Speech Reconstruction from The Larynx Vibration Feature Captured by Laser-Doppler Vibrometer Sensor

Yi-Chieh Lin*, Ji-Yan Han*, Yu-Min Lin*, Wei-Zhong Zheng*, Shuenn-Tsong Young† and Ying-Hui Lai*
*Department of Biomedical Engineering, National Yang Ming Chiao Tung University
†Institute of Geriatric Welfare Technology & Science, MacKay Medical College
E-mail: {veralin.be09; jy.han.be08; yuminlin.be09; s1010654.be07}@nycu.edu.tw; styoung@mmc.edu.tw; yh.lai@nycu.edu.tw

***Abstract*—There are many deep learning (DL)-based models with the contact sensors (e.g., throat microphone, TM) to reconstruct the speech from the vibration signals of the larynx. The TM can obtain robust speech information than an air-conducted microphone (ACM) sensor in noisy environments. However, it needs tight contact with the user's skin, which causes discomfort for users. Therefore, we assume that a non-contact sensor allows users to have a better experience. Following this concept, the DL-based models with a non-contact sensor, a laser-Doppler vibrometer (LDV), are proposed to reconstruct the speech from the vibration signals of the larynx. Notably, the recognition and speech synthesis modules were adopted in the proposed system. The experimental results showed that, on average, the word error rate (WER) of the recognition module in the proposed system achieves similar performance as TM did in both quiet and noisy testing conditions. Furthermore, the listening test showed that the synthesis module's reconstructed speech provided a higher preference rate and naturalness than an original recorded speech of the LDV sensor. These results suggested that the proposed system is a potential approach to reconstruct speech from the vibration signals of the larynx with DL technology, captured by a non-contact LDV sensor.**

I. INTRODUCTIONS

A previous study [1] indicated that the entire human head vibrates while speaking, especially the skin of the larynx. In the meantime, the vibration of the larynx produces essential information regarding the fundamental frequency (F0) of speech [2]. This F0 provides rich data for speech intelligibility [3], including linguistic cues such as the voicing period [4] and lexical boundary [5]. Based on this concept, many contact-based sensors, such as a body-conducted microphone (BCM) and throat microphone (TM), are used to record the vibration of the larynx and reconstruct (or recognize) speech using deep learning (DL) models [6, 7].

The TM sensor detects the voice vibration passed by the bone and skin through the accelerometer (or uses a special mechanism design with an acoustic sensor), in which the signals are less affected by surrounding noise than a conventional air-conducted microphone (ACM) [8]. Several successful applications and studies have been conducted. For instance, Mcbridge et al. proved that the TM could catch the vibration while speaking with the sensor attached to different parts of the head. The results showed that the foreheads and temples could achieve the highest speech intelligibility and quality [9]. Liu et al. used a DL-based approach to transfer recorded speech of TM to acoustic-based speech to enhance speech quality. The results showed that the converted speech could improve the recognition performance of the automatic speech recognition (ASR) system in quiet and noisy conditions [10]. Additionally, because the skin attenuates high-frequency voice components, the accelerometer type TM cannot acquire high-frequency speech information. Therefore, some studies have used bandwidth extension technology to reconstruct high-quality speech from recorded signals of the accelerometer type TM, such as [11]. Conversely, Zheng et al. [6] combined the signal of ACM and the accelerometer type TM that attached to the larynx skin to catch the vibration signal (i.e., robust against external noise). This was to train a deep bidirectional using long short-term memory neural network to enhance these input speeches further. The experiment results showed that the proposed method provided higher speech quality performance than original noisy speech.

Although the TM was shown to have potential benefits for speech signal processing systems in noisy conditions, there is still room for improvement. More specifically, the TM sensor requires tight contact with the user's skin, which may cause discomfort and is easily affected by moisture [9]. Therefore, a non-contact sensor, such as a laser-Doppler vibrometer (LDV), could be used to release the TM sensor to alleviate the above issues. The LDV sensor can measure the vibration frequencies of moving targets based on a non-contact approach, and it has been used to observe the vibration of objects in the industry [12] and medical field [13, 14]. Recently, several studies [15-19] have used LDV to record speech from vibration objects, and the results have proven that the LDV sensor provides advantages of non-contact, long-distance, and noise robustness in speech processing tasks. The major problem of LDV sensors is that the high-frequency part of speech will be missing depending on the object material; hence, it will decrease the speech quality in actual application conditions. Many studies used the DL technologies [16, 20] to reconstruct the high-frequency part to improve the speech quality to alleviate this issue. In addition, Sun et al. and Xie et al. used an LDV sensor to record the vibration signals of speakers' larynx skin like the TM did and

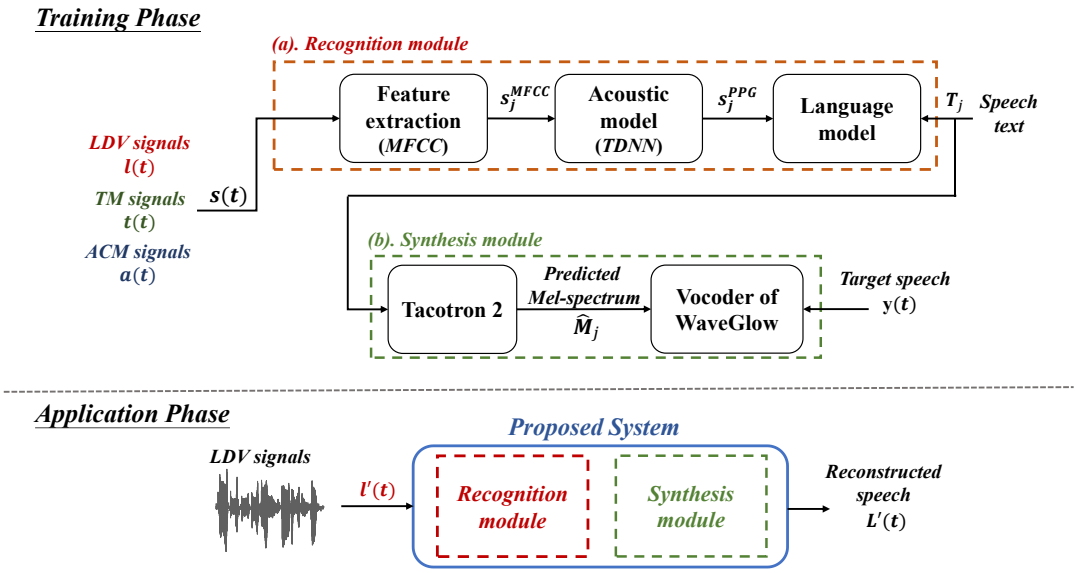

Fig. 1. *Block diagram of the proposed reconstruction system from the vibration of larynx skin, which is recorded via the LDV sensor.* $l(t)$, $t(t)$, *and* $a(t)$ *were the signal recorded by LDV, ACM, TM respectively.* $s(t)$ *was the combined signal of three sensors, and* T_j *was the speech text.* s_j^{MFCC}, s_j^{PPG}, *were the MFCC features and PPG features in recognition module,* $\widehat{M}_j$ *was the predicted Mel-spectrum of* T_j, $y(t)$ *was the target speech, and the component* $l'(t)$ *and* $L'(t)$ *represent the LDV testing set and the output of our proposed system.*

used these recorded signals as auxiliary features for the ASR system. The results showed that the auxiliary features could significantly improve recognition performance under quiet and noisy test conditions [21, 22]. Although the above studies [15-19] proved that the LDV sensor could capture speech signals from rigid objects around us (e.g., a glass window, metal plate, plastic box), there was no research to explore whether the speech could be reconstructed through the vibration signals of the larynx, captured by LDV sensor only. Note that we assume that the signals obtained from the larynx will be more robust than targeting objects around us, because it is the voice source. Therefore, the main purposes of this research are (1) to study the potential to reconstruct speech recorded from the vibration signals of the larynx by a non-contact LDV sensor via DL technology; (2) to compare the performance of a non-contact LDV sensor with the ACM and TM sensor in the quiet and noisy conditions.

The remainder of this paper is organized as follows. First, the proposed system is introduced in Section II. The experimental design and results are presented in Sections III and IV, respectively. Finally, Section V summarizes the findings.

II. Proposed System

The proposed speech reconstruction system is shown in Fig. 1, where $l(t)$ was the vibration signals of the larynx recorded by the LDV sensor, $t(t)$ was recorded by the TM, and $a(t)$ was recorded by the ACM, respectively. For the LDV sensor, the laser beam is divided into an object beam and a reference beam by a beam splitter. The object beam focuses on the surface of the target measurement point (speaker's larynx in this study), and the corresponding backscattered beam with a Doppler shift (f_D) is reflected in the beam splitter. Meanwhile, the reference beam passes through a Bragg cell to produce a frequency shift (f_B). Subsequently, these two beams (i.e., f_B and f_D) are mixed, and the photodetector converts the frequency shift (i.e., f_B+f_D) into voltage signals. Finally, these converted voltage signals were used to calculate the velocity and obtain information on the vibration of the target measurement point. A detailed description of the LDV sensor can be found in [23, 24]. Next, the other two recorded signals, $t(t)$ and $a(t)$, were combined with $l(t)$ to a set of $s(t)$ to train the proposed system. More specifically, the recognition module of the proposed system was trained by $s(t)$ with the speech texts (T_j), where j is the frame index. During training stage, the Kaldi ASR system [25] was used to train the recognition module, and the Kaldi data augmentation methods, such as changing the speed and *VTLN* [26], was used to increase the training set 15 times. Note that there were two main models, the acoustic model (AM) and the language model (LM), included in the ASR system. In AM, we used Mel-frequency cepstral coefficients (s_j^{MFCC}) [27], cepstral mean and variance normalization [28], and the i-vector [29] features to train a GMM-HMM LDA+MLLT system, and it would produce a distribution over from a monophone target to triphone target. Meanwhile, the time-delay neural network (TDNN) [30] model was used to predict the probability of phoneme for the input speech and to find possible words in the pronunciation dictionary, where the 13 layers of the TDNN and 128 sampling points of each layer were used in this study. Subsequently, it built the PPG features [31] (s_j^{PPG}) of the speech based on AM. For the LM, we trained with the lexicon of Taiwan Mandarin hearing in the noise test (MHINT) [32], which consisted of over thousands of Chinese words. It predicts the possibility of the next word from the word predicted by AM,

where the trigram is used in our LM, shown in (1). More specifically, the $p(w_1^3)$ means the conditional probability of three words likely to be together, which is the multiplication of the probability that w_1 occurs ($p(w_1)$), w_2 occurs when w_1 occurs ($p(w_2|w_1)$), and w_3 occurs when both w_1 and w_2 occur $p(w_3|w_1, w_2)$. Finally, the recognition module of the proposed system will predict the text ($\hat{T}_j$) from the input signals $s(t)$.

$$p(w_1^3) = p(w_1) * p(w_2|w_1) * p(w_3|w_1, w_2) \quad (1)$$

Next, the utterances of the target speaker (recorded by ACM sensor) were used to train the synthesis module of the text-to-speech (TTS) system, which the Tacotron 2 [33] technology be used in this study. First, the speakers' ACM utterances $y(t)$ and labeled speech text (T_j) were used as the training data for the Tacotron 2 [33] model. Tacotron 2 used an end-to-end encoder and decoder with an attention mechanism to make the neural network learn the tempo and intonation of the subject from the text. The primary mechanism is illustrated in (2) to (4).

$$\alpha_t(n) = \exp(e_{t,n}) / \sum_{m=1}^{N} \exp(e_{t,m}) \quad (2)$$

$$e_{t,i} = v^T \tanh\ (Wy_{t-1} + Vh_i + Uf_{t,i} + b) \quad (3)$$

$$f_{t,i} = F \times \alpha_{t-1}(n) \quad (4)$$

In the encoder, character embedding sequences are extracted from the input text sequences, and the encoded values are decoded with an attention-based LSTM decoder to predict the Mel-spectrogram ($\hat{M}_j$) features from the input speech text ($\hat{T}_j$). In the attention mechanism, the weight, $\alpha_t(n)$, represents the probability of text energy and is calculated as follows; where t is the timestep, W, V, U, v, and b were trainable attention parameters, $e_{t,i}$ represents the energy, y_{t-1} is the hidden state of the previous decoder layer, h_i is the i-th character embedding from hidden encoder state, $f_{t,i}$ stands for a location-sign calculated by a convolution F and the previous attention weight. Then this TTS system will learn the distribution of the Mel-spectrogram for the character, and we need a vocoder to transform the Mel-spectrogram into speech.

Following, we used the predicted Mel-spectrogram $\hat{M}_j$ to synthesize speech with the vocoder of WaveGlow, which was proven that it could provide fast and more high-quality speech synthesis [34] than other classical waveform synthesis approaches. The WaveGlow vocoder was trained using the same acoustic-based utterances recorded by the target speaker. For the used WaveGlow vocoder in this study, the sampling rate was 16 kHz, the parameters of the Mel-spectrograms filter length was 1024, hop length was 256, and window size was 1024. Next, the 12 steps of flows were used in this study with a transformation function of eight-layer architecture, which was similar to WaveNet [35], to learn the distribution of our target speech. A detailed description of WaveGlow can be found in [34].

In the application phase, the speaker's laryngeal vibration signals $l'(t)$ as the input signals of the proposed system, recorded by the LDV sensor. Next, the trained modules of the proposed system were used to recognize and synthesize speech $L'(t)$ from $l'(t)$ directly.

III. Methods

A. Materials

The same person recorded the training and testing utterances with an air conduction microphone (ACM) (GRAS type 40PH) [36], throat microphone (TM) (GRAS 40LS) [37], and laser-Doppler vibrometer (LDV) (Optomet Vector Star) [38] sensors, respectively. This study protocol was approved by the Institutional Review Board (IRB) (YM110144E) of National Yang Ming Chiao Tung University. The experimental setup was shown in Fig. 2. The LDV was set one meter away from the larynx of the speaker, a TM was attached on the left side of the speaker's throat, and an ACM was set 15 cm away from the speaker's mouth. A background noise source, $n(t)$, was set 1.5 meters away from the speaker for the noise robustness test. Note that the LDV sensor was a class two laser with a wavelength of 632.8 nm, which would not harm the eye unless a person deliberately stared into the beam and had no risk to the skin.

This experiment was conducted in an audiometric room to ensure the credibility of the TM and ACM data. A total of 3840 (320 utterances × 4 times × 3 sensors) utterances as the training set, recorded in quiet condition, where the corpus list was adopted from the Taiwan Mandarin hearing in the noise test [32]. Meanwhile, the other 270 (30 utterances × 3 sensors × 3 conditions) were used as a testing set. These were recorded in three conditions: one was quiet condition, and the other two conditions were in the IEEE speech shape noise (SSN) and constructions noise at 65 dB SPL level. Note that these utterances were randomly selected from the list of the training set and asked the speaker to repeat these utterances. Fig. 3 shows an example of the spectrogram of the above test conditions in quiet and noisy conditions, respectively. Notably, the non-contact LDV sensor obtained suitable vibration signal quality of the larynx and was less affected by the background noise (red circle) than ACM and contact TM sensor. However, the middle-

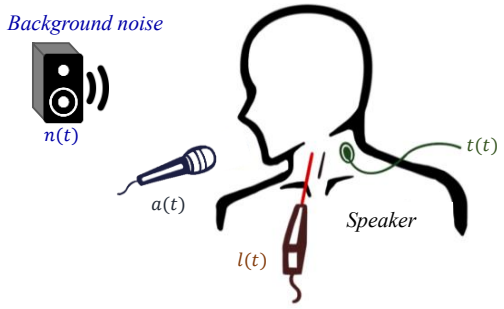

Fig. 2. *The experimental setup of data recording in this study.* $l(t)$, $t(t)$, $a(t)$ *represent the LDV signals, TM signals, ACM signal, and* $n(t)$ *represent the background noise.*

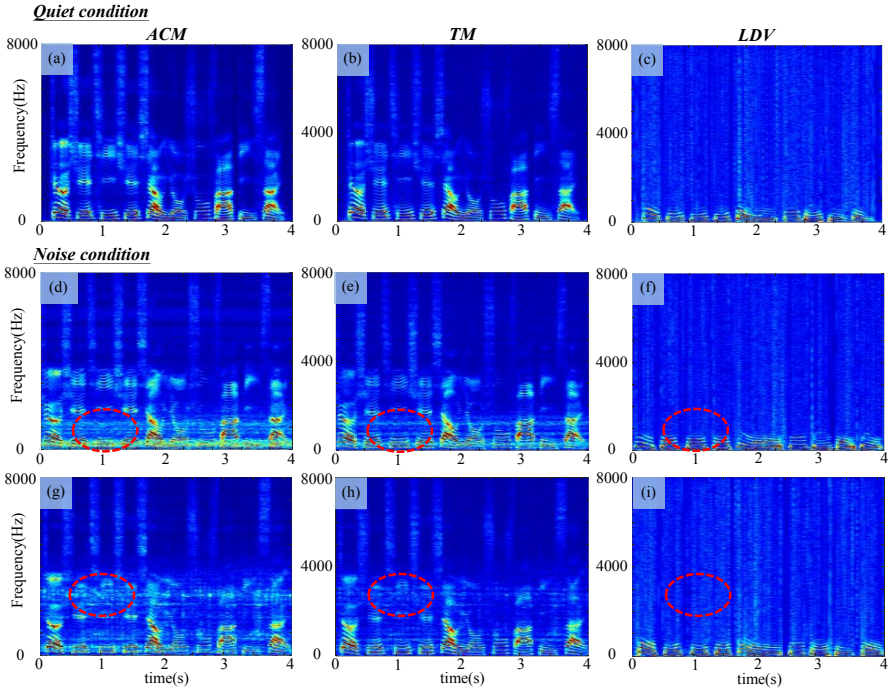

Fig. 3. *An example of the spectrogram of the recorded signal by ACM, TM and LDV sensors. (a) to (c) were recorded by ACM, TM and LDV sensor in quiet conditions, respectively; (d) to (f) were recorded by ACM, TM and LDV sensor in IEEE SSN noisy condition, respectively; and (g) to (i) were recorded by ACM, TM and LDV sensor in constructions noisy condition, respectively. Note that the background noise level was 65 dB SPL in this study.*

to high-frequency regions is empty. Therefore, the DL technology could be needed to reconstruct speech to improve speech quality from the recorded information by LDV sensor.

B. Procedure

This study aimed to ensure the capability of reconstruction speech from the vibration signals of the larynx by LDV sensor with DL technology; meanwhile, the ACM and TM sensor were used as the comparison in quiet and noisy test conditions. Therefore, first, the training set (described in the section of *Materials*) was used to train the recognition and synthesis models in the training phase. The detailed descriptions of training steps were described in Section II. (Proposed System). Subsequently, in the application phase, the trained models of the proposed system were used to reconstruct the speech $L'(t)$ from the speech $l'(t)$ derived from the LDV sensor.

Following this, we conducted the objective and subjective listening test to ensure the performance of the proposed system, in which the ACM and TM sensors were used as the comparison. Notably, the word error rate (WER), which represents the rate of deletion, substitutions and insertions in the number of chinese words in the reference sentences, was used to evaluate the recognition module of the proposed system in objective evaluation. The testing set (described in the Materials section) was used to evaluate the performance of the proposed system in quiet and noisy test conditions. Meanwhile, the ACM and TM sensors were used as the comparison simultaneously. Next, the listening test evaluates the naturalness and speech preference between the synthesis module of the proposed system and the original LDV recorded speech. We recruited six subjects to enter the perceptual listening test, which included 45 audio files of 15 random utterance sets (15 original LDV speeches, 15 reconstructed speeches of the proposed system, and 15 ACM speeches of the target speaker) with the same volume. By each utterance, each subject would listen to the ACM speech to build the first impression. Then, subjects would listen to the other speech in random order and choose the speech with better naturalness and preference between the original LDV and reconstructed speech by the synthesis module. Note that for speech naturalness, we asked, "*Which speech do you think is more natural?*" For speech preference, we asked, "*Which speech do you prefer?*".

IV. Results and Discussion

A. Recognition performance

Table I shows the WER (%) of the recognition module in the proposed system in three test conditions. In quiet test conditions, the performance of the proposed system achieves good performance (0%), similar to the performance of ACM and TM sensors via the same recognition module in duplicate test conditions (speaker repeats the utterances). These results prove that a non-contact LDV sensor can obtain enough speech information for the recognition module to recognize the speech. Moreover, the vibration-based speech signals of the larynx and the effective recorded signals were less with F0 under 1000 Hz. Therefore, it indicated that a lower sampling rate LDV devices

could reduce hardware setup costs for real application scenarios in future studies.

Conversely, Table I shows the WER of the recognition module in IEEE SSN background noise conditions. According to the results, the performance of ACM was 16.59%, TM sensor was 1.46%, and LDV sensor was 3.9% in the SSN background noise, respectively. As for the construction noise, the LDV achieves 7.32%, and TM reaches 5.37%. However, ACM only gets 24.88%. The result above shows that in noise conditions, LDV and TM have greater noise robustness than ACM. The one-way analysis of variance (ANOVA) [39] with Tukey HSD post-hoc comparison [40] was used to analyze the result of the ASR error rate for three sensors under two noise conditions. The ANOVA results confirmed that the error rate was differed significantly among different sensors, with the F=13.872 under SSN and F=8.641 (p<0.05) under construction noise. The post-hoc comparisons value further indicated that the ACM sensor was significantly different from LDV and TM sensors. Meanwhile, the difference between LDV and TM was not significant with the p-value>0.9. In other words, it implies that this non-contact LDV sensor can provide the same performance as the contact TM sensor in noisy conditions. Hence, the LDV sensor could be a promising approach to catch the vibration signals of the larynx to perform the advantage of noise-robust in the future.

TABLE I

The WER (%) of the recognition module of the purposed system in quiet and two noisy test conditions.

Conditions	ACM	TM	LDV
Quiet	0%	0%	0%
SSN	16.59%	1.46%	3.9%
Construction	24.88%	5.37%	7.32%

B. Spectrogram analysis

Fig. 4(c) shows an example of synthesis speech from the proposed system. The input signals are shown in Fig. 4(a). This example indicates that the synthesis speech had a similar speech structure to that of the target speech (i.e., recorded by an ACM), such as red arrow and circle parts in Fig. 4. From this example, the recognized text of the recognition module can be synthesized by the synthesis module to reconstruct more detailed information to restore the speech quality, especially in the middle- and high-frequency regions. Therefore, the proposed system could provide a more natural speech than an original recorded speech from the LDV sensor for listeners.

C. Listening test

The listening test denotes the benefits of our proposed system in speech naturalness and preference. From the average results, 88% of the reconstructed speeches were preferred by all subjects with normal hearing; meanwhile, the subjects also agreed that the reconstructed speech has better naturalness in 82% of

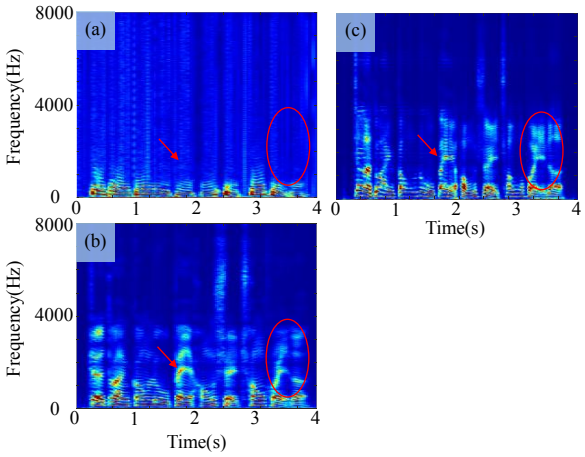

Fig. 4. *Comparison of spectrogram between (a) recorded speech by LDV, (b) recorded speech by ACM, and (c) reconstructed speech by the proposed system in our study. More samples of reconstructed speech are available at* "https://reurl.cc/bXYqXl".

utterances. We think that subjects' preferences might be affected by attenuation in the high frequency of larynx vibration signals, which made subjects unable to properly hear LDV speech sound to affect the speech quality. Compared to the original recorded LDV speech, the higher frequency information (e.g., after 1k Hz) is reconstructed by the DL technology, such as in Fig. 4(c). Hence, it would help listeners hear clearer and detailed speech information to increase their naturalness and preference.

In summary, the above results demonstrate that the proposed system with LDV signals could acquire powerful features from the larynx. Then reconstruct a highly intelligible and suitable speech quality for listeners. The naturalness did not perform perfectly, perhaps due to the TTS system, which still has room for improvement in our future studies. However, this study's primary purpose is to examine the potential to reconstruct speech recorded from the vibration signals of the larynx by a non-contact LDV sensor via DL technology. Thus, fine-tuning synthesis module was not the current focus of this study. Therefore, we can further record more ACM data and fine-tune the synthesis module to improve the synthesis speech quality. Additionally, we believe that the speech naturalness and quality can reach higher performance than the current version in future studies. On the other hand, this proposed system still have some limitations. There is no such a small size and low-cost commercial LDV device currently. Hence, the hardware design of this non-contact optical speech reconstruction system still needs to be studied. However, with the advancement of semiconductor technology in the future, we believe this miniaturized and low cost sensor can be put into practice in speech processing products.

V. CONCLUSIONS

We studied the potential of a speech reconstruction system with the DL technology based on the vibration signals of the larynx, recorded by an LDV sensor; meanwhile, the ACM and

TM sensors were used for comparisons. The experimental results showed that, on average, the recognition module of the proposed system (LDV sensor) provided lower WER performance in quiet and noisy testing conditions. This was similar to the TM sensor in the same recognition module. Conversely, the LDV and TM sensors provided better performance than the ACM sensor in noisy testing conditions (Table I shows the detailed results). Furthermore, the listening test results indicated that the speech synthesis module of the proposed system provided suitable preferences and naturalness rate with the target speaker, compared to the original LDV recorded speech. These findings suggested that the vibration signals of the larynx, recorded by the LDV sensor, are potential speech features for speech reconstruction applications in the future.

ACKNOWLEDGMENT

This study was supported by the Ministry of Science of Technology of Taiwan under the 110-2218-E-A49A-501 project. The authors would like to thank IEA Electro-Acoustic Technology Co., Ltd. for providing the experimental devices.

REFERENCES

1. B. J. Munger and S. L. Thomson, “Frequency response of the skin on the head and neck during production of selected speech sounds, ” *The Journal of the Acoustical Society of America*, vol.124, no.6, pp. 4001-4012, 2008.
2. Understanding How Voice is Produced, Available from: “https://voicefoundation.org/health-science/voice-disorders/anatomy-physiology-of-voice-production/understanding-voice-production/.”
3. C. A. Brown and S. P. Bacon, “Fundamental frequency and speech intelligibility in background noise, ” *Hearing research*, vol.266, no.1-2, pp. 52-59, 2010.
4. L. L. Holt, A. J. Lotto, and K. R. Kluender, “Influence of fundamental frequency on stop-consonant voicing perception: A case of learned covariation or auditory enhancement?,” *The Journal of the Acoustical Society of America*, vol. 109, no.2, pp.764-774, 2001.
5. S. M. Spitzer, J. M. Liss, and S. L. Mattys, “Acoustic cues to lexical segmentation: A study of resynthesized speech, ” *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3678-3687, 2007.
6. C. Zheng, X. Zhang, M. Sun, J. Yang, and Y. Xing, “A novel throat microphone speech enhancement framework based on deep BLSTM recurrent neural networks, ” *2018 IEEE 4th International Conference on Computer and Communication*, 2018.
7. M. A. T. Turan, and E. Erzin, “Improving phoneme recognition of throat microphone speech recordings using transfer learning, ” *Speech Communication*, 2021.
8. S. Lin, T. Tsunakawa, M. Nishida, and M. Nishimura, “DNN-based feature transformation for speech recognition using throat microphone, ” *in Proc. IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference,* 2017.
9. M. McBride, P. Tran, T. Letowski, and R. Patrick, “The effect of bone conduction microphone locations on speech intelligibility and sound quality, ” *Applied ergonomics*, vol. 42, no. 3, pp. 495-502, 2011.
10. H. Liu, Y. Tsao, and C. Fuh, “Bone-conducted speech enhancement using deep denoising autoencoder,” *Speech Communication*, vol. 104, pp. 106-112, 2018.
11. C. Zheng, J. Yang, X. Zhang, T. Cao, M. Sun, and L. Zheng, “Bandwidth Extension WaveNet for Bone-Conducted Speech Enhancement,” *in Proc. of the 7th Conference on Sound and Music Technology ,* Springer, Singapore, 2020.
12. S.J. Rothberga, M.S. Allenb, P. Castellinic, D. Di Maiod, J.J.J. Dirckx, D.J. Ewins, et al, “An international review of laser Doppler vibrometry: Making light work of vibration measurement,” *Optics and Lasers in Engineering*, vol. 99, pp. 11-22, 2017.
13. L. Antognoli, S. Moccia, L. Migliorelli, S. Casaccia, L. Scalise, E. Frontoni, “Heartbeat Detection by Laser Doppler Vibrometry and Machine Learning,” *Sensors*, vol. 20, no. 18, pp. 5362, 2020.
14. H. Tabatabai, D. E. Oliver, J. W. Rohrbaugh, and C. Papadopoulos, “Novel applications of laser Doppler vibration measurements to medical imaging,” *Sensing and Imaging: An International Journal*, vol. 14, no. 1, pp. 13-28, 20130.
15. Y. Deng, “Distant Speech Recognition Using Laser Doppler Vibrometer,” *23rd Iberoamerican Congress on Pattern Recognition*, Available from: “https://www.researchgate.net/publication/330290096_Distant_Speech_Recognition_Using_Laser_Doppler_Vibrometer”
16. C. Cai, K. Iwai,T. Nishiura, and Y. Yamashita, “Speech Enhancement for Optical Laser Microphone With Deep Neural Network,” *in Proc. IEEE 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2020.
17. S. Ueda, K. Iwai, T. Fukumori, T. Nishiura, “Sound quality improvement for speech acquisition based on deep learning and harmonic reconstruction with laser microphone,” *Universitätsbibliothek der RWTH Aachen*, 2019.
18. W. Li, M. Liu, Z. Zhu, T. S. Huang, “LDV remote voice acquisition and enhancement,” *in Proc. IEEE 18th International Conference on Pattern Recognition (ICPR'06)*, vol. 4, 2006.
19. R. Peng, B. Xu, G. Li, C. Zheng, and X. Li, “Long-range speech acquirement and enhancement with dual-point laser Doppler vibrometers,” *2018 IEEE 23rd International Conference on Digital Signal Processing*, 2018.
20. R. Peng, C. Zheng, and X. Li, “Bandwidth extension for speech acquired by laser Doppler vibrometer with an auxiliary microphone,” *in Proc. 2015 10th International Conference on Information, Communications and Signal Processing*, 2015.
21. L. Sun, J. Du, Z. Xie, and Y. Xu, “Auxiliary features from laser-Doppler vibrometer sensor for deep neural network based robust speech recognition,” *Journal of Signal Processing Systems*, vol. 90, no. 7, pp. 975-983, 2018.
22. Z. Xie, J. Du, I. McLoughlin, Y. Xu, F. Ma, and H. Wang, “Deep neural network for robust speech recognition with auxiliary features from laser-Doppler vibrometer sensor,” *in Proc. IEEE 2016 10th International Symposium on Chinese Spoken Language Processing*, 2016.
23. Y. Avargel, T. Bakish, A. Deke, G. Horovitz, Y. Kurtz, and A. Moyal, “Robust speech recognition using an auxiliary laser-doppler vibrometer sensor,” *in Proc. Speech Process, Conf., Tel-Aviv*, Israel, 2011.
24. Y. Avargel, and I. Cohen, “Speech measurements using a laser Doppler vibrometer sensor: Application to speech enhancement,” *in Proc. IEEE 2011 Joint Workshop on*

Hands-free Speech Communication and Microphone Arrays, 2011.

25. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Geol, et al, "The Kaldi speech recognition toolkit. in IEEE 2011 workshop on automatic speech recognition and understanding," *IEEE Signal Processing Society*, 2011.
26. M. B. Sung, B. K. Choi, Y. H. Choi, and H. S. Kim, " VTLN Based Approaches for Speech Recognition with Very Limited Training Speakers," *in Proc. IEEE International Conference on Intelligent Systems, Modelling and Simulation*, 2014.
27. S. Davis, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357-366, 1980.
28. O. Viikki, and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. (1-3), pp. 133-147,1998.
29. N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788-798, 2010.
30. A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *in IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328-339, March 1989.
31. L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," *in Proc. IEEE International Conference on Multimedia and Expo*, pp. 1-6, 2016.
32. L. L. N. Wong, S. D. Soli, S. Liu, N. Han, and M. W. Huang, "Development of the Mandarin hearing in noise test (MHINT) ," *Ear and hearing*, vol.28, no.2, pp. 70S-74S, 2007.
33. J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, et al, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4779-4783, 2018.
34. R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.
35. A. V. D. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, et al, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499* ,2016.
36. GRAS type 40PH. Available from: "https://www.grasacoustics.com/products/special-microphone/array-microphones"
37. GRAS type 40LS. Available from: "https://www.grasacoustics.com/products/special-microphone/surface-microphones/product/192-40ls"
38. Optomet. Vector-Series:Digital Free Beam HeNe Laser Vibrometer. Available from: "https://www.optomet.com/products/single-point-vibrometers/vector-series/"