論文2003-40CI-5-6

잡음환경에서의 바이모달 시스템을 위한 견실한 끝점검출

(Robust Endpoint Detection for Bimodal System in Noisy Environments)

吳 炫 和 *, 權 洪 錫 *, 孫 宗 睦 *, 秦 成 一 *, 裵 建 星 **

(Hyun-Hwa Oh, Hong-Seok Kwon, Jong-Mok Son, Sung-Il Chien, and Keun-Sung Bae)

요 약

음성인식 시스템과 입술독해 시스템을 결합한 하여 음향학적 잡음에 대하여 안정된 성능을 갖는 바이모달(bimodal) 시스템을 구현한다. 바이모달 시스템의 성능은 두 인식 시스템의 성능뿐만 아니라 입력 신호의 끝점검출 성능에도 크게 영향을 받는다. 본 논문에서는 음성신호와 영상신호에서 끝점을 각각 자동 검출하여 입력 음성신호로부터 음성신호에서 추정한 신호대잡음비(signal-to-noise ratio: SNR)로 두 끝점검출 결과를 선택하는 방법을 제안한다. 즉 낮은 SNR에서는 영상신호로부터 검출된 끝점을 선택하고 높은 SNR에서는 음성신호로부터 검출된 끝점을 선택함으로써 음향학적 잡음에 대하여 견실하게 끝점을 검출한다. 제안한 끝점검출 방법이 적용된 바이모달 시스템이 강한 음향학적 잡음에 대하여 만족스러운 인식성능을 나타냄을 실험결과에서 확인할 수 있다.

Abstract

The performance of a bimodal system is affected by the accuracy of the endpoint detection from the input signal as well as the performance of the speech recognition or lipreading system. In this paper, we propose the endpoint detection method which detects the endpoints from the audio and video signal respectively and utilizes the signal-to-noise ratio (SNR) estimated from the input audio signal to select the reliable endpoints to the acoustic noise. In other words, the endpoints are detected from the audio signal under the high SNR and from the video signal under the low SNR. Experimental results show that the bimodal system using the proposed endpoint detector achieves satisfactory recognition rates, especially when the acoustic environment is quite noisy.

Keywords : 끝점검출, 바이모달 시스템, 입술독해, 음성/영상 데이터베이스

I. 서 론

인간의 의사소통은 음향정보와 더불어 화자의 입술 모양이나 표정, 제스처 등의 시각정보가 결합된 바이모달 인식에 의하여 이루어진다. 특히 입술독해(lipreading)는 청각장애자의 의사소통에 있어서 중요한 보조수단이 되며[1], 음향학적 잡음의 영향에 매우 둔감한 특성때문에 잡음이 많이 존재하거나 화자간의 음성

* 學生會員, ** 正會員, 慶北大學校 電子電氣工學部
(School of Electronic and Electrical Engineering, Kyungpook National University)

※ 본 연구는 한국과학재단 목적기초연구(R01-1999-000 00233-0) 지원으로 수행되었습니다.

接受日字:2003年3月19日, 수정완료일:2003年8月11日

간섭(cocktail party effect)이 존재하는 환경에서 의사소통을 위한 유용한 보조수단이 됨은 이미 잘 알려진 사실이다[2]. 그러므로 최근 들어 음향학적 잡음이 많은 환경에서 음성인식 시스템의 성능 저하를 개선시키기 위하여 입술독해를 결합한 바이모달 시스템을 구현하고자 하는 연구가 진행되고 있다[3-6]. 현재까지 바이모달 시스템에 대한 연구의 대부분은 음성 및 영상 특징벡터 추출, 인식기 성능 개선, 그리고 음성인식 및 입술독해 시스템의 결합방법에 집중되어 왔다. 그러나 입력신호의 시작점(beginning point)과 끝점(ending point)을 정확하고 견실하게 검출하는 것이 바이모달 시스템의 전체 성능에 큰 영향을 미침에도 불구하고 이에 대한 연구는 미진하다.

음성인식 시스템에서는 음성신호의 시작점과 끝점을 검출하는 많은 연구가 진행되어 왔다. 대표적인 끝점검출 방법에는 Rabiner 등의 에너지와 영교차율을 이용한 방법[7], Lamel의 레벨등화기(level equalizer)를 이용한 방법[8], 그리고 Teager 에너지를 이용하는 방법[9] 등이 있다. 이러한 방법에서는 SNR이 큰 경우에 우수한 성능을 나타내지만, 음향학적 잡음이 강한 환경에서는 신뢰성 있는 끝점을 검출하기에 불충분하다. 그러므로 다양한 음향학적 잡음이 존재하는 바이모달 시스템에 이와 같은 끝점검출 방법을 직접 적용하는 데에는 한계가 있다. 바이모달 시스템의 입력 영상신호는 음향학적 잡음에 의한 영향에 둔감하므로 닫힌 입술영상을 이용한 끝점검출 방법을 제안한다. 그러나 영상신호에 기반한 끝점검출 방법은 음향학적 잡음이 적은 경우 음성신호에서 검출된 끝점결과와 비교하여 정확성이 떨어지므로 바이모달 시스템의 인식률을 저하시키는 경향이 있다. 그러므로 본 논문에서 구현된 바이모달 시스템에서는 음향학적 잡음에 대하여 견실하고 정확하게 끝점을 검출하기 위하여 두 신호로부터 각각 검출된 끝점을 입력 음성신호로부터 추정된 SNR에 따라 선택하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. II장에서는 바이모달 시스템의 끝점검출 및 인식성능 평가를 위하여 구축된 음성/영상 데이터베이스에 대하여 기술하고, III장에서는 구현된 두 시스템을 결합한 바이모달 시스템에 대하여 서술한다. IV장에서는 음성신호와 영상신호, 그리고 입력 음성신호의 SNR에 따른 끝점검출 방법에 대하여 자세히 설명한다. V장에서는 음성 SNR에 따른 끝점검출 방법을 적용한 바이모달 시스템의 인식 결과를 평가한다. 마지막으로 VI장에서 본 논문의 내용을 요약하고 결론을 맺도록 한다.

II. 음성/영상 데이터베이스

바이모달 시스템의 끝점검출 및 인식 성능을 평가하기 위하여 동기화된 음성 및 영상신호를 획득하여 음성/영상 데이터베이스를 구축하였다. 영상 및 음성신호의 동기를 맞추기 위하여 실험실 환경에서 한 대의 개인용 컴퓨터를 이용하여 두 신호를 동시에 획득하였다. 음성신호는 TMS320C6201 DSP EVM과 THS1206 ADC EVM으로 구성된 4채널 마이크로폰 배열을 이용하여 채널당 100kHz로 샘플링되었다. 이와 동시에 입술 동영상은 CCD카메라와 영상획득 보드를 이용하여 30frames/sec로 획득되었다. 동영상의 각 프레임은 320×240크기의 8-bit 그레이레벨 영상이다. 인식 대상은 음향학적 잡음이 많이 존재하는 역무 자동화에 대한 응용을 고려하여 전국의 20개 주요 기차역명으로 선정하였다. 총 20개 고립단어를 20명의 화자가 각각 10번씩 발음하여 총 4000개의 데이터로 구성된 음성/영상 데이터베이스를 구축하였다. 본 논문에서는 고립단어 발음전후에 나타나는 입술모양 및 움직임의 개인적 편차를 고려하여 입술이 닫혀있는 상태를 입술 동영상에서의 시각적 묵음구간으로 정의한다. 따라서 데이터 획득 과정에서 화자가 입술을 닫은 상태에서 발음을 시작하여 발음 후 입술을 다시 닫도록 하였으며, 1회 발성에 대하여 3초 길이의 음성 및 동영상 데이터를 수집하였다. <표 1>은 데이터베이스의 구성을 나타내며 <그림 1>은 수집된 음성/영상 데이터베이스에서 '서울'을 발음한 한 화자의 입술 동영상과 4채널 음성신호의 예를 나타낸다.

표 1. 음성/영상 데이터베이스의 구성
Table 1. Composition of audio/visual database.

대상단어	총 20 단어: 서울, 영등포, 수원, 천안, 조치원, 대전, 영동, 김천, 구미, 동대구, 밀양, 부산, 논산, 익산, 정읍, 광주, 목포, 전주, 남원, 순천
화자	남자: 11명 여자: 9명
화자당 발성횟수	학습용 발성: 7회 테스트용 발성: 3회

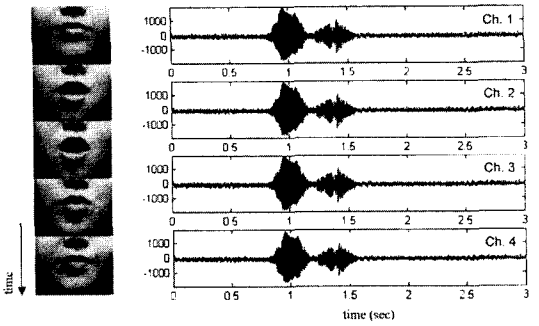


그림 1. '서울'에 대한 그레이레벨 입술 동영상과 4채널 음성신호로 구성된 음성/영상 데이터의 예
Fig. 1. Example of audio/visual data composed of gray-level lip image sequence and 4-channel audio signals for 'Seoul [soul]'.

Ⅲ. 바이모달 시스템

본 논문에서는 영상신호와 4채널의 음성신호를 입력으로 하는 바이모달 시스템을 <그림 2>와 같이 구현한다. 먼저, 음성인식 시스템의 전처리 과정으로써 4채널 음성신호를 이용하여 MMSE-STSA(minimum mean square error short-time spectral amplitude)에 기반한 음성개선기법을 적용한다[10, 11]. 이 과정을 통하여 끝점 검출 및 인식에 오류를 발생시킬 수 있는 음향학적 잡음의 영향을 감소시킬 수 있다. 음성신호는 preemphasis 계수 0.95로 전처리한 후, 25.6ms 길이의 해밍 윈도우를 11.1ms 간격으로 중첩하여 분석한다. 각 구간에서 1차의 에너지와 13차의 멜켑스트럼(mel-frequency cepstral coefficient: MFCC)을 구하고, 현재 구간을 포함한 전후 각 3구간의 정보를 이용하여 1차의 차분 에너지와 13차의 차분 MFCC을 구한다. 따라서 음성신호로부터 총 28차의 음성 특징벡터를 추출하여 음성인식 시스템에서 사용한다.

입술독해를 위한 음성의 시각적 특징은 특정한 시점에서의 입술 모양이나 영상의 명암 정보보다는 발음하는 동안 입술의 동적인 움직임에 많이 포함되어 있다. 그러므로 본 논문에서는 입력영상의 복잡한 공간적 특징을 효과적으로 반영하는 그물 망(mesh)을 이용하여 시간에 따른 입술 움직임 특징을 추출한다[12]. 이를 위하여 화자마다 입술크기가 다른 점과 발음하는 동안 입술이 벌어지는 정도를 고려하여 입술이 포함된 최적의 관심 영역(region of interest: ROI)을 동영상의 각 프레임에서 검출한다. 그물 망 특징벡터는 ROI의 위와 아래 높이를 각각 $n/2$ 등분하고 폭을 n 등분하여 추출된다. 일반적으로 그레이레벨 영상의 그래디언트가 조명에 덜 민감한 특징이 있다. 그러므로 그물 망 내의 그래디언트 크기의 평균을 계산하여 i번째 프레임의 특징벡터 $\mathbf{f}^i$를 식 (1)과 같이 구성한다.

$$\mathbf{f}^i = [\bar{g}_1^i \ \bar{g}_2^i \ \cdots \ \bar{g}_M^i]^T, \quad i = n_b, n_b+1, \cdots, n_e,$$
$$\bar{g}_j^i = \frac{1}{N_j}\sum_{l=1}^{N_j} \| \nabla I_{j\,l}^i \| \tag{1}$$

여기서 n_b와 n_e는 발음구간의 시작점과 끝점을 의미한다. 또한 $\bar{g}_j^i$는 i번째 프레임에서 j번째 그물 망 내의 평균 그래디언트 크기이며, M과 N_j는 ROI 내의 그물 망의 개수와 j번째 그물 망 내의 픽셀 수를 각각 나타낸다. 실험에서는 8×8의 그물 망을 이용하여 추출된 64차 특징벡터를 입술독해 시스템에 사용하였다.

본 논문에서는 <그림 2>에 나타낸 바와 같이 각각 독립적으로 학습된 음성인식 시스템과 입술독해 시스템을 결합하기 위하여 인식 후 결합방식[5]을 사용한다. 두 인식 시스템은 동일한 구조의 간단한 left-to-right 연속 HMM을 기반으로 구현된다. 인식 후 결합방식에서는 음성인식 시스템과 입술독해 시스템의 HMM 출력 확률값 $S_{ia} = \log P(V_a|\Lambda_{ia})$과 $S_{iv} = \log P(V_v|\Lambda_{iv})$을 각각 [0, 1]로 정규화한 확률값 S'_{ia}과 S'_{iv}을 구한다. 그리고 식 (2)와 같이 영상 가중치 α를 S'_{ia}과 S'_{iv}에 적용하여 계산된 S_i가 최대가 되는 클래스 c^*를 최종 인식결과로 결정한다. 여기서 V와 Λ는 특징벡터와 인식모델을 각각 나타내며, 영상 가중치 α는 음성신호의 잡음 정도에 따라 가변적으로 결정된다.

$$S_i = \alpha S'_{iv} + (1-\alpha) S'_{ia}$$
$$c^* = \arg\max_i S_i \tag{2}$$

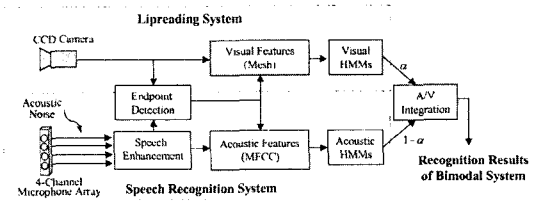


그림 2. 바이모달 시스템의 전체 개요도
Fig. 2. Overall block diagram of bimodal system.

Ⅳ. 바이모달 시스템에서의 끝점검출

1. 음성신호를 이용한 끝점검출

에너지 기반의 끝점검출 방법은 음성신호에서 널리 사용되는 방법 중의 하나이다. Teager 에너지는 신호 크기의 자승뿐만 아니라 진동주파수의 자승을 포함한다[1, 9]. 따라서 기존의 에너지 측정 방법과 달리 신호의 진폭 및 주파수의 변화에 빠른 응답이 가능하다는 특징이 있다. 일반적으로 마찰음과 파열음은 유성음에 비해서 낮은 진폭을 가지지만 높은 주파수에 많은 에너지가 분포한다. 평균자승에너지를 이용하면 이러한 마찰음과 파열음을 검출하기 어려운 반면, Teager 에너지를 이용하면 낮은 진폭을 가지더라도 고주파 특성을 에너지로 변환하게 되어 끝점검출이 용이하다. 그러므로 본 논문에서는 Teager 에너지를 이용하여 음성신호에서 발음구간을 검출한다. 고립단어에 대한 음성신호의 시작부분이 묵음구간이라고 가정하고 처음 6프레임의 Teager 에너지의 평균을 구하여 시작점과 끝점에 대한 문턱치를 정한다. 그리고 각 프레임별로 산출된 Teager 에너지를 문턱치와 비교하여 시작점 n_b^a과 끝점 n_e^a을 결정한다.

2. 영상신호를 이용한 끝점검출

현재까지 입술독해 시스템에서 영상신호를 이용한 끝점검출에 대한 연구는 매우 미진하다. 본 논문에서는 <그림 3>에 나타낸 바와 같이 닫힌 입술영상의 그래디언트 크기를 이용하여 발음구간의 시작점과 끝점을 검출한다.

Ⅱ장에서 서술한 바와 같이 동영상 데이터를 획득하는 과정에서 화자가 입술을 닫은 상태에서 발음을 시작하도록 제약을 두었으므로 동영상의 처음 5프레임까지 입술이 닫혀있는 시각적 묵음구간이 존재한다. 그러므로 입력 동영상의 처음 5프레임에서 추출한 ROI 영상들의 그래디언트 크기를 평균하여 묵음구간에 해당하는 닫힌 입술의 평균 그래디언트 크기 영상(average gradient magnitude image: AGMI)을 식 (3)과 같이 구한다.

$$\|\nabla I_k\|_{avg} = \frac{1}{5}\sum_{i=1}^{5}\|\nabla I_k^i\| \tag{3}$$

여기서, $\|\nabla I_k^i\|$는 i번째 프레임의 ROI 내에서 구한 k번째 픽셀의 그래디언트 크기를 나타낸다. 닫힌 입술의 AGMI와 6번째 이후 프레임에서 추출된 ROI의 그래디언트 크기 영상과의 Euclidean 거리는 두 영상간의 상이한 정도(dissimilarity)를 나타낸다. 그러므로 발음구간에서 산출된 Euclidean 거리 값은 입술이 닫혀있는 묵음구간에서 산출된 값보다 크다. 동영상의 전 프레임에서 산출된 1차원 거리 시퀀스에 median 필터를 적용하여 급격한 변화부분을 제거한 후 거리의 평균을 산출하여 문턱치를 결정한다. 그리고 문턱치를 이용하여 발음구간의 시작점 n_b^v과 끝점 n_e^v을 검출한다.

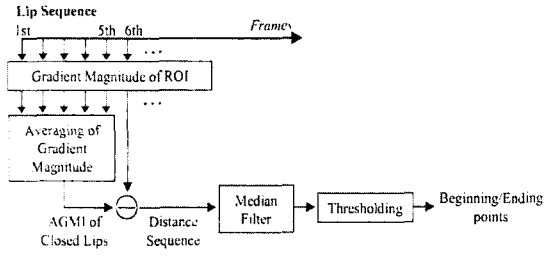


그림 3. 입술 동영상으로부터 발음구간의 시작과 끝점 검출 과정

Fig. 3. Block diagram for endpoint detection from lip image sequence.

3. 음성 및 영상신호에서 검출된 끝점 선택

'동대구'에 대한 잡음 음성신호에 음성개선을 적용한 후의 음성파형과 음성신호 및 영상신호를 이용하여 검출된 끝점을 <그림 4>에 나타내었다. 여기서 n_b^m와 n_e^m는 음성신호에서 수동으로 검출된 시작점과 끝점을 나타내며, 이는 실제 발음구간의 시작점과 끝점에 해당한다. <그림 4(a)>와 같이 음향학적 잡음이 비교적 적은 경우, 음성신호과 영상신호에서 각각 검출된 시작점과 끝점은 실제 발음구간보다 앞쪽과 뒤쪽에 각각 위치하며, 영상신호에서 검출된 시작점과 끝점의 정확성이 음성신호의 경우보다 떨어지는 경향을 나타낸다. <그림 4(b)>와 같이 음향학적 잡음이 15dB로 매우 강한 경우 음성신호를 이용하면 음성개선 방법이 적용되었더라도 시작점과 끝점이 실제 발음구간의 내부에서 검출되는 경향을 나타낸다. 이로 인하여 실제 발음구간 내의 음성신호가 일부 유실되는 문제가 발생하여 음성인식 시스템뿐만 아니라 결과적으로는 바이모달 시스템의 인식성능 저하를 초래하게 된다. 본 논문에서는 음향학적 잡음에 대하여 끝점을 안정적으로 검출하기 위하여 입력 음성신호의 SNR을 추정하고 음성신호와

영상신호에서 각각 검출된 시작점과 끝점을 선택한다. <그림 5>에 그 과정을 간략하게 나타내었다.

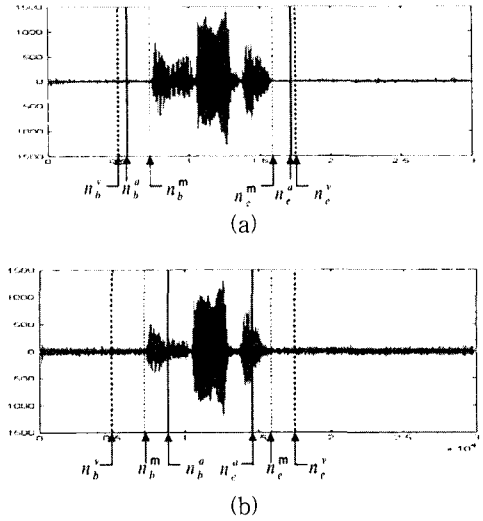


그림 4. '동대구' 음성신호에 대한 음성개선 후의 음성 파형과 음성 및 영상신호로부터 수동 및 자동으로 검출된 끝점 (a) 0dB SNR, (b) -15dB SNR

Fig. 4. Enhanced speech waveform of 'Dongdaegu [dôŋdægu]' and endpoints detected in audio signal manually and automatically and in video signal automatically (a) 0dB SNR, (b) -15dB SNR.

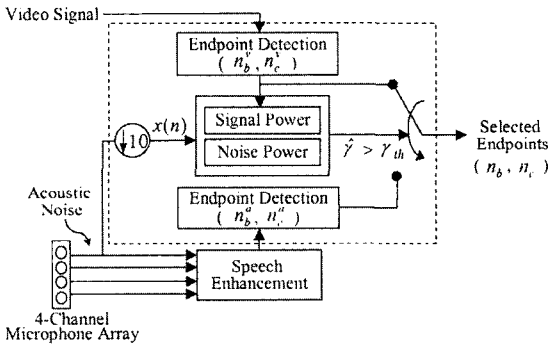


그림 5. 제안한 끝점 선택방법에 대한 개요도

Fig. 5. Block diagram of the proposed endpoint switching method.

먼저 음성/영상 데이터 획득시의 제약조건을 고려하면 입력 음성신호의 첫 부분을 묵음구간으로 볼 수 있다. 따라서 음성신호에서 묵음구간에 해당하는 처음 5 프레임에는 잡음만 존재하므로, 이 구간에서 음성신호의 잡음전력을 추정할 수 있다. 일반적으로 대부분의 고립단어를 발음하는 과정에서 입술의 움직임이 음성보다 앞서 시작된다[3]. 그리고 음성/영상 데이터 획득과정에서 고립단어를 발음한 후 입술을 다시 닫도록 제약을 두었으므로 음성 발성이 끝난 이후에도 대부분의 동영상에서는 입술 움직임이 존재한다. 따라서 영상신호에서 검출된 발음구간 내에 고립단어의 실제 음성 발음구간이 포함된다. 만약 음성신호와 잡음이 서로 상관성이 없다고 가정한다면 입력 음성신호의 SNR $\hat{\gamma}$은 영상신호에서 검출된 시작점과 끝점, 잡음 음성신호 $x(n)$으로부터 식 (4)와 같이 추정할 수 있다.

$$\hat{\gamma}[dB] = 10\log_{10}\left(\frac{\frac{1}{n_e^v - n_b^v + 1}\sum_{n=n_b^v}^{n_e^v}|x(n)|^2}{\frac{1}{5N}\sum_{n=0}^{5N-1}|x(n)|^2} - 1 \right) \quad (4)$$

여기서 N은 음성신호의 한 프레임 길이를 나타낸다. 최종적으로 시작점 n_b와 끝점 n_e는 입력 음성신호로부터 추정된 SNR과 실험적으로 결정된 문턱치 γ_{th}를 비교하여 식 (5)에 의하여 결정된다.

$$n_b = \begin{cases} n_b^a, & \hat{\gamma} \ge \gamma_{th} \\ n_b^v, & \hat{\gamma} < \gamma_{th} \end{cases}, \quad n_e = \begin{cases} n_e^a, & \hat{\gamma} \ge \gamma_{th} \\ n_e^v, & \hat{\gamma} < \gamma_{th} \end{cases} \quad (5)$$

V. 실험결과 및 고찰

음성인식 시스템에 적용되는 끝점 검출기의 성능은 검출된 끝점의 위치를 음성파형에서 시각적으로 비교하는 방법[8, 9]과 시스템의 인식률을 비교하는 방법[13] 등에 의하여 평가된다. 끝점 검출기의 성능이 견실할수록 우수한 음성 인식률을 획득할 수 있으므로 후자의 방법을 통하여 끝점 검출기 성능을 객관적인 평가할 수 있다. 그러므로 본 논문에서는 전술한 끝점검출 방법들을 바이모달 시스템에 적용하여 인식률을 비교함으로써 제안한 끝점검출 방법의 성능을 평가한다.

1. 바이모달 시스템의 인식실험 결과

바이모달 시스템의 인식성능을 평가하기 위하여, 4000개의 음성/영상 데이터로부터 20명 화자에 대한 2800개의 데이터를 HMM 기반의 두 인식기를 독립적으로 학습하는데 사용하였다. 그리고 학습에 참여한 20명의 화자로부터 획득된 1200개의 데이터를 테스트에 사용하였다. 학습에 사용된 음성 및 영상 특징벡터는

음성신호로부터 수동으로 검출된 발음구간 내에서 추출하였다. 음성잡음 환경에서 바이모달 시스템의 인식 성능을 평가하기 위하여 100kHz의 입력 음성신호에 Gaussian random 잡음을 0dB에서 -15dB까지 5dB 간격으로 강제로 첨가하였다. 그리고 잡음이 첨가된 100kHz의 입력 음성신호를 10kHz로 다운샘플링하여 음성인식 시스템의 입력으로 사용한다. 인식 후 결합방식에 기반한 바이모달 시스템을 구현하는 과정에서 입력 음성신호의 SNR에 따라 영상 가중치를 0.1에서 0.9까지 설정하였다.

입력 음성신호의 SNR과 영상가중치, 그리고 끝점검출 방법에 따른 바이모달 시스템의 인식률을 <표 2>에 나타내었다. 음성신호에서 수동 검출된 발음구간 내에서 추출된 특징벡터를 이용한 실험 결과에서는 SNR에 무관하게 최고의 인식률을 나타내었다. 특히 잡음이 -15dB로 매우 높은 경우에 대해서도 96.8%의 인식률을 나타내고 있다. 이 결과로부터 발음구간의 시작점과 끝점의 정확한 검출이 바이모달 시스템의 인식성능에 상당히 큰 영향을 미친다는 사실을 확인할 수 있다.

음성신호에서 검출된 시작점과 끝점을 이용하는 바이모달 시스템은 높은 SNR에서 매우 우수한 인식률을 나타내고 있으나 SNR이 낮아짐에 따라 인식성능이 현저하게 저하되는 경향을 보인다. 영상신호는 음향학적 잡음의 영향에 매우 둔감하므로 입력 음성신호의 SNR이 낮아지더라도 영상신호를 이용한 끝점검출 방법을 적용한 바이모달 시스템에서는 90%이상의 인식률을 획득하였다. 그러나 음성신호의 SNR이 -5dB이상인 경우에서는 음성신호를 이용한 끝점검출을 적용한 바이모

표 2. 음성 SNR에 따른 영상가중치와 끝점 검출 방법에 따른 바이모달 시스템의 인식률

Table 2. Recognition rates of bimodal system with respect to SNR of audio signal and endpoint detection methods.

SNR of 100kHz input audio signal		Clean	0dB	-5dB	-10dB	-15dB
Video weighting factor		0.1	0.3	0.5	0.7	0.9
Endpoint Detectio	In audio signal (manually)	100	99.8	98.2	96.7	96.8
	In audio signal (automatically)	99.2	98.7	96.3	87.3	68.9
	In video signal (automatically)	99.2	97.8	93.9	91.5	91.6
	Proposed method	98.9	98.4	94.6	91.3	91.6

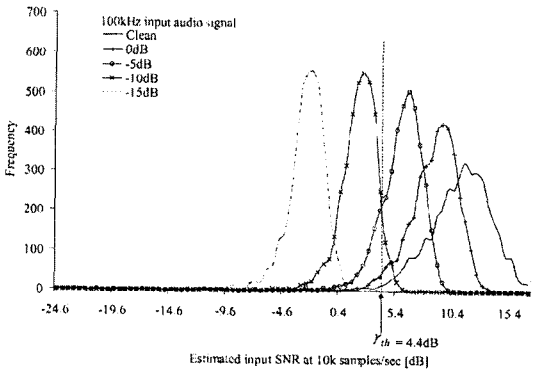


그림 6. 10kHz로 다운샘플링된 음성신호로부터 추정된 SNR 분포

Fig. 6. SNR distribution estimated from down-sampled audio signals at 10kHz.

달 시스템에서 더 높은 인식률을 나타내었다. 제안한 끝점검출 방법을 적용한 바이모달 시스템에서는 clean 음성에 대하여 98.9%, -10dB에서 91.3%의 인식률을 나타내어, 음성신호의 SNR에 무관하게 대부분 만족스러운 인식률을 나타내었다. 여기서 제안된 끝점검출 방법에서 끝점 선택을 위한 문턱치 γ_{th} 는 <그림 6>과 같이 10kHz로 다운샘플링된 음성신호로부터 추정된 SNR분포를 조사하여 4.4dB로 결정하였다. 이상의 결과로부터 음향학적 잡음이 강한 바이모달 시스템에서는 제안한 끝점검출 방법이 발음구간의 시작점과 끝점을 견실하게 검출함을 확인할 수 있다.

2. 실험결과에 대한 고찰

음성신호에서 수작업으로 검출된 시작점과 끝점은 실제 발음구간의 시작과 끝에 해당한다. 그러므로 실험에 사용된 모든 음성/영상 데이터에 대하여 수작업으로 시작점과 끝점을 검출하고 자동 검출된 시작점과 끝점과의 차이(difference)를 다음 식 (6)과 같이 정의한다.

$$d_b = n_b^{\mathrm{a}} - n_b^{\mathrm{m}}$$
$$d_e = n_e^{\mathrm{a}} - n_e^{\mathrm{m}} \qquad (6)$$

여기서 n_b^{a}와 n_e^{a}는 음성 및 영상신호에서 자동으로 검출된 시작점과 끝점을 각각 의미한다. <그림 7>에 시작점에 대한 d_b와 끝점에 대한 d_e의 도수분포도를 나타내었다. 영상신호를 이용한끝점검출은 음향학적 잡음의 영향에 무관하므로 d_b와 d_e의 도수분포도가 음성신호의 SNR에 상관없이 일정하다. 시작점에 대하여 d_b값이 양수인 부분과 끝점에 대하여 d_e값이 음수인

부분을 도수분포도에서 짙은 회색으로 나타내었다. 이 부분은 <그림 4(b)>에 나타낸 바와 같이 자동으로 검출된 시작점과 끝점이 실제 발음구간 내에 존재하는 경우에 해당한다. 음성신호를 이용한 끝점검출은 음향학적 잡음의 영향을 크게 받으므로 d_b와 d_e의 도수분포가 SNR에 따라서 크게 변화됨을 확인할 수 있다. 특히, -10dB와 -15dB의 도수분포에서 짙은 회색부분이 현저히 증가되었다. 이 결과는 잡음이 증가할수록 실제 발음구간 내에서 시작점과 끝점이 오검출될 확률이 높아짐을 의미한다. 이와 같은 오검출로 인한 실제 발음구간의 일부 유실은 이후의 어떤 단계에서도 복원되기 어려우므로 시스템의 인식성능을 저하시키는 심각한 원인이 된다. <표 2>의 실험결과에서 음성신호를 이용한 끝점검출 방법을 적용한 바이모달 시스템에서는 -10dB이하의 잡음 음성신호에 대하여 인식률이 현저히 저하됨을 확인할 수 있었다.

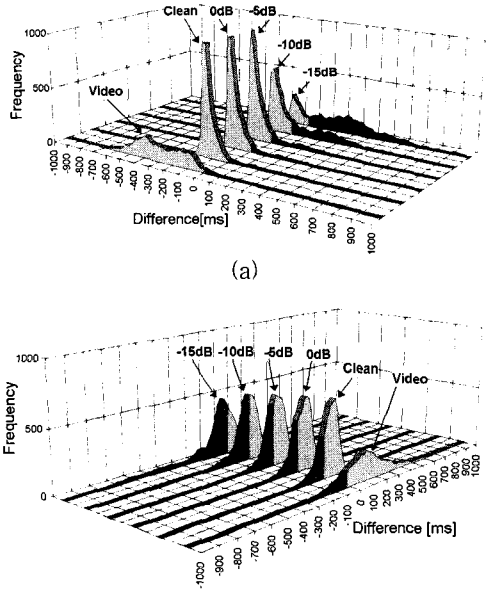


그림 7. (a) 음성신호에서 수작업으로 검출된 시작점과 음성신호 및 영상신호에서 자동으로 검출된 시작점과의 차이에 대한 도수분포도, (b) 음성신호에서 수작업으로 검출된 끝점과 음성신호 및 영상신호에서자동으로 검출된 끝점과의 차이에 대한 도수분포도

Fig. 7. (a) Histogram of differences between beginning points detected manually and ones detected automatically in audio or video signal, (b) Histogram of differences between ending points.

<그림 7>의 밝은 회색은 d_b와 d_e가 각각 음수와 양수인 부분으로 음성신호의 묵음구간에 해당하는 부분이 발음구간에 일부 포함됨을 의미한다. 대부분의 고립단어 발성에서는 화자의 음성이 나오기 전에 입술이 먼저 움직이기 시작하여 음성이 끝난 이후에도 입술 움직임이 어느 정도 진행된다. 그러므로 음성신호의 묵음에 해당하는 넓은 구간이 영상신호를 이용한 끝점검출 방법으로 검출된 발음구간에 포함되게 된다. 따라서 <그림 7>에서 보는 바와 같이 영상신호에서 산출된 도수분포도의 밝은 회색부분이 잡음 음성에 대한 도수분포도에서 보다 넓게 분포한다. 자동으로 검출된 발음구간에 포함된 일부 묵음구간은 HMM기반의 인식기에서 묵음에 대한 상태에 의하여 모델링되어 진다. 그러므로 밝은 회색에 해당하는 묵음구간의 포함정도가 바이모달 시스템의 인식성능 저하에 미치는 영향이 짙은 회색에 해당하는 오검출 결과가 미치는 영향보다 덜 심각함을 <표 2>의 인식결과에서 확인할 수 있다.

VI. 결 론

음성인식에 있어서 끝점검출 성능이 전체 인식시스템의 성능에 큰 영향을 미치는 바와 같이 바이모달 시스템에서도 끝점검출 성능이 시스템의 인식성능에 미치는 영향이 크다. 본 논문에서는 음향학적 잡음이 존재하는 바이모달 시스템에서 견실한 성능을 나타내는 끝점검출 방법을 제안하였다. 먼저 바이모달 시스템 구현 및 끝점검출 성능을 평가하기 위하여 음성신호와 영상신호가 동기화된 음성/영상 데이터베이스를 구축하였다. 바이모달 시스템은 자동 음성인식 시스템과 입술독해 시스템을 인식 후 결합방식으로 결합하여 구현되었다. 음성신호에서 Teager 에너지를 이용하여 검출된 시작점과 끝점은 높은 SNR에서는 우수한 성능을 나타내지만, 음향학적 잡음이 증가함에 따라 오검출율이 증가하게 되어 바이모달 시스템의 인식률을 저하시키는 주요 원인이 되었다. 영상신호에서 검출된 시작점과 끝점은 음향학적 잡음의 영향에는 무관하므로 비교적 우수한 검출율률을 나타내었다. 그러나 높은 SNR에서는 음성신호에서 검출된 시작점과 끝점의 정확성보다 떨어지므로 인식 성능이 다소 저하되는 경향을 보였다. 최종적으로 입력 음성신호에서 추정된 SNR에 따라서 음성신호와 영상신호에서 검출된 시작점과 끝점을 선택하도록 제안된 끝점검출 방법은 음향학적 잡음이 존

재하는 바이모달 시스템에서 안정된 성능을 나타내므로 입력 잡음음성의 SNR에 무관하게 만족스러운 바이모달 음성 인식률이 획득됨을 실험결과에서 확인할 수 있었다. 향후 연구에서는 제안된 끝점검출 방법을 적용한 바이모달 시스템을 조명변화와 음향학적 잡음이 다양한 실제 환경에 적용하여 그 성능을 평가하고자 한다.

참 고 문 헌

[1] H. Kaplan, C.J. Bally, and C. Garretson, Speechreading: A Way to Improve Understanding, Gallaudet University Press, Washington D.C., 1999.

[2] B. Dodd and R. Campbell, Hearing by Eye: The Psychology of Lip-reading, Lawrence Erbaum Press, Hillsdale NJ, 1987.

[3] C. Bregler and Y. Konig, "Eigenlips for Robust Speech Recognition," in Proc. of IEEE Int'l Conf. on Acoustics, Speech and Signal Processing, vol. 2, pp. 669-672, 1994.

[4] M.E. Hennecke, K.V. Prasad, and D.G. Stork, "Automatic Speech Recognition System Using Acoustic and Visual Signals," in Proc. of 29th Asilomar Conf. on Signals, Systems and Computers, vol. 2, pp. 1214-1218, 1995.

[5] 박병구, 김진영, 최승호, "바이모달 음성인식의 음성정보와 입술정보 결합방법 비교," 한국음향학회지, 제 18권 제 4호, pp. 31-37, 1999.

[6] S. Dupont and J. Luettin, "Audio-Visual Speech Modeling for Continuous Speech Recognition," IEEE Trans.on Multimedia, vol. 2, no. 3, pp. 141-151, 2000.

[7] L.R. Rabiner and M.R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Uttrances," Bell Syst. Tech. J., vol. 54, no. 2, pp. 297-315, 1975.

[8] L.F. Lamel, L.R. Rabiner, A.E. Rosenberg, and J.G. Wilpon, "An Improved Endpoint Detector for Isolated Word Recognition," IEEE Trans. Acoust., Speech, and Signal Processing, vol. 29, no. 4, pp. 777-785, 1981.

[9] G.S. Ying, C.D. Mitchell, and L.H. Jamieson, "Endpoint Detection of Isolated Utterances Based on a Modified Teager Energy Measurement," in Proc. of IEEE Int'l Conf. on Acoustics, Speech and Signal Processing, pp. 732-735, 1993.

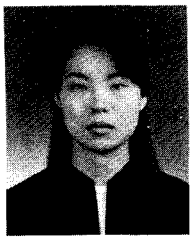
[10] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator," IEEE Trans. on Acoustic, Speech and Signal Processing, vol. ASSP-2, no. 6, pp. 1109-1121, 1984.

[11] H.-S. Kwon, J.-M. Son, S.-Y. Jung, and K.-S. Bae, "Speech Enhancement Using Microphone Array with MMSE-STSA Based Post-Processing," in Proc. of Int'l Conf. on Electronics, Information and Communications, pp. 186-189, Ulaanbaatar, Mongolia, Jul. 2002.

[12] H.-H. Oh, Y.-M. Jeoun, and S.-I. Chien, "A Set of Mesh Features for Automatic Visual Speech Recognition," in Proc. of IARP Workshop on Machine Vision Applicaions, pp. 488-491, Nara, Japan, Dec. 2002.

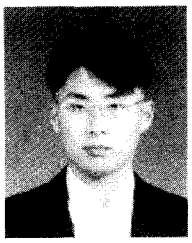
[13] S. Bou-Ghazale and K. Assaleh, "A Robust Endpoint Detection of Speech for Noisy Environments with Application to Automatic Speech Recognition," Proc. IEEE Int'l Conf. On Acoustics, Speech and Signal Processing, pp. IV-3808 - IV-3811, Orlando, Florida, May 2002.

저 자 소 개



吳 炫 和(學生會員)

1998년 2월 : 경북대학교 전자공학과 졸업. 2000년 2월 : 경북대학교 전자공학과석사. 2000년 3월~현재 : 경북대학교 전자공학과 박사과정. <주관심분야 : 컴퓨터비전, 패턴인식, 영상처리 등>



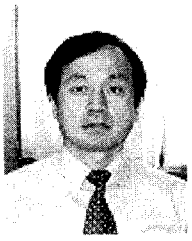
孫 宗 睦(學生會員)

1997년 2월 : 경북대학교 전자공학과 졸업. 1999년 2월 : 경북대학교 전자공학과 석사. 1999년 3월~현재 : 경북대학교 전자공학과 박사과정. <주관심분야 : 음성신호처리, 음성인식, 웨이브렛이론 등>

權 洪 錫(學生會員)

1997년 2월 : 경북대학교 전자공학과 졸업. 1999년 2월 : 경북대학교 전자공학과 석사. 1999년 3월~현재 : 경북대학교 전자공학과 박사과정. <주관심분야 : 음성신호처리, 디지털신호처리, 어레이프로세싱 등>

秦 成 一(正會員) 第39卷 SP編 第4號 參照



裵 建 星(正會員)

1977년 2월 : 서울대학교 전자공학과 졸업. 1979년 2월 : 한국과학기술원 전기및전자공학과 석사. 1989년 5월 : University of Florida 공학박사. 1979년 3월~현재 : 경북대학교 전자전기공학부 교수. <주관심분야 : 음성신호처리, 디지털신호처리, 디지털통신, 웨이브렛이론, 오디오신호처리 등>