IMPROVED F0 MODELING AND GENERATION IN VOICE CONVERSION

Aki Kunikoshi[*,1,2], *Yao Qian*[1], *Frank Soong*[1], *Nobuaki Minematsu*[2]

[1]Microsoft Research Asia, Beijing, China
[2]The University of Tokyo, Japan
{kunikoshi, mine}@gavo.t.u-tokyo.ac.jp, {yaoqian, frankkps}@microsoft.com

ABSTRACT

F0 is an acoustic feature that varies largely from one speaker to another. F0 is characterized by a discontinuity in the transition between voiced and unvoiced sounds that presents an obstacle to GMM modeling for use in voice conversion. A Multi-Space Distribution (MSD) [5] can be used to model unvoiced and voiced F0 regions in a linearly weighted mixture. However, the use of two incompatible probabilistic spaces, for example a continuous probability density for voiced observations, and a discrete probability for unvoiced observations, may result in an imprecise voiced/unvoiced (v/u) conversion in a maximum likelihood (ML) sense. In this paper we propose to use voicing strength, characterized by the normalized correlation coefficient magnitude, as calculated from F0 feature extraction, as an additional feature for improving F0 modeling and the v/u decision in the context of voice conversion. The proposed method was evaluated on male-to-female voice conversion tasks in both Mandarin and English. Objective tests showed that the approach is effective in reducing the Root Mean Square Error, while the results for subjective metrics including AB preference and ABX speaker similarity tests also showed gains.

Index Terms— Voice Conversion, v/u decision model, F0 generation, Voicing Strength

1. INTRODUCTION

Voice conversion (VC) is a technique, which modifies a source speaker's speech to be perceived as if a target speaker had spoken it [2]. One of the typical spectral conversion frameworks is based on the Gaussian mixture model (GMM). It was proposed more than 10 years ago [1, 2] and has been widely applied, not only for speech, but also for various other media. [2] estimated GMM parameters with the joint vectors and generated speech in the least squares sense. [3, 4] used the minimum likelihood estimation instead of the least squares and used not only static but also dynamic feature statistics for realizing the appropriate converted spectrum sequence. These methods effectively realize a continuous mapping of the spectral parts. However, F0 sequences are usually converted by a simple linear function. One of the reasons is that F0 sequence is a piecewise continuous trajectory, no value can be observed in unvoiced region, it is not easy to treat voiced and unvoiced regions in the same framework. As a result special models have been proposed for F0 modeling in HMM-based speech synthesis. MSD-HMM models F0 with a discrete subspace for the unvoiced regions and a continuous subspace for the voiced F0 contours [5]. GTD-HMM assumes that F0 still exists in unvoiced regions and it is distributed according to an underlying globally tied continuous probability distribution field [6]. We proposed to use voicing strength as an additional feature in F0 modeling and for v/u decision in [8]. The MSD-HMM approach has already been extended to VC for simultaneously modeling spectrum and F0 features [7]. In this paper, we apply voice strength to improve F0 modeling by GMM and v/u decision in voice conversion. Experimental results demonstrate that our proposed method is more effective than conventional linear conversion. Furthermore, it works better for non-tonal language voice conversion than for tonal language voice conversion.

In Section 2, the typical framework for the GMM-based conversion is described. Section 3 describes our proposed algorithm considering voicing strength. Then, experimental evaluations are described in Section 4. Finally, this paper is summarized in Section 5.

*An intern in the Speech Group, Microsoft Research Asia

2. VOICE CONVERSION BASED ON GMM

The typical framework for the GMM-based spectral conversion [4] is as follows: Denote that the spectral features of a source speaker is $\boldsymbol{X}$ and that to a target speaker is $\boldsymbol{Y}$. Let a vector $\boldsymbol{Z}_t = [\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top$ be a joint feature vector of the source one $\boldsymbol{X}_t$ and the target one $\boldsymbol{Y}_t$ at time t. In the GMM-based voice conversion, the vector sequence $\boldsymbol{Z} = [\boldsymbol{Z}_1^\top, \boldsymbol{Z}_2^\top, ..., \boldsymbol{Z}_T^\top]$ is modeled by GMM $\boldsymbol{\lambda} = \{\omega_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i | i = 1, 2, ..., M\}$. The output probability $P(\boldsymbol{Z}|\boldsymbol{\lambda})$ can be computed as follows:

$$p(\boldsymbol{Z}|\boldsymbol{\lambda}) = \prod_{t=i}^{T} \sum_{i}^{M} \omega_i \mathcal{N}(\boldsymbol{Z}_t | \boldsymbol{\mu}_i^{(Z)}, \boldsymbol{\Sigma}_i^{(Z)}), \tag{1}$$

$$\boldsymbol{\mu}_i^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_i^{(X)} \\ \boldsymbol{\mu}_i^{(Y)} \end{bmatrix}, \boldsymbol{\Sigma}_i^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(XX)} & \boldsymbol{\Sigma}_i^{(XY)} \\ \boldsymbol{\Sigma}_i^{(YX)} & \boldsymbol{\Sigma}_i^{(YY)} \end{bmatrix}, \tag{2}$$

where M is the number of mixtures, ω_i is the mixture weight of the i-th component, $\mu_i^{(.)}$ is the mean vector and $\Sigma_i^{(.)}$ is the covariance matrix.

2.1. Maximum likelihood spectral conversion

In the maximum likelihood spectral conversion [4], the optimal sequence of the target feature vectors $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, ..., \boldsymbol{Y}_T^\top]^\top$ given a source feature vectore sequence $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, ..., \boldsymbol{X}_T^\top]^\top$ is obtained by maximizing the following conditional distribution:

$$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) = \prod_{t}^{T} \sum_{i}^{M} p(m_t = i | \boldsymbol{X}_t, \boldsymbol{\lambda}) p(\boldsymbol{Y}_t | \boldsymbol{X}_t, m_t = i, \boldsymbol{\lambda}), \tag{3}$$

where $\boldsymbol{m} = (m_1, m_2, ..., m_T)$ is a mixture index sequence. The conditional distribution given $\boldsymbol{X}$ also becomes a GMM and its output probability distribution can be written as follows:

$$p(\boldsymbol{Y}_t | \boldsymbol{X}_t, m_t = i, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{Y}_t | \boldsymbol{E}_i(t), \boldsymbol{D}_i) \tag{4}$$

and

$$\boldsymbol{E}_i(t) = \boldsymbol{\mu}_t^{(Y)} + \boldsymbol{\Sigma}_i^{(YX)}{\boldsymbol{\Sigma}_i^{(XX)}}^{-1}(\boldsymbol{X}_t - \boldsymbol{\mu}_i^{(X)}), \quad (5)$$

$$\boldsymbol{D}_i = \boldsymbol{\Sigma}_i^{(YY)} - \boldsymbol{\Sigma}_i^{(YX)}\boldsymbol{\Sigma}_i^{(XX)^{-1}}\boldsymbol{\Sigma}_i^{(XY)}. \quad (6)$$

2.2. F0 conversion

In the conventional method, F0 is converted linearly using the following equation:

$$p_t^{(Y)} = \frac{p_t^{(X)} - \mu^{(X)}}{\sigma^{(X)}} \times \sigma^{(Y)} + \mu^{(Y)}, \quad (7)$$

where $p_t^{(X)}$ and $p_t^{(Y)}$ are input and converted F0 values, respectively. $\mu^{(\cdot)}$ and $\sigma^{(\cdot)}$ are the same mean and the standard deviation of F0, respectively.

3. OUR PROPOSED METHOD FOR F0 CONVERSION

F0 is a highly variable acoustic feature. Speaker difference in F0 could be determined by a variety of factors, e.g. age, gender, dialectal background, health condition, education and personal style. However, the discontinuity of F0 between voiced and unvoiced transition has traditionally been a hurdle in building a GMM for F0 conversion. As mentioned in Section 2.2, it assumes that the F0 has a single Gaussian distribution and converted F0 has a same distribution as the target speaker in the conventional voice conversion approach. This assumption is not appropriate for F0 conversion in converting the characteristics of source speaker to target speaker. In addition, voiced/unvoiced (v/u) mismatch for some phones uttered by source and speakers are totally ignored. The v/u errors can cause degradation in the converted speech quality or intelligibility, especially for tonal language like mandarin. In [7], it proposed a simultaneous modeling of spectrum and F0 for voice conversion, where the multi-space distribution (MSD) models unvoiced region and continuous voiced F0 contour in a linearly weighted mixture. However, incompatible two probabilistic spaces, the continuous probability density for voiced observations or the discrete probability for unvoiced observations, may incur an imprecise v/u conversion in maximum likelihood (ML) sense. Recently, we propose to use voicing strength as an additional feature for F0 modeling and v/u decision in HMM-based TTS [8]. It can significantly decrease the v/u decision error in F0 generation. We extend this approach to F0 conversion. Voicing Strength (VS) is characterized by the normalized correlation coefficient (NCC) magnitude, which is calculated during F0 feature extraction on a short-time basis by applying the Robust Algorithm for Pitch Tracking (RAPT)[10]. The NCC magnitude is described in the following formula,

$$\phi_{i,k} = \frac{\sum_{j=m}^{m+n-1} s_j s_{j+k}}{\sqrt{c_m c_{m+k}}}, \quad (8)$$

where

$$c_m = \sum_{l=m}^{m+n-1} s_l^2 \quad (9)$$

and s_j is a sampled speech signal: $i = 0, 1, ..., M-1$ represents a frame index; $k = 0, 1, ..., K-1$ is the lag; n is the sample number in an analysis window; $m = iz$ and z represents the sample number in a frame.

The procedure of our approach to voice conversion is as follows: In the training phase, F0s in unvoiced regions are firstly interpolated by the spline function. The entire F0 sequence after interpolation and VS sequence extracted by eq. (8) are then smoothed with a low-pass filter. Finally, a GMM is trained with both continuous F0 features (F0 and its first order time derivatives) and VS features (NCC and its first derivatives) as well as spectral features. In other words, the source and target feature vectors, $\boldsymbol{X}$ and $\boldsymbol{Y}$ in section 2, both contain F0, VS and spectral features and these three features are simultaneously modeled by GMM. The optimal number of mixtures for the spectral part and the F0 and VS parts are calculated independently. In F0 conversion phase, both F0 and VS trajectories are firstly generated in the maximum likelihood sense. Due to the over-smoothing problem of this method, the range of generated trajectory by this method is often smaller than the original target trajectory. In order to solve this problem, Global Variance (GV) proposed by Toda and Tokuda [9] is applied for the generated F0 as follows:

$$v_d^{(w)} = \sqrt{\frac{\sigma^{(Y)}}{\sigma^{(X)}}}(v_d^{(v)} - \mu^{(X)}) + \mu^{(Y)}, \quad (10)$$

$$y_i^{(w)} = \sqrt{\frac{v_d^{(w)}}{\sigma^{(v)}}}(y_i^{(v)} - \mu^{(v)}) + \mu^{(Y)}, \quad (11)$$

where $v_d^{(w)}$ is the predicted global variance of the converted sentence, $v_d^{(v)}$ is the global variance of the generated sentence, $\mu^{(X)}$ and $\mu^{(Y)}$ are the means of source and target sentence's global variances over all training data respectively, $\sigma^{(X)}$ and $\sigma^{(Y)}$ are the variance of source and target sentence's global variances respectively. $y_i^{(w)}$ is the predicted F0 value, $\mu^{(v)}$ is the mean F0 value of generated sentence. Here, $\mu^{(Y)}$ is assumed the same to $\mu^{(w)}$, the mean of the predicted F0 value.

The generated voicing strength for each frame indicates the probability that a frame is voiced or not. Frames with larger values are more likely to be voiced. According to a preset threshold, voiced or unvoiced decisions can be made consequently. The threshold value can be fixed regardless of the source data to be the optimal value obtained by Brute force method. However appropriate threshold value of the target data is expected to depend on the source data. In order to calculate corresponded threshold value, the optimal threshold (OT) is introduced. The optimal threshold is defined for each sentence such that over 99% of voiced frames are decoded correctly from the interpolated F0 when the (voicing strength) value is larger than the optimal threshold. In these experiments the optimal threshold was calculated for all the training data and modeled using a Gaussian distribution. The optimal threshold for target speech was then estimated from the source speech using the following formula:

$$OT^Y = \frac{OT^X - \mu^X}{\sigma^X} * \sigma^Y + \mu^Y, \quad (12)$$

where $\mu^{(\cdot)}$ is the mean of OT for source/target speech and $\sigma^{(\cdot)}$ is the standard deviation of OT for source/target speech.

4. EXPERIMENTS AND RESULTS

4.1. Experimental setups

Voice Conversion experiments on Mandarin were conducted.Speech databases, recorded in Microsoft Research Asia, were used. In the database, one native male and female speaker are chosen as a source and target speaker, respectively. Every corpus consists of 100 training and 20 testing sentences. Speech signals were sampled at 16kHz, using a 25ms window with a 5ms shift. The feature set comprised

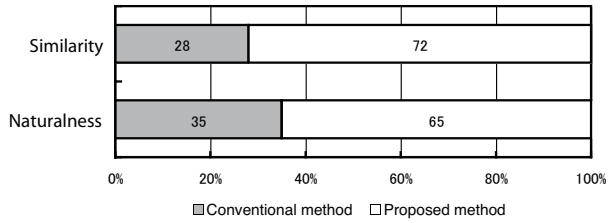

Fig. 3. Subjective comparison between the conventional method and our framework in Mandarin M2F conversion

24^{th} order LSP, log F0, VS and their first order time derivatives. The F0 and VS were extracted using the ESPS robust pitch tracking algorithm [10].

4.2. The optimal mixture number and the v/u decision threshold

To obtain the optimal mixture number and the v/u decision threshold, average correlation between predicted F0 and target F0 sequences is calculated for a grid search, shown in in Fig.1. While Root Mean Square Error (RMSE) and correlation are two common metrics for evaluating F0 model performance objectively, we use correlation as the sole criterion in the grid search since correlation is more relevant to the subjective quality of generated F0. According to the figure, the optimal number of mixtures is 32 and the optimal v/u decision threshold value is 0.76. The horizontal line indicates the values by the conventional method, linear conversion in eq. (7).

4.3. Comparison with the conventional method

The Results of objective comparison between the conventional method and our methods in Mandarin M2F conversion are shown in Fig. 2. GMM indicates our method when performed GV for F0 and set threshold value to 0.76. GMM-VS(GV) is our method when performed GV for F0 and VS then set threshold value to 0.76. GMM-VS+OT is our method when performed GV for F0 and used threshold value predicted by the source data. GMM-VS+OT(GV) is our method when performed GV for F0 and VS and used threshold value predicted by the source data. RMSE, average correlation, v $\rightarrow$ u error and u $\rightarrow$ v error between predicted F0 sequences and target F0 sequences were used for the evaluation. The objective results show that our framework performs better than the conventional method in terms of all four evaluations. In those four proposed methods, GMM shows the best result. The detailed numbers of the comparison between the conventional and our best method (GMM) for Mandarin male to female voice conversion are shown in Table 1. The effectiveness of our approach was further evaluated subjectively through two listening tests. One is an AB preference test, in which subjects select the preferred one from a pair of sentences in term of naturalness. The other test is an ABX similarity test, which measures the perceptual distance from source to target speaker. X is the original sentences from target speakers. Subjects were provided with two candidate sentences, A and B, and asked to determine which one was closest in terms of speaker similarity to the original, X. 6 subjects participated in this test. To isolate the spectral effect, we synthesize 20 sentences by using original target speaker's spectrum and converted F0s by our approach and conventional approach for AB and ABX tests. The results are shown in Fig. 3, where indicates our approach can significantly outperform the conventional approach in both synthesis naturalness and speaker similarity.

Our method is further extended to the voice conversion in non-tonal languages, such as English. CMU ARCTIC databases [11] were used in our experiments. Speaker bdl and clb are chosen as a source and a target speaker. Experimental settings are the same to 4.1. According to a grid search, 64-mixtures GMM and threshold 0.83 are the optimal values. Table 2 shows the comparison between the conventional method and our method for English Voice Conversion. For English, not all four evaluations are improved. Our method significantly improves performance evaluation matrices: RMSE, correlation and u $\rightarrow$ v error rate, while v $\rightarrow$ u error is slightly degraded. Mandarin is known as a syllabically paced tonal language. Compared with English, Mandarin has a more restricted pitch contour pattern due to its lexical meaning. The variation of F0 contour among different speakers is less than that in English. We think this is the possible reason that the objective measure improvement of our method in English is larger than Mandarin.

5. CONCLUSION

This paper presents an approach to treat voiced and unvoiced regions of F0 in the same framework for voice conversion. Voicing strength is introduced and one GMM is trained with the F0 and VS as well as spectral features. Furthermore, the improvement for non-tonal languages is bigger than that for tonal languages.

6. REFERENCES

[1] Y. Stylianou, O. Cappe and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion, *IEEE Trans. Speech Audio Proc.*, vol.6, pp.131–142, Mar. 1998.

[2] A. Kain and M.W.Macon, "Spectral voice conversion for text-to-speech synthesis," *Proc. ICASSP1998*, vol.1, pp.285-288, 1998.

[3] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a Trajectory Model by Imposing Explicit Relationships between static and Kynamic Feature Vector Sqnces," *Computer Speech & Language*, vol. 21, no. 1, pp.153–173, 2007.

[4] F T. Toda, A. W. Black and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory, *IEEE Trans. Audio Speech Language Proc.*, vol.15, pp.2222–2235, Nov. 2007.

[5] T. Masuko, K. Tokuda, N. Miyasaki and T. Kobayashi, "Pitch pattern generation using multi-space probability distrinutioin HMM," *IEICE Trans.*, vol. J83-D-II, no. 7, pp. 1600–1609, 2000.

[6] K. Yu, T. Toda, M. Gasic, S. Keizer, F. Mairesse, B. Thomson and S. Young, "Probablistic modelling of f0 in unvoiced regions in HMM based speech synthesis," *Proc. ICASSP2009*, 2009.

[7] K. Yutani, Y. Uno, Y. Nankaku, A. Lee and K. Tokuda, "Voice Conversion based on Simultaneous Modeling of Spectrum and F0, *Proc. ICASSP2009*, pp.3897–3900, 2009.

[8] Q. Zhang, F. Soong, Y. Qian, J. Pan and Y. Yan, "Improved Modeling for F0 Generation and V/U Decision in HMM-based TTS, *Proc. ICASSP2010*, pp.4606–4609, 2010

[9] T. Toda and K. Tokuda, "A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis," *IEICE Trans. Information and Systems*, E90-D(5), pp.816–824, May. 2007

[10] D. Talkin, W. Kleijn and K. Paliwal, "A robust algorithm for pitch tracking (RAPT) in Speech Coding and Synthesis," *Eds. Elsevier*, pp. 495–518, 1995

[11] CMU ARCTIC databases : http://festvox.org/cmu_arctic/

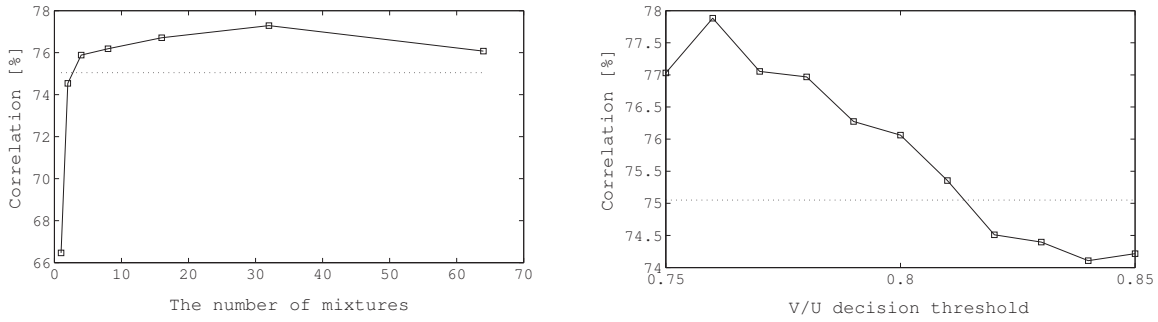

Fig. 1. The optimal number of mixtures, and v/u threshold.

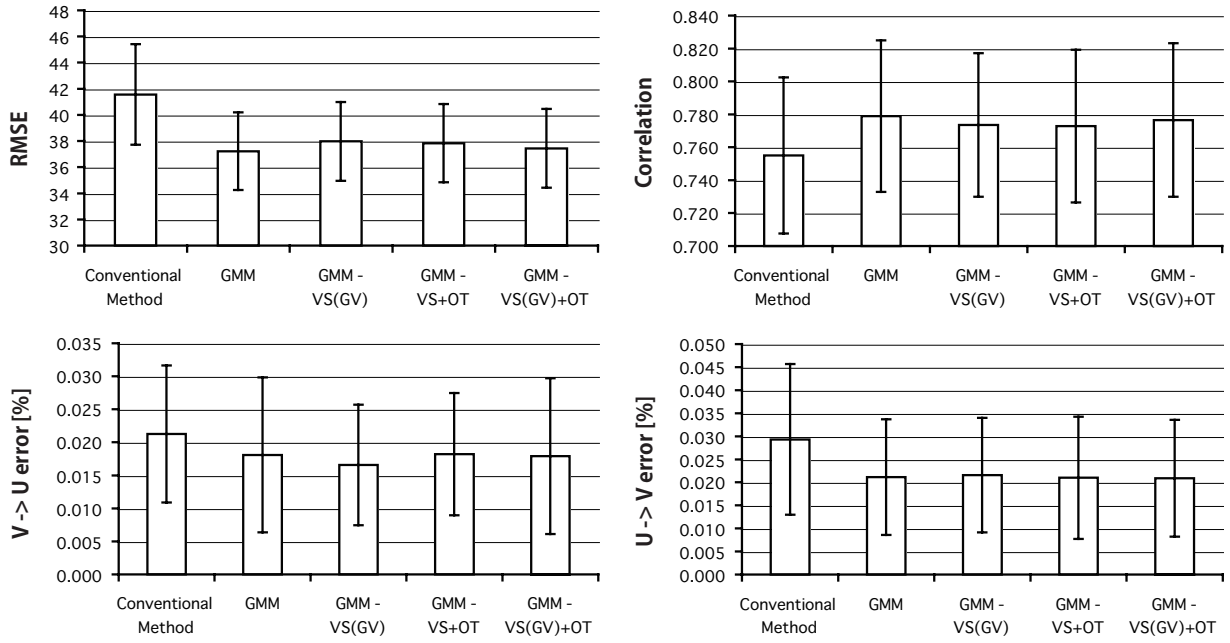

Fig. 2. Objective comparison of the conventional method and our proposed methods in Mandarin M2F conversion

Table 1. Comparison with the conventional method and our method for Mandarin VC

	RMSE	CorrCoef [%]	v → u error [%]	u → v error [%]
Conventional	41.6	75.5	2.13	2.93
Proposed	37.2	77.9	1.81	2.11
Improvement rate	10.4	3.15	14.9	27.9

Table 2. Comparison with the conventional method and our method for English VC

	RMSE	CorrCoef [%]	v → u error [%]	u → v error [%]
Conventional	14.2	75.8	1.10	3.35
Proposed	22.6	77.9	1.25	1.89
Improvement rate	18.9	2.79	-13.1	43.4