한국음향학회지 제35권 제1호 pp. 74~82 (2016)
The Journal of the Acoustical Society of Korea Vol.35, No.1 (2016)
http://dx.doi.org/10.7776/ASK.2016.35.1.074

pISSN : 1225-4428
eISSN : 2287-3775

초음파 도플러를 이용한 음성 인식

Automatic speech recognition using acoustic doppler signal

이기승†

(Ki-Seung Lee†)

건국대학교 전자공학과

(Received July 21, 2015; accepted September 17, 2015)

초 록: 본 논문에서는 음성 신호 대신 초음파 도플러 신호를 이용하여 음성을 인식하는 새로운 음성 인식 방법을 제안하였다. 제안된 방법은 주변 잡음에 대한 강인성과 무 접촉식 센서 사용에 따른 사용자의 불편함 감소를 포함하는 기존의 음성/무음성 인식 방법에 비해 몇 가지 장점을 갖는다. 제안된 방법에서는 40 kHz의 주파수를 갖는 초음파 신호를 입 주변에 방사하여, 반사된 신호를 취득하고, 취득된 신호의 도플러 주파수 변화를 이용하여 음성 인식을 구현하였다. 단일 채널 초음파 신호를 사용하는 기존의 연구와 달리, 다양한 위치에서의 취득된 초음파 신호를 음성 인식에 사용하기 위해 다채널 취득 장치를 고안하였다. PCA(Principal Component Analysis)특징 변수를 사용한 음성 인식에는 좌-우 모델을 갖는 은닉 마코프 모델을 사용하였다. 제안된 방법의 검증을 위해 60개의 한국어 고립어에 대해 6명의 화자로부터 취득된 초음파 도플러 신호를 인식에 사용하였으며, 기존 음성기반 음성인식 기법과 비교할 만한 수준의 인식율을 얻을 수 있었다. 또한 실험 결과 제안된 방법은 기존의 단일 채널 음성 인식 방법과 비교하여 우수한 성능을 나타내었으며, 특히 잡음 환경에서도 90 % 이상의 인식율을 얻을 수 있었다.
핵심용어: 음성인식, 초음파 도플러, 무음성 인터페이스, 강인 음성인식

ABSTRACT: In this paper, a new automatic speech recognition (ASR) was proposed where ultrasonic doppler signals were used, instead of conventional speech signals. The proposed method has the advantages over the conventional speech/non-speech-based ASR including robustness against acoustic noises and user comfortability associated with usage of the non-contact sensor. In the method proposed herein, 40 kHz ultrasonic signal was radiated toward to the mouth and the reflected ultrasonic signals were then received. Frequency shift caused by the doppler effects was used to implement ASR. The proposed method employed multi-channel ultrasonic signals acquired from the various locations, which is different from the previous method where single channel ultrasonic signal was employed. The PCA(Principal Component Analysis) coefficients were used as the features of ASR in which hidden markov model (HMM) with left-right model was adopted. To verify the feasibility of the proposed ASR, the speech recognition experiment was carried out the 60 Korean isolated words obtained from the six speakers. Moreover, the experiment results showed that the overall word recognition rates were comparable with the conventional speech-based ASR methods and the performance of the proposed method was superior to the conventional signal channel ASR method. Especially, the average recognition rate of 90 % was maintained under the noise environments.
Keywords: Speech recognition, Ultrasonic doppler signals, Silent speech interface, Robust speech recognition

PACS numbers: 43.72.Ne, 43.72.Kb

I. 서 론

음성은 정보 전달의 중요 수단으로서, 화자의 발성음이 공기 매질을 통해 청취자의 귀에 전달됨으로써 의사소통이 이루어진다. 이와 같은 음성 전달은

†**Corresponding author:** Ki-Seung Lee (kseung@konkuk.ac.kr)
Department of Electronic Engineering, Konkuk University, 120 Neangdong-ro, Gwangjin-gu, Seoul 05029, Republic of Korea
(Tel: 82-2-450-3489, Fax: 82-2-3437-5235)

정상적인 환경에서만 원활하게 이루어지는 반면, 주변 잡음이 극심한 경우, 주변 사람들에게 전달되는 음성을 은닉해야 하는 경우, 주변 사람들에게 시끄러움 등의 불쾌함을 느끼지 않도록 해야 하는 상황 등에서는 의사전달이 어려워진다. 이러한 특수한 상황에서 적용 가능한 음성 정보 전달 방법으로 무음성 전달(silence speech interface)[1]이 제안되었다. 무음성 전달 방법으로서, 입주변에서 발생하는 근전도 신호를 이용하는 방법,[2] NAM(Non-Audible Microphone)을 입주변에 부착하여 음성을 취득하는 방법,[3] 자석과 자계센서를 이용하는 방법,[4] 구강 및 비강의 초음파 영상을 이용한 방법,[5] GHz microwave를 이용하는 방법,[6] 초음파 신호를 이용하는 방법[7] 등을 들 수 있다.

본 논문에서는 무음성 전달 방법의 하나로서, 초음파 도플러 신호를 이용한 인식 방법을 한국어 고립어 인식에 적용하여 결과를 살펴보았다. 초음파를 이용한 인식은 초기에 손이나 발의 동작 인식에 사용되었고[9,10] 이후 음성을 발성하고 있는 화자의 입주변에 초음파를 방사하고 반사되어 돌아오는 신호의 도플러 신호를 분석하는 연구가 진행되었다.[11-14] 초음파를 이용한 동작 및 음성 분석은 움직임이 있는 손, 발, 입주변 근육에 초음파를 방사하면, 각 부위의 변위로 인한 속도변화가 도플러 현상을 발생시킨다는 사실에 바탕을 두고 있다. 이때 발생되는 도플러 주파수가 특정 움직임 패턴에 의해 각기 다르게 나타난다고 가정하면, 도플러 주파수 패턴을 이용하여 손발의 움직임과 음성을 인식할 수 있게 된다.

Kalgaonkar *et al*.[8]의 연구에서는 간단한 손동작을 초음파 도플러를 이용하여 인식하였을 때 평균 88.4 %의 인식율[10]이, 보행 패턴을 인식하는 경우 91.7 %의 인식율을 얻는 것으로 보고하였다.[9] 초음파 도플러를 이용한 음성인식은 초음파 신호가 가청 음역대의 잡음에 영향을 받지 않으며, 신호의 취득 시 기존의 무음성 전달 방법이 주로 접촉식 센서를 사용하는 것과 비교하여 비접촉식 센싱 방법을 사용하여 사용자에게 불편감을 주지 않는다는 장점을 지닌다. 이러한 장점으로 초기에는 음성과 묵음을 구분(Voice Activity Detection, VAD)하여 초음파 도플러를 이용하였는데,[11] 잡음의 크기와 종류에 무관하게 92 ~ 97 %의 인식율을 나타내었으며, 이는 잡음에 의해 손상된 음성을 사용하는 경우 단지 10 % 인식율이 얻어지는 것과 비교하여 잡음 환경에 매우 적합한 방법임을 나타내었다. 이후 초음파 도플러 신호를 화자인식,[12] 음소인식,[13] 음성인식[7] 및 음성합성[14]에 적용하였는데, 화자 인식율 81 %, 음소 인식율 30 ~ 60 %, 음성인식은 10개 숫자를 인식하는 경우 37 %의 단지 인식율이 얻어지는 것으로 보고되는 등, 음성신호를 단독으로 사용하는 경우보다 다소 낮은 성능이 얻어지는 것으로 나타났다.

기존 방법 들은 단일 채널의 초음파 도플러 신호를 사용하였는데 본 논문에서는 음성 인식의 성능을 향상시키기 위해 다채널 센서를 사용하였으며 센서의 개수와 위치, 조합에 따른 인식율을 정량적으로 평가하였다. 한국어 음소를 골고루 포함하는 60개의 고립어에 대해 다양한 조건에서 초음파 신호를 취득하고 인식율을 관찰하였다. 또한 극심한 잡음 환경에서 기존의 음성을 이용한 음성 인식 기법과 성능을 비교하여 초음파 도플러를 이용한 음성 인식 방법의 유용성을 평가하였다.

II. 초음파 반사 신호 분석

속도 v로 움직이고 있는 물체에 주파수 f를 갖는 정현파 신호를 쏘게 되면, 본래의 주파수와는 다른 주파수를 갖는 신호가 반사되어 돌아오는데, 이때 변이된 주파수 $\hat{f}$는 다음과 같다.

$$\hat{f}=\frac{v_s+v}{v_s-v}f, \tag{1}$$

여기서 v_s는 정현파 신호의 속도로서, 초음파 신호의 경우 공기 중에서 $v_s=(331.3+0.606\,T)\,m/s$ (T는 섭씨온도) 로 나타낼 수 있다. v_s가 물체의 속도에 비해 충분히 크다면 위의 식은 아래와 같이 나타낼 수 있다.

$$f\approx\left(1+\frac{2v}{v_s}\right). \tag{2}$$

Fig. 1과 같이 발성하고 있는 사람의 입주변에 일

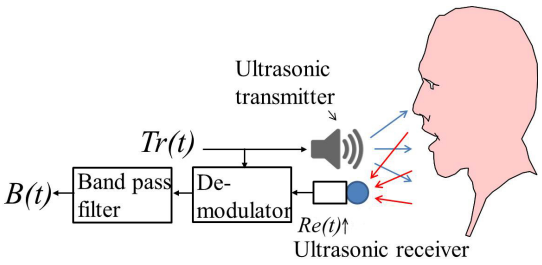

Fig. 1. Block diagram of the speech ultrasonic doppler system.

정한 주파수를 갖는 초음파 신호를 방사하게 되면, 근육과 입의 움직임 및 목 주변의 떨림 등으로 인하여 반사면의 속도 변이가 나타나며, 도플러 현상에 의한 주파수 변이가 일어난다. 방사된 신호 $T_r(t)$가 다음 식과 같이 주파수 f_c, 크기가 A_T, 위상이 ψ_T인 정현파 신호라면,

$$T_r(t) = A_T \cos(2\pi f_c t + \psi_T). \tag{3}$$

반사되어 돌아오는 신호는 입술, 뺨, 턱과 같은 다양한 부위의 움직임에 따라 개별적으로 발생하는 도플러 신호의 합으로 주어지게 된다. 도플러 현상을 일으키는 부위가 총 M개 이고, 각각은 $v_i(t)$의 속도를 갖는다면, 수신된 신호는 아래와 같이 나타낼 수 있다.

$$R_e(t) = \sum_{i=1}^{M} A_T k_i \cos(\phi_i + \psi_T),$$
$$\phi_i = 2\pi f_c \left[t + \frac{2}{v_s} \int_o^t v_i(\tau) d\tau \right] + \psi_i, \tag{4}$$

여기서 ψ_i는 i-번째 성분에 대한 위상변이를 나타낸다. $R_e(t)$에 대해 복조(demodulation)를 수행하고 DC 값과 f_c 이상의 주파수 성분을 차단하는 대역통과필터를 통과한 신호 $B(t)$는 다음과 같이 나타낼 수 있다.

$$B(t) = BPF\left[A \cos(2\pi f_c t + \psi_T) R_e(t) \right]$$
$$\approx \sum_{i=1}^{M} k_i' \sin\left[\frac{4\pi f_c}{v_s} \int_0^t v_i(\tau) d\tau \right]. \tag{5}$$

$v_i(t)$는 조음기관의 움직임에 의한 변이를 나타내며, 수 10 ms 구간에서는 상수 v_i로 근사화될 수 있다.[15] 따라서 $B(t)$에 창함수(window function)을 곱하여 얻어지는 단구간 신호 $B_w(t)$는 다음과 같다.

$$B_w(t) \approx \sum_{i=1}^{M} k_i' \sin\left(\frac{4\pi f_c}{v_s} v_i t \right)$$
$$= \sum_{i=1}^{M} k_i' \sin\left[4\pi \frac{\triangle x_i(t)}{\lambda} \right], \tag{6}$$

여기서 $\triangle x_i(t)$는 i-번째 부위에 대한 변위를 나타낸다. Eq.(6)으로부터 $B_w(t)$를 구성하는 각 정현파 신호의 위상은 각 부위의 변위에 따라 결정되며, 음성신호에 따라 각기 다르게 나타나는 것으로 가정할 수 있다.

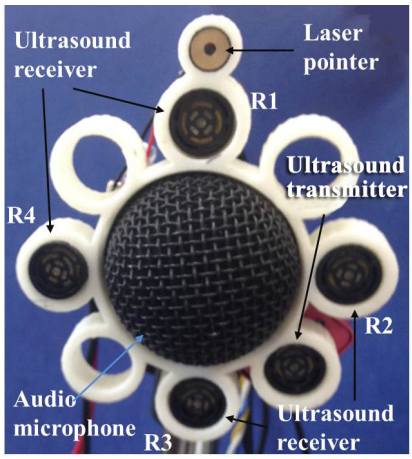

Fig. 2. The photography of the prototype ultrasound doppler acquisition equipment (front).

Fig. 3. The prototype acquisition equipment integrated with audio microphone.

III. 초음파 도플러 취득 장치 제작

Figs. 2와 3에 본 연구에서 제작한 prototype 초음파 도플러 취득 장치를 나타내었다. 초음파 도플러를 이용한 이전의 연구[7-14]에서는 송신용과 수신용 각각 1개의 센서가 사용되었는데, 본 연구에서는 보다 다양한 방향에서 반사되는 초음파 신호를 취득하기 위해 최대 8개의 센서를 장착할 수 있는 센서 고정 장치를 제작하였다. 실제 실험에서는 1개의 송신 센서와 4개의 수신 센서를 사용하였으며, 센서간 거리는 53 mm였다. 송신용 초음파 센서는 4사분면에 부착하였는데, 이는 수차례의 실험을 통해 가장 높은 인식율을 나타내는 위치를 경험적으로 찾은 것이다.

사용된 센서는 40 kHz의 중심 주파수를 갖는 송/수신 겸용의 초음파 센서(AW8TR40, Audiowell, China) 송신모드에서 음압레벨은 115 dB, 수신감도는 −65 dB 로서, 2 kHz의 대역폭을 갖는다.

Prototype 취득 장치에는 초음파 신호와 음성 신호 간의 상관 분석 및 음성 인식 결과의 비교를 위해 가청 주파수 대역의 오디오 신호 취득이 가능한 마이크로폰(AKG880, AKG, Austria)이 함께 장착되어 있다. 또한 취득할 때 마다 각 센서와 피시험자간의 상대적인 위치, 방향, 거리가 변동되는 것을 방지하기 위해 상단에 레이저포인터를 장착하였다. 신호 취득 전 피시험자는 정면에 장착된 웹캠을 통해 현재 레이저 포인트 위치를 보고, 포인트 위치가 항시 코와 윗입술의 중간 지점에 오도록 자세를 움직여 항시 일정한 지점에서 신호가 취득되도록 하였다.

초음파 송신센서에는 함수발생기(33250A, Agilent, USA)에서 발생된 40 kHz 정현파 신호를 광대역 오디오 증폭기(LM386N, National Semiconductor, USA)를 통해 증폭한 신호를 입력하였다. 4개의 초음파 센서에서 수신된 각 신호 및 마이크로폰에서 취득된 음성 신호를 동시에 디지털값으로 변환하기 위해 다채널 오디오인터페이스(Fireface 800, RME, Germany)를 사용하였으며, 초음파 신호 및 음성 신호 모두 샘플링 주파수는 192 kHz, 양자화 비트수는 16비트로 설정하였다. 마이크로폰 및 각 초음파 센서의 감도차에 따른 신호의 크기 차이를 보상하기 위해, 샘플링 전 아날로그값을 각기 다른 이득으로 증폭하였다.

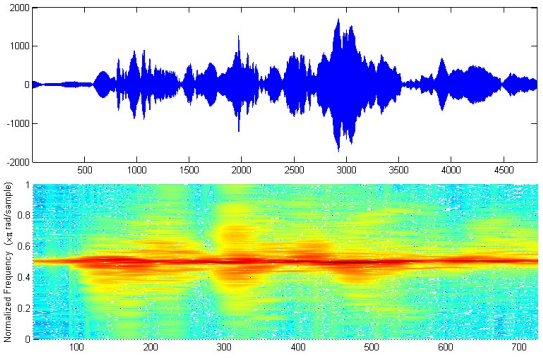

Fig. 4. Top: The received ultrasound waveform for the word “mother” [ʌ-mʌ-ni]. Bottom: Spectrogram.

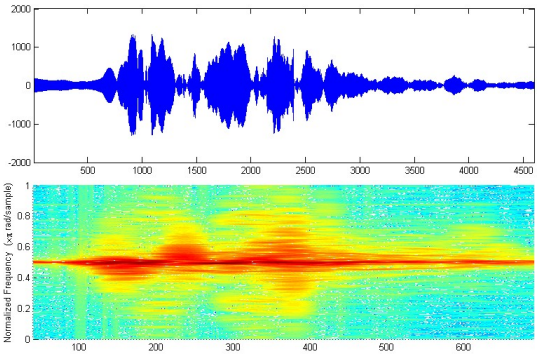

Fig. 5. Top: The received ultrasound waveform for the word “eye” [nun-do ŋ-ja], Bottom: Spectrogram.

즉, 음성 취득 전 센서 전면에 초음파 신호를 잘 반사시키는 재질의 판재를 설치하고, 초음파 신호를 방사하여 각 센서로부터 반사되어 돌아오는 초음파 신호를 취득하고 각 센서 별 취득된 신호가 레벨을 유지하도록 오디오인터페이스의 이득 값을 조정하였다.

Figs. 4와 5에 제작된 초음파 도플러 취득 장치를 이용하여 저장된 음성 “어머니” 와 “눈동자”에 대한 초음파 신호 와 해당 스펙트로그램을 제시하였다. 그림의 초음파 신호는 Fig. 2의 R3센서에서 취득된 신호로, 중앙 주파수($f_N/2$, f_N = Nyquist 주파수)가 40 kHz에 위치하도록 복조된 신호이다. 시간 영역의 파형을 살펴보면, 유성음 구간에서 큰 크기의 신호가 관찰되는 전형적인 음성 신호의 특성을 따르지는 않지만, 두 음성 간의 차이가 비교적 뚜렷하게 나타나며, 스펙트로그램상으로도, 초음파 송신주파수(40 kHz)를 중심으로 에너지가 시간적으로 다르게 분포하는 것을 알 수 있다. 이러한 특성은 초음파 신호를 마코프 랜덤 신호로 모델링 할 수 있으며, 은닉 마코

프 모델(Hidden Markov Model, HMM)과 같은 기존의 음성 인식에 널리 사용된 통계 모델을 이용하여 음성 인식을 구현할 수 있음을 의미한다.

IV. 초음파 도플러를 이용한 음성인식

4.1 특징 변수

음성 신호에 대한 특징 변수로 멜 주파수 켑스트럼 계수(Mel-Frequency Cepstral Coefficient, MFCC)가 주로 사용된다.[15] 기존의 초음파 도플러 기반 음성 인식 방법[7]에서는 신호를 기저 대역신호(baseband signal) 로 복조하고, 푸리어 변환을 통해 주파수 도메인 신호로 바꾼 후 크기 스펙트럼에 대한 주요성분 분석(Principal Component Analysis, PCA)를 통해 특징 변수를 구성하였다.

초음파 도플러 신호의 경우, 음성과 유사한 패턴으로 진동하는 목 부위에서 주파수 변위가 일어날 수 있으며, 취득된 초음파 신호도 음성과 유사한 주파수 특성을 보일 것으로 예상된다. 따라서 복조된 초음파 신호에 대해 음성 신호와 마찬가지로 MFCC를 사용하여 음성 인식이 가능하나, 실험 결과를 통해 MFCC를 사용하는 경우 PCA특징 변수를 사용하는 경우와 비교하여 인식율이 다소 저하되는 것을 관찰하였다. 이는 초음파 도플러 신호가 목의 진동 외에 입 주변 근육의 움직임 등 다양한 요인에 의해 발생되는데 원인이 있는 듯하다.

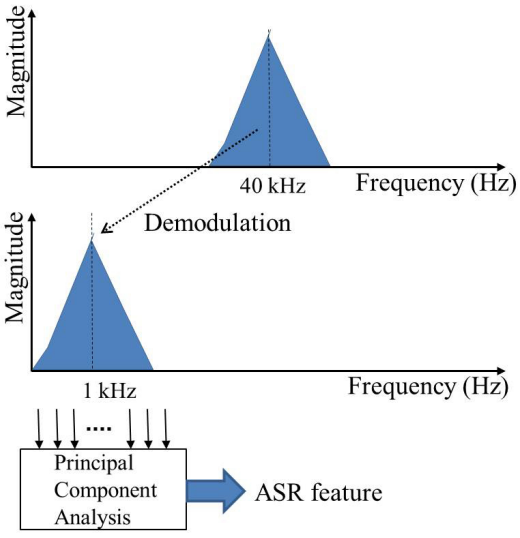

Fig. 6. Procedure for extracting feature parameter.

본 연구에서는 이러한 실험 결과를 반영하여 PCA 특징 변수를 음성인식을 위한 특징 변수로 사용하였다. PCA특징 변수를 얻는 과정을 Fig. 6에 설명하였다. 사용된 초음파 센서의 대역폭 2 kHz를 고려하여 복조 주파수는 39 kHz로 설정하였으며, 중심 주파수 1 kHz로 이동된 크기 스펙트럼에 대해 PCA를 수행하였다. PCA 계수 중 상위16개를 특징변수로 사용하였다.

4.2 채널별 특징변수의 조합

본 연구와 같이 여러 채널에서 취득된 신호로부터 인식을 수행하는 경우, 각 채널에서 얻어진 각 특징 변수를 어떤 식으로 조합할 것인가 하는 조합 방법을 정해야 한다. 고려 가능한 첫 번째 방법으로서, 각 채널의 특징 변수를 단순히 연결하여 하나의 벡터로 표현하는 것으로서, 인식에 사용되는 특징 벡터는 다음과 같이 나타낼 수 있다.

$$O = o_1 \oplus o_2 \oplus o_3 \oplus o_4 , \tag{9}$$

여기서 $\oplus$는 벡터 간 연결(concatenation)을 나타낸다. 본 연구에서 같이 채널 당 16개 특징 변수×4채널을 사용하는 경우, 인식에 사용되는 특징 변수는 64차원 벡터 형태를 갖는다. 이와 같은 방법은 인식을 위한 우도(likelihood)값을 얻기 전에 특징 변수를 조합시키므로 Early Integration(EI) 방법이라 한다. 또 다른 조합 방법으로, 각 채널의 특징 변수들로 독립적인 HMM을 생성하고, 각 채널별 사후 확률(posterior probability)를 구하고, 이들 확률을 조합하는 방법으로, 입력 벡터 $O = \{o_1, o_2, o_3, o_4\}$의 모델 $\Lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$에 대한 우도는 다음과 같다.

$$p(O|\Lambda) = p(o_1|\lambda_1)^{w_1} p(o_2|\lambda_2)^{w_2} p(o_3|\lambda_3)^{w_3} p(o_4|\lambda_4)^{w_4}, \tag{10}$$

여기서 $\{w_k\}_{n=1}^{4}$는 각 채널 우도에 대한 가중치를 나타낸다. 이와 같은 채널 조합 방법을 Late Integration (LI) 방법이라 한다. LI방법은 EI방법과 비교하여 채널별 중요도를 반영하여 인식을 수행할 수 있고 높

은 차원의 벡터 사용에 따른 singular matrix 문제를 감소시킬 수 있다는 장점이 있는 반면, 채널 간 상관 성을 반영할 수 없다는 단점을 갖는다. 본 연구에서 사용된 초음파 센서의 빔 각도가 55°이고 화자와 거리 40 cm로서, 센서 당 41×41 cm^2 영역에서 반사된 신호가 취득된다. 각 센서간 거리가 7 cm 임을 고려하면, 각 센서가 신호를 취득할 수 있는 영역은 많은 영역이 공유됨을 의미한다. 이러한 사실을 반영하여 본 연구에서는 EI방법을 사용하였다.

V. 실험 및 결과

초음파 도플러 신호를 이용한 음성 인식 기법의 유효성을 검증하기 위하여 Table 1에 제시된 60개 한국어 고립어에 대한 음성 인식 실험을 수행하였다. 실험은 6명의 피실험자(남성 5명, 여성 1명 모두 20대)로부터 5개의 서로 다른 실험 조건에서 단어 당 50회 반복 취득하였다. 이 중 30개는 학습에, 나머지 20개는 검증에 사용하였다. 신호 취득은 비교적 조용한 환경에서 제작된 취득 장치를 이용하여 초음파 신호와 음성 신호를 동시에 취득하였다.

음성 인식을 위한 모델로서 HMM이 사용되었으며 상태 천이는 좌-우 모델(left-right model)이 적용되었다. 각 단어별 HMM은 화자마다 독립적으로 학습되었으며, 모든 단어에 대해 상태 수는 5개, 각 상태의 관찰 확률 밀도 함수는 5개의 가우시안 함수를 포함하는 혼합 가우시안 모델(mixture gaussian model)을 사용하였다. 각 가우시안 함수의 공분산 행렬은 대각 행렬로 표현하였으며, HMM의 학습에는 Baum-Welch 알고리즘이 사용되었다.

다양한 조건에서의 음성 인식 유효성을 검증하고 인식율을 비교하기 위해 다음과 같은 5가지 조건에서 신호를 취득하였다.

[조건1] 센서-피실험자간 거리 31 cm, 피실험자의 얼굴 위치 고정, 초음파 신호는 4사분면에서 방사(Fig. 2에 제시된 것과 동일).

[조건2] 센서-피실험자간 거리 31 cm, 피실험자의 얼굴 위치 고정, 초음파 신호는 4개 초음파 센서의 중심부에서 방사.

[조건3] 센서-피실험자간 거리 41 cm, 피실험자의 얼굴 위치 고정, 초음파 신호는 4사분면에서 방사.

[조건4] 센서-피실험자간 거리 41 cm, 피실험자의 얼굴 위치가 다소 움직이는 것을 허용, 초음파 신호는 4사분면에서 방사.

[조건5] 센서-피실험자간 거리 31 cm, 피실험자의 얼굴 위치 고정, 초음파 신호는 4사분면에서 방사(Fig. 2에 제시된 것과 동일), 주변에 배경잡음.

성능평가에는 각 화자별 인식율을 합산한 평균 인식율이 사용되었으며, 센서의 조합, 각 조건, 동적인

Table 1. Word list.

Meaning	Pronunciation	Meaning	Pronunciation	Meaning	Pronunciation	Meaning	Pronunciation	Meaning	Pronunciation
Wind	[ba-ram]	Fly	[ʃ-pari]	Baby	[æ-ki]	Place	[ja-ri]	Cave	[doŋ-gul]
Lip	[ip-sul]	Time	[si-gan]	Liberation	[hæ-baŋ]	Now	[i-che]	Reason	[k*a-dak]
Butterfly	[na-bi]	Bush	[s*a-ri]	Herb	[ʧwi-na-mul]	Soup	[ʧi-gæ]	Knife	[kal]
Washing	[p*al-ræ]	Seed	[s*i-al]	Taste	[ʧwi-hjaŋ]	First	[ʧʌ-ŭm]	Nest	[duŋ-ji]
Flute	[pi-ri]	Sky	[ha-nŭl]	Consolation	[wi-mun]	Sound	[so-ri]	Moon	[dal]
Lag	[da-ri]	East sea	[doŋ-hæ]	Back	[dwi]	Calendar	[dal-rjʌk]	Copper	[gu-ri]
Give	[bat-go]	Heart	[ma-ŭm]	Repetition	[doe-pul-i]	Neighbor	[i-ut]	Mother	[ʌ-mʌ-ni]
Wave	[pa-do]	Lily	[na-ri]	Soybean	[doeŋ-jaŋ]	Haircut	[i-bal]	Fever	[jʌl]
Daughter	[t*al]	Letter	[gŭl]	Road	[oe-gil]	Buckwheat	[me-mil]	Health	[gʌŋ-gaŋ]
Frame	[tŭl]	recent	[gŭn-sæ]	Foreign	[oe-kuk]	World	[se-saŋ]	Snow man	[nun-saram]
Fall	[ga-ŭl]	Food	[jaŋ-sik]	Sea	[ba-da]	Old boy	[no-in]	Finish	[wan-sʌŋ]
Color	[sæk-doŋ]	Six or seven	[jenil-gob]	Talk	[mal-s*ŭm]	Eye	[nun-doŋ-ja]	Doctor	[ŭi-sa]

특성의 반영여부, 음성신호 기반의 인식 방법과 인식율을 비교하였다.

5.1 센서의 조합에 따른 인식율 비교

Table 2에 각 조건, 센서 조합별 평균 인식율이 제시되어있다. 센서의 조합에 따른 인식율을 살펴보면 센서의 개수가 증가함에 따라 인식율도 대체적으로 증가함을 알 수 있다. 센서를 1개만 사용한 경우, 각 센서에 따라 인식율의 차이가 비교적 뚜렷하게 나타나며, 조건3의 경우, R2센서 하나만을 사용한 경우, 3개 또는 4개의 센서를 사용한 경우보다 우수한 인식율이 얻어짐을 알 수 있다. 이는, 센서의 개수 뿐이 아니고 센서의 위치도 인식율에 중요한 영향을 끼치고 있음을 의미한다. 실험 결과는 대체적으로 센서의 개수를 3개 이상 증가하면서 정체된 인식율을 나타내었다.

5.2 각 조건에 따른 인식율 비교

조건에 따른 인식율의 차이는 적은 수의 센서를 사용하는 경우 차이가 비교적 크게 나타나는 반면, 다수의 센서를 사용하는 경우 조건5를 제외하고 비교적 적게 나타났다. 이는 복수의 센서를 사용함으로써 각 조건에 따른 인식율의 차이를 보상할 수 있음을 의미한다.

Table 2. Average word recognition rate (%) for the various sensor combinations and acquisition conditions. When static features were employed.

Combinations	Cond.1	Cond.2	Cond.3	Cond.4	Cond.5
R1	96.83	91.83	88.08	88.25	79.58
R2	96.58	89.17	96.42	88.92	74.00
R3	93.67	91.50	93.83	95.83	88.00
R4	95.25	89.83	89.58	93.83	74.33
R1+R2	98.00	94.33	95.08	92.92	84.25
R1+R3	98.42	95.17	94.58	94.33	90.00
R1+R4	98.83	94.92	93.50	93.92	86.17
R2+R3	98.25	95.25	97.17	97.50	88.42
R2+R4	98.42	93.83	95.92	95.67	80.83
R3+R4	98.17	95.00	95.00	97.83	88.42
R1+R2+R3	99.17	96.00	95.25	96.75	90.58
R1+R2+R4	99.17	96.33	95.92	96.58	87.58
R1+R3+R4	99.42	97.08	95.17	96.67	91.33
R2+R3+R4	99.00	96.50	96.42	98.25	89.00
ALL	99.33	97.33	96.17	97.42	90.92

평균적으로 가장 높은 인식율을 얻은 경우는 조건1로서, 센서와 피실험자의 거리가 비교적 가깝고, 취득 시 움직임을 허용하지 않았기 때문인 것으로 판단된다. 이와 동일한 조건이면서 방사되는 초음파의 위치만 다른 조건2의 인식율은 조건1보다 다소 낮게 나타났다. 이와 같은 인식율 저하는 방사되는 초음파가 수신 센서들의 중심에 위치할 경우, 4개 센서의 반사 영역은 많은 부분이 공유되며, 센서별 취득 신호의 차이가 크게 나타나지 않는 것에 원인이 있는 듯하다. 조건1과 같이, 중심에서 벗어난 위치에서 초음파 를 방사하는 경우, 4개의 센서가 공유하는 반사면적이 줄어들고, 센서 간 변별력이 증가하면서 인식율이 향상 되는 것으로 판단된다.

초음파 도플러 신호는, 반사면의 특성과 입사/반사 각도 등의 영향을 받을 수 있으며 따라서 취득 시 사용자가 몸을 움직이는 경우, 안정적인 신호 취득이 어려워지고 인식율이 저하될 수 있다. 그러나 사용자에게 고정된 자세를 강요하는 것은 큰 불편감으로 작용할 수 있는데 조건 4는 이와 같은 상황을 고려한 것이다. 피실험자의 움직임을 어느 정도 허용한 조건4의 경우 동일한 조건에서 움직임을 제한한 조건3과 유사한 인식율을 나타내었는데, 센서의 수가 3개 이상인 경우에는 오히려 더 높은 인식율을 나타내었다. 이는 어느 정도의 움직임을 허용하더라도 높은 인식율이 유지됨을 나타낸다.

5.3 잡음 환경에서의 인식율

조건5에서는 학습을 위한 초음파/음성 신호는 조용한 환경에서 취득하고, 테스트를 위한 신호는 별도의 스피커를 통해 백색 잡음, 음악, 전투기 이륙음, 음성 등 다양한 소음을 출력시킨 상태에서 취득하였다. 이와 같은 조건에서 평균 신호 대 잡음 비(SNR: Signal to Noise Ratio)는 0 dB였으며, 음성 신호를 이용한 인식율은 11.42 %로서 매우 낮은 인식율을 보인 반면, 초음파를 이용한 인식율은 잡음이 없는 조건에 비해서는 전반적으로 낮은 인식율을 나타내었지만, R1+R3, R1+R2+ R3, R1+R3+R4, R1+R2+R3+R4 센서 조합의 경우 90 %를 상회하는 인식율을 얻을 수

있었다.

배경 잡음 발생용으로 사용한 스피커는 가청 주파수 대역만을 재생하기 때문에 배경 잡음이 초음파 센서에 직접적으로 유입될 가능성은 희박하다. 그럼에도 잡음 환경에서 낮은 인식율을 나타낸 것은, 스피커 진동면에서 초음파 반사가 일어나면서 도플러 신호가 발생되거나, 스피커에서 발생한 음파가 얼굴의 피부에 반사되면서 2차적인 진동을 일으키는 것에 원인이 있는 듯하다. 실제로 스피커의 방향과 초음파 센서의 방향이 서로 직교하는 경우에는 Table 2에 제시된 값보다 높은 인식율을 나타내었다.

신호 대 잡음 비에 따른 인식율을 살펴보면 0 dB 이상의 잡음에 대해서는 인식율이 선형적으로 증가하는 반면, 0 dB 이하에서는 인식율의 변화가 크게 관찰되지 않아 Table 2에 제시한 잡음 환경에서의 낮은 인식율은 주로 0 dB 이하의 잡음에 의한 것임을 알 수 있었다.

5.4 동적 특성을 반영한 음성 인식

Table 3은 Table 2와 동일한 조건에서 특징 변수에 동적인 특성을 반영한 델타-특징 변수를 추가한 경우의 평균 인식율을 나타낸 것이다. Table 2의 인식율과 비교하면 센서 개수, 조건, 특징 변수에 관계없이 전반적으로 높은 인식율을 나타내고 있다. 동적 특성을 추가함으로써 가장 높은 인식율의 증가를 보이는 경우는 조건3에서 R1센서를 사용하는 경우로서, 5.58 %의 인식율 향상이 관찰되었는데 이를 제외한 나머지 경우는 1 ~ 2 %의 인식율 향상이 얻어졌다.

Table 3. Average word recognition rate (%) for the various sensor combinations and acquisition conditions, when both static features and dynamic features were employed.

Combinations	Cond.1	Cond.2	Cond.3	Cond.4	Cond.5
R1	97.50	94.17	93.00	88.92	80.17
R2	97.33	91.17	96.25	91.17	77.42
R3	94.92	92.33	95.75	95.50	89.42
R4	96.42	90.58	91.58	94.58	76.67
R1+R2	98.83	96.33	97.50	94.42	85.75
R1+R3	98.58	96.33	96.17	96.42	91.92
R1+R4	99.42	96.00	94.75	95.42	88.50
R2+R3	98.25	96.17	98.25	98.00	89.92
R2+R4	99.00	95.42	97.42	97.25	84.58
R3+R4	98.08	96.08	95.08	98.25	89.75
R1+R2+R3	99.17	96.42	97.92	97.75	92.50
R1+R2+R4	99.42	97.25	97.25	97.08	88.83
R1+R3+R4	99.25	97.17	95.50	97.42	94.08
R2+R3+R4	99.17	97.17	97.50	98.92	90.33
ALL	99.58	97.58	97.75	98.33	93.08

Table 3에는 동적인 특성으로 가속 특성이 포함된 경우의 인식율이 제시되어 있지 않은데, 실험적으로 얻은 결과를 살펴보면 Table 3에 제시된 결과와 비슷하거나 약간 감소된 인식율을 보였다.

5.5 음성을 이용한 음성 인식과의 비교

13차 멜 켑스트랄 계수와 델타 및 델타-델타 변수를 포함하는 총 39차원 특징 벡터를 함께 취득한 음성 신호로부터 추출하여 음성 인식 실험을 수행하였다. 특징 변수는 초음파 신호와 동일하게 25 ms의 길이를 갖는 창함수를 10 ms 만큼 이동하여 추출하였고, 똑같은 조건의 HMM을 화자마다 독립적으로 학습하여 인식에 사용하였다. 결과를 살펴보면 조건5를 제외하고 모두 100 %, 조건5에서는 11.42 % 인식율이 얻어졌다.

VI. 결 론

한국어 음성에 대한 초음파 도플러 기반 음성 인식 결과를 제시하였다. 영어 숫자음에 대한 기존의 연구 결과가 30 ~ 60 %의 낮은 인식율을 보인 것과 비교하여 100 %에 근접한 인식율을 얻었는데, 이는 센서의 개수 증가, 취득 방법의 향상 등에 기인된 것으로 판단되며, 한국어 고립어 음성에 대한 초음파 도플러 신호의 변별 정도가 높음을 의미하는 것으로 볼 수 있다. 제안된 초음파 도플러 기반 음성인식 방법은 개당 $4 이하의 비교적 저렴한 가격으로 구현이 가능하며, 기존 마이크로폰을 이용한 음성 인식 방식과 비교하여 주변 잡음의 영향을 덜 받고, 속삭임(whispering)과 같은 작은 목소리만으로 음성 인식이 가능하여 주변 사람들에게 피해를 주지 않는 새로운 음성 인터페이스의 구현이 가능하다는 장점을 갖는다.

또한 기존의 무음 음성 인터페이스 방식은 전극이

나 센서 등을 피부에 붙이는 접촉식 방법이나 제안된 초음파 방식은 센서를 신체에 부착하지 않는 방식으로, 취득 시 어느 정도의 움직임이 허용 가능한 사용자에게 편리한 방식임을 알 수 있었다.

감사의 글

이 논문은 2015년도 건국대학교 KU학술연구비 지원에 의한 논문임.

References

1. B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," Speech Comm. **52**, 270-287 (2010).
2. K.-S. Lee, "EMG-based speech recognition using Hidden Markov Models with global control variables," IEEE Trans. on Biomed. Eng. **55**, 930-940 (2008).
3. T. Toda and K. Shikano, "NAM-to-speech conversion with gaussian mixture models," in Proc. Interspeech, 1957-1960 (2005).
4. R. Hope, S. R. Ell, M. J. Fagan, J. M. Gilbert, P. D. Green, R. K. Moore, and S. I. Rybchenko, "Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing," Speech Comm. **55**, 22-32 (2013).
5. T. Hueber, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Continuous-speech phone recognition from ultrasound and optical images of the tongue and lips," in Proc. Interspeech, 658-661 (2007).
6. M. Jiao, G. Lu, X. Jing, S. Li, Y. Li, and J. Wang, "A novel radar sensor for the non-contact detection of speech signals," Sensors **10**, 4622-4633 (2010).
7. S. Srinivasan, B. Raj, and T. Ezzat, "Ultrasonic sensing for robust speech recognition," in Proc. ICASSP, 5102-5105 (2010).
8. K. Kalgaonkar and B. Raj, "An acoustic doppler-based front end for hands free spoken user interfaces," in Proc. SLT, 158-161 (2006).
9. K. Kalgaonkar and B. Raj, "Acoustic doppler sonar for gait recognition," in Proc. 2007 IEEE Conf. Advanced Video and Signal Based Surveillance, 27-32 (2007).
10. K. Kalgaonkar and B. Raj, "One-handed gesture recognition using ultrasonic doppler sonar," Proc. ICASSP, 1889-1892 (2009).
11. K. Kalgaonkar, R. Hu, and B. Raj, "Ultrasonic doppler sensor for voice activity detection," IEEE Signal Process. Lett. **14**, 754-757 (2007).
12. K. Kalgaonkar and B. Raj, "Ultrasonic doppler sensor for speaker recognition," in Proc. ICASSP, 4865-4868 (2008).
13. K. Livescu, B. Zhu, and J. Glass, "On the phonetic information in ultrasonic microphone signals," in Proc. ICASSP, 4621-4624 (2009).
14. A. R. Toth, B. Raj, K. Kalgaonkar, and T. Ezzat, "Synthesizing speech from doppler signals," in Proc. ICASSP, 4638-4641 (2010).
15. L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition* (Prentice-Hall, New Jersey, 1993), pp. 69-83.

저자 약력

▸ 이 기 승 (Ki-Seung Lee)

1991년 2월: 연세대학교 전자공학과 학사
1993년 2월: 연세대학교 전자공학과 석사
1997년 2월: 연세대학교 전자공학과 박사
2000년 9월: AT&T Labs-Research, Senior techical staff member
2001년 8월: 삼성전자(주)종합기술원
2001년 9월 ~ 현재: 건국대학교 전자공학과 교수