한국음향학회지 제35권 제2호 pp. 134~142 (2016)
The Journal of the Acoustical Society of Korea Vol.35, No.2 (2016)
http://dx.doi.org/10.7776/ASK.2016.35.2.134

pISSN : 1225-4428
eISSN : 2287-3775

초음파 도플러 신호를 이용한 음성 합성

Speech synthesis using acoustic Doppler signal

이기승†

(Ki-Seung Lee†)

건국대학교 전자공학과

(Received October 8, 2015; revised October 30, 2015; accepted December 10, 2015)

초 록: 본 논문에서는 40 kHz 초음파 신호를 입 주변에 쏘고, 되돌아오는 초음파 신호를 이용해 음성신호를 합성하는 방법을 소개하고 성능을 평가하였다. 발성하고 있는 입주변에 초음파를 방사하게 되면, 입술, 턱, 뺨 등의 움직임으로 인한 변위로 도플러 현상이 발생하고, 이에 따라 반사 신호에는 본래의 주파수 성분과는 다른 도플러 주파수가 관찰되는데, 본 논문에서는 이러한 도플러 주파수를 이용하여 음성 파라메터를 추정하도록 하였다. 음성합성에 앞서서 초음파 도플러 신호와 음성 신호 간의 상관관계를 각 주파수 별로 분석하였으며, 이로부터 초음파 도플러 신호를 이용한 음성 신호의 합성 가능성을 살펴보았다. 변환에는 초음파 도플러의 정적, 동적 특성을 함께 반영한 특징 변수를 사용하였으며 결합-혼합 가우시안 기법을 이용하여 음성 파라메터로 변환하였다. 5명의 피 실험자를 이용한 음성 합성 실험에서 필터뱅크 에너지 값을 초음파신호의 특징변수로, LPC(Linear Predictive Coefficient) 켑스트럼 계수를 음성 변수로 사용하는 경우 가장 우수한 변환 성능을 나타내었다. 음성신호에서 추출한 여기신호를 이용하여 합성음을 생성하고, 이를 청취하였을 때 72.2 %의 평균 인식율이 얻어짐을 확인할 수 있었다.
핵심용어: 음성합성, 초음파 도플러 신호, 무음성 인터페이스, 음성변환

ABSTRACT: In this paper, a method synthesizing speech signal using the 40 kHz ultrasonic signals reflected from the articulatory muscles was introduced and performance was evaluated. When the ultrasound signals are radiated to articulating face, the Doppler effects caused by movements of lips, jaw, and chin observed. The signals that have different frequencies from that of the transmitted signals are found in the received signals. These ADS (Acoustic-Doppler Signals) were used for estimating of the speech parameters in this study. Prior to synthesizing speech signal, a quantitative correlation analysis between ADS and speech signals was carried out on each frequency bin. According to the results, the feasibility of the ADS-based speech synthesis was validated. ADS-to-speech transformation was achieved by the joint Gaussian mixture model-based conversion rules. The experimental results from the 5 subjects showed that filter bank energy and LPC (Linear Predictive Coefficient) cepstrum coefficients are the optimal features for ADS, and speech, respectively. In the subjective evaluation where synthesized speech signals were obtained using the excitation sources extracted from original speech signals, it was confirmed that the ADS-to-speech conversion method yielded 72.2 % average recognition rates.
Keywords: Speech synthesis, Ultrasonic Doppler signals, Silence speech interface, Voice conversion

PACS numbers: 43.72.Ja, 43.72.Kb

I. 서 론

인간은 음성발성이 가능한 유일한 존재로서, 음성을 의사 전달의 중요한 수단으로 사용한다. 공기를 매질로 하는 전통적인 음성 전달 방법은 주변에 소음이 없는 경우 원활한 의사소통이 가능하지만, 주변 소음이 매우 큰 경우 음성을 통한 대화가 불가능하다. 사회가 복잡해지면서 사람과 사람 간에 접촉 기회가 증가함에 따라 서로간의 에티켓이 중요시 되

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

고 있는데, 예로서 도서관, 영화관과 같은 공공장소에서 큰 소리로 대화를 하거나 통화를 하는 것은 예의에서 벗어난 행동으로 간주되고 있다. 또한 대화 내용이 다른 사람들에게 노출되지 않아야 하는 상황에서는 공기로 전파되는 음성을 인위적으로 차단할 필요가 있다. 이러한 특수 상황에서 음성을 이용한 의사전달 방법으로 무음성 인터페이스(silence speech interface)[1] 방법이 제안되었다.

무음성 인터페이스에서는 기본적으로 음성을 발성시키기 위한 동작과 행위만 취하되, 실제 소리는 내지 않아도 의사전달이 가능한 방법이다. 예로서 입 주변에 근전도 취득을 위한 전극을 붙이고, 취득된 근전도 신호로부터 음성 신호를 추정하는 방법,[2]

작은 목소리로 속삭일 때 얼굴의 근육부위에서 음성의 미세한 진동을 취득하는 NAM(Non-Audible Microphone)[3]을 사용하는 방법, GHz microwave를 입 주변에 쏘고, 되돌아오는 신호의 도플러를 이용한 방법,[4] 초음파 도플러를 이용한 방법[5] 등이 있다.

근전도, NAM을 이용한 방법은 센서가 피부에 항시 부착되어 있어야 하며 이와 관련된 감염, 알러지 등이 발생할 수 있으며 외관상의 문제가 있다. GHz 전파를 사용한 방법은 접촉식 센서를 사용하지 않아 사용자에게 편리한 반면 다른 센서에 비해 크기와 무게, 그리고 비용 측면에서 불리한 방식이다. 초음파 도플러를 이용한 방법은 비접촉식 방법의 장점을 지니면서 소형, 휴대가 간편하고 저렴한 가격으로 제작이 가능하다. 본 논문에서는 초음파센서를 이용하여 무 음성 인터페이스를 구현하였다.

초음파 도플러는 변위가 있는 물체에 반사되어 돌아오는 초음파신호가 본래와는 다른 주파수를 갖는 현상이다. 복잡한 움직임이 있는 경우, 각 부위의 변위에 따라 개별적인 도플러신호가 발생되고, 수신된 초음파신호에는 각 도플러신호가 중첩되어 나타난다. 이때 도플러신호는 움직임 패턴에 따라 의존적으로 나타나는데, 이를 이용하여 인간의 동작을 인식하는 연구가 Kalgaonkar *et al.*[6-8]에 의해 진행되었다. 결과를 살펴보면 간단한 손동작은 평균 88.4 %, 보행패턴을 인식하는 경우 91.7 %의 인식율이 얻어지는 것으로 보고되어 기존의 영상기반 동작인식 방법을 대치할 수 있는 가능성을 보여주었다. 이러한 연구 결과에 힘입어 초음파 도플러를 이용해 음성을 인식하고자 하는 연구가 Srinivasan *et al.*,[9] Livescu *et al.*[10] 등에 의해 시도되었으나 기존의 음성인식에 비해 매우 낮은 인식율(30 ~ 60 %)을 나타내었다. Toth *et al.*[11]은 음성 변환에 사용되는 혼합가우시안 모델(Gaussian Mixture Model, GMM)을 이용하여 초음파 도플러신호를 음성신호로 변환하였는데, 청취 시 인지 가능한 합성음을 얻을 수 있다고 보고하였다.

기존 연구는 모두 영어를 대상으로 하였는데 본 연구에서는 한국어 음소를 골고루 포함하는 60개의 한국어에 대해 음성합성 실험을 수행하였다. 이로부터 초음파 도플러 기반 음성 합성 방법이 한국어에 적용 가능 여부를 살펴보았다. 기존의 초음파 음성 합성 방법이 단일채널 초음파 신호만을 사용한 것과 비교하여, 본 연구에서는 4채널 초음파 신호를 이용, 다양한 방향에서 수신되는 초음파 신호를 이용하여 음성을 합성하였다. 본 논문에서는 음성 합성에 앞서서 초음파 도플러 신호와 음성 신호간의 상관성을 정량적으로 분석하였으며, 이로부터 음성 합성의 가능성을 살펴보았다. 음성 합성의 유효성은 음성 파라메터의 왜곡을 이용한 객관적 검증과 청취테스트를 통한 주관적인 검증을 통해 확인하였다.

II. 초음파 신호 분석 및 변수 추출

Fig. 1과 같이 발성하고 있는 사람의 입주변에 일정한 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 주파수 f_c, 크기가 A_T, 위상이 ψ_T인 정현파 신호라면,

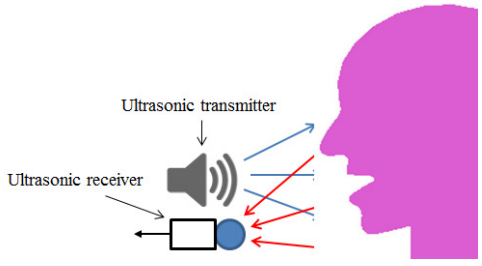

Fig. 1. Speech acquisition using ultrasound Doppler effects.

$$T_r(t) = A_T \cos(2\pi f_c t + \psi_T). \quad (1)$$

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 속도로 움직이고 있다면, 수신된 신호는 아래와 같이 나타낼 수 있다.

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

여기서 ψ_i는 i-번째 성분에 대한 위상변이를 나타낸다. $R_e(t)$에 대해 복조(demodulation)를 수행하고 DC값과 f_c 이상의 주파수 성분을 차단하는 대역통과필터를 통과시키면 발성과 관련된 성분을 추출할 수 있다.[9]

기존의 초음파 도플러를 이용한 음성합성 기법에서는 초음파 신호에 대한 특징변수로 음성 신호의 특징 변수로 널리 사용되는 멜 주파수 켑스트럼 계수(mel-frequency cepstral coefficient)가 사용되었다.[11] 켑스트럼 계수는 인간의 청각 특성을 반영한 멜 주파수 대역별 에너지 값을 이산 여현 변환(discrete cosine transform)하여 얻는다.

$$D_k = \sum_{i=1}^{N_B} Y_i \cos\left[\frac{\pi}{N_B} \left(i + \frac{1}{2} \right) k \right], \quad (3)$$

여기서 $1 \le k \le N_C$, D_k는 k-번째 켑스트럼 계수를 나타내며, N_C는 전체 켑스트럼 계수의 수, 그리고 Y_i와 N_B는 $i-$번째 멜 주파수의 에너지 와 전체 멜 주파수 계수를 나타낸다.

초음파 도플러의 주 근원(primary source)이 발성시 근육의 움직임과 목 주변의 진동이라 가정하면, 취득된 초음파 신호는 각 근원의 특성과 유사한 특성을 지닐 것으로 예측할 수 있다. 얼굴 부위 근전도 신호를 이용한 음성 인식의 연구[2]에서 멜 주파수 단위로 대역 통과된 신호를 특징 변수로 사용했을 때 가장 높은 인식율이 얻어지는 것으로 나타났으며, 목 주변의 진동은 음성 파형과 유사한 모양을 나타내는 것으로 알려져 있다.[4] 이는 취득된 초음파 신호가 송신 초음파 신호의 주파수를 중심주파수로 변조된 음성신호 및 근전도 신호와 매우 유사함을 나타낸다. 이를 고려하여 본 논문에서는 아래와 같은 멜 밴드별 log 에너지 값을 특징변수로 사용하였다.

$$Y_k = \log \left[\sum_{m=1}^{M} |X(m)| H_k(m) \right], \quad (4)$$

여기서 Y_k는 k-번째 멜 밴드에 대한 로그 에너지 값을 나타내며, $X(m)$은 수신된 초음파 신호의 푸리어 변환 값, $H_k(m)$은 k번째 멜 밴드의 대역통과 특성을 나타낸다. 각 멜 밴드별 대역통과 특성을 Fig. 2에 나

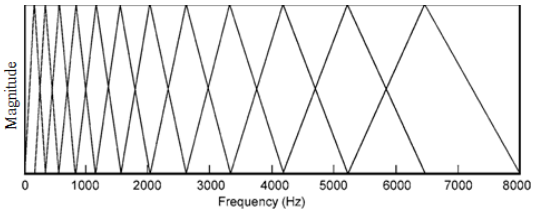

Fig. 2. Band transfer characteristics for each mel band.

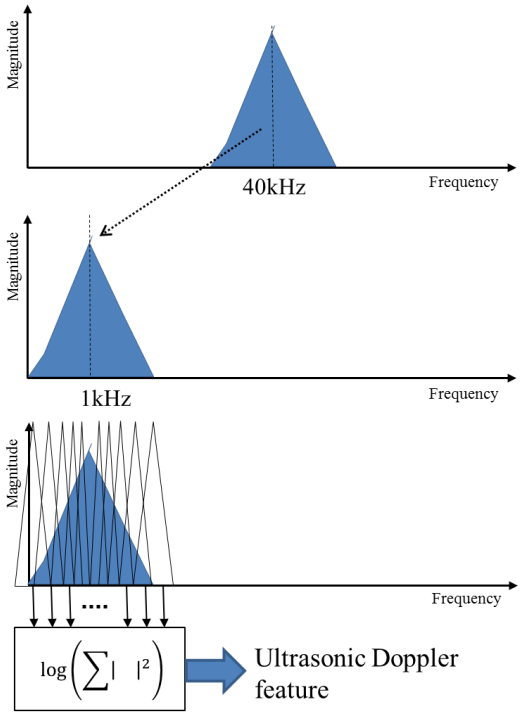

Fig. 3. Procedure for extracting feature parameter.

타내었다. 취득된 초음파 신호는 40 kHz의 주파수를 중심으로 분포하므로 Fig. 3에 설명된 과정을 거쳐 특징 변수를 얻도록 하였다.

III. 초음파 도플러신호의 음성변환

3.1 상관분석

초음파 도플러 신호를 음성신호로 변환하기에 앞서, 두 신호간의 상관관계를 분석하였다. 상관 분석 시 초음파 신호는 II장에서 제시한 필터뱅크 에너지 값과 델타 값을 특징 변수를 사용하였다. 음성 신호는 주파수별로 상관관계를 살펴보기 위해 푸리어 변환 후 각 멜 밴드의 에너지 값을 특징 변수로 사용하였다. 사용된 상관계수는 아래와 같다.[12]

$$R^2(i)=1-\frac{\sum_{t=1}^{T}\left[y_t(i)-\hat{y_t}(i)\right]^2}{\sum_{t=1}^{T}\left[y_t(i)-\overline{y_t}(i)\right]^2}, \quad (5)$$

여기서 $y_t(i)$는 t번째 음성특징변수의 i번째 성분을 나타내며 T는 각각 전체 데이터의 수를 나타낸다. $\hat{y_t}(i)$는 초음파 도플러 신호에 의해 예측된 특징 변수로서, 아래와 같이 주어진다.

$$\hat{y_t}(i)=\sum_{j=0}^{J-1}b_i(j)x_t(j), \quad (6)$$

여기서 $x_t(j)$는 초음파 신호의 j번째 특징변수를 나타내며, $b_i(j)$는 선형예측 계수로서, $y_t(i)$와 $\hat{y_t}(i)$의 자승오차합이 최소화되도록 얻어진다. Eq.(5)의 R^2 값은 두 신호 간 상관관계가 높을수록 1에 가까운 값을 갖는다.

Table 1에 각 특징변수에 따른 평균 상관계수 값을 제시하였다. 이 값은 5명의 피 실험자로부터 취득한 300개단어의 음성 및 초음파 신호로부터 $R^2(i)$를 얻고, 모든 주파수성분 i에 대해 평균과 표준편차를 구한 값이다. 표에서 N-Neighboring은 현재 특징벡터를 중심으로 좌, 우 N개의 특징벡터를 포함시킨다는 것을 의미하며 따라서 $N=8$인 경우 좌, 우 각각 4개의 특징벡터를 포함하여 새로운 특징벡터를 구성함을 나타낸다. 표에서 보면 동적인 특성을 반영하지 않고 필터뱅크에너지 단독으로 사용한 경우 상관 값은 작고 표준편차는 큰 것으로 나타났으며, 인접된 값을 많이 포함시킬 수 록, 상관 값이 증가됨을 알 수 있다. 따라서 되도록 다수의 인접 특징 벡터를 포함시키는 것이 합성음이 실제 음성과 가까워질 것으로 판단된다. 그러나 많은 특징벡터를 포함시킬 경우 특징벡터의 차원수가 크게 증가하고, 이는 변환규칙 생성 시 특이성(singularity)과 관련된 문제가 발생할 수 있다.

Fig. 4는 주파수 별로 평균 상관 계수 값과 표준 편차 값을 도시한 것으로($N=4$인 경우), 1000 Hz까지 평균 상관 값이 증가하고, 이후 8000 Hz까지 0.7 이상의 상관값이 유지됨을 알 수 있다. 이로부터 Table 1에 제시한 $N=4$인 경우의 평균 상관 값 0.7을 넘지 못한 것은 1000 Hz 이하 대역의 작은 상관 값 때문인 것으로 판단된다. 음성 신호의 주파수 대역을 4 kHz로

Table 1. Average and standard deviation of the correlation coefficients over all bands, according to the ultrasonic features.

Features of ultrasonic signal	AVG	STD
FBANK	0.6105	0.0272
FBANK+△FBANK	0.6424	0.0234
FBANK+△FBANK+△²FBANK	0.6692	0.0211
FBANK+2-Neighboring FBANK	0.6559	0.0225
FBANK+4-Neighboring FBANK	0.6825	0.0198
FBANK+6-Neighboring FBANK	0.7022	0.0180
FBANK+8-Neighboring FBANK	0.7180	0.0170

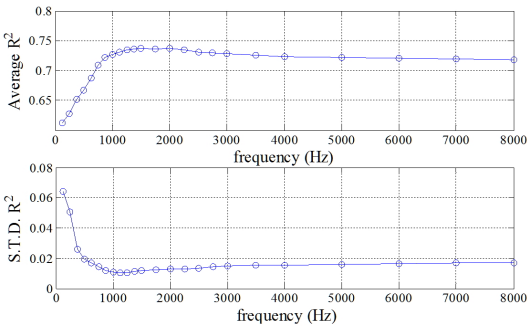

Fig. 4. Average and standard deviation of the corrlea-tion coefficients for each mel frequency.

가정하였을 때, 음성 신호의 상당수 대역이 초음파 도플러 신호와 0.7 이상의 높은 상관관계를 가짐을 알 수 있다. 또한 화자별 고유 음색과 관련이 있는 4 kHz의 고주파 대역에서도 높은 상관값을 보였는데, 이는 화자 고유 특성과 초음파 도플러신호 간에도 높은 상관성이 존재하여, 초음파 도플러를 이용해 음성을 합성할 경우, 화자의 음색도 유지될 수 있음을 의미한다.

3.2 초음파도플러-음성 간 변환

본 논문에서는 Nakamura *et al.*[5]에 의해 제안된 GMM 기반의 궤적 변환 기법을 초음파 특징 변수와 음성 특징 변수간의 변환 규칙을 생성하는데 사용하였다. 앞 절의 실험결과에 따르면 인접된 특징 변수도 함께 고려할 때 초음파 신호와 음성 신호 간 상관관계가 증가하는 것을 알 수 있었다. 이를 고려하여 변환을 위한 입력 특징 변수는 다음과 같이 나타내었다.

$$X_t = (\mathbf{x}_{t-N}^T \cdots \mathbf{x}_t^T \cdots \mathbf{x}_{t+N}^T). \quad (7)$$

본 연구에서는 $N=2$로 설정하였다. 음성으로부터 추출한 특징 변수 $\mathbf{y}_t = [y_t(0) \cdots y_t(J)]^T$를 포함시킨 벡터 $Z_t = (\mathbf{x}_{t-N}^T \cdots \mathbf{x}_t^T \cdots \mathbf{x}_{t+N}^T \mathbf{y}_t)^T$를 구성하면 Z_t는 M개 가우시안을 포함하는 혼합가우시안 모델 λ로 나타낼 수 있다.

$$P(Z_t|\lambda) = \sum_{m=1}^{M} \alpha_m N[Z_t; \mu_m^{(z)}; \Sigma_m^{(z)}], \quad (8)$$

여기서 α_m은 m 번째 가우시안의 가중치로서 $\sum_{m=1}^{M} \alpha_m z = 1$이고, $N[\bullet; \mu_m^{(z)}, \Sigma_m^{(z)}]$은 평균벡터 $\mu_m^{(z)}$, 공분산행렬 $\Sigma_m^{(z)}$을 갖는 가우시안 함수를 나타낸다. $\mu_m^{(z)}$ 와 $\Sigma_m^{(z)}$은 각각 다음과 같이 나타낼 수 있다.

$$\mu_m^{(z)} = \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{bmatrix}, \quad \Sigma_m^{(z)} = \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix}, \quad (9)$$

여기서 $\mu_m^{(x)}$와 $\mu_m^{(y)}$는 각각 X_t와 Y_t에 대한 평균벡터를, $\Sigma_m^{(x)}$와 $\Sigma_m^{(y)}$는 각각 X_t와 Y_t에 대한 공분산 행렬을 나타내며, $\Sigma_m^{(xy)}$와 $\Sigma_m^{(yx)}$ 상호공분산 행렬을 나타낸다. 본 논문에서 $\Sigma_m^{(x)}$와 $\Sigma_m^{(y)}$는 대각행렬로 나타내었다.

모델 λ를 구성하는 각 파라메터들은 학습데이터를 이용하여 EM-알고리즘을 이용하여 얻을 수 있다.

변환 식 $F(X_t)$는 Y_t와 $\hat{Y}_t = F(X_t)$의 평균자승오차가 최소화되도록 얻어지며 이때의 $F(X_t)$는 다음과 같다.

$$F(X_t) = \hat{Y}_t = E(Y|X_t) = \sum_{m=1}^{M} h_m(X_t)\left[\mu_m^{(y)} + \Sigma_m^{(yx)}\Sigma_m^{(xx)^{-1}}(X_t - \mu_m^{(x)})\right], \quad (10)$$

$$h_m(X_t) = \frac{\alpha_m N(X_t; \mu_m^{(x)}, \Sigma_m^{(xx)})}{\sum_{m=1}^{M} \alpha_m N(X_t; \mu_m^{(x)}, \Sigma_m^{(xx)})}. \quad (11)$$

IV. 실험 및 결과

초음파 신호의 방사를 위한 센서로 40 kHz의 중심주파수를 갖는 소형 초음파트랜스듀서(AW8TR40, Audiowell, China, 음압레벨 115 dB)가 사용되었고 수신용으로는 광대역 특성을 갖는 초소형 MEMS 센서(SPM0404UD5, Knowles Acoustic, Japan)를 사용하였다. 수신감도는 -47 dB (10 kHz ~ 65 kHz)이며 수신부에 사용된 센서의 사진을 Fig. 5의 왼쪽에 제시하였다. 음성을 동시 취득하기 위해 가청주파수 대역의

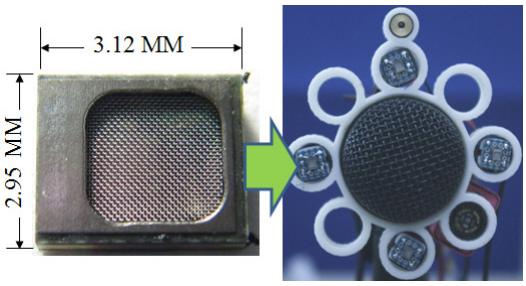

Fig. 5. Photography of the employed ultrasonic sensor (left) and frontal photography of the mircophone - integrated ultrasonic/audio acquisition apparatus (right).

마이크로폰(AKG880, AKG, Austria)이 사용되었으며, 각 센서와 레이저 포인터 및 마이크로폰은 별도 제작된 홀더를 이용하여 Fig. 5의 오른쪽 사진과 같이 배치하였다. 이러한 배치는 수차례의 실험을 통해 상관계수가 가장 크게 나타나는 경우를 경험적으로 찾은 것이다.

40 kHz 초음파 신호는 함수발생기(33250A, Agilent, USA).를 사용하여 발생시켰으며 4개의 초음파 센서에서 수신된 각 신호 및 마이크로폰에서 취득된 음성 신호를 동시에 디지털 값으로 변환하기 위해 다채널 오디오인터페이스(Fireface 800, RME, Germany)를 사용하였다. 초음파 신호 및 음성 신호 모두 샘플링 주파수는 192 kHz, 양자화 비트수는 16비트로 설정하였으며 각 채널 이득은 동일하게 맞추었다.

초음파 도플러 신호를 이용한 음성 합성 기법의 유효성을 검증하기 위하여 Table 2에 제시된 60개 한국어 고립어에 대해 실험을 수행하였다. 실험은 5명의 피실험자(남성 5명 M1-M4, 여성 1명 F, 모두 20대)로부터 각 단어 당 50회 반복 취득하였다. 이 중 30개는 변환규칙을 생성하는데 사용하고, 나머지 20개는 테스트에 사용하였다. 신호 취득은 비교적 조용한 환경에서 제작된 취득 장치를 이용하여 초음파 신호와 음성 신호를 동시에 취득하였다.

음성 합성을 위한 다양한 파라메터로 본 논문에서는 LPC 켑스트럼 계수(LPC Cpetrum Coefficient, LPCC), 쌍선형 계수(Line Spectrum Pair, LSP), 멜 켑스트럼 계수(Mel Cepstral Coefficients, MCC)가 사용되었으며, 각 파라메터에 대한 객관적인 성능을 비교, 평가하고 가장 우수한 성능을 나타내는 파라메터를 선택하여 합성음을 생성하는데 사용하였다.

4.1 객관적 성능평가

객관적인 척도로서 아래 식으로 주어지는 평균 거리 감소율(average distance reduction ratio)과 평균 스펙트럼왜곡(average spectral distorion)이 사용되었다.

$$\overline{\mathrm{DRR}}=\left(1-\frac{\sum_{t=1}^{T}\|\mathbf{y}_t-\hat{\mathbf{y}}_t\|^2}{\sum_{t=1}^{T}\|\mathbf{y}_t-\bar{\mathbf{y}}\|^2}\right), \tag{12}$$

$$\mathrm{SD}=\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\log_{10}\frac{P(\omega)}{\hat{P}(\omega)}\right]^2, \tag{13}$$

여기서 $P(\omega)$와 $\hat{P}(\omega)$는 각각 음성파라메터 $\mathbf{y}_t$와 $\hat{\mathbf{y}}_t$를 통해 얻어진 스펙트럼을 나타낸다. $\overline{\mathrm{DRR}}$은 원래 신호가 갖고 있는 분산 값과 비교하여 상대적인 거리 감소 정도를 나타내는 척도로서 100에 가까울수록 높은 성능을 나타내며, SD는 절대왜곡으로서 작을수록 좋은 성능을 나타낸다.

Table 3에 피 실험자별 평균거리 감소율과 평균 스팩트럼 왜곡을 음성 파라메터별로 제시하였다. 모든 실험자에 대한 평균값을 살펴보면 LPC 켑스트럼 계

Table 2. 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]

Table 3. Distance Reduction Ratio (DRR) and Spectral Distance (SD) for each subjects, for each speech parameters and the averages and standard deviations.

subject	LPCC		LSP		MCC	
	DRR	SD	DRR	SD	DRR	SD
M1	32.45	35.59	34.93	35.30	35.75	46.74
M2	26.62	43.44	23.47	44.47	26.32	56.98
M3	31.38	37.72	30.57	38.91	28.36	52.72
M4	28.21	36.93	26.93	38.41	22.75	51.97
F	36.58	36.29	34.33	37.94	33.00	48.71
AVG.	31.05	37.99	30.05	39.01	29.24	51.42
STD.	3.88	3.14	4.88	3.36	5.19	3.94

수가 평균 거리 감소율 및 스펙트럼 왜곡에 있어서 가장 우수한 성능을 나타내었다. 또한 표준 편차값도 LPC 켑스트럼이 가장 작게 나타나 화자 간 변환 성능 차이 면에서도 우수한 파라메터임을 알 수 있다.

Table 3에 제시된 값은 일반적인 음성 부호화기 에서 관찰되는 왜곡 값과 비교하여 매우 큰 값이나, 과거 무 음성 인터페이스 연구에서 보고된 왜곡 값과는 대체적으로 비슷한 값을 나타내었다. MCC는 상대적으로 저하된 성능을 나타내었는데, 이는 LPC계열의 파라메터가 초음파 도플러를 이용한 음성합성에 적합한 파라메터임을 나타낸다. 이와 같은 결과에 따라 합성음 생성에는 LPC켑스트럼을 사용하였다.

4.2 주관적 성능평가

변환 규칙을 이용하여 얻어진 LPC켑스트럼 계수를 이용하여 음성을 합성하고 합성음을 평가하였다. 청취 가능한 음성을 합성하기 위해서는 추정된 LPC 켑스트럼 외에 선형예측 잔차 신호(Linear Prective residual; LP-residual) 신호가 필요하다. 본 논문에서는 유성음 구간에서는 주기적인 임펄스 열, 무성음 구간에서는 백색 잡음을 여기신호로 사용하거나 본래의 음성에서 얻어진 선형예측 잔차 신호(natural excitation)을 그대로 사용하여 음성을 합성하였다.

Fig. 6에 음성 "나비"에 대한 본래 음성과 초음파 도플러 신호를 이용해 합성한 음성의 파형과 스펙트로그램을 나타내었다. 합성된 음성의 스펙트로그램을 살펴보면 본래 음성의 포만트와 유사한 궤적을 보이고 있으나, 확대해서 살펴보면 본래의 포만트 궤적과 비교하여 번짐 현상(blurring effects)이 다소 관찰

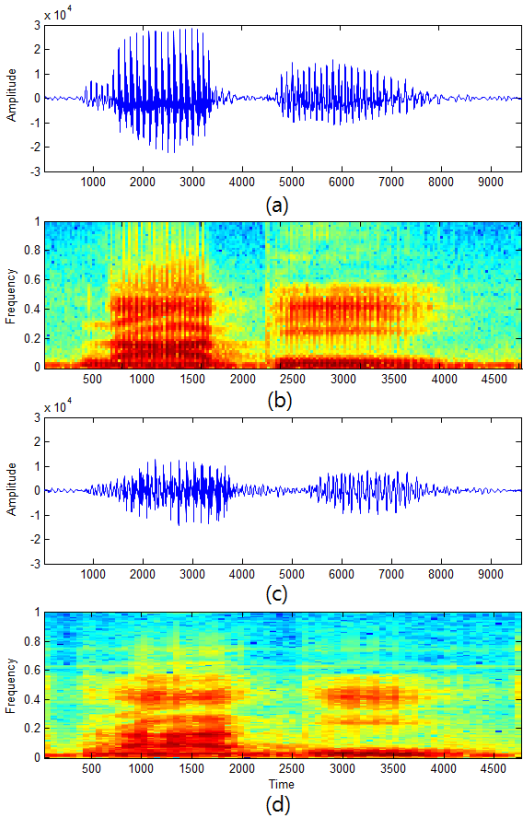

Fig. 6. An example of the speech waveforms: (a) original speech waveform [na-bi], (b) spectrogram of original speech, (c) reconstructed speech waveform, (d) Bottom: spectrogram of reconstructed speech.

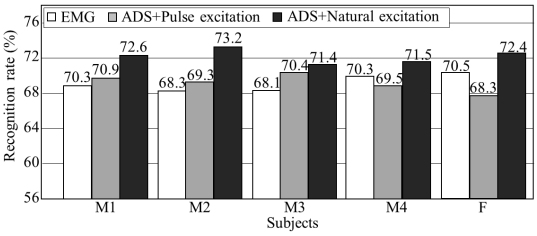

Fig. 7. Correct word identification ratios for each speaker, for each method (EMG:EMG-to-Speech, ADS+Pulse excitation: Acoustic Doppler-to-Speech using pulse excitation, ADS+Natural excitation: Acoustic Doppler-to-Speech using natural excitation).

되었다. 이는 본 논문에서 사용한 파라메터 변환 방법이 혼합가우시안 기반 방법으로서, Eqs.(9)와 (10)과 같이 선형조합에 의해 음성파라메터를 얻는 것에 원인이 있는 듯하다. 즉, 여러 개의 파라메터들이 더해지면서 평균현상(averaging effects)이 발생하게 되고 이로 인해 뚜렷한 포만트 특성이 손실되는데, 이

로 인해 청취 시 머플링 현상(muffling artifacts)이 합성음에서 감지되었다.

Fig. 7은 청취 테스트의 결과로서, 근전도 기반 음성 합성방법[2]과 비교하였다. 본 논문에서 사용된 주관적 척도인 인식율은 합성음을 청취자에게 들려주었을 때, 올바르게 인식된 단어수를 전체 단어수로 나눈 값(%)이다. 청취테스트에는 15명의 정상청력을 가진 피 실험자가 참여하였으며, 일부 참가자는 이전에 유사한 실험에 참여한 경험이 있었다. 결과를 살펴보면 실제 LP-residual을 사용한 경우 근전도를 이용한 합성방법에 비해 모든 화자의 음성에 대해 우수한 성능을 나타내었다. 청취자들은 초음파를 통해 합성된 음성이 상대적으로 잡음이 적고, 자음 부분이 좀 더 명료하게 들려 전반적으로 음질이 우수하다는 의견을 표시하였다.

유/무성음 구간에 따라 주기적인 임펄스 열과 백색잡음을 사용한 경우는 잔차 신호를 사용한 경우와 비교하여 모든 화자에 걸쳐 성능저하가 관찰되었으며, 근전도 음성합성 기법과 비교하여 약간 우수하거나(M1, M2, M3), 저하된 인식율을 나타내었다(M4, F). 이는 LP-잔차 신호에도 성도전달함수의 특성이 다수 포함되어 잔차 신호를 사용하여 합성된 음성의 인식율이 증가된 것에 원인이 있는 듯하다.

테스트결과를 살펴보면, 모음에 대해서는 대체로 높은 인식율(96 %)을 보이는 반면, 초성자음에 대해서는 낮은 인식율(82 %)을 나타내어 초성자음의 합성음 품질이 인식율에 주된 영향을 끼치는 것으로 관찰되었다. 이는 유성자음(예: “바람”을 “사람”으로, “근세”를 “뜬세”로 인식)과 무성자음(예:“찌게”를 “희게”로, “틀”을 “풀”로 인식)의 구분 없이 공통적으로 나타내어 초성자음에 대한 음질 향상이 요구된다고 판단된다.

V. 결 론

한국어 음성에 대한 초음파 도플러 기반 음성 합성 결과를 제시하였다. 제안된 초음파 도플러 기반 음성합성 방법은 비교적 저렴한 비용으로 구현이 가능하며, 기존 마이크로폰을 이용한 음성 취득 방식과 비교하여 주변 잡음의 영향을 덜 받고, 속삭임(whispering)과 같은 작은 목소리만으로 상대방에게 의사 전달이 가능한 방법이다. 이러한 특정을 이용하여 주변 사람들에게 피해를 주지 않는 새로운 음성 인터페이스의 구현이 가능할 것으로 기대된다.

본 논문에서는 한국어 고립어에 대한 실험을 수행하고 성능을 평가하였는데, 보다 실용적이 되기 위해서는 기본주파수, 여기신호와 같은 음성 변수들을 어떻게 추정할 것인지에 대한 연구가 추후에 진행되어야 할 것으로 판단된다.

감사의 글

이 논문은 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (과제번호: 2015R1D1A1A01059626).

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, “Prediction of acoustic feature parameters using myoelectric signals,” IEEE Trans. on Biomed. Eng. **51**, 1587-1595 (2010).
3. T. Toda and K. Shikano, “NAM-to-Speech conversion with Gaussian Mixture Models,” in Proc. Interspeech, 1957-1960 (2005).
4. S. Li, J. Q. Wang, M. Niu, T. Liu, and X. J. Jing, “The enhancement of millimeter wave conduct speech based on perceptual weighting,” Progress in Electromagnetics Research B, **9**, 199-214 (2008).
5. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech,” Speech Comm. **54**, 134-146 (2012).
6. K. Kalgaonkar and B. Raj, “An acoustic Doppler-based front end for hands free spoken user interfaces,” in Proc. SLT, 158-161 (2006).
7. 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).
8. K. Kalgaonkar and B. Raj, “One-handed gesture recognition using ultrasonic Doppler sonar,” Proc. ICASSP, 1889-1892 (2009).
9. S. Srinivasan, B. Raj, and T. Ezzat, “Ultrasonic sensing for robust speech recognition,” in Proc. ICASSP, 5102-5105

(2010).
10. K. Livescu, B. Zhu, and J. Glass, “On the phonetic information in ultrasonic microphone signals,” in Proc. ICASSP, 4621-4624 (2009).
11. A. R. Toth, B. Raj, K. Kalgaonkar, and T. Ezzat, “Synthesizing speech from Doppler signals,” in Proc. ICASSP, 4638-4641 (2010).
12. I. Almajai and B. Milner, “Visually derived Wiener filters for speech enhancement,” IEEE Trans. on Audio, Speech, and Lang. Process. **19**, 1642-1651 (2011).

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