논문-03-08-2-10

음성/영상 정보를 이용한 새로운 끝점추정 방식에 기반을 둔 음성인식 시스템

이 동 근*, 김 성 준*, 계 영 철*

A Speech Recognition System based on a New Endpoint Estimation Method jointly using Audio/Video Informations

Dong-Keun Lee*, Seong-Jun Kim*, Young-Chul Kay*

요 약

본 논문에서는 멀티미디어 데이터에 존재하는 입술의 움직임(영상언어)과 음성을 함께 이용하여 음성의 끝점을 정확히 추정하는 방법과 이를 기반으로 한 음성인식 시스템을 제안한다. 잡음 섞인 음성의 끝점추정 방법은 다음과 같다. 각 테스트 단어에 대하여 영상언어를 이용한 끝점과 깨끗한 음성을 이용한 끝점을 각각 구한 후 이것들의 차이를 계산한다. 이 차이에 영상언어 끝점을 더하여 잡음 섞인 음성의 끝점으로 추정한다. 이와 같은 끝점(즉, 음성구간)의 추정방법을 인식기에 적용한다. 동일한 구간의 음성이 인식기의 각 단어모델에 입력되는 기존의 인식 방법과는 달리, 새로운 인식기에서는 각 단어별로 추정된 서로 다른 구간의 음성이 각 해당 단어모델에 입력된다. 제안된 방식을 모의실험 한 결과, 음성잡음의 크기에 관계없이 정확한 끝점을 추정 할 수 있었으며, 그 결과 약 8% 정도의 인식률 향상을 이루었다.

Abstract

We develop the method of estimating the endpoints of speech by jointly using the lip motion (visual speech) and speech being included in multimedia data and then propose a new speech recognition system (SRS) based on that method. The endpoints of noisy speech are estimated as follows: For each test word, two kinds of endpoints are detected from visual speech and clean speech, respectively. Their difference is made and then added to the endpoints of visual speech to estimate those for noisy speech.

This estimation method for endpoints (i.e. speech interval) is applied to form a new SRS. The SRS differs from the conventional one in that each word model in the recognizer is provided an interval of speech not identical but estimated respectively for the corresponding word. Simulation results show that the proposed method enables the endpoints to be accurately estimated regardless of the amount of noise and consequently achieves 8 % improvement in recognition rate.

I. 서 론

종래의 음성인식 시스템에서 음성신호의 정확한 끝점(endpoints) 검출은 인식 시스템의 성능을 크게 좌우한다. 끝점 검출은 입력 신호로부터 음성구간을 검출하는 과정이다. 음성구간의 검출을 위하여 현재 사용중인 끝점 검출 알고리즘의 대부분은 Rabiner가 제안한 프레임 에너지와 영교차율(Zero-Crossing Rate)의 조합을 바탕으로 하고 있다[1]. 그러나 이 알고리즘은 신호 대 잡음비(SNR)가 작을 경우나 에너지가 큰 비음성을 포함하는 경우 신뢰성 있는 경계점을 얻기에 불충분하다. 또한, 음성의 시작이나 끝 부

* 홍익대학교 전자공학과
Hongik University, Dept. of Electronic Engineering

분에 존재하는 파열음이나 마찰음의 경우는 신호의 에너지가 유성음 구간에 비해 작기 때문에 잡음환경에서 검출되기가 용이하지 않으며 끝점 검출 실패의 주요한 원인 중 하나가 된다.

이러한 문제점을 해결하기 위하여 멀티미디어 데이터에 존재하는 영상정보와 음성정보를 함께 이용하는 방법을 제안한다. 입술영상의 움직임을 분석하면 발음된 단어의 시작과 끝뿐만 아니라 그 의미도 인지할 수 있다. 본 논문에서는 이러한 입술영상의 움직임을 영상언어(visual speech)라고 한다. 먼저, 각 테스트 단어들에 대하여, 영상언어에서 검출한 끝점과 깨끗한 음성으로부터 검출한 끝점의 차이 값을 각각 계산하여 테이블을 작성한다. 본 논문에서는 이 테이블을 DIS 테이블이라 칭한다.

잡음 섞인 음성신호의 끝점을 추정하기 위하여 영상정보로부터 검출한 끝점에 DIS 테이블에 저장되어있는 (영상끝점과 음성끝점의 평균차이인) 평균 끝점거리를 더한다. 이러한 끝점(즉, 음성구간) 추정방법을 인식기에 적용한다. 입력영상과 DIS 테이블을 이용하여 각 단어별 예상 음성구간을 추정한 후, 인식기의 각 단어모델에 해당 구간의 음성을 입력하여 모델별 인식 스코어를 구한다. 출력된 인식 스코어들 중 가장 높은 값을 갖는 모델을 인식 결과로 출력한다.

Ⅱ. 본 론

정확한 인식을 위해서는 전처리로서 음성의 구간을 정확히 검출하여야한다. 본 논문에서는 음성구간의 정확한 검출을 위하여 음성 끝점검출 방식과 입술모양의 영상 변화로부터 음성구간을 검출하는 영상언어 끝점검출 방식을 함께 이용한다. 음성 끝점 검출을 위한 알고리즘으로 Rabiner가 제안한 프레임 에너지와 영교차율을 이용하였다[1][5].

1. 영상 언어의 끝점 검출

입력 영상으로부터 입술영역을 판별한 후 입술영역에서 16개의 경계점을 추출하고 이것들을 이용하여 입술을 포물선으로 모델링 한다. 본 논문에서는 입술 포물선의 높이를 음성구간 검출을 위한 특징으로 사용한다[2][3][4].

입술의 높이는 입을 다문 상태에서도 사람에 따라 그리고 영상의 크기에 따라 변하므로 이의 비교를 위해서는 정규화(normalization)을 하여야한다. 입술 동영상의 첫 번째 프레임은 입술이 닫혀있는 상태라고 무리없이 가정할 수 있으므로, 첫 번째 프레임에서 검출한 입술의 높이로 나머지 프레임의 높이들을 정규화한다. 따라서 정규화된 높이가 1이면 입술이 닫혀있는 상태이고, 1보다 커지면 열려있는 상태라고 가정할 수 있다. 실험결과 문턱치(threshold)가 1.1 정도이면 영상언어의 발음 구간을 검출할 수 있음을 확인하였다. 그림 1은 영상 프레임의 진행에 따른 입술높이의 변화를 나타낸다. 약 8번째 프레임에서 시작점이, 그리고 59번째 프레임에서 끝점이 검출되었으며, 그 사이를 영상언어 구간이라 할 수 있다.

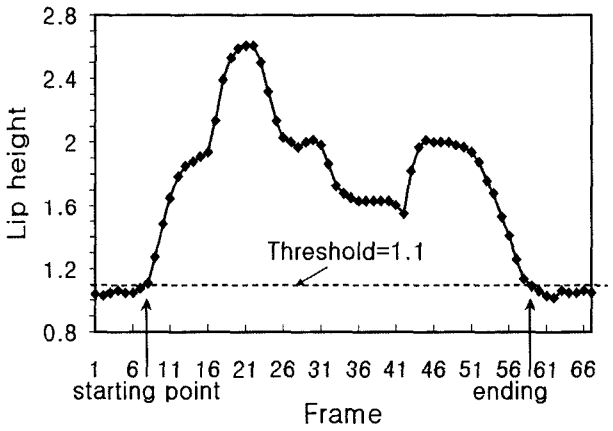

그림 1. 영상언어의 끝점검출
Fig. 1. Endpoints detection of visual speech

2. 새로운 음성구간 검출 방식

음성에 잡음이 심한 경우에는 음성의 끝점검출이 상당히 부정확하게 된다고 알려져 있다[1]. 본 절에서는 영상정보를 이용하여 이러한 문제를 극복하는 방법을 제시한다.

2.1 음성과 영상언어 사이의 평균 끝점거리 *DIS* 추정

잡음 없는 상태에서 녹음된 테스트 단어들에 대해 기존의 음성끝점 검출방식에 의하여 얻은 음성의 시작점과 끝점을 각각 A_{Start}와 A_{End}라 하고, Ⅱ.1절의 영상언어의 끝점검출 방식에 의하여 얻은 시작점과 끝점을 각각 V_{Start}

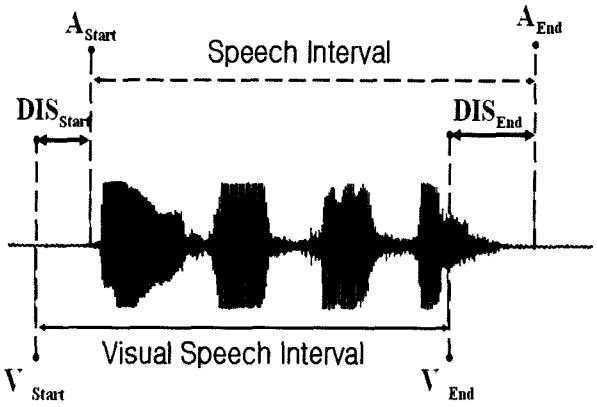

그림 2. 음성과 영상으로부터 검출된 음성구간
Fig 2. Speech intervals determined from audio and video

와 V_{End}라고 한다. 그림 2는 테스트 단어 '0483' 에 대한 이들 사이의 관계를 보여주고 있으며, 이들 관계는 테스트 단어에 따라 달라진다.

식 (1)을 이용하여 음성과 영상언어의 시작점 사이의 거리 DIS_{Start}와 끝점 사이의 거리 DIS_{End}를 구하고 학습용 테스트 단어들을 이용하여 평균 거리를 구한다.

$$DIS_{Start} = A_{Start} - V_{Start}$$
$$DIS_{End} = A_{End} - V_{End} \quad (1)$$

그림 3은 DIS를 구하는 블록도를 나타내고 있다. 선형보간에 대한 설명은 Ⅲ.1절에서 다룬다.

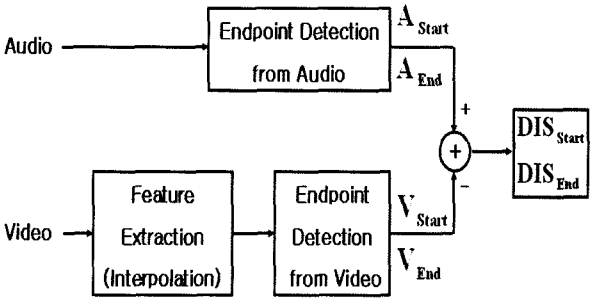

그림 3. DIS 결정을 위한 구성도
Fig 3. Block Diagram for DIS

각 테스트 단어별 DIS를 미리 구한 후 이것들을 시스템에 저장하여 음성끝점을 추정하는데 사용한다. 표 1은 본 연구에서 사용한 10개의 단어에 대한 DIS_{Start}와 DIS_{End}를 나타낸다.

표 1. DIS TABLE
Table. 1. (단위:ms)

Test Words	DIS_{Start}	DIS_{End}
0483	200.4	204.9
1354	96.2	-44
1843	99.1	224.1
4037	93.3	-68.6
4152	99.4	-42.7
5738	194.8	-11.5
6374	121.1	-48.3
7351	147.3	-39.7
8649	27.9	-18
9208	242.6	-17.6

2.2 예상 음성단어의 끝점 추정

음성의 SNR이 낮은 경우에는 음성으로부터 구한 끝점검출값이 상당히 부정확하므로 이것 대신에 영상정보를 기준으로 하여 끝점을 추정하여 사용한다[그림 4].

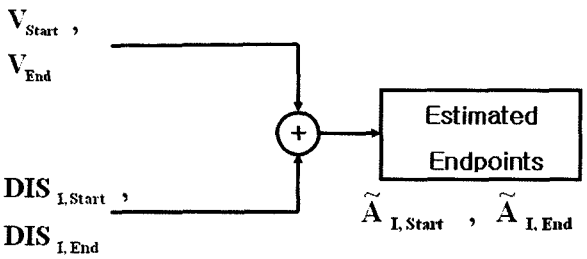

그림 4. 예상 단어 I의 끝점 추정
Fig 4. Estimation of Endpoints for an expected Word I

음성과 영상으로 구성된 단어가 입력되면 영상으로부터 영상언어의 끝점검출 과정(Ⅱ.1절)을 거쳐 V_{start}와 V_{end}를 구하고, 미리 테이블로 만들어져있는 예상단어 I에 대한 DIS_I를 여기에 더하여 $A_{I,start}$와 $A_{I,end}$를 추정한다. 그러나 이러한 방법을 적용하기 위해서는 단어 I가 정확히 예상되어야 하므로 끝점검출 목적만으로는 이것을 사용할 수 없다. 따라서 본 논문에서는 이러한 방법을 인식기에 적용하여 인식률을 향상시키는 방법을 제안한다.

3. 제안된 음성인식 시스템

그림 5는 기존의 음성인식 시스템을 나타내고 있다. 입력음성으로부터 끝점검출에 의한 음성구간을 검출한 후, 검출된 구간만의 음성이 각 단어모델에 동시에 전달되어 인식되도록 되어 있다. 이 경우 잡음에 의하여 음성구간이 부정확하게 검출되면 인식결과에 오류가 생기게된다.

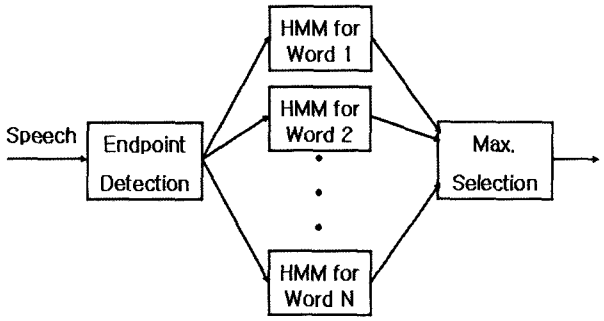

그림 5. 기존의 음성인식 시스템
Fig. 5. Conventional Speech Recognition System

그림 6은 제안된 인식 시스템으로서 이와 같은 문제를 극복할 수 있다. 영상으로부터 영상언어 구간을 검출한 후 이것을 기반으로 하여 각 단어에 대한 예상 음성구간을 추정한다(그림 4). 음성입력이 각 단어의 HMM 모델에 입력되면 단어별로 추정된 예상 음성구간에 한하여 인식을 행한다. 각 단어의 HMM 모델에 연결되는 음성구간의 추정치가 서로 다름에 유의한다. 만일 입력단어가 I 라면, 단어모델 I에 대한 음성구간 추정값은 정확할 것

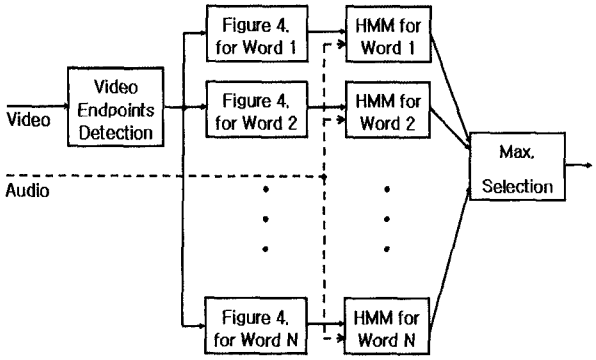

그림 6. 제안된 음성인식 시스템
Fig. 6. Proposed speech recognition system

이고 따라서 높은 인식 스코어가 나올 것이다. 반면에 그 외의 단어모델에 대해서는 음성구간 추정이 잘못되어 도리어 낮은 인식 스코어가 출력될 것이다. 결과적으로 인식스코어의 최대값과 기타값들의 차이가 커져 안정적으로 인식이 될 수 있다.

Ⅲ. 실험 및 결과

1. 실험 조건

디지털 캠코더를 이용하여 음성과 영상신호를 동시에 획득하였으며, 10명의 화자가 총 4자리 연속 숫자 10가지를 각각 10번씩 반복 발음하여 총 1000개로 구성된 음성-영상 데이터베이스를 구축하였다. 음성은 실험실 환경에서 16bit 양자화, 16kHz 샘플링 주파수로 녹음되었으며, 12차 LPC-켑스트럼 계수를 사용하였다. 영상은 디지털 캠코더로 촬영한 320×240 픽셀, 30 프레임/s, 24-bit RGB 컬러 이미지이며, 영상언어의 끝점을 좀더 정확히 검출하기 위해 60 프레임/s로 선형 보간하여 사용하였다[5]. 인식 알고리즘은 코드북 사이즈 256, state수 8인 discrete HMM을 이용하였다.

2. 실험 결과

입력단어가 정확히 예상되는 단어모델 I 에서 음성과 영상을 함께 이용한 끝점 추정의 성능을 비교하기 위해, 식 (2)와 같이 알고리즘으로 추정한 끝점(그림 4)에서 음성신호로부터 수작업으로 검출한 끝점을 뺀 차이를 오차로 사용하였다.

$$\text{오차} = \text{알고리즘 추정 끝점} - \text{수작업 검출 끝점} \quad (2)$$

그림 7과 8은 제안된 끝점추정방식, 기존의 음성 끝점 검출방식 (clean,20dB,10dB,0dB), 그리고 영상만을 이용한 끝점검출 방식을 적용하였을 경우에 시작점과 끝점에 발생하는 오차의 분포를 각각 나타낸다. 가로축은 오차를, 그리고 세로축은 그 오차가 나타난 학습단어들의 수를 의미한다. 기존의 음성 끝점 검출 알고리즘을 이용하면 시작점을

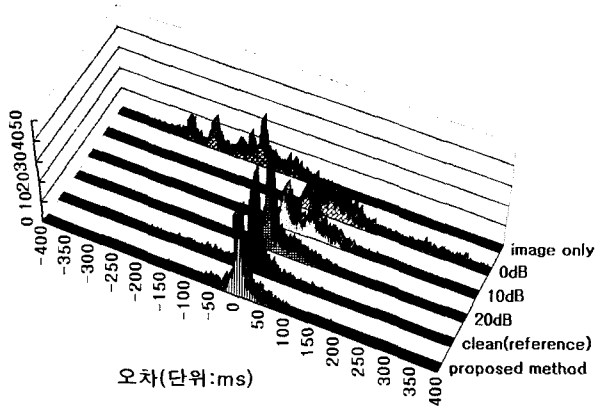

그림 7. 시작점 오차 분포도
Fig. 7. Distribution of errors in starting points

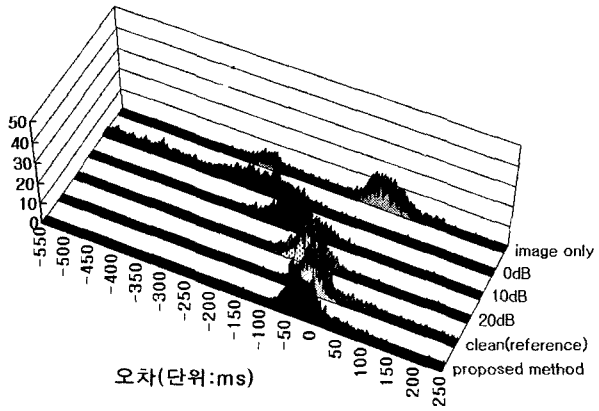

그림 8. 끝점 오차 분포도
Fig. 8. Distribution of errors in ending points

검출하는 경우 SNR이 감소함에 따라 기준이 되는 깨끗한 음성에 비해 시작점이 (+)방향으로 검출되며, 끝점의 경우 (-)방향으로 검출됨을 알 수 있다. 영상만을 이용하는 경우에는 시작점은 기준값(참값) 보다 앞에서 분포되는 반면 (그림 7), 끝점은 앞과 뒤 두 영역으로 나누어져서 분포된다(그림 8). 앞쪽으로 분포되는 끝점은 실험에서 사용된 테스트 단어 '0483'과 '1843'에 의한 것으로 숫자 '3'을 발음하면 발음이 끝나기전에 입술이 먼저 닫히기 때문이다.

그림 7과 8에서 확인하였듯이, 기존의 끝점 검출 알고리즘을 이용하여 끝점을 검출하는 경우, 잡음이 심한 환경에서의 음성구간 검출이 정확하지 못함을 알 수 있다. 그러나, 제안된 방법을 사용하면 참값에 비하여 거의 오차없이 신뢰성 있게 음성구간을 검출함을 알 수 있다.

앞서 언급한 음성+영상 끝점검출 방식이 인식기의 성능에 미치는 영향을 분석하기 위하여 음성의 SNR을 변화시키면서 기존 인식기와 제안된 인식기의 성능을 비교하였다(표 2). SNR이 높은 경우에는 인식률에 차이가 없으나, 25dB, 20dB 에서는 약 8% 정도의 인식률이 향상되었다. 그러나 그 이하의 SNR에서는 음성구간의 정확성 보다는 음질이 성능을 좌우하므로 인식성능의 향상을 얻지 못하였다.

표 2. 인식률의 비교
Table 2. Comparison of recognition rates (단위 :%)

	clean	40dB	30dB	25dB	20dB
기존의 인식방법	100	100	95	68	36.2
제안된 인식방법	100	100	95	76.7	45

Ⅳ. 결 론

본 논문에서는 음성정보와 영상정보를 함께 이용하는 새로운 끝점추정 방법에 기반을 둔 음성인식 시스템을 제안하였다. 실험결과 제안된 음성+영상 끝점 추정 방법은 음성신호만을 이용한 끝점검출 방법에 비하여 잡음크기에 무관하게 음성구간을 검출할 수 있음을 보였다. 이러한 방법으로 추정한 음성구간을 제안된 인식시스템에 적용하였을 경우 인식률이 약 8% 정도 향상됨을 알 수 있었다.

본 논문에서는 고립단어의 인식만을 다루었다. 연결단어나 가변어휘의 경우에 대한 본 방식의 적용에 관한 연구가 현재 진행 중이다.

참 고 문 헌

[1] L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances," Bell Syst. Tech. J., Vol. 54, No.2, February 1975

[2] G. Potamianos, H.P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading", Proc. Int. Conf. Image Process., Chicago, 1998.

[3] Juergen Luettin, Neil A. Thacker and S.W.Beet, "Locating and Tracking Facial Speech Features", Proceedings of ICPR'96 1996.

[4] 이철우, 계영철, 고인선, "강인한 음성인식을 위한 이중모드 센서의 결합방식에 관한연구", 한국음향학회 논문지, 제 20권, 제 6호, PP. 51-56, 2001.

[5] T. Wark and S. Sridharan, " A syntactic approach to automatic lip feature extraction for speaker identification", Proceedings of the IEEE. 1998.

저 자 소 개

이 동 근

- 2001년 2월 : 호서대학교 제어계측공학과 학사
- 2003년 2월 : 홍익대 전자공학과 석사
- 2003년 7월~현재 : GPS Korea 근무
- 주관심분야 : 음성 및 영상인식, 디지털 신호처리

김 성 준

- 1998년 2월 : 동신대학교 전기전자공학과 학사
- 2000년 2월 : 홍익대학교 전자공학과 석사
- 2000년 2월~현재 : 홍익대학교 전자공학과 박사과정
- 주관심분야 : 음성 및 영상인식, 화자인식, 디지털 신호처리

계 영 철

- 1980년 2월 : 서울대학교 전자공학과 학사
- 1982년 2월 : 한국과학기술원 전기 및 전자공학과 석사
- 1991년 5월 : Univ. of Southern California, Electrical Eng. Ph.D.
- 1991년 9월~현재 : 홍익대학교 전자전기공학부 부교수
- 주관심분야 : 디지털 신호처리, 음성 및 영상인식, 로봇 비젼