계에 있다는 것을 나타내고 있다. 즉 음성으로 인식이 어려운 부분은 입술영상으로 인식이 되는 부분이 있고 입술영상으로 인식하기 힘든 패턴은 음성으로 인식이 가능한 패턴들이 있다는 것을 나타낸다. 그러한 상호 보완작용에 의해서 음성과 입술영상이 가진 인식률보다 더 좋은 결과를 보이는 시각가중치 구간이 존재한다.

표 1. 잡음정도에 따른 인식률 변화

$\alpha(=k_c)$ (시각가중치)	SNR= 0dB	5dB	10dB	15dB	20dB	clean
1.0	46.96	46.96	46.96	46.96	46.96	46.96
0.9	50.72	52.46	54.78	57.97	61.74	74.20
0.8	49.28	54.20	60.87	68.12	72.46	88.41
0.7	44.06	51.88	62.90	73.04	80.00	93.33
0.6	36.23	46.96	58.26	71.88	83.77	94.78
0.5	30.72	39.42	53.62	68.99	84.93	96.23
0.4	23.48	33.33	44.64	64.06	85.22	96.81
0.3	16.81	24.93	37.10	57.97	83.19	96.81
0.2	13.23	17.39	28.99	48.41	74.20	96.81
0.1	10.14	13.33	20.00	37.68	67.54	96.23
0.0	7.25	10.14	14.49	26.38	54.78	96.23

Ⅳ. 결론

잡음환경에서의 입술정보를 이용한 바이모달(Bi-modal) 음성인식이 실험결과를 통해 기존의 잡음환경에서의 음성만을 이용한 음성인식의 한계를 극복하기 위한 좋은 방법이 될 수 있다는 것을 보였다. 특히 음성과 입술영상을 각각의 인식에 어떻게 가중치를 주느냐에 따라서 더 좋은 인식률을 보이는 구간을 실험을 통해서 살펴보았다. 즉, 바이모달 음성인식은 입술영상의 추가로 입술영상의 인식률만큼의 잡음환경에서 향상이 있고 추가로 입술영상과 음성영상의 상호보완 관계를 이용 더 좋은 인식률을 얻을 수 있다는 것을 알 수 있다. 이러한 관계는 인간의 인지구조를 흉내낸 것으로 음성과 입술영상인식을 상호 독립적이면서 보완관계로서 사용이 가능하다는 것을 보였다. 또한 입술영상뿐만 아니라 잡음환경에서 견인한 음성인식시스템을 위해서 다른 형태의 정보이용에 대한 연구가 계속 되어야 할 것이다.

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

참고문헌

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

[2] K. Venkatesh Prasad, David G. Stork and Gregory J. Wolff "Preprocessing video images for neural learning of lipreading", Ricoh California Research Center, Technical Report CRC-TR-93-26, 1993.

[3] Paul Duchnowski, Uwe Meier and Alex Waibel "SEE ME, HRER ME: INTEGRATING AUTOMATIC SPEECH RECOGNITION AND LIP-READING", Proceedings of the International conference on Spoken Language Processing, Yokoha Japan, Setember 1994.

[4] Lawrence Rabiner, Biing-Hwang Juang "FUNDAMENTALS OF SPEECH RECOGNITION", PTR Prentice-Hall, 1993.

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