Mandarin electrolaryngeal voice conversion with combination of Gaussian mixture model and non-negative matrix factorization

Conference Paper · December 2017

DOI: 10.1109/APSIPA.2017.8282244

Ming Li

Duke Kunshan University

80 PUBLICATIONS **828** CITATIONS

Mandarin Electrolaryngeal Voice Conversion with Combination of Gaussian Mixture Model and Non-negative Matrix Factorization

Ming Li*†, Luting Wang*, Zhicheng Xu*, Danwei Cai*†
* SYSU-CMU Joint Institute of Engineering, School of Electronics and Information Technology,
Sun Yat-Sen University, Guangzhou, China
† SYSU-CMU Shunde International Joint Research Institute, Guangdong, China
E-mail: liming46@mail.sysu.edu.cn Tel/Fax: +86-20-39943593

***Abstract*—Electrolarynx (EL) is a speaking-aid device that helps laryngectomees who have their larynx removed to generate voice. However, the voice generated by EL is unnatural and unintelligible due to its flat pitch and strong vibration noise. Targeting these challenges, previous works show that the electrolaryngeal speech can be enhanced using Gaussian Mixture Model (GMM) based voice conversion (VC). Although effective in improving the naturalness, it degrades the intelligibility of the converted speech. To address this issue, we propose a hybrid approach using both Non-negative Matrix Factorization (NMF) and GMM methods. For better intelligibility, we apply the NMF to estimate the high quality spectral features. For better naturalness, we use the GMM with dynamic trajectory constraint to recover a smoothed F_0. Additionally, to suppress the EL vibration noise, we include the 0^{th} MCC coefficient in the GMM-based VC. The proposed method significantly increases the F_0 dynamic range, reduces vibration noise, and improves both speech naturalness and intelligibility. One hundred pairs of the normal and electrolaryngeal speech in daily mandarin are recorded as our evaluation data. Experimental results show that our proposed hybrid method reduces the mel-cepstral distortion by 7.1 dB and increases the F_0 correlation coefficient to 0.54.**[1]

I. Introduction

Speech is one of the most convenient and important media for people to communicate with each other. However, patients undergoing laryngectomy who get their vocal folds removed still have difficulties to speak [1]. They need to use an auxiliary device to generate voice without vocal folds vibration. Electrolarynx (EL) is one of the most commonly used device for voice rehabilitation [2]. When a laryngectomee tries to speak, he holds the EL against his low jaw, and as EL vibrates, mechanical energy will be transmitted to the speaker's oral cavity and make the inside air flow. Then as tongue tip, tongue body and mouth articulate, voice is generated. Although EL is effective for laryngectomees to generate voice, there are multiple points that could be further improved.

First, EL drives vibrator with the same cycle, so the pitch of EL speech is static, which makes the generated voice unnatural. Second, the strong vibration noise generated by EL affects the speech signal to noise ratio. Third, laryngectomees lose their unique voice, and are only able to produce non-human-like voice.

To address these challenges, multiple works have been done on electrolaryngeal speech enhancement [3-7]. Among these proposed methods, there are two main approaches. One is to improve the voice quality by noise reduction [3-4], and the other one is to modify the voice's spectral parameters by statistical voice conversion (VC), such as Gaussian Mixture Model (GMM) [5-7]. The former one could not significantly improve the naturalness since it keeps F0 contour unchanged. On the other hand, although VC technique improves naturalness, it suffers degradation in intelligibility, which is a common problem for GMM-based VC [5].

In order to improve the naturalness of EL speech while minimizing intelligibility degradation, we propose a hybrid approach using both Non-negative Matrix Factorization (NMF) and GMM methods. To improve the naturalness, we adopt GMM with dynamic trajectory constraint to estimate a smoothed F_0. And to minimize the intelligibility degradation, we apply NMF to estimate the high quality spectral features. Furthermore, to suppress the EL vibration noise, we also include the 0^{th} MCC coefficient in the GMM-based VC.

We conduct both objective and subjective experiments for evaluation, and the results show that the proposed hybrid method outperforms the GMM VC baseline in both naturalness and intelligibility of the EL speech.

II. Methods

In this section, we first introduce two typical VC approaches: GMM-based and NMF-based VC. Then we propose our hybrid method building on top of these two techniques. The schematic diagram of our method is shown in Figure 3.

A. GMM-based Voice Conversion

1) Basic conversion process: GMM is one of the most widely used statistical approaches in voice conversion [8-11].

[1] This research was funded in part by the National Natural Science Foundation of China (61401524,61773413), Natural Science Foundation of Guangdong Province (2014A030313123), Natural Science Foundation of Guangzhou City (201707010363), Science and Technology Development Foundation of Guangdong Province (2017B090901045), National Key Research and Development Program (2016YFC0103905).

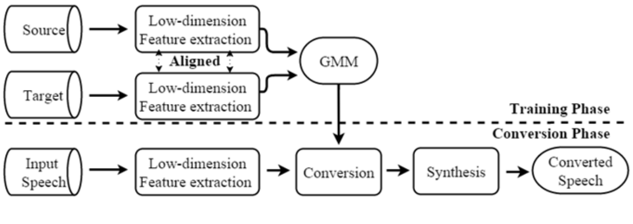

Fig. 1. The basic framework of GMM-based VC

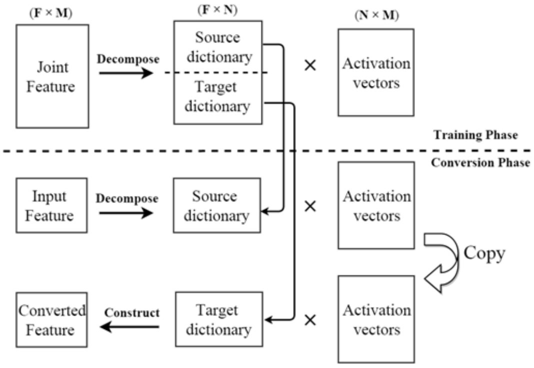

Fig. 2. The basic framework of NMF-based VC

It includes two phases: training phase and conversion phase, as shown in Figure 1.

In the training phase, two speakers utter the same sentences as the parallel training data. Then the low-dimensional feature sequence pair is time-aligned by dynamic time warping (DTW) to construct the joint vectors. After that, the joint probability density of the source and target feature vectors are trained by GMM.

In the conversion phase, given the input source feature, the converted feature is determined by maximizing the likelihood of the conditional probability density of the target feature.

2) Challenges: While GMM is good at estimating the output target feature vectors even if the input source feature vectors are not included in the training data, it discards spectral details to model the average properties of the spectra [5]. Since the estimated features tend to be over-smoothed, intelligibility degradation occurs in the converted voice. Moreover, since low-resolution feature such as mel-cepstrum coefficients (MCCs) are usually used to reduce the computational complexity in statistical VC, spectral details are further reduced.

B. NMF-based Voice Conversion

1) Basic conversion process: In recent years, exemplar-based sparse representation has attracted great interest in signal processing [11-12], and non-negative matrix factorization [13], which is based on this idea, has shown significantly success among voice conversion techniques [14]. The basic idea is to represent the spectral features as a liner combination of a set of exemplars, namely dictionary, as shown in Figure 2.

In the training phase, source and target exemplars are extracted from the aligned parallel training data. In the conversion phase, the input data is decomposed with the source dictionary and corresponding activation vectors. Then the activation vectors are copied as the target activation vectors and multiplied by the target dictionary to generate the converted data.

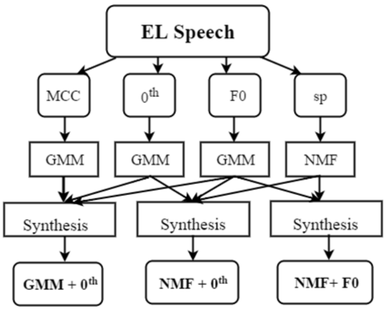

Fig. 3. The schematic diagram of the hybrid method

2) Challenges: Although NMF-based VC allows to manipulate the high-dimensional spectrum features directly without any spectral information loss, it converts the F_0 through simple normalization based on the mean and covariance of the source and target F_0 [11]. However, Chinese Mandarin is characterized as a tonal language, in which pitch matters in semantic understanding. Since EL can only produce flat pitch due to its fixed vibration frequency, which makes the Mandarin EL speech difficult to understand and sound like non-human.

C. Our Hybrid Method

1) NMF-based VC for accurate spectral parameters: To maintain the intelligibility of converted speech, we use NMF to extract the spectrum features of the EL speech. To build the source and target dictionaries, it contains the following five steps: Obtain the 24-dimensional MCCs (exclude the 0th coefficient) from the source and target speaker base on the mel-cepstral analysis [15]. Perform DTW on the MCCs and record the alignment. Extract the 513-dimensional spectrum by STRAIGHT [16]. Align the source and target spectrum based on the alignment. Apply NMF to the aligned source and target spectrum features and obtain their dictionaries.

During conversion, we decompose the input data based on the source dictionary. Since the source dictionary and target dictionary are time-aligned, we can assume that the activation vectors are approximately equivalent. So we copy the source activation vectors to the target activation vectors and multiply them with the target dictionary $X \approx WH$, where X is the converted spectrum; W is the target dictionary and H is the activation vectors shared by source dictionary and target dictionary.

2) GMM-based VC for smoothed F0 contour: In the basic NMF-based VC, the converted F_0 is just a shift and scale of the source F_0 based on its mean and variance. In EL speech, however, since the F0 is not dynamic, simple shift and normalization of F0 doesn't work. So we consider a GMM-based technique with dynamic trajectory constraint [8] for a smoothed and dynamic F_0 estimation. The training data is a joint vector of source dynamic MCCs and target dynamic F_0 ,

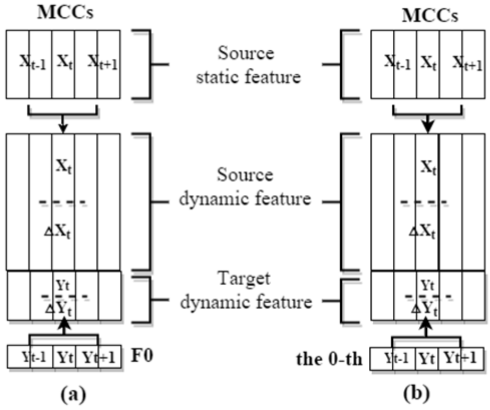

Fig. 4. The construction of features to extract F_0 and the 0^{th} MCC coefficient

as shown in Figure 4(a). Then, the joint probability density of the source and target feature parameters are trained by GMM.

During conversion, the input feature vectors are constructed in the same way, and through maximizing the likelihood of the conditional probability density of the target feature, we can estimate a varied F_0 contour. Although the proposed hybrid method of NMF for spectrum and GMM for F_0 is better than the basic GMM-based VC for MCCs and F_0, the improvement of speech intelligibility is not obvious. It is because the EL speech contains vibration noise which is resident in the spectrum. So to further improve the speech intelligibility, we proposed another technique to reduce the EL vibration noise based on VC, as described below.

3) Considering the 0^{th} MCC coefficient: The 0^{th} MCC coefficient is usually discarded in GMM-based VC because it contains the energy information and may bring adverse effect to the model [8][11]. However, in EL speech, the EL will generate vibration noise, which is captured by the 0^{th} MCC coefficient. Therefore, to correctly restore the EL speech, we cannot just adopt this coefficient of the input data as we always do in basic GMM-based VC. So we proposed to estimate the 0^{th} MCC coefficient by an additional GMM. The training data construction is shown in Figure 4(b). It is similar to F_0 extraction construction, where the only difference is that we use the target's 0^{th} MCC coefficient instead of the target F_0. As for the NMF-based VC that we propose before, we further suggest to normalize the spectrum with this coefficient. The modified NMF-based VC is evaluated in the experiments.

III. Experimental Results

A. Experiments Setup

In our experiments, the source speaker and the target speaker are the same person who is a Chinese female graduate student with 23 years old. She records one hundred mandarin daily sentences both in a normal way and in a special alaryngeal way with a EL device. In an alaryngeal way we mean that she learns to use EL to generate voice without vocal fold vibration. The EL is manufactured by Huzhou Tianchou medical machinery Co. Ltd. We conduct 10-fold cross validation tests where 90 utterance pairs are used for training while the remaining 10 utterance pairs are used for validation. The sampling frequency is set to 16kHz.

The mel-cepstral coefficients and the F0 contours are used as features in the GMM-based model, while the spectrum and the F0 contour are used in the NMF-based model. STRAIGHT analysis and synthesis method [16] is performed to extract the spectrum from the input voice and convert back to voice given a new spectrum. Mel-cepstral analysis [15] is applied to obtain the 25 mel-cepstral coefficients, while the 0th coefficient which captures the vibration noise is considered separately with other 24 mel-cepstral coefficients. The window length and shift length are set to 25 ms and 5ms separately. The number of mixture components of the GMMs used to estimate spectral parameters is 128. The size of NMF dictionary is set to 100.

$$MCD(dB) = \frac{1}{T}\Sigma_{t=1}^{T}\frac{10\sqrt{2\Sigma_{i=1}^{24}(c_i{}^{ref} - c_i{}^{cov})^2}}{ln10} \quad (1)$$

In the objective evaluation, the Mel-Cepstral Distortion (MCD) is used to measure the MCC conversion accuracy. where $c_i{}^{ref}$ and $c_i{}^{cov}$ are the i^{th} coefficient of the reference target and the converted MCCs; T is the total number of frames. To evaluate the accuracy of F_0 and the 0^{th} MCC coefficient, we use the correlation coefficient between the reference target parameters and the converted parameters.

In the subjective evaluation, 6 listeners independently evaluate the voice quality in terms of naturalness, intelligibility and similarity, using 5-point scale: 1-bad, 2-poor, 3-fair, 4-good, and 5-excellent. Seven tests are given: 1) original EL speech (EL), 2) the converted speech based on basic GMM (GMM+F0) 3) the converted speech based on GMM with considering the 0th MCC coefficient (GMM+F0+0th), 4) the converted speech based on basic NMF with simple scale normalization of F0 (NMF+F0 norm), 5) the converted speech based on NMF for spectrum and GMM for F0 (NMF+F0), 6) the converted speech based on NMF with considering the energy (NMF+F0+0th), 7) the laryngeal target speech (TG).

B. Objective results

Table I shows the results of the mel-cepstral distortion (MCD). As you can see from the table, the basic GMM-based VC can decrease the MCD of the input EL speech by 6.28 (dB). After considering the 0^{th} MCC coefficient, the MCD is further decreased by 0.82 (dB). As for the basic NMF-based VC, the MCD is bigger than that of basic GMM-based VC. It is reasonable because GMM directly takes MCCs as features while NMF takes spectrum as features [11]. Although it is incomparable in terms of MCD, you can see a trend in the NMF-based method that NMF with considering the 0th MCC coefficient is better than the basic NMF approach.

Moreover, the F_0 and the 0^{th} MCC correlation coefficients against the reference are also shown in Table I. We can see that the original F_0 correlation coefficient is 0.0035 since there is little correlation between the constant pitch of EL speech and dynamic normal speech. It shows that the traditional F_0 conversion approach with liner normalization is not quite

TABLE I
MEL-CEPSTRAL DISTORTION, F_0 CORRELATION COEFFICIENT AND $0^{th}MCC$ CORRELATION COEFFICIENT OF DIFFERENT METHODS

Method	MCD(dB)	F_0	$0^{th}MCC$
No Conversion	13.62	0.0035	0.46
Basic GMM+F0	7.34	\	\
GMM + F0+0th	6.52	\	\
Basic NMF+F0 Nor	10.49	\	\
NMF + 0th	9.11	\	\
F0 scale normalization	\	0.074	\
GMM-based F0/0th conversion	\	0.5432	0.8451

TABLE II
THE RESULT OF MEAN OPINION TEST ON NATURALNESS, INTELLIGIBILITY AND SIMILARITY

System	Naturalness	intelligibility	similarity
EL	1.42	2.75	0.75
GMM+F_0	2.47	2.61	2.7
GMM+F_0+$0^{th}MCC$	**2.69**	2.70	**3.28**
NMF+F_0norm	2.32	2.52	2.64
NMF+F_0	2.38	2.59	2.75
NMF+F_0+$0^{th}MCC$	**2.63**	**2.85**	**3.01**
Target	5	5	5

useful in EL speech since the correlation coefficient is only 0.0749, and our proposed GMM-based F_0 estimation can obtain a 0.54 correlation coefficient. As for the 0^{th} MCC, we improve the correlation coefficient from 0.46 to 0.84. This high correlation significantly helps to remove the vibration noise in the EL speech.

C. Subjective results

Table II shows the mean opinion score (MOS) for each method. Results show that all methods yield improvements in terms of naturalness compared to the original EL speech, among which the highest relative improvement is 47%.

The proposed GMM-based and NMF-based VC considering the 0th MCC coefficient (GMM+F0+0th, NMF+F0+0th) both outperform the basic GMM and NMF methods (GMM+F0, NMF+F0 norm), while GMM+F0+0th slightly outweighs NMF+ F0+0th. Also, the proposed NMF+F0 is better than the basic NMF+F0 norm, which shows that a dynamic F0 contour contributes to better naturalness. The results indicates that the proposed GMM+ F0+0th is highly effective for improving the naturalness of EL speech.

Furthermore, the proposed NMF+ F0+0th outperforms all other methods in terms of intelligibility. There is a notable intelligibility improvement when considering the 0^{th} MCC coefficient than the F_0 contour, which means that the vibration noise has more significance to EL speech intelligibility than the dynamic F0 contour. Table II also shows that all methods greatly improve the similarity than the original EL speech. The GMM-based VC is better than NMF-based VC in terms of similarity. It is also observed that both F_0 contour and the 0^{th} MCC coefficient can affect the speech similarity.

The MOS results demonstrate that the proposed GMM+F0+0th improves the naturalness and similarity but degrades the intelligibility; while the proposed NMF+F0+0th improves naturalness, intelligibility as well as similarity.

IV. CONCLUSION AND FUTURE WORKS

In this paper, we propose a hybrid method for mandarin electrolaryngeal voice conversion based on both GMM-based and NMF-based VC. The experimental results show that when using the NMF-based VC to extract high quality spectrum information, adopting the GMM-based VC to estimate a dynamic F_0 contour, and considering the energy information in the 0^{th} MCC coefficient, we can improve the naturalness as well as the intelligibility in EL speech. In the future, we will try to convert EL speech from different speakers, and to improve the accuracy of F^0 and spectrum estimation for better naturalness, intelligibility and similarity.

REFERENCES

[1] I. Hocevarboltezar, M. Zargi, "Communication after laryngectomy," *Radiology and Oncology*, vol. 35, no. 4, 2001.

[2] H. Liu, M. L. Ng, "Electrolarynx in voice rehabilitation." *Auris Nasus Larynx*, vol. 34, no. 3, 2007, pp. 327-332.

[3] H. Liu, Q. Zhao, M. Wan, S. Wang, "Application of spectral subtraction method on enhancement of electrolarynx speech," *The Journal of the Acoustical Society of America*, vol. 120, no. 1, 2006, pp. 398-406.

[4] H. Liu, Q. Zhao, M. Wan, S. Wang, "Enhancement of electrolarynx speech based on auditory masking," *IEEE Transactions on Biomedical Engineering*, vol. 53, issue. 5, 2006, pp. 865-874.

[5] K. Nakamura, T. Toda, H. Saruwatari,K. Shikano, "Electrolaryngeal speech enhancement based on statistical voice conversion," in *Proceeding of INTERSPEECH*, 2009.

[6] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "An evaluation of alaryngeal speech enhancement methods based on voice conversion techniques," in *Proceeding of ICASSP*, 2011, pp. 5136-5139.

[7] K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, 2012, pp 134-146.

[8] T. Toda, A. W. Black, K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, issue. 8, 2007, pp 2222-2235.

[9] A. Kain, M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceeding of ICASSP*, 1998, pp. 285-288.

[10] Y. Stylianou, C. Olivier, E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, 1998, pp. 131-142.

[11] Z. Wu, T. Virtanen, E. S. Chng, H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, issue. 10, 2014, pp. 1506-1521.

[12] J. F. Gemmeke, T. Virtanen, A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, issue. 7, 2011, pp. 2067-2080.

[13] D. Lee, H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, 2001, pp. 556-562.

[14] R. Aihara, T. Nakashika, T. Takiguchi, Y. Ariki, "Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 2014, pp. 7894-7898.

[15] K. Tokuda, T. Kobayashi, T. Masuko, S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in *proceeding of ICSLP*, 1994, pp. 1043-1045.

[16] H. Kawahara, I. Masuda-Katsuse, A. D. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, 1999, pp. 187-207.

[17] S. Behnke, "Discovering hierarchical speech features using convolutional non-negative matrix factorization," in *Proceedings of the International Joint Conference on Neural Networks*, 2003.