Text-Independent F0 Transformation with Non-Parallel Data for Voice Conversion

Zhi-Zheng Wu[1], *Tomi Kinnunen*[2], *Eng Siong Chng*[1], *Haizhou Li*[1,2,3]

[1]School of Computer Engineering, Nanyang Technological University, Singapore
[2]School of Computing, University of Eastern Finland, Joensuu, Finland
[3]Human Language Technology Department, Institute for Infocomm Research , Singapore
{wuzz,ASESChng}@ntu.edu.sg, tomi.kinnunen@uef.fi, hli@i2r.a-star.edu.sg

Abstract

In voice conversion, a simple frame-level mean and variance normalization is typically used for fundamental frequency (F0) transformation, which is text-independent and requires no parallel training data. Some advanced methods transform pitch contours instead, but require either parallel training data or syllabic annotations. We propose a method which retains the simplicity and text-independence of the frame-level conversion while yielding high-quality conversion. We achieve these goals by (1) introducing a text-independent tri-frame alignment method, (2) including delta features of F0 into Gaussian mixture conversion and (3) reducing the well-known GMM over-smoothing effect by F0 histogram equalization. Our objective and subjective experiments on the CMU-Arctic corpus indicate improvements over both the mean/variance normalization and the baseline GMM conversion.

Index Terms: voice conversion, F0 transformation, GMM, histogram equalization

1. Introduction

Voice conversion [8] is the task of converting one's voice(*source*) so that it sounds as if spoken by another person (*target*). Voice conversion systems operate on two independent phases, *training* and *conversion* phases. In the training phase, a conversion function between the vocal spaces of the two speakers is established by using a set of training utterances. In the conversion phase, an unseen utterance is presented to the system; the parameters of this utterance are then converted using the learned conversion function and passed to a *vocoder* which finally reconstructs an audible speech signal. For the conversion function, the *de facto* method is Gaussian mixture modeling of the joint probability of the source and target features [9].

The context of the present work is *prosody transformation* in voice conversion, in particular transformation of the fundamental frequency or F0, the acoustic correlate of the vocal folds' vibration frequency. While conversion of the spectrum has been extensively studied [9, 10], the number of F0 transformation studies in voice conversion is surprisingly small (see Table 1). The most common approach, given in the first row of Table 1, is to transform the mean and variance of the (log-)F0 distribution of the source speaker to match the target speaker's mean and variance. This is implemented by a straightforward linear transformation of the frame-level (or instantaneous) F0 values. Extensions of this approach, but still operating on instantaneous F0, include higher-order polynomial [2], GMM-based mapping [1] and piecewise linear transformation based on hand-labaled intonational target points [3].

Transformation methods for the instantaneous F0 are simple and work well for speakers with "similar" intonation. For speakers with drastically different intonation patterns, however, it might be advantageous to convert the F0 contours (*intonation contours*) instead [1,2,5,6]. In these methods, the prosodic segments (e.g. syllables or entire utterances) are represented either as variable-length sequences processed with dynamic time warping (DTW) [2] or, alternatively, by parameterizing each prosodic segment by a fixed-dimensional parameter vector [5,6] which is computationally more feasible. For an extensive objective and subjective comparison of five different F0 conversion methods, including both instantaneous and contour-based methods, refer to [1].

Even though the contour-based conversion may outperform the instantaneous conversion methods [1], care must be taken: since the intonation contour depends on both lexical factors (e.g. interrogative vs declarative sentence) and various paralinguistic factors (e.g. language and speaker's mood), it is difficult to isolate only the speaker-dependent component for conversion purposes. Consequently, if the training data and the utterance under conversion do not match in the lexical and paralinguistic attributes, the converted utterance is expected to sound unnatural. Additionally, some of the methods require syllable-level annotation, and, importantly, majority of them requires a *parallel* training corpus. That is, corpus where the source and target speaker read the same utterances. Note that this is *not* the case for the baseline mean and variance conversion method which enjoys complete text-independency. The only prosody conversion not relying on parallel data that we are aware of is [5]. In that study, the authors used syllable-level F0 and duration features in a maximum likelihood linear regression (MLLR) conversion method.

In this paper, we propose a system for F0 conversion that is completely text-independent: it requires neither parallel training data nor any phonetic or syllable-level transcriptions as hinted in Table 1. The method is thus more practical when adapting a voice conversion system to new speakers, new languages or for cross-language conversion [11]. To achieve these requirements, we combine three independent ideas. Firstly, new frame alignment method is proposed to improve frame alignment; secondly, *delta* features of F0 are incorporated with GMM-based conversion to improve naturalness; thirdly, *histogram equalization* (HEQ) is used for converting the entire F0 distribution and reducing the well-known over-smoothing prob-

The work of T. Kinnunen was supported by the Academy of Finland (project no 132129). The work of Haizhou Li was supported by a grant from Nokia foundation.

Table 1: A few approaches for F0 modification in voice conversion. The methods have been grouped according to the domain of conversion (Frame-level, local contour, utterance contour), whether they require source and target speakers to speak the same utterances in the training phase or not (parallel/non-parallel). Any additional data (in addition to the speech signal itself) is also indicated. (DCT = discrete cosine transform, CART = classification and regression tree.)

Approach	Conversion domain	Parallel data required?	Additional data
Linear conversion (mean and std) [1]	Frame-level	No	-
Polynomial conversion [1, 2]	Frame-level	Yes	-
GMM conversion [1]	Frame-level	Yes	-
Intonation marks + piecewise linear mapping [3]	Frame-level	Yes	Intonation marks
Contour codebook + DTW [1, 2]	Utterance contour	Yes	-
Weighted contour codebook [1, 4]	Local contour	Yes	-
Syllable features + MLLR adaptation [5]	Local contour	No	Syllable marks
Syllable DCT codebook + CART [6]	Local contour	Yes	Syllable marks
Multi-Space Probability Distribution HMM + $\Delta F0$ [7]	Utterance contour	Yes	-
Frame alignment + GMM (F0, ΔF0) + HEQ **[Proposed]**	Frame-level	No	-

lem [12] introduced by GMM-based conversion.

2. Baseline F0 Conversion

The simplest F0 conversion is to equalize the means and variances of the source and target F0 distributions. Denoting the F0 value of a single frame of the source speaker by x, the converted value x' is obtained as,

$$x' = \frac{\sigma_y}{\sigma_x}(x - \mu_x) + \mu_y, \tag{1}$$

where μ_x and μ_y are the means and σ_x and σ_y are the standard deviations of the training data for the source and the target speakers, respectively. This method only changes the global F0 level and dynamic range while retaining the shape of the source contour. Note that the source and target speaker distributions are modeled independently of each other. Another approach, originally developed for spectral conversion [9] but also applied for prosody conversion [1], is to model the joint distribution of the source and target feature vectors by a GMM. The conversion function is given by,

$$\mathbf{x}' = F(\mathbf{x}) = \sum_{k=1}^{K} p_k(\mathbf{x}) \cdot \left[\boldsymbol{\mu}_k^y + \boldsymbol{\Sigma}_k^{yx} (\boldsymbol{\Sigma}_k^{xx})^{-1} (\mathbf{x} - \boldsymbol{\mu}_k^x)\right], \tag{2}$$

where $p_k(\mathbf{x}) = \alpha_k \cdot N(\mathbf{x}, \boldsymbol{\mu}_k^x, \boldsymbol{\Sigma}_k^{xx}) / \sum_{l=1}^{K} \alpha_l \cdot N(\mathbf{x}, \boldsymbol{\mu}_l^x, \boldsymbol{\Sigma}_l^{xx})$ is the posterior probability of vector $\mathbf{x}$ belonging to the kth Gaussian, and $\boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_k^x \\ \boldsymbol{\mu}_k^y \end{bmatrix}$, $\boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_k^{xx} & \boldsymbol{\Sigma}_k^{xy} \\ \boldsymbol{\Sigma}_k^{yx} & \boldsymbol{\Sigma}_k^{yy} \end{bmatrix}$ are the mean vectors and covariance matrices for the kth Gaussian of the joint distribution. For this method we are required to have paired source and target training vectors. The pairing can be established via parallel training or, as in this paper, by a text-independent frame-alignment procedure.

3. Proposed F0 Conversion System

The proposed F0 conversion system (Fig. 1) consists of three independent sub-components. Firstly, we relax the requirement of parallel training data by using a text-independent frame alignment procedure. Secondly, we incorporate delta coefficients into the conversion so as to improve naturalness, and finally, we address the GMM oversmoothing problem [12] by a histogram-based post-processing technique.

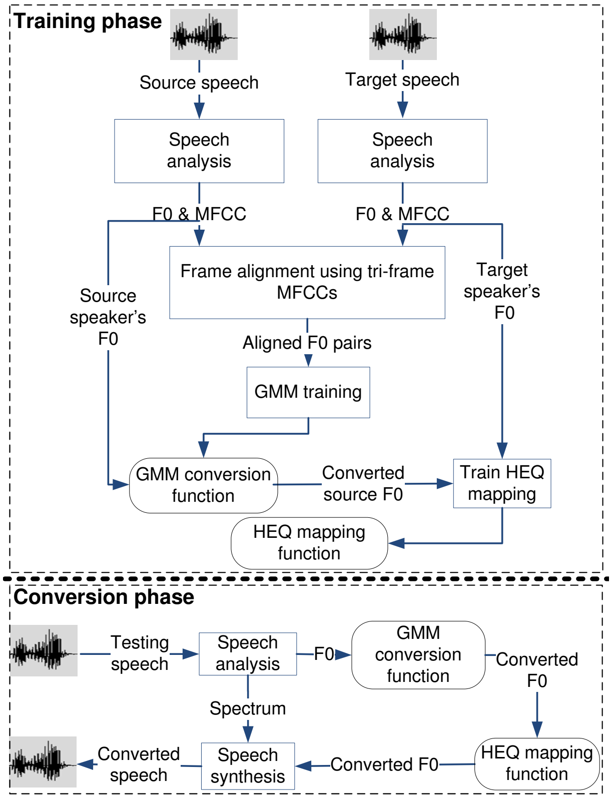

Figure 1: *Proposed text-independent F0 conversion system.*

3.1. Frame Alignment for Non-Parallel Data

The most common approaches for voice conversion rely on parallel training data for the source and the target speakers which is not feasible in some applications. Although automatic speech recognition (ASR) techniques could be used for pairing frames for non-parallel training data, this is both complex and subject to ASR errors. A text-independent frame alignment algorithm is proposed in [13] for spectral features, where the conversion function and frame alignment are jointly optimized in an iterative manner. In this research, utterance-level cepstral mean and variance normalization (CMVN) is included to normalize speaker effect. However, each mel-frequency cepstral coefficients (MFCC) vector only captures immediate local information, then by considering longer time context beyond the conventional $delta$ coefficients, the 12 dimension MFCC (without energe) vector with $delta$ is expanded with left and right acoustic context. We call the expanded vectors as *tri-frame*. The tri-frame alignment procedure is carried out only for voiced frames since F0 is undefined for unvoiced segments. A source MFCC vector is paired up with its nearest neighbor (target MFCC vector) in Euclidean distance sense.

3.2. Delta Features of F0 for Naturalness

It appears that contextual features are useful not only for robust frame alignment but also for the naturalness of the converted prosody. In the baseline GMM-based conversion (2), each frame is converted independently from each other but here we advocate the inclusion of the local time derivative features or *delta* parameters of F0. The delta features have been used for spectrum conversion [10] with excellent results and, recently, in F0 conversion as well [7]. We are, however, unaware of the approach being used in a non-parallel training scenario which is the theme of the current paper. We follow the same approach as in [10] which we shortly summarize in the following.

To use delta features, the F0 values are appended with their delta coefficients, followed by joint density GMM training as in the conventional method [9]. In the conversion phase, given the source speaker's F0 sequence appended with the deltas, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, and the joint GMM density model λ, the optimized GMM mixture sequence $\mathbf{m} = (m_1, \ldots, m_N)$ can be determined by maximizing the likelihood $p(\mathbf{X}|\mathbf{m}, \lambda)$. Having the optimized sequence, the converted F0 values are determined by maximizing the (log-)likelihood $p(\mathbf{X}'|\mathbf{X}, \mathbf{m}, \lambda)$ with respect to $\mathbf{X}'$. The solution is given by $\mathbf{X}^{'} = (\mathbf{W}^{\mathrm{T}}\mathbf{D}_m^{-1}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{D}_m^{-1}\mathbf{U}_m$, where $\mathbf{W}$ is the matrix for computing the static and delta features [10] and

$$\begin{aligned}
\mathbf{U}_m &= \big[\mathbf{U}_1(m_{k_1}), \mathbf{U}_2(m_{k_2}), \ldots, \mathbf{U}_N(m_{k_N})\big] \\
\mathbf{D}_m^{-1} &= \mathrm{diag}\big[\mathbf{D}(m_{k_1})^{-1}, \mathbf{D}(m_{k_2})^{-1}, \ldots, \mathbf{D}(m_{k_N})^{-1}\big] \\
\mathbf{U}_n(m_k) &= \boldsymbol{\mu}_k^y + \boldsymbol{\Sigma}_k^{yx}(\boldsymbol{\Sigma}_k^{xx})^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_k^x) \\
\mathbf{D}(m_k) &= \boldsymbol{\Sigma}_k^{yy} - \boldsymbol{\Sigma}_k^{yx}(\boldsymbol{\Sigma}_k^{xx})^{-1}\boldsymbol{\Sigma}_k^{xy}.
\end{aligned}$$

3.3. Postprocessing by Histogram Equalization

Originally used in image processing to automatically balance image contrast, *histogram equalization* (HEQ) is a method for converting the histogram of any random variable to match a given distribution. Due to the statistical average of GMM-based conversion, the converted F0 contours tend to be *oversmoothing* [12]. In this study, we apply HEQ to reduce oversmoothing effect. Specifically, we apply HEQ as a post-processing method after the GMM-based conversion with delta features, to equalize the converted and target F0 distribution.

For the source speaker's F0 sequence $X = \{x_1, x_2, ..., x_N\}$, we first sort X and find the minimum ($x_{\min}$) and the maximum ($x_{\max}$). The range $[x_{\min}, x_{\max}]$ is then divided in to L bins uniformly: $x_{\min} = a_1 < a_2 < \ldots < a_{L+1} = x_{\max}$ with intervals $A_i = [a_i, a_{i+1})$. Based on these bins, histogram and the corresponding cumulative distribution function (CDF) are then constructed as

$$p_x(i) = \frac{n_i}{N} \quad \text{and} \quad f_x(i) = \sum_{j=1}^{i} \frac{n_j}{N},$$

where n_i is the count of values in bin A_i. Using the same method, we find the bins B_i, the CDF $g_y(i)$ and the histogram $(B_i, g_y(i))$ from the target speaker's training data Y. With equal increments of CDF $f_x(i)$ and $g_y(i)$, a mapping (A_i, B_i) can be established. In the conversion phase, the converted F0 value after the GMM-based conversion, x, is further converted using the mapping

$$x' = \frac{b_{i+1} - b_i}{a_{i+1} - a_i}(x - a_i) + b_i, \tag{3}$$

where a_i is the nearest bin to x and b_i is the corresponding target speaker's histogram. HEQ is a nonparametric and nonlinear transformation.

4. Experiments

4.1. Experimental Setup

We conduct voice conversion experiments on the *CMU arctic corpus* [14]. Subsets of RMS, AWB (Scottish English accent) and SLT speakers are used, each subset consists of 70 utterances from which 50 are used for training and 20 for conversion. We conduct RMS to AWB (RMS→AWB) and AWB to SLT (AWB→SLT) conversion experiments. RMS→AWB is male-to-male, standard English to accented English conversion, and AWB→SLT is male-to-female, accented English to standard English conversion.

Both objective and subjective evaluation are conducted to assess the performance of the proposed approach. For the objective evaluation, Pearson correlation coefficient is used to measure similarity between target F0 contours and converted F0 contours over all voiced frames. A higher correlation shows more similar between target and converted F0 contours. Since the converted F0 contour is not time-aligned with the target F0 contour, dynamic time warping (DTW) is performed to find the frame alignment prior to correlation computation.

For the GMM-based conversion, we use $K = 4$ Gaussians and for the HEQ-based post-processing we use $L = 30$ histogram bins which were set in preliminary informal experiments. In comparison, in [6], 8 Gaussians were used for 90 training utterances.

4.2. Objective Evaluation Results

Mean/variance conversion is used as one of two baseline approaches. And results using this method for both RMS→AWB and AWB→SLT conversion are presented in table 2 as a reference.

Table 2: Results for mean/variance conversion.

	RMS→AWB	AWB→SLT
Method	correlation	correlation
Mean/var	0.638	0.584

Table 3: Results for mono-frame and tri-frame GMM conversion (no deltas).

	RMS→AWB	AWB→SLT
Method	correlation	correlation
Mono-fr. GMM	0.623	0.576
Tri-fr. GMM	0.626	0.594

Then we compare the proposed tri-frame alignment to alignment using only one frame context (mono-frame alignment). The results are given in Table 3. From the results we can see tri-frame can achieve good performance by slightly increasing the correlations for both RMS→AWB and AWB→SLT conversion. So in the rest of the experiments, we use the tri-frame based alignment. The mono-frame alignment GMM conversion is the other baseline approach.

We next study the effects of adding F0 deltas and the HEQ-based postprocessing for the GMM conversion. The results shown in Table 4 indicate the importance of delta coefficients. Regarding HEQ post-processing, it increases the correlation measure for both RMS→AWB and AWB→SLT conversion. Comparing with the two baseline approaches, using our proposed approach, tri-frame alignment GMM conversion with delta feature and HEQ post-processing, the correlation coefficients are improved from 0.638 or 0.623 to 0.655 for RMS→AWB conversion, and 0.584 or 0.576 to 0.618 for AWB→SLT conversion.

Table 4: Results for GMM with deltas and/or HEQ.

		RMS→AWB	AWB→SLT
ΔF0	HEQ	correlation	correlation
No	No	0.626	0.594
No	Yes	0.639	0.594
Yes	No	0.647	0.612
Yes	Yes	0.655	0.618

4.3. Subjective Evaluation

For the subjective evaluation part we conducted a number of ABX tests. The ABX tests are conducted as follows: we first presented the original target speech as a reference, then the subject listened to two versions of speech, A and B, which had been converted using two different methods. Subjects were asked to choose whether A or B sounded more similar to the target speech, or choose X when they could not hear difference. The order of the listening trials and the method pairs were randomized. The subjects were recruited from our colleagues and fellow students and they were naive to the given task; we did not tell ask them to pay special attention to prosody.

We compared our proposed method with two baseline approaches on RMS→AWB and AWB→SLT conversion. In each test, 10 listeners participated and 10 sentence pairs were used . From table 5 and 6, the results indicate and confirm that the proposed method achieves better F0 transformation than baseline approaches.

Table 5: Tri-frame GMM+delta+HEQ comparing with mono-frame GMM ABX test result

	GMM+Δ+HEQ	GMM	equal
RMS→AWB	55%	18%	27%
AWB→SLT	53%	17%	30%

Table 6: Tri-frame GMM+delta+HEQ comparing with mean/var conversion methods ABX test result

	GMM+Δ+HEQ	mean/var	equal
RMS→AWB	52%	17%	31%
AWB→SLT	47%	21%	32%

5. Conclusions

In this paper, we proposed a text-independent F0 transformation system which does not require neither parallel training data nor any phonetic information. Our objective evaluation indicated that including F0 deltas helps to create better mimics of the target F0 contours. The experiments also show the proposed methods achieved better performance than the conventional method.

6. Acknowledgements

The authors would like to thank Dr. Minghui Dong for useful discussion, Dr. Xiong Xiao for helpful discussion on HEQ implementation and proofreading this manuscript, Mr. Hui Liang for providing web pages for subjective listening tests and all the listeners for their helps in participating the evaluation tests.

7. References

[1] Z. Inanoglu, "Transforming pitch in a voice conversion framework," Master's thesis, St. Edmund's College, University of Cambridge, Cambridge, July 2003.

[2] D. Chappel and J. Hansen, "Speaker-specific pitch contour modeling and modification," in *ICASSP*, vol. 2, Seattle, Washington, USA, May 1998, pp. 885–888.

[3] B. Gillett and S. King, "Transforming F0 contours," in *Eurospeech*, Geneva, Sept. 2003, pp. 101–104.

[4] O. Turk and L. Arslan, "Voice conversion methods for vocal tract and pitch contour modification," in *Eurospeech*, Geneva, Sept. 2003, pp. 2845–2848.

[5] D. Lovive, N. Barbot, and O. Boeffard, "Pitch and duration transformation with non-parallel data," in *Speech Prosody 2008*, Campinas, Brazil, May 2008, pp. 111–114.

[6] E. Helander and J. Nurminen, "A novel method for prosody prediction in voice conversion," in *ICASSP*, vol. 4, Honolulu, Hawaii, USA, April 2007, pp. 509–512.

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

[8] Y. Stylianou, "Voice transformation: A survey," in *ICASSP*, Taipei, Taiwan, April 2009, pp. 3585–3588.

[9] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE T. Speech, Audio & Lang. Proc.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[10] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE T. Audio, Speech & Lang. Proc.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[11] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, and J. Hirschberg, "Text-independent cross-language voice conversion," in *ICSLP*, 2006, pp. 2262–2265.

[12] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model withdynamic frequency warping of STRAIGHT spectrum," in *ICASSP*, vol. 2, pp. 841–844.

[13] D. Erro and A. Moreno, "Frame alignment method for cross-lingual voice conversion," in *Interspeech*, Antwerp, Belgium, August 2007, pp. 1969–1972.

[14] J. Kominek and A. Black, "The CMU Arctic speech databases," in *5th ISCA Workshop on Speech Synth.*, 2004.