Mandarin Electro-Laryngeal Speech Enhancement based on Statistical Voice Conversion and Manual Tone Control

Zhaopeng Qian[*, ‡], Haijun Niu[*, †], Li Wang[§], Kazuhiro Kobayashi[‡], Shaochuan Zhang[*, †] and Tomoki Toda[‡]
[*] School of Biological Science and Medical Engineering, Beihang University, Beijing, China
E-mail: hjniu@buaa.edu.cn Tel: +86-13466641948
[†] Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University. Beijing, China
[‡] Information Technology Center, Nagoya University, Nagoya Aichi, Japan
[§] Beijing Research Center of Urban System Engineering. Beijing, China

Abstract— **Electro-Larynx can help the laryngectomees re-pronounce the voice, while the Electro-Laryngeal (EL) speech has a poor intelligibility and naturalness. Recently, voice conversion (VC) has been applied to enhance the EL speech, which achieves a good result. However, the complicated tone variation rule of continuous Mandarin EL speech takes a new challenge into enhancement of EL speech by VC. In this paper, a novel framework combining manual tone control (MTC) and statistical VC is proposed to enhance the continuous Mandarin EL speech. As statistical VC methods, GMM-based VC and CLDNN-based VC are implemented for the proposed framework. The objective and subjective evaluations are designed to validate the proposed framework. The experimental results have demonstrated that 1) the combination of MTC and statistical VC yields significant improvements in both naturalness and intelligibility of the enhanced Mandarin EL speech, 2) the word perception error rates of the enhanced Mandarin EL speech is decreased from 11.35% of Mandarin EL speech with MTC to 5.61% by using statistical VC, and 3) the proposed framework achieves the average tone accuracy of 26.59% higher than that of original continuous Mandarin EL speech.**

I. Introduction

Electro-Larynx is one of the most commonly used assistive devices for speech recovery of laryngectomees. It has many advantages such as simple operation and continuous pronunciation. However, Electro-Laryngeal (EL) speech is limited in its intelligibility and naturalness due to its defects of flatten fundamental frequency (F0), mechanical sound radiation noise, especially for tonal languages [1-3].

Many methods have proposed to improve the quality of EL speech. Generally, conventional methods are proposed to improve the F0 or to improve the whole quality of the EL speech using voice signal processing technologies. To improve the F0, researchers proposed several methods based on the physiological signal controlled such as vocal air pressure signal [4], neck strap muscle electromyographically activity signal [5, 6], finger pressure signal [7-9], and finger sliding signal [10, 11]. Especially, Wang [11] designed the electro-larynx based on capacity touch technology to allow laryngectomees to pronounce the four tones of Mandarin EL speech by controlling F0 of an excitation signal of the electro-larynx with the capacity touch board. However, the above methods for improving the F0 are limited by the complexity of continuous speech pronunciation mechanism. To improve the whole quality of the EL speech, several methods including noise cancelling based on adaptive filter and spectral subtraction are proposed. Typically, Espy-Wilson et al. [12, 13] and Niu et al. [14] proposed the noise cancelling methods based on adaptive filter to enhance the EL speech; Cole [15], Pandey [16] and Liu et al. [17] proposed noise cancelling based on spectral subtraction to enhance the quality of EL speech. The above methods cannot process EL speech frame by frame, due to their fixed parameter settings. Therefore, the effect of EL speech enhancement is still limited.

Voice Conversion (VC) is a powerful method combining the machine learning and speech signal processing. VC can convert the source speaker speech to the target effectively [18, 19]. One of the most commonly used VC method for enhancing EL speech is based on Gaussian Mixture Model (GMM) [20-24]. Recently, Kobayashi et al. [25] proposed CLDNN-based VC for enhancing the EL speech. The CLDNN was originally proposed as an acoustic model in automatic speech recognition, which enabled significant improvements in speech recognition accuracy [26]. The original CLDNN consisted of convolutional neural network (CNN) layers, long short-term memory (LSTM) recurrent layers, and fully connected (FC) layers with two skipped connections. Results of reference [25] illustrate that the CLDNN-based VC can further improve the intelligibility and naturalness of EL speech.

However, the studies about enhancement of Mandarin EL speech are still insufficient. Mandarin speech is a kind of typical tonal language which is different from English. The complicated tone variation is one characteristic of the continuous Mandarin speech. Even if using the F0-controlled electro-larynx based on the capacity touch technology, the errors of tone cannot be avoided absolutely, because the moving finger of speakers are not flexible enough on the touch

This study was supported by the Beijing Natural Science Foundation (No. 4194079), China; Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No. VRLAB2018B06), China; and partly supported by JST CREST (No. JPMJCR19A3), Japan.

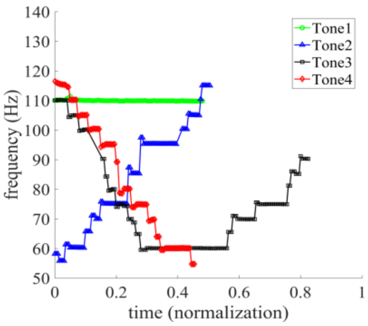

Fig. 1 Four Tones of Mandarin EL Speech (pronunciation of monosyllable “ma”) by MTC. The green line represents the Tone1; the blue line represents Tone2; the black line represents Tone3; the red line represents the Tone4.

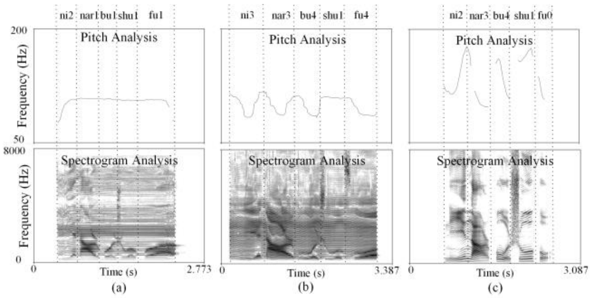

Fig. 2 Comparison of acoustic characteristic analysis for continuous Mandarin EL speech and normal speech. (a) EL-FT; (b) EL-VT; (c) normal speech. The top shows the pitch analysis, and the bottom shows the spectrogram analysis.

board. However, tone plays an important role in the Mandarin speech, or even influences the understanding of the semantic information from the Mandarin speech. Therefore, the tone errors badly affect the intelligibility and naturalness of the Mandarin EL speech.

In this paper, we propose the combined framework including the statistical VC and the manual tone control (MTC) to address the problems that the tone errors occur in the continuous Mandarin EL speech. The contribution of this paper includes 1) the combined framework is proposed to enhance the continuous Mandarin EL speech 2) the proposed combination yields significant improvements in both naturalness and intelligibility of the enhanced Mandarin EL speech; 3) a comparison of GMM-based VC and CLDNN-based VC is given to evaluate which method can better enhance the continuous Mandarin EL speech.

II. MTC for Continuous Mandarin EL speech

Figure 1 shows the MTC methodology for pronouncing the Mandarin EL speech with four tones. A speaker can achieve tone variation of continuous Mandarin EL speech by moving finger on the touch board of electro-larynx. Moving finger on the touch board can conveniently change the vibration frequency of the electro-larynx. The results of reference [11] illustrate that the Mandarin EL speech pronounced by handhold electro-larynx based on touch capacity technique performs excellent on single Chinese character and Chinese words. The intelligibility of Mandarin EL speech with variable tone (EL-VT, pronounced by touch-capacity electro-larynx) is higher than that of Mandarin EL speech with fixed tone (EL-FT, pronounced by finger-pressed electro-larynx). Similarly, the speaker can also pronounce the continuous Mandarin EL speech using this kind of electro-larynx by MTC. However, the tone errors would occur in the continuous Mandarin EL speech, if speaker is not very familiar with this kind of electro-larynx, due to the tone variation of Mandarin speech has a great complicated rule.

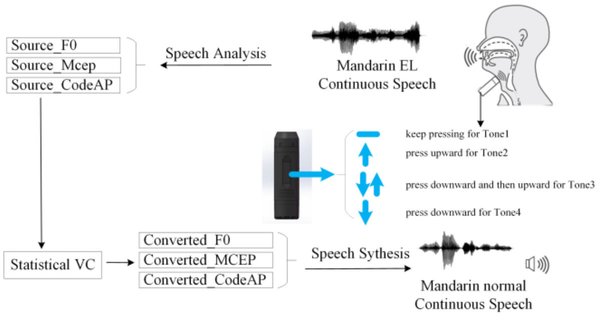

Fig. 3 An overview of the combined framework including the MTC and statistical VC. This framework includes two parts: the pronunciation part and the speech enhancement part. The pronunciation part is responsible to generate the continuous Mandarin EL speech with four tones. The enhancement part is responsible to further improve the intelligibility and naturalness of the continuous Mandarin EL speech.

An example of acoustic characteristics of the continuous Mandarin EL speech is shown in Fig. 2. In EL-FT as shown in Fig. 2 (a), the pitch increases at the beginning of an utterance, it keeps horizontal straight, and then, it decreases at the end. This pitch pattern is usually generated by pressing the button at the beginning of EL speech production and releasing it at the end. Consequently, this EL-FT speech sample sounds "ni2 nar1 bu1 shu1 fu1." On the other hand, in EL-VT as shown in Fig. 2 (b), a more natural pitch pattern is generated by MTC. Consequently, this EL-VT speech sample sounds "ni3 nar3 bu4 shu1 fu4", which is close to "ni2 nar3 bu4 shu1 fu0" of a natural Mandarin speech sample as shown in Fig. 2 (c), Note that a few tone errors are still observed in this EL-FT speech sample. This kind of tone errors easily occur in the continuous Mandarin EL speech due to complicated tone variations of continuous Mandarin speech, which are difficult to correctly produce by MTC. In addition, the first syllable of this utterance should be changed Tone2, because two syllables with Tone3 are right next to each other. The EL speech has no zero tone, therefore, in this paper we would not discuss the Tone0.

III. Proposed Framework Combining MTC and Statistical VC

A. An Overall of the Combined Framework

The proposed framework consists of the MTC and statistical VC shown in Fig. 3. This combination is different from the conventional methodology that the statistical VC is applied to enhance the intelligibility and naturalness of EL-FT. The EL-VT generated using the handhold electro-larynx by MTC can extremely decrease the difficulty of the VC task, because the difference between EL-VT and normal speech is much smaller than the difference between El-FT and normal speech.

B. GMM-based method

During the training process of GMM-based VC, the *t*-th frame of the input segmental feature can be represented as X_t; the *2D*-dimensional static and delta feature of the normal speech can be represented as $\mathrm{Y}_t = [y_t^T, \Delta y_t^T]^T$, where the y_t^T represents the *D*-dimensional static feature, and the Δy_t^T represents the dynamic feature at frame *t*. The joint features can be represented as $Z_t = [X_t^T, Y_t^T]^T$, where the T denotes the transpose of the vector. $\mathcal{N}(Z_t; \mu_m^{(Z)}, \Sigma_m^{(Z)})$ means that the joint Z_t obeys the normal distribution, where the mean is $\mu_m^{(Z)}$ and the variance is $\Sigma_m^{(Z)}$. $\lambda^{(Z)}$ is the set including the mixture-component weight α_m, mean and variance parameters of m-th mixture component. The mean and variance of m-th mixture component. During the conversion process, the parameters of GMM is calculated by Expectation-Maximization (EM) [27] algorithm.

C. CLDNN-based method

In the proposed combined framework, the CLDNN-based VC is implemented in replace of the GMM-based VC. For the inputs of the first CNN layer, a one-dimensional feature $X = \{x_1, x_2, \ldots, x_n\}$ at frame *t* is transformed into a two-dimensional feature by concatenating several preceding and succeeding frames to capture contextual information from the spectral envelope sequence of the Mandarin EL speech, which is essential to achieve the conversion from unnatural speech features into natural speech features in Mandarin EL speech enhancement. To add the original input feature vector at frame *t* through a skipped connection, dimension reduction is performed using a linear layer with outputs from the CNN layers. During the procedure in CNN layers, the random pooling function is used to get the active features in this paper. The pooling strategy is the non-overlapping max pooling. A pooling size of 3 is used for the first layer, and no pooling is done in the second layer of CNN. Then the resulting outputs are fed into the LSTM layers. The LSTM layers are used to model the dynamic characteristics of speech parameters. Finally, the fully connected layers are used to model nonlinear mappings of speech features between Mandarin EL speech and normal speech.

During the training process, three CLDNNs are used to model the segmental feature (including MCEP and CodeAP), continuous *log* F0 and U/V symbols, separately. During the conversion process, the mel-cepstrum extracted from Mandarin EL speech is converted into U/V symbols, a continuous F0, the mel-cepstrum and the aperiodicities by separate CLDNNs. For F0, the estimated continuous F0 sequence is masked using the estimated U/V symbols. Finally, the converted acoustic features are used to generate the enhanced Mandarin EL speech.

IV. EXPERIMENTAL EVALUATIONS

In the proposed framework, the EL-FT and EL-VT are used to test the effect of enhancement. Besides, GMM-based VC and CLDNN-based VC are separately implemented in the proposed framework, and the results include GMM-FT, GMM-VT, CLDNN-FT and CLDNN-VT. The objective and subjective evaluation are used to test the performance of proposed framework. During the procedure of enhancement, the sampling frequency was set 16000 Hz, logarithm F0 (*log* F0, 1 dimension), Mel-Cepstrum Parameter (MCEP, 25 dimensions) and Code Aperiodic Parameter (CodeAP, 1 dimension) are extracted by WORLD [28]. Besides, the unvoiced/voiced (U/V) symbols calculated according to the F0 are also used to train the VC. In this paper, two conventional VC methods (GMM-based [21] and CLDNN-based [25]) are used to enhance the continuous Mandarin EL speech, separately.

A. Experiment Conditions

(1) Listeners

In this paper, 10 Chinese listeners are enrolled. The average age of them are 25, where 5 of them are females and 5 of them are males. All of the listeners have good listening ability and they are all native Chinese Mandarin speakers. Pinyin with tone is written by listener according to his listening content. And the listener gave the MOS of intelligibility and naturalness according to the audio files.

(2) Data Preparation

In this paper, 250 utterances of EL-FT, 250 utterances of EL-VT and 250 utterances of normal speech are recorded according to 250 utterances of daily speaking materials, where 200 pairs of the speech are used to train VC and 50 utterances of Mandarin EL speech (including EL-FT and EL-VT) are used for statistical tests (including objective and subjective evaluations).

(3) Objective Evaluation Conditions

During the objective test, the Mel-Cepstrum Distortion (MCD) and root mean square error (RMSE) of CodeAP (CodeAP RMSE) are used to evaluate the segmental features; correlation coefficient of F0 (F0 CC), RMSE of log F0 (log F0 RMSE) are used to evaluate the F0 pattern of enhanced speech. MCD can be calculated by (1)

$$\mathrm{MCD[dB]} = \frac{10}{ln10}\sqrt{2\sum_{d=1}^{24}(c_d - {c'}_d)^2} \tag{1}$$

where c_d is the MCEP of target speech, and ${c'}_d$ is the MCEP of predicted speech. The dimension of MCEP is set as 24 during test.

(4) Subjective Evaluation Conditions

Listeners gave MOS for the intelligibility and naturalness of the Mandarin EL speech (including EL-FT and EL-VT), GMM-based enhanced speech (including GMM-FT and GMM-VT), and CLDNN-based enhanced speech (including CLDNN-FT and CLDNN-VT) according to their listening content. Please note that the sequence list of the audio is played randomly, and the listeners did not know the content of the audio previously. Moreover, the subjective evaluation is according to the MOS standard, where 5 means excellent; 4 means good; 3 means common; 2 means poor and 1 means bad.

B. Results of Objective Evaluation

(1) Evaluation Results of Acoustic Characteristic

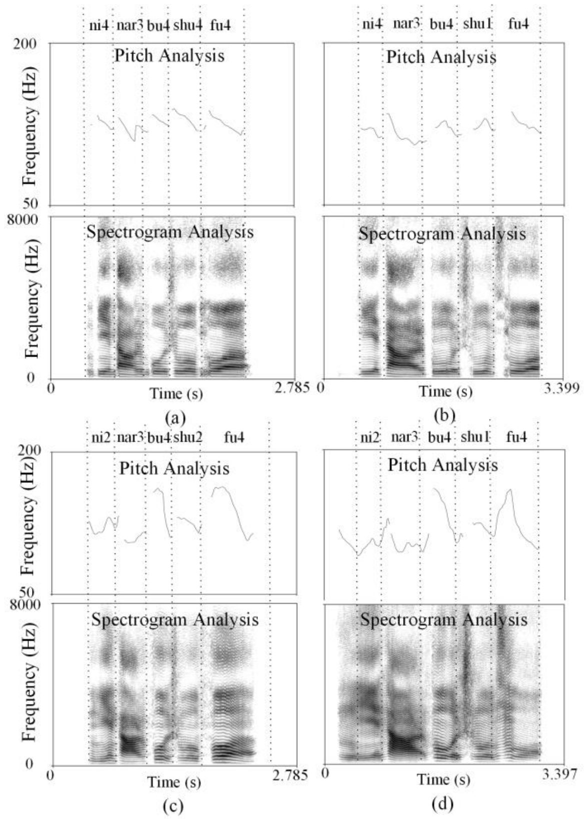

Fig. 4 Comparison of acoustic characteristic analysis for enhanced speech. (a) GMM-FT; (b) GMM-VT; (c) CLDNN-FT; (d) CLDNN-VT.

Figure 4 (a) shows the enhancement effect of EL-FT converted by GMM-based VC. All tones except of the second syllable are modified to Tone4, the tone of the second syllable is correctly modified to Tone3. Figure 4 (b) shows the results of EL-VT enhanced by GMM-based VC. The enhancement effect of GMM-VT is better than GMM-FT. The results show that more tone has been correctly modified. Figure 4 (c) shows the results of EL-FT enhanced by CLDNN. Except of the fourth syllable, the other tones are all correct. Especially, the first tone has been correctly modified to Tone2. Figure 4 (d) shows the EL-VT enhanced by CLDNN. This result shows that all tones are modified correctly. The above results illustrate that the proposed framework including CLDNN-based VC and MTC has the best performance in application of enhancing the continuous Mandarin EL speech.

(2) Evaluation Results of F0 Patterns

Table I shows that two aspects of result can be obtained. Firstly, the voiced prediction accuracy of GMM-based VC is close to that of CLDNN-based VC. Secondly, the unvoiced prediction accuracy of CLDNN-based VC is better than that of GMM-based VC.

Table II shows the log F0 RMSE and F0 CC of the converted speech. Obviously, the CLDNN-based VC performs better than GMM-based VC in enhancing the F0 of Mandarin EL speech. And the converted F0 of VT speech is better than that of FT speech. This is because the differences between EL-VT speech and normal speech are less than the differences between the EL-FT and normal speech. Combining the results in Table I and Table II, the results illustrate that the improvement would be limited by the difficulty of VC task.

Table I. U/V Analysis for continuous Mandarin EL speech based on different VC methods (%)

Method	U/V Type	U_{est}	V_{est}
GMM-FT	U_{tar}	75.93	24.07
	V_{tar}	2.64	97.36
CLDNN-FT	U_{tar}	81.94	18.06
	V_{tar}	3.99	96.01
GMM-VT	U_{tar}	78.70	21.30
	V_{tar}	3.40	96.60
CLDNN-VT	U_{tar}	80.25	19.75
	V_{tar}	3.39	96.61

Table II. Objective Evaluations of Converted F0 pattern

	F0 RMSE	F0 CC
GMM-FT	0.1779	0.6564
CLDNN-FT	0.1778	0.6783
GMM-VT	0.1569	0.7645
CLDNN-VT	0.1564	0.8040

Table III. Objective Evaluations of Segmental Features

	MCD (dB)	BAP RMSE (dB)
GMM-FT	6.7285	4.5047
CLDNN-FT	6.3702	4.5162
GMM-VT	6.2311	4.3731
CLDNN-VT	5.9153	4.3266

Table IV. The tone accuracy of different VC-based unenhanced/enhanced continuous Mandarin EL speech (%)

Speech Types	Tone1	Tone2	Tone3	Tone4	Mean	STD
EL-FT	95.44	31.37	35.61	43.14	51.39	25.78
GMM-FT	73.49	43.37	55.18	72.24	61.07	12.52
CLDNN-FT	67.27	56.19	60.07	73.13	64.17	6.53
EL-VT	88.64	68.49	57.00	87.20	75.33	13.46
GMM-VT	81.15	57.27	68.21	79.25	71.47	9.57
CLDNN-VT	78.00	70.07	78.52	85.33	77.98	5.41

(3) Evaluation Results of Segmental Features

Table III shows the MCD and CodeAP RMSE of the test utterances. For the segmental features, the CLDNN-based VC outperforms the GMM-based VC. The above results of Table I, Table II and Table III have statistical significance.

C. *Results of Subjective Evaluation*

(1) Evaluation Results of Tone Accuracy

Table IV shows the statistical result that the Tone2, Tone3 and Tone4 accuracy of EL-FT do no equal zero. Indeed, it is difficult to avoid that the listeners can correct the tone according to his listening content. The Tone1 accuracy of EL-FT decreases, while the accuracy of other tones increases. Comparing with the EL-FT, for the GMM-FT, the accuracy of Tone2 and Tone3 increases, while the accuracy of other tones decreases. In addition, the average tone accuracy of the GMM-FT increases nearby 10%. Comparing with the results of GMM-FT, the average tone accuracy of CLDNN-FT increases nearby 13%. In addition, the average tone accuracy of EL-VT is much higher than that of EL-FT. However, the average tone accuracy of GMM-VT is lower than that of EL-VT. The CLDNN-VT has the best average tone accuracy that the

Table V. The statistical WER of unenhanced/enhanced continuous Mandarin EL speech

Speech Types	WER (%)	STD
EL-FT	6.74	0.0683
GMM-FT	13.36	0.1148
CLDNN-FT	10.11	0.1076
EL-VT	11.35	0.0940
GMM-VT	9.40	0.1093
CLDNN-VT	7.69	0.0888

accuracy arrives at 77.98%, which increases above 26% comparing with the EL-FT. This result illustrates that our proposed framework has a great performance in application of enhancing the continuous Mandarin EL speech.

(2) Evaluation Results of Word Error Rate

Table V shows the WER of unenhanced/enhanced continuous Mandarin EL speech. The WER is calculated according to the comparison of listening and writing results from 10 listeners and the real content. Please note that the WER means the syllable error rate here, while it does not include the tone error. The WER of the EL-VT is higher than the WER of the EL-FT. Because the speech with wrong tone variation is more difficult understandable than the speech even with fixed tone. This is the typical characteristic of the tonal language. In addition, the statistical VC can effectively decrease the WER of the EL-VT, where the CLDNN-based VC performs better than the GMM-based VC.

(3) Evaluation Results of Intelligibility and Naturalness

Figure 5 shows the experimental results for the perceptual speech intelligibility of the test speech. Obviously, the difference between GMM-FT and CLDNN-FT is not significant. Moreover, the intelligibility of EL-FT speech is higher than the EL-VT speech because the variable tone error more easily misleads the listener. The intelligibility of GMM-based enhanced speech is bit lower than continuous Mandarin EL speech. The CLDNN-VT speech have a higher intelligibility than the EL-VT. Apparently, the proposed framework including CLDNN-based VC and MTC can effectively improve the intelligibility of continuous Mandarin EL speech.

Figure 6 shows the experimental results for the naturalness of the enhanced speech. The naturalness of EL-FT and EL-VT are close. The naturalness of CLDNN-based enhanced speech is higher than the GMM-based. The naturalness of continuous Mandarin EL speech enhanced only by MTC cannot be improved effectively. However, the proposed framework including MTC and statistical VC can improve the naturalness of EL-FT speech significantly. Moreover, the naturalness of enhanced speech with variable tone is higher than the enhanced speech with fixed tone. Another important point is what the proposed framework not only improved the tone accuracy but also decreased the WER makes the enhanced test speech more natural. Therefore, this point also plays an important role in enhancing the naturalness of the continuous Mandarin EL speech. However, the improvement for naturalness of EL-FT is limited by the tone accuracy.

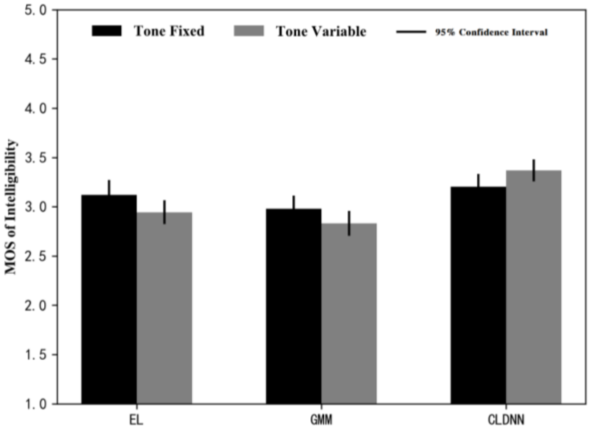

Fig. 5 The intelligibility MOS of unenhanced/enhanced continuous Mandarin EL speech. In this figure, EL represents the intelligibility MOS of continuous Mandarin EL speech. GMM represents the intelligibility MOS of GMM-based enhanced continuous Mandarin EL speech. CLDNN represents the CLDNN-based. The black histogram represents the FT speech. The grey histogram represents the VT speech. Confidence interval is 95%.

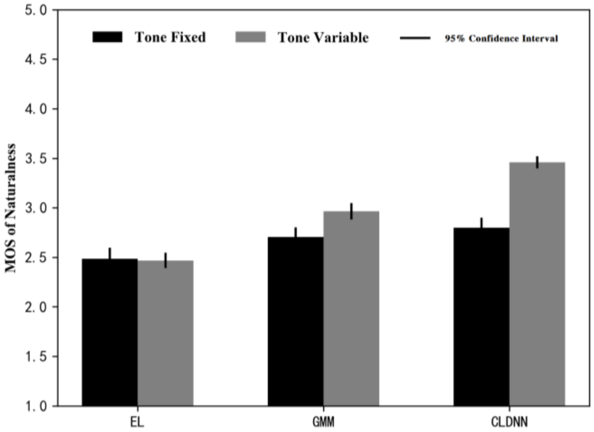

Fig. 6 The naturalness MOS of unenhanced/enhanced continuous Mandarin EL speech. In this figure, EL represents the naturalness MOS of continuous Mandarin EL speech. GMM represents the naturalness MOS of GMM-based enhanced continuous Mandarin EL speech. CLDNN represents the CLDNN-based.

Combined the objective evaluation and subjective evaluation, the results illustrate the naturalness and intelligibility of continuous Mandarin EL speech can be improved by proposed framework. Higher the performance of the method is, better the effect would be. The CLDNN-based VC performs better than GMM-based VC in enhancing the continuous Mandarin EL speech. The improvement effect of the EL-VT speech is much better than the EL-FT speech.

V. Conclusions

In this paper, a novel framework combined the MTC and statistical VC is proposed to enhance the continuous Mandarin EL speech. The objective and subjective evaluations are designed to analyze the enhancement effect. The results illustrate that the proposed framework can effectively improve the intelligibility and naturalness of the continuous Mandarin EL speech. Especially, the average of tone accuracy has been improved nearby 27%.

References

[1] H. Liu, and M. L. Ng, “Electrolarynx in voice rehabilitation,” *Auris Nasus Larynx*, vol. 34, no. 3, pp. 327-332, SEP. 2007.

[2] P. J. Watson, and R. S. Schlauch, “Fundamental frequency variation with an electrolarynx improves speech understanding: A case study, ” *American Journal of Speech-Language Pathology*, vol. 18, no. 2, pp. 162-167, MAY. 2009.

[3] L. Guo, K. F. Nagle, and J. T. Heaton, “Generating tonal distinctions in Mandarin Chinese using an electrolarynx with preprogrammed tone patterns,” *Speech Communication*, vol. 78, pp. 34-41, APR. 2016.

[4] N. Uemi, T. Ifukube, M. Takahashi, J. Matsushima, “Design of a new electrolarynx having a pitch control function,” *Proceedings of 1994 3rd IEEE International Workshop on Robot and Human Communication, IEEE*, Nagoya, Japan, JAN. 18-20, 1994, pp. 198-203.

[5] E. A. Goldstein, J. T. Heaton, J. B. Kobler, G. B. Stanley, and R. E. Hillman, “Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity,” *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 2, pp. 325-332, FEB. 2004.

[6] E. A. Goldstein, J. T. Heaton, C. E. Stepp, and R. E. Hillman, “Training Effects On Speech Production Using a Hands-Free Electromyographically Controlled Electrolarynx,” *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 2, pp. 335-351, APR. 2007.

[7] H. Takahashi, M. Nakao, T. Okusa, Y. Hatamura, Y. Kikuchi, and K. Kaga. “Pitch control with finger pressure for electrolaryngeal or intra-mouth vibrating speech,” *The Japan Journal of Logopedics and Phoniatrics*, vol. 42, no. 1, pp. 1-8, JAN. 2001.

[8] H. S. Choi, Y. J. Park, S. M. Lee, and K. M. Kim, “Functional Characteristics of a New Electrolarynx "Evada" Having a Force Sensing Resistor Sensor,” *Journal of Voice*, vol. 15, no. 4, pp. 592-599, DEC. 2001.

[9] L. Wang, Y. Feng, Z. Yang, and H. Niu, “Development and evaluation of wheel-controlled pitch-adjustable electrolarynx,” *Medical and Biological Engineering and Computing*, vol. 55, no. 8, pp. 1463-1472, AUG. 2017.

[10] C. Wan, E. Wang, L. Wu, S. Wang, and M. Wan, “Design and Evaluation of an Electrolarynx with Tonal Control Function for Mandarin,” *Folia Phoniatr Logop*, vol. 64, no. 6, pp. 290-296, 2012.

[11] W. Li, Q. Zhaopeng, F. Yijun, and N. Haijun. “Design and Preliminary Evaluation of Electrolarynx With F0 Control Based on Capacitive Touch Technology,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 3, pp. 629-636, MAR. 2018.

[12] C. Y. Espy-Wilson, V. R. Chari, J. M. MacAuslan, and M. J. Walsh, “Enhancement of electrolaryngeal speech by adaptive filtering, ” *Journal of Speech, Language, and Hearing Research*, vol. 41, no. 6, pp. 1253-1264, DEC. 1998.

[13] C. Y. Espy-Wilson, V. R. Chari, and C. B. Huang, “Enhancement of alaryngeal speech by adaptive filtering,” *Proceeding of 4 th International Conference on Spoken Language Processing. ICSLP'96. IEEE*, Philadelphia, PA, USA, 3-6 OCT. 1996, vol. 2. pp. 764-767.

[14] H. J. Niu, M. X. Wan, S. P. Wang, and H. J. Liu. “Enhancement of electrolarynx speech using adaptive noise cancelling based on independent component analysis,” *Medical and Biological Engineering and Computing*, vol. 41, no. 6, pp. 670-678, NOV. 2003.

[15] D. Cole, S. Sridharan, M. Moody, and S. Geva, “Application of noise reduction techniques for alaryngeal speech enhancement,” *TENCON'97 Brisbane-Australia. Proceedings of IEEE TENCON'97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications (Cat. No. 97CH36162), IEEE*, Brisbane, Queensland, Australia, 4-4 DEC. 1997, vol. 2, pp. 491-494.

[16] P. C. Pandey, S. M. Bhandarkar, G. K. Bachher, and P. K. Lehana, “Enhancement of alaryngeal speech using spectral subtraction,” *2002 14 th International Conference on Digital Signal Processing Proceedings. DSP 2002 (Cat. No. 02TH8628). IEEE*, Santorini, Greece, 1-3 JUL. 2002, vol. 2, pp. 591-594.

[17] H. Liu, Q. Zhao, M. Wan, and S. Wang. “Enhancement of electrolarynx speech based on auditory masking,” *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 5, pp. 865-874, APR. 2006.

[18] A. Kain, M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), IEEE*, Seattle, WA, USA, 15-15 May. 1998, vol. 1, pp. 285-288.

[19] A. B. Kain. “High resolution voice transformation,” *Ph.D. dissertation, computer science and engineering*, Oregon Health & Science University, Oregon, USA, 2001.

[20] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano. “Electrolaryngeal Speech Enhancement Based on Statistical Voice Conversion,” *10th Annual Conference of the International Speech Communication Association (Interspeech 2009 – Eurospeech)*, Brighton, United Kingdom, SEP 6-10, 2009 pp.1431–1434

[21] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “Speaking-Aid Systems Using Gmm-Based Voice Conversion for Electrolaryngeal Speech,” *Speech Communication*, vol. 54, no.1, pp.134-146, JUL, 2012

[22] T. Moriguchi, T. Toda, M. Sano, H. Sato, G. Neubig, S. Sakti, and S. Nakamura, “A Digital Signal Processor Implementation of Silent/Electrolaryngeal Speech Enhancement based on Real-Time Statistical Voice Conversion,” *14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, Lyon, France, AUG 25-29, 2013, pp. 3072-3076

[23] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, “Alaryngeal Speech Enhancement Based on One-to-Many Eigenvoice Conversion,” *IEEE/ACM Transactions on Audio Speech & Language Processing*, vol. 22, no. 1, pp. 172-183, OCT, 2014

[24] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. “A Hybrid Approach to Electrolaryngeal Speech Enhancement Based On Noise Reduction and Statistical Excitation Generation,” *IEICE Transactions on Information and Systems*, vol. 97, no. 6, pp. 1429-1437, JUN, 2014,

[25] K. Kobayashi, and T. Toda, “Electrolaryngeal Speech Enhancement with Statistical Voice Conversion based on CLDNN,” *2018 26 th European Signal Processing Conference (EUSIPCO). IEEE*, Rome, ITALY. 03-07 SEP. 2018. pp. 2115-2119.

[26] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, “Convolutional, long short-term memory, fully connected deep neural networks,” *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, Brisbane, QLD, Australia, 19-24 APR. 2015, pp. 4580-4584.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal*

of the royal statistical society, Series B (methodological), vol. 39, no. 1, pp. 1-38, 1977.

[28] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877-1884, JUL. 2016.