PAPER

Silent Speech Interface Using Ultrasonic Doppler Sonar

Ki-Seung LEE[†a)], ***Member***

SUMMARY Some non-acoustic modalities have the ability to reveal certain speech attributes that can be used for synthesizing speech signals without acoustic signals. This study validated the use of ultrasonic Doppler frequency shifts caused by facial movements to implement a silent speech interface system. A 40kHz ultrasonic beam is incident to a speaker's mouth region. The features derived from the demodulated received signals were used to estimate the speech parameters. A nonlinear regression approach was employed in this estimation where the relationship between ultrasonic features and corresponding speech is represented by deep neural networks (DNN). In this study, we investigated the discrepancies between the ultrasonic signals of audible and silent speech to validate the possibility for totally silent communication. Since reference speech signals are not available in silently mouthed ultrasonic signals, a nearest-neighbor search and alignment method was proposed, wherein alignment was achieved by determining the optimal pair of ultrasonic and audible features in the sense of a minimum mean square error criterion. The experimental results showed that the performance of the ultrasonic Doppler-based method was superior to that of EMG-based speech estimation, and was comparable to an image-based method.

key words: *silent speech interface, ultrasonic Doppler, deep neural networks*

1. Introduction

In conventional voice communication, voices are spoken by a speaker (generation), speech signals are transmitted through the air (transmission), and the messages are delivered through the recipient's ears (reception). When specific problems occur or some constraints should be considered in each voice communication step, voice messages may not be delivered correctly. A silent speech interface (SSI) techniques [1] has been proposed to cope with these situations. This technique can be applied to specific situations where speech communication suffers from several inherent problems and constraints. SSI is useful for persons with speaking impairment, for example, who have undergone a laryngectomy caused by an accident or laryngeal cancer. SSI technology allows these patients to speak in their own voice using assistive devices. The SSI technique is also applicable when an acoustic signal is no longer useful for communication among humans (e.g., very noisy environments). SSI allows persons to speak to one another in a public place, such as a library or conference room without disturbing others. Moreover, it can be useful in a situation where private information should not be audibly exposed to others. An important application of SSI is "Silent Call" service, which was created to facilitate communication with emergency responders by callers who could not speak with them due to the nearby presence of criminals who might seek to terminate the call and or harm the caller if they became aware that their victim is summoning help.

The non-acoustic modalities adopted for SSI should be highly correlated with the corresponding audio speech signals and not affected by high levels of ambient audio interferences [2]. The modalities satisfying these conditions include Doppler frequency shifts caused by GHz microwaves [3]–[7], the ultrasound images (UI) of a vocal tract [8]–[10], the visual shape of the mouth [11]–[16], acoustic Doppler sonic signals [17]–[20], recorded signals by a non-audible microphone (NAM) [21], [22], and an electromyogram (EMG) [23]–[27].

For the EMG- and NAM-based methods, the non-acoustic signals associated with speech are directly acquired through facial expressions. Accordingly, this method has the advantage of robustness against the environmental interferences that are commonly found in vision-based methods. However, the use of contact sensors can be uncomfortable for users and could cause skin allergies. Methods that utilize visual information employ the images captured from a speaker's mouth region as a secondary source of speech information. Since it is unnecessary for a user to have electrodes attached to the face, the vision-based methods may be more convenient than the methods that require contact sensors. Capturing images and windowing the mouth region is not a trivial task [14], however, and overall performance may be more or less affected by the techniques associated with image processing.

The Doppler effect is the change in frequency of a wave when an observer moves relative to the source of the wave. When a fixed frequency wave is incident to the moving object, the echoes are Doppler shifted, which creates components at other frequencies, that are proportional to their velocity relative to the sensor. If multiple objects are moving with different velocities, the reflected signal will contain multiple frequencies, one for each object. This principle applies to the recognition of various human motions [28]–[30] and it can be used to detect speech signals [18]–[20]. With Doppler-based speech processing techniques, the underlying assumption is that the primary sources of Doppler frequency shifts are vocal vibrations of the body surface caused by sound. In the Doppler-based methods, the de-

Manuscript received August 4, 2019.
Manuscript revised November 25, 2019.
Manuscript publicized May 20, 2020.

[†]The author is with the Department of Electronic Engineering, Konkuk University, Seoul, 143–701 Korea.
a) E-mail: kseung@konkuk.ac.kr

DOI: 10.1587/transinf.2019EDP7211

tection of speech signals is achieved by means of non-contact sensing. Hence, the problems associated with contact sensing, found in the EMG- and NAM-based methods, can be avoided. Moreover, since image-related information is unnecessary to detect a speech signal, the Doppler-based techniques are free from image-related problems. An approach based on microwave Doppler, which is referred to as *"radar-microphone"* was proposed in the literature for detecting speech [3], [4] and for the detection of vocal cords vibrations [7]. Such an approach has the advantage of a high level of directional sensitivity with penetration and a long detection distance. However, the radar speech itself has several serious shortcomings, which include artificial quality, reduced intelligibility, and poor audibility [5], [6]. The requirement of large and complicated hardware is another drawback of radar microphone techniques. Doppler frequency shifts are also observed in the reflections of tone emitted by an ultrasonic transmitter. A method that uses an ultrasound transducer has several advantages over the other methods: lighter in weight, smaller in size, less expensive, and allows for non-contact sensing. Accordingly, ultrasonic Doppler signals (UDS) are used for various speech processing fields, such as voice activity detection [19], speech recognition [20], and speech synthesis [17].

The purpose of this study was to verify the usefulness of Doppler frequency shifts generated by single tone ultrasonic signals, particularly for synthesizing audible speech. One straightforward way to implement UDS speech synthesizers is to use both an automatic speech recognizer and a text-to-speech (TTS) system. For such a method, it is possible to synthesize unlimited utterances by using phoneme-level ASR, and high-quality speech can be obtained by employing a corpus-based waveform concatenating TTS. However, the intelligibility of the resultant speech is significantly affected by the accuracy of the adopted ASR. Moreover, there are some problems associated with the usage of TTS, including long delay and voice personality/prosody mismatches.

In the present study, we employed a method of estimating the speech parameters from UDS features and synthesizing audible speech using those estimations. The underlying principle relies on a certain degree of correlation between UDS and audible speech. The feasibility of UDS for the estimation of speech can be explained by what has been learned from previous work in EMG-based speech recognition [24] and synthesis [25]. In EMG-based speech recognition/synthesis, the movement of articulatory facial muscles is assumed to be closely related to corresponding speech signals and can be measured via surface EMG sensors. Since ultrasonic Doppler frequency shifts are also caused by articulatory facial movements, the same degree of performance obtained by the EMG-based methods can be achieved by adopting UDS. Previous studies have shown that the quality of synthesized speech signals from EMG is reasonable in both naturalness and intelligibility [24]. Nevertheless, the results were mostly obtained from EMG signals of audible speech, rather than from those of silent speech. The UDS of whispered or silent speech was not taken into consideration in previous UDS-based speech synthesis schemes [17]. Since articulatory movements differ between silently articulated and normally spoken speech [15], investigation into discrepancies between audible and silent speech would be highly desirable for implementing silent communication. In the previous study, deterioration of recognition accuracy caused by discrepancies between audible and silent speech was investigated and a method for compensating for such discrepancies was proposed to improve the performance of EMG-based automatic speech recognition [26].

For a speech synthesis system that uses silent UDS, no audible speech is available to construct the mapping rules between USD and speech. To deal with such problems, we proposed a nearest-neighbor search and alignment (NNSA) scheme that determined the appropriate audible speech feature for a given UDS feature. We also investigated the differences between two speech-synthesized signals, one from the estimation rules using silent-UDS and the other from those using audible-UDS. An investigation into the discrepancies between silent and audible UDS in terms of the quality of synthesized speech would be helpful for implementing a full SSI system.

This paper is organized as follows. In Sect. 2, the acoustic Doppler sensor is described and a procedure for UDS and audio data acquisition is explained. Section 3 presents a procedure for extracting the features from UDS for audible speech estimation. The estimation of the speech spectra, which includes the NNSA method, is explained in Sect. 4. Experimental results are presented in Sect. 5. Finally, concluding remarks are provided in Sect. 6.

2. Ultrasonic Doppler Sensor

A photograph of a subject wearing the developed device is shown in Fig. 1. There are two sensors A and B to detect the Doppler shifts in various directions, as shown in Fig. 2. Sensor A is used to detect Doppler shifts in the mouth and cheek area and is attached to the wireframe of the headset microphone. Sensor B is located at the front of the throat, which is used to detect Doppler sonar in the jaw and neck region. A hairband-shaped holder is employed to fix sensor B to the neck. The holder is made of plastic material and is well warped and fitted according to the shape of the user's neck. Although sensor holder is attached to the body, the sensor itself is not in direct contact with the skin surface of the body. This allows the continuous acquisition of signals in a non-contact manner, which is different from the conventional contact sensing methods, such as sEMG and NAM. Each sensor is composed of an ultrasonic transmitter that emits a continuous ultrasonic tone at 40kHz and ultrasonic sensors that are tuned to receive signals around at 40kHz. The ultrasonic transmitter used in this study was product model number MA40H1S-R, that is a surface-mounted device (SMD)-type ultrasonic transducer. The center frequency of the transmitter is 40kHz. The employed ultrasonic sensor is product model number SPM0404UD5.

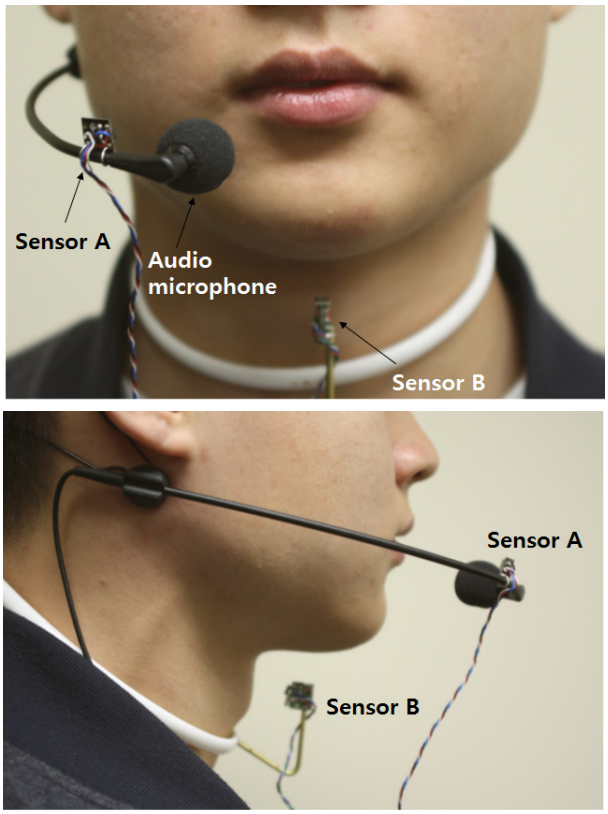

Fig. 1 Two sensors, A and B attached to a subject. Top: front. Bottom: left.

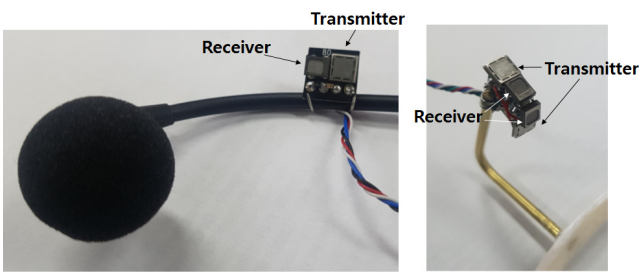

Fig. 2 Photograph of the ultrasonic sensors. Left: sensor A. Right: sensor B.

The effective frequency range of the sensor is 10 to 65 kHz. Sensor B uses two transmitter-receiver pairs, one for detecting vocal cord vibration and the other for jaw movements. Since very weak vocal cord vibrations are normally detected in silent UDS signals, the sensor for detecting vocal cord vibration was not used for silent UDS-based speech synthesis. To simultaneously acquire audio-frequency range signals (just for audible UDS), an audio microphone (Model: SHURE BETA 53) was also employed. By attaching the UDS sensors to the subject's neck and head, it was possible to prevent the undesirable motion artifacts caused by a subject's head movements.

To emit a continuous ultrasonic tone at 40 kHz, a function generator (Model: 33250A, Agilent) was employed and the tone signal was amplified using a custom-made audio frequency range power amplifier. The amplified 40 kHz sinusoidal signal was inputted to ultrasonic transducers. The effective beam angle of the employed ultrasonic transducer was ±40°, and the average distance between the transducer and face skin was 3 cm. Hence, the effective radiation area was 19.9 cm^2, which was sufficient to detect the subject's articulatory movements. The signals from the sensor were digitized at a rate of 192 kHz with 16 bits using a multichannel digital audio interface (Model: Fireface 800, RME). Digital data were transmitted over an IEEE 1394 serial bus to a desktop PC with an i7-6700k processor. To compensate for the differences in the sensitivity of the employed ultrasonic sensors, gain adjustments were carried out on each channel amplifier before recording.

Table 1 Length information of the recorded signals.

Length (sec)	Audible speech		Silent speech	
	All	Valid	All	Valid
Shortest	1.206	0.336	0.992	0.112
Longest	1.847	0.880	2.047	1.216
Average	1.513	0646	1.354	0.610

A total of five subjects participated in the recording, one female speaker and four male speakers, between 22 and 50 years of age. The purpose of this study was to develop a UDS-based speech synthesizer for a specific application, such as a "silent call". Accordingly, synthesizable utterance was limited to a few isolated words that were essential in a specific application, rather than long sentences. In the future, we will extend this work to long sentences. This study used a 60-Korean word vocabulary [25] that was phonetically balanced, which means it has speech sounds, or phonemes, that occur as often as they would in a normal conversation. For each speaker, the words recorded in the audible and silent speaking mode were identical. We recorded each word 14 times for silent speech and 20 times for audible speech. During the silent speech recording, each subject was asked to make as few sounds as possible. If a voice signal over a certain level was detected through the headset microphone, the corresponding UDS signal was deleted and re-recorded. Most subjects were advised to have sufficient practice time before recording because they were not familiar with pronouncing silent speech. Length information of the recorded signals is summarized in Table 1 where "valid" indicates a length excluding the non-speech regions. This showed that the average length of the recorded silent speech was slightly shorter than that of audible speech.

3. Feature Extraction from UDS

Assuming that the ultrasonic transmitter emits a continuous tone with a frequency f_c, the transmitted signal is given by

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

where A_T and Ψ_T are the amplitude and phase of the underlying sinusoid, respectively. When the ultrasonic tone is reflected on an articulating face, the Doppler frequency shifts

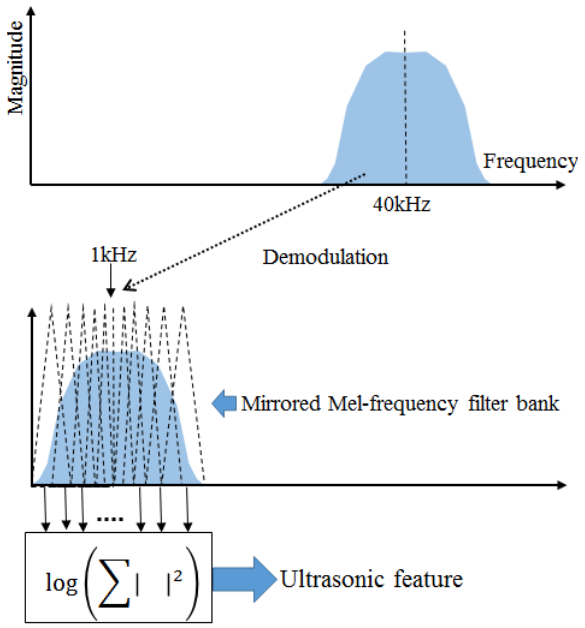

Fig. 3 Procedure for extracting features from ultrasonic signal.

are found in the reflected signals, due to the movements of several articulatory organs. Assuming that total M objects are engaged with the Doppler frequency shift, and the instantaneous velocity of the m-th object at time t is given by $v_m(t)$, the reflected signal is given by

$$R_e(t) = \sum_{m=1}^{M} A_T k_m \cos(\phi_m + \Psi_T),$$
$$\phi_m = 2\pi f_c \left[t + \frac{2}{v_s} \int_0^t v_m(\tau) d\tau \right] + \Psi_m \quad (2)$$

where k_m and Ψ_m are the attenuation coefficient and phase shift of the m-th object at frequency f_c, respectively. v_s is the speed of sound.

The procedure for extraction of the ultrasonic feature is explained in Fig. 3 where demodulation is first performed and low pass filtering/decimation is subsequently carried out on the demodulated signal,

$$x_u(n) = \mathrm{LPF}_{\downarrow}\left[R_e(t) \cos(2\pi f_d t) \right] \quad (3)$$

where $\mathrm{LPF}_{\downarrow}[\cdot]$ denotes low pass filtering followed by decimation and f_d is the demodulation frequency. Although the received ultrasonic signal was a real value signal, the experimental result showed that its magnitude spectrum was not exactly symmetrical with the 40kHz frequency bin. Moreover, the experimental results showed that maintaining the left sideband by setting the demodulated frequency as $f_c - \Delta f$ yielded slightly better results than in the case of $f_d = f_c$. The Δf represents the portion of the effective bandwidth of UDS that is most closely related to the maximum frequency of articulatory movements. Previous studies associated with EMG-based speech recognition have suggested that a sampling rate of 1kHz ∼ 2kHz is sufficient for representing the articulatory movements [23], [24]. Accordingly, Δf was set at 1kHz in this study.

Since the Doppler signals are related to the underlying speech signals, the features that have been adopted in speech-related research were considered in this study. Mel-scale filter bank analysis is widely employed in many speech processing techniques [31]. Although the Doppler frequency shifts are not perceived by the human ear, the usefulness of mel-scale filter bank analysis has been confirmed in several speech recognition tasks that use articulatory muscle movements [24], [25]. Hence, mel-frequency filter bank energy was used as a feature for the ultrasonic signal. The k-th mel-band energy is given by

$$\mathbf{x}_{u,k} = \sum_{i=1}^{N_k} |X_u(i)| H_k(i) \quad (4)$$

where $X_u(i)$ is the i-th discrete Fourier coefficient of $x_u(n)$, $H_k(i)$ is the corresponding magnitude response of the k-th mel band, and N_k is the number of Fourier coefficients in the k-th mel-bands. In the present study, the demodulated ultrasonic signal covers 0∼1kHz. The number of mel-bands within such a frequency range is 8. Note that mel-frequency filter bank analysis was applied to both left- and right-side bands as shown in Fig. 3. This resulted in a total of 32 UDS features per frame.

4. Speech Estimation Using Silent UDS

The overall procedure for constructing a speech estimation rule using silent-UDS is explained in Fig. 4. Three corpora were built prior to construction of the estimation rule; an audible UDS, audible speech, and a silent UDS databases. In summary, an initial speech estimation rule was first constructed using the audible UDS and audible speech database. The speech estimation rules for silent UDS was then iteratively refined using NNSA and re-estimation procedures.

4.1 Nearest Neighbor Search and Alignment

One of the fundamental problems associated with silent UDS-based speech synthesis is that there is no reference speech for given UDS that is essential for constructing the speech estimation rules. In this study, a method of NNSA was proposed to deal with this problem where the speech estimation rule was iteratively found using the pairs of a time-aligned UDS feature stream and an audible speech feature stream. The underlying principle is that for a given silent UDS sample, its corresponding audible speech sample is chosen by minimizing the overall errors between the time-aligned version of an estimated speech stream and those of a selected audible speech stream, while silent UDS samples and audible speech samples have the same word context. A graphical explanation of the NNSA scheme is shown in Fig. 5. The initial speech estimation rule was constructed using the pairs of audible UDS and speech signals. Let $\{\mathbf{X}_i^{(A)}\}_{i=1}^{N_A}$ and $\{\mathbf{Y}_i\}_{i=1}^{N_A}$ denote the sets of the audible UDS samples and the corresponding audible speech

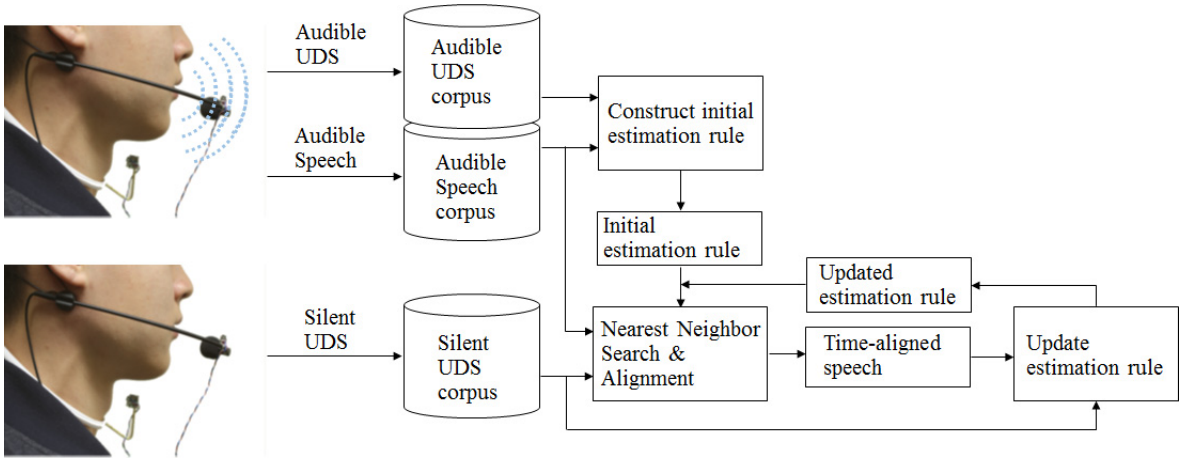

Fig. 4 Procedure for construction of the speech estimation rule.

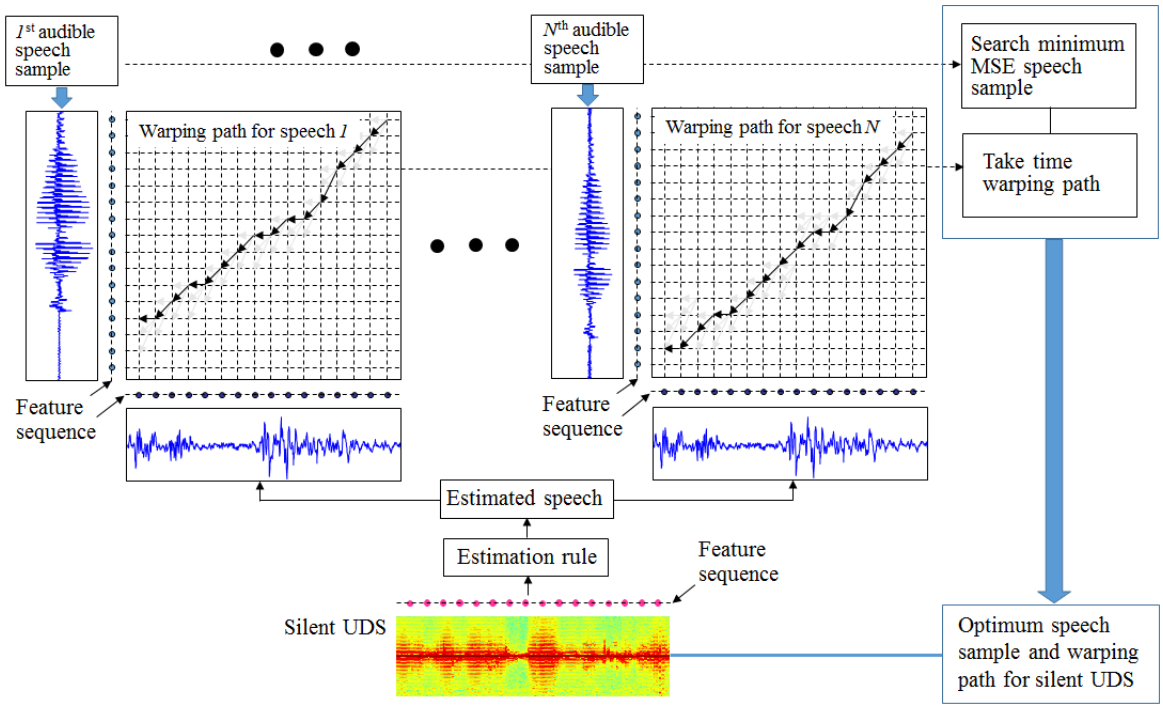

Fig. 5 Procedure of nearest-neighbor search and alignment.

samples, respectively, where $\mathbf{Y}_i = \{\mathbf{y}_i(0), \ldots, \mathbf{y}_i(N_i - 1)\}$, $\mathbf{X}_i^{(A)} = \{\mathbf{x}_i^{(A)}(0), \ldots, \mathbf{x}_i^{(A)}(N_i - 1)\}$, and N_i is the total number of feature vectors for the i-th UDS/speech sample. The initial estimation rule is then given by

$$\mathcal{F}^{(A)} = \arg\min_{\mathcal{F}} \sum_{i=1}^{N_A} \|\mathbf{Y}_i - \mathcal{F}\{\mathbf{X}_i^{(A)}\}\|^2 \tag{5}$$

where $\mathcal{F}\{\cdot\}$ is a transformation function that maps UDS feature to audible speech feature and N_A is the total number of the pairs of audible UDS samples and corresponding speech samples.

To build the speech estimation rule for silent UDS, we first defined the following error measurement between the feature stream of silent UDS and that of audible speech.

$$\epsilon(\mathcal{M}, \mathcal{W}, \mathcal{F}) = \sum_{i=1}^{N_{NA}} \|\mathbf{Y}_{M_i} - \mathcal{F}\{\tilde{\mathbf{X}}_{i,W_i}^{(NA)}\}\|^2 \tag{6}$$

where M_i is a sample map function that maps the i-th silent UDS sample to the M_i-th audible speech sample and N_{NA} is the total number of silent UDS samples in the training corpus. $\tilde{\mathbf{X}}_{i,W_i}^{(NA)}$ is a time-warped version of the i-th silent UDS sample, which is given by

$$\tilde{\mathbf{X}}_{i,W_i}^{(NA)} = \{\mathbf{x}_i^{(NA)}(w(0)), \ldots, \mathbf{x}_i^{(NA)}(w(N_i - 1))\} \tag{7}$$

where $W_i = \{w_i(0), \ldots, w_i(N_i - 1)\}$ is a set of time warping functions that align the sequence of silent UDS feature

vectors with that of audible speech feature vectors, $\mathbf{Y}_{M_i}$.

The optimal transformation function for silent-UDS is given by

$$\mathcal{F}^{(NA)*} = \arg\min_{\mathcal{F}} \left\{ \min_{\mathcal{M},\mathcal{W}} \epsilon(\mathcal{M}, \mathcal{W}, \mathcal{F}) \right\} \tag{8}$$

This implies that the optimal estimation rule is constructed using the feature vectors derived from the selected audible speech sample $\mathbf{Y}_{M_i}$ and the time-warped feature vector stream from the silent UDS sample. Such pairs are obtained by minimizing the overall distance between the estimated speech feature stream from the time-warped feature stream of silent UDS and that of the selected audible speech. Note that context information of the i-th silent UDS sample is identical to that of the selected audible speech sample $\mathbf{Y}_{M_j}$. Since simultaneous minimization (8) cannot be achieved by a closed form solution, an iterative method was employed in this study, where NNSA and updating of the speech estimation rules are iteratively performed to minimize the overall error ϵ, as shown in Fig. 4. In each NNSA step, for a given $\mathcal{F}$, a set of sample map functions $\mathcal{M}$ and a set of time warping functions $\mathcal{W}$ are found by minimizing ϵ.

$$\mathcal{M}^{(j)},\ \mathcal{W}^{(j)} = \arg\min_{\mathcal{M},\mathcal{W}} \epsilon(\mathcal{M}, \mathcal{W}, \mathcal{F}^{(j-1)}) \tag{9}$$

where (j) is the iteration index. Minimization of ϵ with respect to $\mathcal{W}$ can be achieved by a dynamic programming procedure [32]. With previously obtained $M^{(j)}$ and $W^{(j)}$, the estimation rule $\mathcal{F}$ is found by minimizing ϵ.

$$\mathcal{F}^{(j)} = \arg\min_{\mathcal{F}} \epsilon(\mathcal{M}^{(j)}, \mathcal{W}^{(j)}, \mathcal{F}) \tag{10}$$

$\mathcal{M}^{(j)}$, $\mathcal{W}^{(j)}$, and $\mathcal{F}^{(j)}$ are then used for the next iteration $(j+1)$ and the process is repeated until an acceptable convergence threshold is reached. Note that the initial estimation rule $\mathcal{F}^{(0)}$ is given by $\mathcal{F}^{(A)}$, which can be obtained using audible UDS features.

4.2 DNN-Based Speech Feature Estimation

Deep neural networks (DNN) were employed to estimate the relationship between the input (UDS feature) and the output (audible speech feature). In the training stage, a regression DNN model was trained from a training corpus, that consisted of pairs of speech features and those of ultrasound signals. The audible speech represented by the log magnitude spectra was the target output of the DNN. In the estimation stage, the feature parameters extracted in the training stage, were derived from the incoming ultrasound signals. The feature parameters were then inputted to the trained DNN.

To obtain the initial network, a deep generative model of input features was adopted by a stacking of multiple restricted Boltzmann machines (RBMs) [33]. A back-propagation algorithm with the minimum mean square error (MMSE) criterion was employed to train the DNN. The objective function was given by a mean square error between the estimated log magnitude spectrum and that of audible speech. A stochastic gradient descent algorithm was performed in mini-batches with multiple epochs to improve the learning convergence.

Multiple frames of the ultrasonic signal were used as DNN input. By using this configuration, the DNN captures the acoustic context information along the time axis [34]. When N_u multiple frames, which is given as an odd number, are used, the input of the DNN is given by

$$\mathbf{X}_{N_u}(t) = \{\mathbf{x}(t - N_u/2), \ldots, \mathbf{x}(t), \ldots, \mathbf{x}(t + N_u/2)\} \tag{11}$$

Note that the sample index and audible/silent marks are omitted for simplicity. It was also reported that performance improvements were achieved by estimating multiple speech features over time [16]. Accordingly, N_s neighboring feature vectors of speech signals were simultaneously estimated using DNN, and a method of overlap-and-add using a triangular window was adopted to form the output vector sequence. It is generally accepted that increasing the number of frames leads to an increase in the DNN performance since DNNs have sufficient acoustic contents. However, as the distance between the two frames is increased, the correlation between the underlying two frames is reduced. In this study, the numbers of neighboring frames for speech and UDS were determined by maximizing the performance in terms of speech enhancement, and hence, the results for the various values of N_u and N_s were obtained and will be shown in the subsequent results section.

4.3 Speech Synthesis

The final step in UDS-based speech synthesis is to synthesize audible speech signal from the estimated speech parameters. There are two ways to synthesize speech signals from the estimated spectral parameters. One is based on a linear prediction (LP) model, where a voice is generated by filtering the excitation source through an all-pole filter that reflects the vocal tract transfer function [31]. The all-pole filters can be represented by a linear prediction coefficient (LPC), LPC cepstrum (LPCC), Log Area Ratio (LAR), and line spectrum pairs (LSP). In this study, the LSP coefficient and LPCC coefficient were employed as LP feature variables. These two feature variables are used in a wide variety of speech processing techniques, and have the advantage of being free from the problems associated with instability [31]. The LP-based synthesis approach requires additional estimation rules for an excitation source and a pitch period for voiced speech. These two parameters are closely related to vocal cord vibration, which, however, cannot be detected with silent speech. An alternative way was adopted in visual-based SSI [16] wherein random noise was used as an excitation source. Although this approach has reportedly produced intelligible speech, the tone of the voice was lost and the resultant sound was equivalent to a whisper.

An approach of short-time Fourier transform (STFT) is another method for speech synthesis. In this case, the magnitude of the spectrum of windowed short-time speech is regarded as a speech parameter, which is estimated using a

DNN. Continuous waveforms were obtained by concatenating the short-time speech signals obtained by inverse Fourier transform. Since the phase spectrum was not available, a method of the least square error estimation of modified short time Fourier transform magnitude (LSEE-MSTFTM) [35] was employed in this study, where the squared error between the STFT of a signal and the estimated magnitude spectrum is decreased iteratively. The results of both LP-synthesis and the STFT-based method will be presented and compared in the following section.

4.4 Harmonic Enhancement

Harmonic structure is a unique characteristic of voiced speech signals. Harmonic enhancement [36] is a method for preserving the harmonic structure of a corrupted speech signal so that the intelligibility is improved. The harmonic-enhanced signal is given by

$$\tilde{y}_{pe}(t) = y(t) + \alpha_{he} \times y(t - P(t)) \tag{12}$$

where α_{he} denotes the harmonic enhancement factor and $P(t)$ is the pitch period at time t. A reliable estimation of the pitch period is essential for harmonic enhancement. In most of the currently available pitch estimation algorithms, the pitch period is estimated from the audible speech that is not available in silent UDS. In the present study, experiments were carried out to confirm the possibility of pitch estimation using silent UDS. The pitch estimation method using UDS was proposed in this study, where a DNN was trained using a set of the pairs of UDS features and pitch periods detected by audible speech. The procedure of pitch estimation was similar to the previous DNN-based pitch estimation method [37] in which DNN provided likelihoods of each candidate pitch periods, a sequence of the optimal pitch periods was obtained by a Viterbi-trellis search. Instead of using spectral feature vectors derived from noisy speech [37], UDS feature vectors were inputted to DNN. The number of the candidate pitch periods was set at 67, and the weights for the posterior probability and the transition probability were set to 0.7 and 0.3, respectively.

The experimental results showed that the average pitch error of silent UDS-based estimation was about 10 times higher than that of speech-based estimation. Those results show that it is questionable whether harmonic enhancement with silent UDS-based pitch estimation can improve the quality of reconstructed speech signals. According to the experimental results, however, the harmonic enhancement method has some allowance for pitch error. Therefore, it is more meaningful to examine the improvement of speech quality by harmonic enhancement than the absolute error of the pitch estimation. In the following sections, results from the adoption of harmonic enhancement will be shown and its usefulness will be verified.

5. Experimental Results

5.1 Experimental Setup

The features were extracted from the windowed speech and from the windowed UDS. A 32 msec length Hanning window was commonly used to compute and extract the feature parameters at 16 msec intervals. For LP-synthesis, the order of LSP was 20, which is identical to that of LP-analysis. The order of LPCC was set as 30. For STFT-based synthesis, the log-magnitude of the Fourier transform (FT) coefficients was used where the length of the fast Fourier transform (FFT) was set at 256. Accordingly, the dimension of the log-power spectral feature vector was 128. The same FFT length was adopted to compute the UDS feature.

To compare with the visual- and EMG-modality speech synthesis schemes, the video and EMG were simultaneously recorded. In the video recording, a frame rate of 30Hz was adopted for an image size of 640×480 pixels. Semi-automatic segmentation was carried out on the captured images to extract the speaker's mouth region. The resultant mouth image was 176×144 pixels. The 2D-DCT was applied to the cropped images, and lower frequency components were established as visual features. The dimensions of the visual features were determined heuristically at 32. Since the frame rate of speech was different from the video frame rate, a cubic spline interpolation was carried out on the DCT coefficients.

During the EMG acquisition, surface myoelectric signals (MES) were collected from the three articulatory facial muscles; *levator anguli oris*, *depressor anguli oris*, and *zygomaticus major*. Each MES was collected using pairs of Ag-AgCl button electrodes (3M, 2258). The electrodes were 3.3 cm in diameter (including the foam adhesive patch). The reference electrode was located at the back of the neck. A Mel frequency cepstral coefficient (MFCC) was used as the feature variable since its usefulness has been confirmed in several speech-related applications [24], [25]. The dimension of the MFCC was set at 5, and hence, a total of 15 EMG features per frame were used for speech prediction. For both visual- and EMG-based speech estimations, a DNN was also used to predict the log-power spectral speech features. Multiple frames were also considered in the EMG- and vision-based estimations. N_e and N_v are the number of multiple frames for EMG and vision, respectively. The maximum number of the frames was set at 11, which was identical to the UDS cases.

Since it is very difficult to mathematically determine the optimum number of hidden layers, we performed several experiments to investigate the relationship between the number of hidden layers and objective performance. According to the experimental results, the best performance was obtained when the DNN contained three hidden layers, and the number of the nodes in the hidden layer was set to $[1.5 \times N_i]$. (where $[x]$ is the nearest integer value of x and N_i is the number of input nodes, which is $32 \times N_u$, $32 \times N_v$,

and $15 \times N_e$ for the DNNs of UDS, vision, and EMG, respectively.) Except for the top layer, the sigmoid activation function was adopted. The momentum constant α of the sigmoid active function was set to 0.7. The linear function was used at the top layer.

The number of RBM pre-training epochs in each layer was 100. The learning rate of the RBM training was set at 0.0005. A fixed learning rate of 0.001 was applied for the fine-tuning of the baseline. The total number of epochs at the fine-tuning stage was 200. For both RBM pre-training and fine-tuning, the momentum was set at 0.05 for the first five epochs, then maintained at 0.07 thereafter. Mean and variance normalization was applied to the input and target feature vectors of the DNN. The performance of the DNN was expected to be improved by dropout regularization [38]. The experimental results also showed clear differences between when a dropout was adopted and when it was not, particularly for the test data. Hence, dropout regularization was adopted in the present study where a keep probability of 0.8 was employed. The DNN was trained using 70% of the total features in the training stage, and the remaining features were used for evaluation. In the following, all presented results were obtained from evaluation data.

Three objective measures were applied in the experiments, including the prediction gain, the perceptual evaluation of speech quality (PESQ) [39], and the short-time objective intelligibility measure (STOI) [40]. The prediction gain is given by

$$G_p = -10 \log_{10} \frac{\sum_{i=1}^{N} \sum_{n=0}^{N_i-1} \|\mathbf{y}_i(n) - \hat{\mathbf{y}}_i(n)\|^2}{\sum_{i=1}^{N} \sum_{n=0}^{N_i-1} \|\mathbf{y}_i(n) - \bar{\mathbf{y}}\|^2} \tag{13}$$

where $\hat{\mathbf{y}}$ and $\bar{\mathbf{y}}$ denote the estimated and mean target feature vectors, respectively. Higher prediction gain corresponds to good estimation performance. The PESQ has a high correlation with subjective evaluation scores [39]. PESQ was mostly used as a compressive objective measure and was also employed to evaluate the perceptual aspect of the synthesized speech from visual information [16]. PESQ is calculated by comparing the enhanced speech with a sample of clean reference speech, and it ranges from -0.5 to 4.5. The STOI [40] was proposed as a correlation-based measurement to evaluate the speech intelligibility degradation caused by speech enhancement solutions.

5.2 Comparison of Prediction Gains

The underlying assumption of the present study is that the speech estimation rules from the pairs of audible UDS and corresponding speech are not well matched with those from the silent UDS and corresponding speech pairs. To verify this assumption, two speech estimation rules were built. The first uses the pairs of audible UDS and audible speech, and then, speech estimation was carried out on silent UDS (method *A-S*). The second uses the pairs of silent UDS and audible speech, and speech was estimated using silent UDS (method *S-S*). Since reference audible speech was not available in the method *S-S*, the audible speech signal selected using NNSA was alternatively used as a reference for audible speech. For comparison, speech estimation was also carried out on audible UDS using the rule constructed by the pairs of audible UDS and audible speech (method *A-A*). Note that since audible UDS was employed in both the rule construction and synthesis stages, the method *A-A* cannot be adopted for silent speech interface. The prediction gain for each method was investigated for each number of the neighboring feature vectors so that the optimal number of the frames in terms of prediction performance was determined.

The results are presented in Table 2 where the values in bold type indicate its maximum prediction gain for each method. As expected, method *A-A* yielded the best performance among the three methods. This was due to obtaining the signals under the same conditions (audible condition) for both rule construction and synthesis. However, performance degradation in terms of prediction gain was observed under mismatched conditions. For method *A-S*, the difference of maximum prediction gain from the method *A-A* was 2.714. This was reduced to 2.181 by employing the method *S-S* where NNSA was adopted to obtain the reference audible speech signals. This indicated that mismatches between audible and silent UDSs were partially compensated for by employing NNSA. It is interesting to note that the acoustic context information was more useful for the matched conditions (method *A-A*). Under these conditions, the highest prediction gain was obtained when the number of the neighboring features was given by 11 frames. Whereas the best performance was obtained with no neighboring UDS features in case of the method *A-S*.

We also evaluated the validity of UDS-based speech estimation by comparing the prediction gains for other modalities (sEMG and vision). The results are shown in Table 2 for each modality. In the following, the results of UDS modality are considered to be obtained from method *S-S*. The maximum prediction gain was obtained for the UDS-modality when the numbers of the neighboring feature vectors were 9 and 3 for UDS and speech, respectively. The average prediction gain for the UDS modality was also remarkably higher than other modalities. This indicates that ultrasonic Doppler is useful for an estimation of the speech signal. For all combinations of *Ns* and *Nu*, the prediction gains of UDS-modality were consistently higher than other modalities. The differences from UDS-based estimation in maximum prediction gain were 1.189 and 0.347 for sEMG- and vision-based estimation, respectively.

The maximum and average prediction gains of vision-based estimation were higher than those of sEMG. For vision-modality, higher prediction gains were obtained when the larger number of the image frames was adopted. In the previous vision-based speech synthesis method [16], more than 11 frames improved the quality of the reproduce speech signals. However, performance improvements were

Table 2 The prediction gains in dB for each modality, according to the number of the frames (N_s denotes for speech. N_u, N_e, and N_v denote for UDS, sEMG, and image, respectively. The bold values indicate the maximum for each modality.)

(a) Audible UDS (method *A*-*A*)

$N_u \setminus N_s$	1	3	5	7	9	11
1	5.511	5.477	5.157	4.736	4.639	4.425
3	6.677	6.483	6.230	6.080	5.790	5.528
5	6.911	6.809	6.646	6.457	6.213	6.113
7	7.129	7.047	6.998	6.745	6.576	6.361
9	7.316	7.366	7.282	6.943	6.735	6.707
11	**7.550**	7.459	7.295	7.227	7.001	6.843
13	7.281	7.247	7.203	7.181	6.820	6.663

(b) Silent UDS (method *A*-*S*)

$N_u \setminus N_s$	1	3	5	7	9	11
1	**4.836**	4.374	3.911	3.401	3.572	3.336
3	4.650	4.439	4.290	4.269	4.141	4.246
5	4.626	4.637	4.345	4.334	4.245	4.275
7	4.469	4.613	4.419	4.187	4.214	4.151
9	4.257	4.524	4.170	4.054	4.066	3.774
11	4.206	4.362	4.266	4.183	3.847	4.002
13	4.138	4.210	4.130	4.110	3.657	3.982

(c) Silent UDS (method *S*-*S*)

$N_u \setminus N_s$	1	3	5	7	9	11
1	4.540	4.650	4.436	4.191	4.029	3.754
3	4.860	5.051	4.978	4.988	4.707	4.580
5	5.061	5.054	5.293	5.194	4.954	4.866
7	5.190	5.293	5.233	4.973	5.068	4.891
9	4.923	**5.369**	5.259	5.176	4.907	5.017
11	4.829	5.299	5.136	5.134	5.147	4.911
13	4.834	5.017	5.186	5.132	5.140	4.908

(d) sEMG

$N_e \setminus N_s$	1	3	5	7	9	11
1	3.124	3.165	2.842	2.508	2.304	1.787
3	3.898	3.564	3.029	3.009	2.796	2.633
5	3.905	3.887	3.610	3.009	2.971	2.709
7	4.043	3.862	3.680	3.389	3.239	3.007
9	4.084	3.899	3.790	3.397	3.389	3.039
11	**4.176**	3.902	3.896	3.589	3.298	3.152
13	4.180	3.982	3.870	3.433	3.178	3.056

(e) Vision

$N_v \setminus N_s$	1	3	5	7	9	11
1	2.295	2.302	2.159	2.170	2.078	2.089
3	3.023	3.024	3.191	3.058	2.748	2.649
5	3.378	3.531	3.188	3.135	3.090	2.996
7	4.098	3.770	3.502	3.409	3.325	3.088
9	4.753	4.357	4.295	3.998	3.378	3.276
11	**5.022**	4.604	4.739	4.377	3.844	3.974
13	5.018	4.556	4.650	4.379	3.875	3.978

not clearly observed in this study, even when the number of frames were increased by 11 or more. This was mainly due to the relatively short stimuli employed in the present study.

For all modalities, the prediction gain for a given number of speech frames (N_s) has an increasing trend as the number of the neighboring input feature vectors N_u increased. The increasing rates differed according to the number of neighboring speech feature vectors N_s and modality. For UDS, the maximum of the correlation between the prediction gain and N_u was 0.8915, when N_s was 3, whereas the maxima of the correlation coefficients were 0.9510 and 0.9936 for sEMG- and vision-based speech estimation, respectively. This indicates that longer input feature vectors are desirable to increase the speech estimation performance for sEMG- and vision-modalities. Similar results were reported in a recent vision-based speech estimation study [16].

The prediction gain was inversely proportional to the number of neighboring speech feature vectors N_s. This was commonly observed for other modalities. Accordingly, the maximum prediction gains were obtained when a relatively small N_s (3, 1, and 1 for UDS, sEMG, and vision, respectively) was employed. Such a poor performance for a longer N_s can be explained by the effects of over smoothing, which is caused by an averaging of many spectral feature vectors. Hence, a method of alleviating the over-smoothing effects is highly desirable, particularly when longer spectral feature vectors are employed. Formant enhancement [4] and harmonic enhancement [36] are possible solutions to the over-smoothing effects.

5.3 Comparison of Perceptual Quality

Although the prediction gain is a good indicator of performance in speech estimation, the DNN configuration with the highest prediction gain did not always correspond to the best performance in terms of PESQ and STOI. Hence, five DNNs were selected that yielded the top five prediction gains.

The results are shown in Table 3. Similar to the results of prediction gain, method *A*-*A* yielded the highest PESQs and STOIs. A remarkable level of performance degradation was also seen when the estimation rules constructed using audible UDS were driven by silent UDS (method *A*-*S*). This was commonly observed for all DNNs. These results provided more proof that articulatory movements differ between silently articulated and normally spoken speech. The results also indicated that PESQs were slightly improved by employing harmonic enhancement, even though the pitch periods were estimated using silent UDS. For estimating speech from silent UDS using method *S*-*S*, the maximum improvement in PESQ was 0.013. In this case, the harmonic enhancement factor was 0.7. It is also noteworthy that audible UDS provided more accurate pitch periods than silent UDS. This was confirmed by the fact that PESQ was maximally increased by 0.117, when the pitch periods were estimated from audible UDS using method *A*-*A* in the case of $(N_u, N_s) = (9, 3)$ and $\alpha_{he} = 0.5$. Such improvement was remarkably higher than in the case of silent UDS-based pitch estimation.

The results in terms of PESQ and STOI are also presented in Table 3 when sEMG and image were used to synthesize speech. Note that only audible EMG and audible image were considered in the experiments and that NNSA was not adopted for these two modalities. When harmonic enhancement was employed, the pitch period was estimated using each modality (EMG and image).

Table 3 Average PESQ and STOI results for different DNN configurations and different harmonic enhancement factors. "No HE" denotes no harmonic enhancement.

(a) The results of estimating speech from audible-UDS using method *A-A*.

DNN		No HE		$\alpha_{he} = 0.3$		$\alpha_{he} = 0.5$		$\alpha_{he} = 0.7$	
N_u	N_s	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
9	1	2.138	0.9092	2.194	0.9088	2.232	0.9048	2.229	0.9004
9	3	2.120	0.9077	2.184	0.9066	2.237	0.9028	2.225	0.8995
11	1	2.201	0.9182	2.255	0.9173	2.263	0.9123	2.262	0.9092
11	3	2.169	0.9175	2.242	0.9163	2.243	0.9109	2.252	0.9077
11	5	2.155	0.9159	2.241	0.9148	2.246	0.9096	2.249	0.9068
Average		2.157	0.9137	2.223	0.9128	2.244	0.9081	2.262	0.9047

(b) The results of estimating speech from silent-UDS using method *A-S*.

DNN		No HE		$\alpha_{he} = 0.3$		$\alpha_{he} = 0.5$		$\alpha_{he} = 0.7$	
N_u	N_s	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
1	1	1.394	0.8282	1.424	0.8280	1.406	0.8275	1.382	0.8224
3	1	1.530	0.8513	1.526	0.8509	1.522	0.8505	1.518	0.8500
5	1	1.515	0.8574	1.496	0.8570	1.494	0.8568	1.526	0.8560
5	3	1.651	0.8578	1.637	0.8565	1.631	0.8552	1.651	0.8539
7	3	1.532	0.8182	1.552	0.8164	1.555	0.8146	1.544	0.8123
Average		1.524	0.8426	1.527	0.8418	1.522	0.8409	1.524	0.8389

(c) The results of estimating speech from silent-UDS using method *S-S*.

DNN		No HE		$\alpha_{he} = 0.3$		$\alpha_{he} = 0.5$		$\alpha_{he} = 0.7$	
N_u	N_s	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
9	3	1.860	0.8821	1.868	0.8800	1.862	0.8768	1.850	0.8736
11	3	1.901	0.8957	1.909	0.8936	1.908	0.8905	1.904	0.8879
7	3	1.947	0.8744	1.945	0.8716	1.948	0.8689	1.947	0.8664
5	5	1.902	0.8630	1.912	0.8612	1.914	0.8587	1.915	0.8567
9	5	1.889	0.8715	1.878	0.8693	1.882	0.8659	1.880	0.8638
Average		1.900	0.8773	1.902	0.8751	1.903	0.8722	1.899	0.8697

(d) The results of estimating speech from sEMG.

DNN		No HE		$\alpha_{he} = 0.3$		$\alpha_{he} = 0.5$		$\alpha_{he} = 0.7$	
Ne	Ns	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
5	1	1.628	0.7537	1.618	0.7524	1.620	0.7500	1.618	0.7466
7	1	1.663	0.7634	1.676	0.7610	1.646	0.7584	1.653	0.7555
9	1	1.682	0.7842	1.683	0.7823	1.678	0.7793	1.680	0.7758
9	1	1.674	0.7690	1.661	0.7673	1.654	0.7640	1.612	0.7588
11	1	1.712	0.8134	1.700	0.8105	1.703	0.8066	1.688	0.8054
Average		1.683	0.7767	1.668	0.7747	1.660	0.7717	1.650	0.7682

(e) The results of estimating speech from image.

DNN		No HE		$\alpha_{he} = 0.3$		$\alpha_{he} = 0.5$		$\alpha_{he} = 0.7$	
Ni	Ns	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
11	1	2.059	0.8648	2.077	0.8681	2.058	0.8632	2.067	0.8613
9	1	2.037	0.9250	1.996	0.9226	1.978	0.9204	1.997	0.9183
11	5	1.771	0.8027	1.801	0.8002	1.792	0.7947	1.780	0.7887
11	3	1.869	0.8939	1.886	0.8900	1.904	0.8869	1.873	0.8838
11	7	1.698	0.7904	1.709	0.7889	1.707	0.7821	1.707	0.7805
Average		1.887	0.8554	1.894	0.8540	1.888	0.8495	1.885	0.8465

The overall performance of EMG-based speech estimation was worse than that of silent-UDS speech estimation (method *S-S*). The maximum difference in average PESQ between silent-UDS and EMG was 0.249 when an harmonic enhancement factor of 0.7 was adopted. For STOI, the maximum difference was 0.1015. Considering that audible-EMG was used for EMG-based speech estimation, UDS was more useful in implementing a silent speech interface. For EMG-based speech estimation, both PESQ and STOI were decreased when harmonic enhancement was employed. Moreover, the performance was further decreased as the harmonic enhancement factor was increased. This indicates that employing harmonic enhancement was not a good choice for improving the quality of EMG-based speech synthesis. According to the experimental results, errors in pitch estimation were significantly higher, compared with silent UDS-based pitch estimation. This means that a major cause of such bad performance for EMG-based speech synthesis with

harmonic enhancement is the low accuracy of EMG-based pitch estimation.

The maximum PESQ of the synthesized speech signals from the image were slightly higher than that of silent UDS-based speech synthesis. The maximum PESQ exceeded 2.0, which could not be obtained by other modalities when no neighboring spectral features were considered (N_s=1). In the case of $N_s > 1$, the quality of the synthesized speech signals from the image was worse than that from silent UDS-based speech synthesis. Although the maximum STOI was found in image-based estimation, the average STOI of the method *S-S* is higher than that of the image-based method regardless of α_{he}. Consequently, the quality of silent UDS-based speech estimation is comparable to image-based estimation. The harmonic enhancement was in part effective to improve the quality of synthesized speech from image features. When a relatively smaller enhancement factor ($\alpha_{he} = 0.3$) was adopted, the average PESQ was increased by 0.007. However, no remarkable improvements were observed when $\alpha_{he} \geq 0.5$. Such results are also associated with the accuracy of image-based pitch estimation. It is generally known that the source components of speech signals including pitch period and excitation are not well predicted using an image captured from the speaker's mouth region [16]. However, our results showed that the quality of synthesized speech from images could be improved somewhat by adjusting the degree of harmonic emphasis when the pitch period is estimated from an image.

5.4 MOS Test Results

A subjective listening test was designed to evaluate the absolute quality of reproduced speech signals using the MOS (Mean Opinion Score) test. In this test, 18 listeners (12 males, 6 females; ages ranging from 21-52 years, mean age 26) participated and were asked to score the quality of the reproduced speech signals. All of them had normal hearing ability. The quality rating scale for each factor is as follows: Excellent=5/Good=4/Fair=3/Poor=2/Bad=1. The test data set consisted of 10 pairs of isolated words that were randomly taken from the database. None of the listeners had significant prior knowledge of the contents of the test sentences. Quality evaluation was carried out on the speech signals reproduced by the following three schemes: (1) audible-UDS synthesis (method *A-A*), (2) silent-UDS synthesis (method *A-S*), (3) silent-UDS synthesis (method *S-S*), (4) sEMG-based synthesis, and (5) image-based synthesis. For each method, the DNN configuration and the harmonic enhancement factor with the highest average PESQ were chosen according to Table 3.

The results are shown in Fig. 6 where the average MOSs are presented for each method. Similar to the results of PESQ and STOI, audible UDS-based synthesis (method *A-A*) yielded the highest MOS. The listeners indicated that utterances reconstructed using the method *A-A* sounded clearer and were less noisy than those produced using the other methods. This method, however, cannot be applied

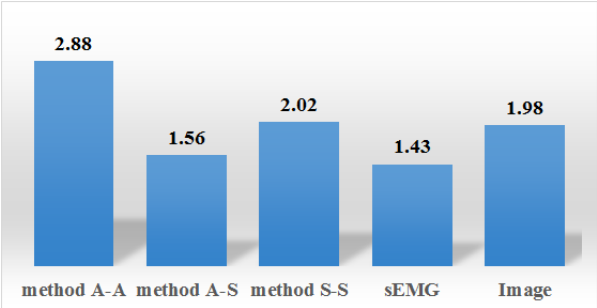

Fig. 6 MOS test results.

to implement a true SSI system, since audible UDS should be used for both the construction of synthesis rules and the synthesis procedure. Except for the synthesis method *A-A*, the *S-S* method showed the highest average MOS. The listeners indicated the quality of the synthesized speech by the method *S-S* was remarkably superior to both the method *A-S* and sEMG-based method in terms of intelligibility and naturalness. Compared with the image-based method, listeners were not able to discriminate differences easily, but indicated that the synthesized speech by the method *S-S* was slightly better. Consequently, the UDS-based speech synthesis scheme can be adopted for implementation of SSI that provides reasonable quality with the advantage of low-cost, no image-based preprocessing, and non-contact sensing.

6. Conclusions

In this study, a method for synthesizing speech using ultrasonic Doppler caused by articulatory movements was proposed and the effectiveness of the proposed method was verified. With the proposed method, the mapping rules between silent UDS and audible speech were taken into consideration. These were not addressed by previous UDS-based speech synthesis methods. The experimental results showed a clear difference between silent UDS and audible UDS. Speech synthesis was also carried out on other modalities. The quality of synthesized speech from silent UDS was remarkably better than that of speech synthesized from EMG signal. The overall quality of the speech synthesized from silent UDS was comparable to speech synthesized from an image.

UDS-based speech synthesis uses equipment that is non-contact, lightweight, and inexpensive, and it yields synthesized speech that compares to existing image-based speech synthesis methods. The rules for speech synthesis under completely silent conditions can be obtained by using the NNSA method proposed in this work. This would be helpful for implementing full silent speech interface systems. Future work will focus on determining the sensor locations and feature variable associated with silent UDS. We will also study a multi-modality silent speech interface system in which the use of both UDS and image complement one another to improve the perceptual quality of synthesized speech signals.

Acknowledgments

This paper was supported by Konkuk University in 2018.

References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert, and J.S. Brumberg, "Silent speech interfaces," Speech Communication, vol.52, no.4, pp.270–287, 2010.
[2] T.F. Quatieri, K. Brady, D. Messing, J.P. Campbell, W.M. Campbell, M.S. Brandstein, C.J. Weinstein, J.D. Tardelli, and P.D. Gatewood, "Exploiting nonacoustic sensors for speech encoding," IEEE Trans. Audio, Speech, Lang. Process., vol.14, no.2, pp.533–544, 2006.
[3] 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, vol.10, no.5, pp.4622–4633, 2010.
[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, vol.9, pp.199–214, 2008.
[5] S. Li, Y. Tian, G. Lu, Y. Zhang, H. Lv, X. Yu, H. Xue, H. Zhang, J. Wang, and X. Jing, "A 94-GHz milimeter-wave sensor for speech signal acquisition," Sensors, vol.13, no.11, pp.14248–14260, 2013.
[6] S. Li, Y. Tian, G. Lu, Y. Zhang, H. Xue, J.-Q. Wang, and X.-J. Jing, "A new kind of non-acoustic speech acquisition method based on millimeter wave radar," Progress in Electromagnetics Research B, vol.130, pp.17–40, 2012.
[7] C.-S. Lin, S.-F. Chang, C.-C. Chang, and C.-C. Lin, "Microwave human vocal vibration signal detection based on Doppler radar technology," IEEE Trans. Microw. Theory Techn., vol.58, no.8, pp.2299–2306, 2010.
[8] B. Denby and M. Stone, "Speech synthesis from real time ultrasound images of the tongue," Proc. IEEE International Conference on Acoustic, Speech and Signal Processing, pp.685–688, 2004.
[9] B. Denby, Y. Oussar, G. Dreyfus, and M. Stone, "Prospects for a silent speech interface using ultrasound imaging," Proc. IEEE International Conference on Acoustic, Speech and Signal Processing, pp.365–368, 2006.
[10] T. Hueber, G. Aversano, G. Chollet, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel, and M. Stone, "Eigentongue feature extraction for an ultrasound-based silent speech interface," Proc. IEEE International Conference on Acoustic, Speech and Signal Processing, pp.1245–1248, 2007.
[11] I. Almajai and B. Milner, "Visually derived Wiener filters for speech enhancement," IEEE Trans. Audio, Speech, Lang. Process., vol.19, no.6, pp.1642–1651, 2011.
[12] L. Girin, L. Varin, G. Feng, and J.L. Schwartz, "Audiovisual speech enhancement: New advances using multi-layer perceptrons," Proc. IEEE 2nd Workshop on Multimedia Signal Processing, pp.77–82, 1998.
[13] S. Deligne, G. Potamianos, and C. Neti, "Audio-visual speech enhancement with AVCDCN (Audio-Visual Codebook Dependent Cepstral Normalization)," Proc. Internationl Conference on Spoken Language Processing, pp.1449–1452, 2002.
[14] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audiovisual speech," Proceedings of the IEEE, vol.91, no.9, pp.1306–1326, 2003.
[15] V.-M. Florescu, L. Crevier-Buchman, B. Denby, T. Hueber, A. ColazoSimon, C. Pillot-Loiseau, P. Roussel, C. Gendrot, and S. Quattrocchi, "Silent vs vocalized articulation for a portable ultrasound-based silent speech interface," Proc. Interspeech, pp.450–453, 2010.
[16] T.L. Cornu and B. Milner, "Generating intelligible audio speech from visual speech," IEEE Trans. Audio, Speech, Lang. Process., vol.25, no.9, pp.1447–1457, 2017.
[17] A.R. Toth, K. Kalgaonkar, B. Raj, and T. Ezzat, "Synthesizing speech from doppler signals," Proc. IEEE International Conference on Acoustic, Speech and Signal Processing, pp.4638–4641, 2010.
[18] K. Livescu, B. Zhu, and J. Glass, "On the phonetic information in ultrasonic microphone signals," Proc. IEEE International Conference on Acoustic, Speech and Signal Processing, 4621-4624, 2009.
[19] K. Kalgaonkar, R. Hu, and B. Raj, "Ultrasonic doppler sensor for voice activity detection," IEEE signal processing Letters, vol.14, no.10, pp.754–757, 2007.
[20] K. Kalgaonkar and B. Raj, "Ultrasonic doppler sensor for speaker recognition," Proc. IEEE International Conference on Acoustic, Speech and Signal Processing, pp.4865–4868, 2008.
[21] T. Toda and K. Shikano, "NAM-to-speech conversion with Gaussian mixture models," Proc. INTERSPEECH, pp.1957–1960, 2005.
[22] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," IEEE Trans. Audio, Speech, Lang. Process., vol.20, no.9, pp.2505–2517, 2012.
[23] M. Janke, M. Wand, K. Nakamura, and T. Schultz, "Further investigations on EMG-to-speech conversion," Proc. IEEE International Conference on Acoustic, Speech and Signal Processing, pp.365–368, 2012.
[24] K.-S. Lee, "Prediction of acoustic feature parameters using myoelectric signals," IEEE Trans. Biomed. Eng., vol.51, no.7, pp.1587–1595, 2010.
[25] K.-S. Lee, "EMG-based speech recognition using Hidden Markov Models with global control variables," IEEE Trans. Biomed. Eng., vol.55, no.3, pp.930–940, 2008.
[26] M. Wand, M. Janke, and T. Schultz, "Tackling speaking mode varieties in EMG-based speech recognition," IEEE Trans. Biomed. Eng., vol.61, no.10, pp.2515–2526, 2014.
[27] M. Janke and L. Diener, "EMG-to-Speech: Direct generation of speech from facial electromyographic signals," IEEE Trans. Audio, Speech, Lang. Process., vol.25, no.12, pp.2375–2385, 2017.
[28] B. Raj, K. Kalgaonkar, C. Harrison, and P. Dietz, "Ultrasonic doppler sensing in HCI," IEEE Pervasive Computing, vol.11, no.2, pp.24–29, 2012.
[29] K. Kalgaonkar and B. Raj, "Acoustic doppler sonar for gait recognition," Proc. IEEE Conference Advanced Video and Signal Based Surveillance, pp.27–32, 2007.
[30] K. Kalgaonkar and B. Raj, "One-handed gesture recognition using ultrasonic doppler sonar," Proc. IEEE International Conference on Acoustic, Speech and Signal Processing, pp.1889–1892, 2009.
[31] L.R. Rabiner and R.W. Schafer, Digital Processing of Speech Signals, Prentice Hall, 1978.
[32] G. White and R.B. Neely, "Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming," IEEE Trans. Acoustic Speech and Signal Processing, vol.24, no.2, pp.183–188, 1976.
[33] L. Deng, M.L. Seltzer, D. Yu, et al, "Binary coding of speech spectrogram using a deep auto-encoder," Proc. Interspeech, pp.1692–1695, 2010.
[34] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE Trans. Audio, Speech, Lang. Process., vol.23, no.1, pp.7–19, 2015.
[35] D. Griffin and J. Lim, " Signal estimation from the modified short-time fourier transform," IEEE Trans. Acoustic Speech and Signal Processing, vol.32, pp.236–243, 1984.
[36] W. Jin, X. Liu, M.S. Scordilis, and L. Han, "Speech enhancement using harmonic emphasis and adaptive comb filtering," IEEE Trans. Audio, Speech, Lang. Process., vol.18, no.2, pp.356–368, 2010.
[37] K. Han and D. Wang, "Neural network based pitch tracking in very noisy speech," IEEE Trans. Audio, Speech, Lang. Process., vol.22, no.12, pp.2158–2168, 2014.
[38] G.E. Hinton, "Training products of experts by minimizing contrastive divergence," Neural Computation, vol.14, no.8, pp.1711–1800, 2002.
[39] ITU-T, Rec. P. 862, Perceptual evaluation of speech quality (PESQ):

An objective method for end-to-end speech quality assessment of narrow band telephone networks and speech codecs, International Telecommunication Union-Telecommunication Standardisation Sector, 2001.

[40] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," IEEE Trans. Audio, Speech, Lang. Process., vol.19, no.7, pp.2125–2136, 2011.

Ki-Seung Lee received the B.S., the M.S. and the Ph.D. degrees in electronics engineering from Department of Electronics, Yonsei University, Seoul, Korea, in 1991, 1993 and 1997, respectively. In February 1993, he joined CSPR (Center for Signal Processing Research) in Yonsei University. From October 1997 to September 2000, he had been with AT&T Labs-Research, Florham Park NJ, USA. From November 2000 to August 2001, he had been with SAIT (Samsung Advanced Institute of Technology), Suwon, Korea. Since August 2001, he has been with the faculty position of Konkuk University, Seoul, Korea. His interests include bio-signal modelling, speech signal processing, and audio signal processing.