F0 CONTOUR ESTIMATION USING PHONETIC FEATURE IN ELECTROLARYNGEAL SPEECH ENHANCEMENT

Zexin Cai[1]*, Zhicheng Xu*[2]*, Ming Li*[1]

[1]Data Science Research Center, Duke Kunshan University, Kunshan, China
[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong

ABSTRACT

Pitch plays a significant role in understanding a tone based language like Mandarin. In this paper, we present a new method that estimates $F0$ contour for electrolaryngeal (EL) speech enhancement in Mandarin. Our system explores the usage of phonetic feature to improve the quality of EL speech. First, we train an acoustic model for EL speech and generate the phoneme posterior probabilities feature sequence for each input EL speech utterance. Then we employ the phonetic feature for $F0$ contour generation rather than the acoustic feature. The experimental results indicate that the EL speech is significantly enhanced under the adoption of the phonetic feature. Experimental results demonstrate that the proposed method achieves notable improvement regarding the intelligibility and the similarity with normal speech.

Index Terms— Electrolaryngeal Speech, Voice Conversion, Phonetic Feature, Fundamental Frequency

1. INTRODUCTION

Each year, thousands of people take laryngectomy surgeries as a treatment of laryngeal cancer. Hence, people lose the capability to produce a normal voice due to the removal of the entire larynx in that treatment. Given that voice is essential in self-expression and communication with others, the device named Electrolarynx (EL) was designed for laryngectomees to rehabilitate their voice [1, 2]. By placing the EL device against our neck as an electromechanical vibrator, EL speech is generated under the combination of electronic sound source produced by EL and the human vocal tract.

However, the EL speech does not sound like human-produced speech in several ways: 1) the sound quality degrades due to the noise caused by the continuous vibration of the EL; 2) EL speech sounds unnatural because it is generated by the mechanical excitation signals; 3) the intelligibility is limited since the EL produces monotonous speech. A study shows that EL speech can be improved by removing the EL noise and providing proper pitch information [3]. To address these issues, different approaches have been adopted [4, 5, 6, 7, 8, 9]. One of the methods is to employ voice conversion technique in EL speech enhancement [7, 8, 9]. This method can improve speech naturalness through conversion in the feature domain. Nevertheless, the generation of the fundamental frequency ($F0$) pattern is still regarded as the most challenging issue in EL speech enhancement, where the intelligibility of the converted EL speech is constrained by $F0$ contour, especially in a tone based language like Mandarin. In Mandarin, each syllable contains one of four basic tones (plus a fifth, neutral one) that make use of $F0$ to differentiate the meaning of words with the same sound pattern. Yet there are EL devices capable of modulating $F0$ with preprogrammed pitch patterns [10], and the EL speakers need further training and practice to manage the EL. Moreover, there is an inherent confusion between, for example, 'p' and 'b' (which contrast in voicing), due to the continuous sound generated by the EL.

Modeling $F0$ contour without linguistic information is difficult [11]. Unfortunately, EL speech, as well as the acoustic feature extracted from it, do not contain sufficient linguistic information since EL produces only monotonous speech. However, studies indicate that the phonetic feature extracted by the acoustic model can provide phonetic information of the speech [12, 13]. This motivates us to explore the linguistic information by adopting the phonetic feature in $F0$ contour generation under the parallel voice conversion framework. Briefly, we train an acoustic model with EL speech for phonetic posterior probabilities (PPP) feature extraction, then train a Gaussian mixture model (GMM) for $F0$ contour estimation with the joint vectors of dimension-reduced phonetic features and the ground truth $F0$ labels. The objective and subjective evaluations show that the phonetic feature outperforms the acoustic feature in predicting $F0$ contour for EL speech enhancement.

This paper is organized as follows. Section 2 describes the parallel conversion system framework for EL speech enhancement. Section 3 presents the experimental set up in this study. Finally, section 4 presents our evaluation methods and the statistical results. Conclusions are provided in section 5.

2. PROPOSED CONVERSION FRAMEWORK

Our proposed parallel conversion framework is shown in figure 1. In the training phase, three models are trained with features extracted from the EL speech and the parallel normal

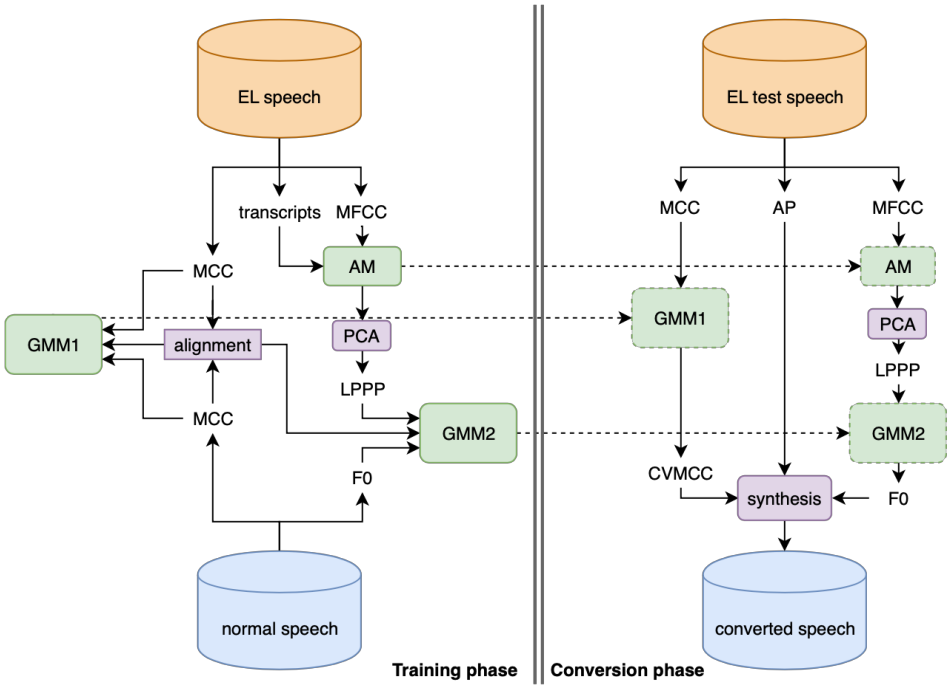

Fig. 1. The proposed parallel conversion framework

speech respectively. These three models are used to obtain the features associated with the synthesis phase in voice conversion. GMM1 is trained for obtaining the converted Mel Cepstral Coefficients (CVMCC), while the acoustic model (AM) and GMM2 are trained for generating the $F0$ contour.

2.1. Training Phase

MCC and Mel-frequency cepstral coefficient (MFCC) are extracted from EL speech after noise reduction. Similarly, MCC and $F0$ are extracted from parallel normal speech. We employ Dynamic Time Warping (DTW) to automatically align segmental frames between EL speech and the normal speech. Models were trained in association with these features and alignments.

2.1.1. GMM Training

The joint vector of time t in one utterance can be represented by $[X_t^\top, Y_t^\top]^\top$, where $X_t = [x_t{}^\top, \Delta x_t{}^\top]^\top$ and $Y_t = [y_t{}^\top, \Delta y_t{}^\top]^\top$. x denotes the feature sequence of the EL speech, y denotes the feature sequence of the normal speech, Δ denotes the dynamic feature and $\top$ denotes transposition of the vector. Conventionally, the GMM is trained to model the probability densities of the joint vectors as follows:

$$P(X_t, Y_t|\lambda) = \sum_{m=1}^{M} w_m \mathcal{N}([X_t^\top, Y_t^\top]^\top; \mu_m^{(X,Y)}, \Sigma_m^{(X,Y)}) \tag{1}$$

$$\mu_m^{(X,Y)} = \begin{bmatrix} \mu_m^{(X)} \\ \mu_m^{(Y)} \end{bmatrix}, \Sigma_m^{(X,Y)} = \begin{bmatrix} \Sigma_m^{(XX)} & \Sigma_m^{(XY)} \\ \Sigma_m^{(YX)} & \Sigma_m^{(YY)} \end{bmatrix} \tag{2}$$

where $\mathcal{N}(\cdot; \mu, \Sigma)$ denotes a Gaussian distribution with the mean vector μ and the covariance matrix Σ, w denotes the weight vector of all components in parameter set λ, $\mu_m^{(X)}$ and $\mu_m^{(Y)}$ are the mean vector of the m_{th} mixture component. The matrices $\Sigma_m^{(XX)}$, $\Sigma_m^{(XY)}$, $\Sigma_m^{(YX)}$ and $\Sigma_m^{(YY)}$ are the covariance matrices and the cross-covariance matrices of the m_{th} mixtrue component regarding the EL source feature and that for the target normal feature [14].

2.1.2. Acoustic model training

The phonetic acoustic model is trained for recognizing the phone sequence given the input feature sequence of EL speech after noise reduction. However, it is used as a phonetic feature extractor in this study. We adopt conventional Hidden Markov Model and Deep Neural Networks (HMM-DNN) speech recognition training process as follows: 1) extract acoustic feature from EL speech and preprocess the phone transcription; 2) based on a Mandarin phonetic lexicon, train a mono-phone HMM-GMM; 3) train a tri-phone HMM-GMM based on the label sequence aligned by the mono-phone model; 4) train the time delay neural network (TDNN) in association with the label sequence aligned by the tri-phone model. The TDNN model is regarded as the phonetic acoustic model.

The output vector of the TDNN is defined as phonetic posterior probabilities (PPP). Since the dimensionality of PPP is too high, the principal component analysis (PCA) algorithm is employed to reduce the dimensionality, which in result obtain the low-dimension phonetic posterior probabilities (LPPP).

2.2. Conversion Phase

The converted speech parameters are estimated by employing maximum likelihood estimation upon the trained GMM given the parameters of EL test speech as input:

$$\begin{aligned} \hat{y} &= \underset{y}{argmax}\ P(Y|X,\lambda) \\ &subject\ to\ Y = Wy \end{aligned} \tag{3}$$

where X is the feature vector sequence of EL test speech, Y is the output vector sequence. $\hat{y} = [\hat{y}_1^\top, \hat{y}_2^\top, \cdots, \hat{y}_T^\top]^\top$ is the converted static feature sequence that would be used to synthesize the converted speech. For example, Y is the CVMCC when employing estimation upon GMM1 but $F0$ contour upon GMM2. Here, matrix W is used for transforming the static feature vector to the joint static and dynamic feature vector. λ is the GMM parameter set that consists of weights, mean vectors and covariance matrices.

The MFCC feature vector sequences are fed into the TDNN acoustic model (AM) to obtain the PPP vectors. Then the PCA matrix is applied to reduce the dimension of PPP vectors and generate LPPP vectors. Similarly, $F0$ contour is estimated by GMM2 given the input LPPP sequences. Finally, we synthesize the CVMCC, aperiodic parameter (AP) and $F0$ to the converted speech.

3. EXPERIMENTAL SETUP

3.1. Data Description

Our database contains five hours of parallel EL speech and normal speech respectively. The EL speech and normal speech each contains 3206 mandarin utterances recorded by one Chinese female speaker. 2669 utterance pairs are used for training and 310 utterance pairs for evaluation. The sampling frequency of all speech utterances is 16kHz.

3.2. Conversion Setup

All EL speech utterances were preprocessed by Adobe Audition to remove constant hiss and crackle. The WORLD analysis and synthesis vocoder was employed [15] for feature extraction and speech synthesis, respectively. SPROCKET was employed [16] for GMM training. The window size and window shift size were 25ms and 5ms respectively. The dimension of MCC is 25. The 0_{th} dimension of MCC denotes the energy of that frame thus it will not be used in GMM training. The number of mixture Gaussian components is 64. The number of iteration in training is 100. The $F0$ values were normalized to scale 0-1 when training.

KALDI [17] was used for our phonetic acoustic model training. Our lexicon contains 38 phonemes including silence and unknown phones. The GMM-HMM training procedure in the TIMIT example of KALDI was used for EL speech training. The frameshift is also set to 5ms in MFCC extractor configuration. The 40-dimension high-resolution MFCC is used as the MFCC feature. Our DNN parameters in training are as follows: the frames subsampling factor is 1; the network has 6 TDNN layers, each with 625 components. The phone error rate (PER) on the EL speech test set is 18.01%. The PCA matrix is trained by the PPP feature vector sequence of 1000 utterances in the training set to reduce the dimensionality to 100.

In this paper, two systems were trained to generate the $F0$ contour individually:

LPPP: use joint vectors of LPPP feature vectors and $F0$ sequence of the normal speech for GMM2 training.

CVMCC: use joint vectors of CVMCC feature vectors of the EL training set and $F0$ sequence of the normal speech for GMM2 training.

4. EVALUATION AND RESULT

Figure 2 presents four spectrograms representing the EL speech, EL speech after noise reduction, converted speech of the proposed system and normal speech. We can see that vocal information remains after noise reduction. And after synthesizing the converted speech with CVMCC and the $F0$ contour generated by the proposed system, the spectrograms of converted speech looks close to the spectrograms of normal speech. The two different systems we have mentioned in 3.2 were evaluated under both objective and subjective evaluations.

4.1. Objective Evaluations

We use Mel-Cepstral Distortion (MCD) as one objective evaluation measure between the CVMCC sequence and the MCC sequence of normal speech:

$$MCD[dB] = \frac{1}{T}\sum_{t=1}^{T}\frac{10\sqrt{2\sum_{i=1}^{24}(c_i - c_i^{cov})^2}}{\ln 10} \tag{4}$$

where T denotes the sequence length, c denotes one MCC feature vector of normal speech and c^{cov} denotes the aligned CVMCC feature vector.

We employed three types of error metrics in $F0$ evaluation, which are correlation coefficient, Voicing Decision Error (VDE) and Gross Pitch Error (GPE) [18]:

$$VDE = \frac{N_{V\to U} + N_{U\to V}}{N} \times 100\% \tag{5}$$

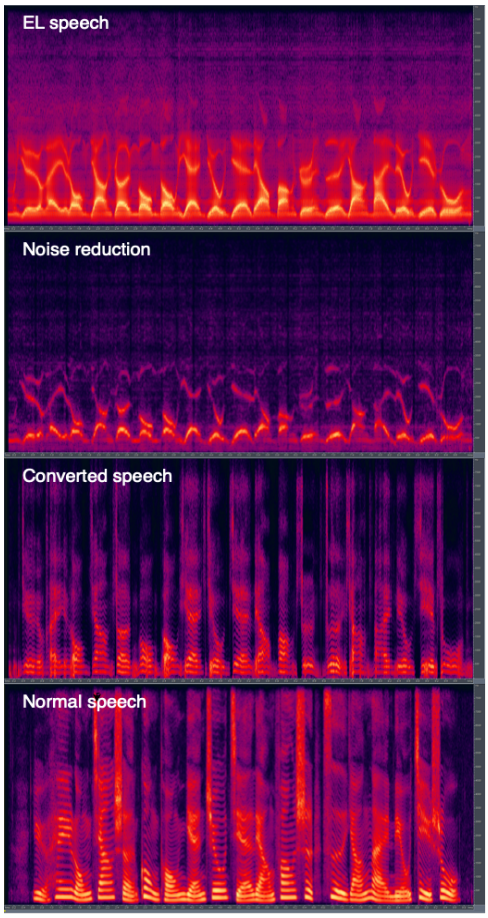

Fig. 2. Spectrograms of four speeches with same content

$$GPE = \frac{N_{F0E}}{N_{VV}} \times 100\% \quad (6)$$

where N is the number of the frames in the utterance, $N_{U \to V}$ is the number of unvoiced frames classified as voiced frames, $N_{V \to U}$ is the number of voiced frames classified as unvoiced frames, N_{VV} is the number of frames which both the predict $F0$ and the ground truth consider to be voiced and N_{F0E} denotes the number of frames for which

$$|\frac{F0_{i,estimated}}{F0_{i,reference}} - 1| > 20\% \quad (7)$$

Table 1. The objective result of the converted systems

	EL	LPPP	CVMCC
MCD	$11.56dB$	$8.0dB$	
VDE	-	0.1225	0.1219
GPE	-	**0.6514**	0.8045
$F0$ correlation coefficient	0.0088	**0.606**	0.4606

It is shown in Table 1 that using phonetic feature LPPP for generating $F0$ has better performance. The GPE of the LPPP system is lower than that in the CVMCC system. However, the LPPP system does not affect the performance in predicting voice frames and unvoiced frames since the VDE of both systems are approximately 0.122. The MCD has reduced $3.56dB$ after conversion. In the case of the $F0$ correlation coefficient, the LPPP system outperforms significantly in comparison to the CVMCC system.

4.2. Subjective Evaluations

We ask 22 subjects to score the converted speech regarding naturalness, intelligibility, and similarity. The term naturalness is used to present a score that indicates how much the speech like human speaking speech. The term intelligibility provides a score that indicates how much the speaker can understand the converted speech, and similarity denotes the sound similarity between converted speech and normal speech.

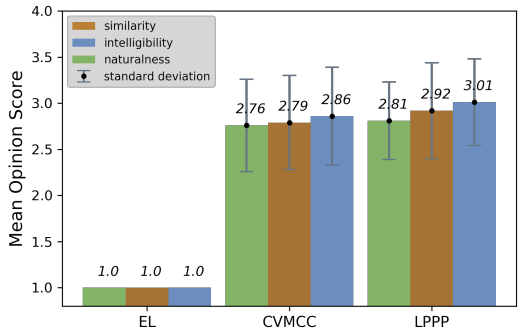

Fig. 3. The mean opinion score of subjective evaluations

The mean opinion score (MOS) [19] is shown in figure 3. It is obvious that the intelligibility and the similarity of converted speech in the LPPP system outperform those in the CVMCC system. However, LPPP yields a tiny improvement in terms of naturalness.

5. CONCLUSION

This paper presents a new conversion framework in electrolaryngeal (EL) speech enhancement. By developing a phonetic acoustic model for extracting phonetic feature that used to estimate $F0$ contour, the quality of the converted speech is further improved. The performance of the system has been evaluated through both subjective and objective measurements. The result shows that the proposed system yields significant improvement in continuous $F0$ estimation. Moreover, the converted EL speech is further enhanced when adopting the phonetic feature for $F0$ estimation rather than acoustic feature regarding the intelligibility and the similarity with the normal speech.

6. REFERENCES

[1] Hanjun Liu and Manwa L Ng, "Electrolarynx in voice rehabilitation," *Auris Nasus Larynx*, vol. 34, no. 3, pp. 327–332, 2007.

[2] Ehab A Goldstein, James T Heaton, James B Kobler, Garrett B Stanley, and Robert 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, 2004.

[3] Geoffrey S Meltzner and Robert E Hillman, "Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech," *Journal of Speech, Language, and Hearing Research*, vol. 48, no. 4, pp. 766–779, 2005.

[4] Norihiro Uemi, Tohru Ifukube, Makoto Takahashi, and Jun'ichi Matsushima, "Design of a new electrolarynx having a pitch control function," in *1994 3rd IEEE International Workshop on Robot and Human Communication*, 1994, pp. 198–203.

[5] Carol Y Espy-Wilson, Venkatesh R Chari, Joel M MacAuslan, Caroline B Huang, and Michael J Walsh, "Enhancement of electrolaryngeal speech by adaptive filtering," *Journal of Speech, Language, and Hearing Research*, vol. 41, no. 6, pp. 1253–1264, 1998.

[6] Hanjun Liu, Qin Zhao, Mingxi Wan, and Supin Wang, "Enhancement of electrolarynx speech based on auditory masking," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 5, pp. 865–874, 2006.

[7] Kou TANAKA, Tomoki TODA, Graham NEUBIG, Sakriani SAKTI, and Satoshi NAKAMURA, "A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation," *IEICE Transactions on Information and Systems*, vol. E97.D, no. 6, pp. 1429–1437, 2014.

[8] Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano, "Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[9] Ming Li, Luting Wang, Zhicheng Xu, and Danwei Cai, "Mandarin electrolaryngeal voice conversion with combination of gaussian mixture model and nonnegative matrix factorization," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2017, pp. 1360–1363.

[10] Bicky Shakya, Vishal Bharam, and Alexander Merchen, "Development of an electrolarynx capable of supporting tonal distinctions in mandarin," 2014.

[11] Ling Hui Chen, Li Juan Liu, Zhen Hua Ling, Yuan Jiang, and Li Rong Dai, "The ustc system for voice conversion challenge 2016: Neural network based approaches for spectrum, aperiodicity and f0 conversion," in *Interspeech*, 2016, pp. 1642–1646.

[12] Ming Li, Jangwon Kim, Adam Lammert, Prasanta Kumar Ghosh, Vikram Ramanarayanan, and Shrikanth Narayanan, "Speaker verification based on the fusion of speech acoustics and inverted articulatory signals," *Computer speech & language*, vol. 36, pp. 196–211, 2016.

[13] Ming Li, Lun Liu, Weicheng Cai, and Wenbo Liu, "Generalized i-vector representation with phonetic tokenizations and tandem features for both text independent and text dependent speaker verification," *Journal of Signal Processing Systems*, vol. 82, no. 2, pp. 207–215, 2016.

[14] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[15] Masanori Morise, Fumiya Yokomori, and Kenji 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, 2016.

[16] Kazuhiro Kobayashi and Tomoki Toda, "sprocket: Open-source voice conversion software," in *Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 203–210.

[17] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[18] Wei Chu and Abeer Alwan, "Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," 2009.

[19] Min Chu and Hu Peng, "Objective measure for estimating mean opinion score of synthesized speech," 2006.