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 (NTU), Singapore
[2]School of Computing, University of Eastern Finland (UEF), Joensuu, Finland
[3]Human Language Technology Department, Institute for Infocomm Research (I^{2}R), Singapore

{wuzz,ASESChng}@ntu.edu.sg, tomi.kinnunen@uef.fi, hli@i2r.a-star.edu.sg

Abstract

In voice conversion, 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 model (GMM) conversion and (3) reducing the well-known GMM oversmoothing 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, text-independence

1. Introduction

Voice conversion [10] 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 reconstructs an audible speech signal. For the conversion function, the *de facto* method is Gaussian mixture modeling (GMM) of the joint probability of the source and target features [6].

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 [11, 14], 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 [1], GMM-based mapping [5] 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 [1,4,5,9]. In these methods, the prosodic segments (e.g. syllables or entire utterances) are represented either as variable-length sequences processed by dynamic time warping (DTW) [1] or, alternatively, by parameterizing each prosodic segment as a fixed-dimensional vector [4, 9] which is computationally more feasible. For an extensive objective and subjective comparison of five different F0 transformation methods, including instantaneous and contour-based methods, refer to [5].

Even though the contour-based conversion may outperform the instantaneous conversion methods [5], 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. In [9], a non-parallel training via maximum likelihood linear regression (MLLR) conversion was proposed but the method still requires syllable annotation.

In this paper, we propose a system for F0 transformation 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 for adapting a voice conversion system to new speakers and languages or for cross-language conversion [12]. To achieve these requirements, we combine three independent ideas. Firstly, a new method is proposed for improving frame alignment for non-parallel data; secondly, *delta* features of F0 are incorporated

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

Table 1: Approaches for F0 modification in voice conversion. The methods have been grouped according to the conversion domain and whether they require parallel training or any additional data (DCT = discrete cosine transform, CART = classification and regression tree, HEQ = histogram equalization).

Approach	Conversion domain	Parallel data required?	Additional data
Mean/var conversion [5]	Frame-level	No	-
Polynomial conversion [1, 5]	Frame-level	Yes	-
GMM conversion [5]	Frame-level	Yes	-
Intonation marks + piecewise linear mapping [3]	Frame-level	Yes	Intonation marks
Contour codebook + DTW [1, 5]	Utterance contour	Yes	-
Weighted contour codebook [5, 16]	Local contour	Yes	-
Syllable features + MLLR adaptation [9]	Local contour	No	Syllable marks
Syllable DCT codebook + CART [4]	Local contour	Yes	Syllable marks
Multi-space prob. distrib. HMM + $\Delta F0$ [17]	Utterance contour	Yes	-
Tri-frame GMM + ΔF0 + HEQ **[Proposed]**	Frame-level	No	-

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 problem in GMM-based conversion [15].

2. Baseline F0 Transformation

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,

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

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 [6] but also applied for prosody conversion [5], is to model the *joint distribution* of the source and target feature vectors by a GMM. The conversion function is given by,

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

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 Transformation System

The proposed F0 transformation 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 to improve naturalness, and finally, we address the GMM oversmoothing problem [15] by a histogram-based post-processing technique.

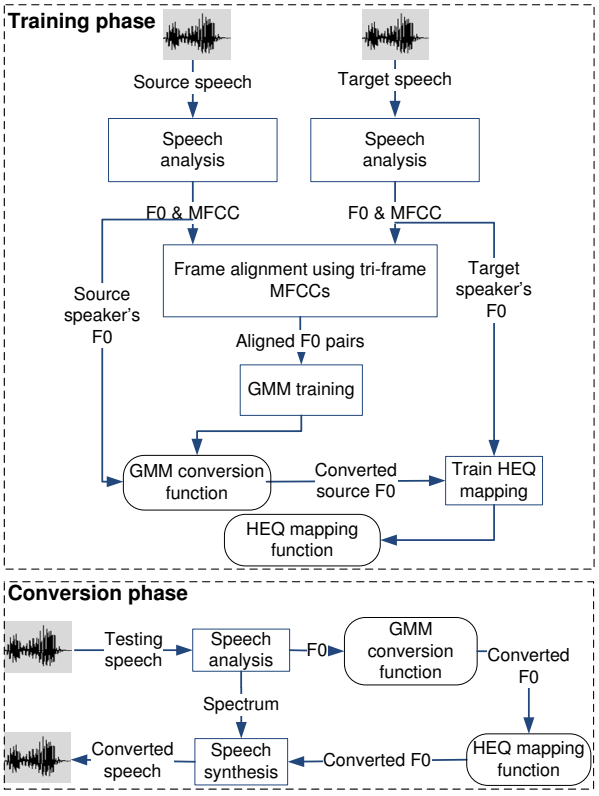

Figure 1: *Proposed text-independent F0 transformation system.*

3.1. Tri-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. In this study, we adopt a text-independent frame alignment proposed in [2] for spectral features, where the conversion function and frame alignment are jointly optimized in an iterative manner. We enhance the method by using cepstral mean and variance normalization (CMVN) for speaker normal-

ization and by using contextual information to help in alignment. Contextual features are routinely used in speech and speaker recognition applications to improve robustness, therefore we expand the cepstral vectors (12 MFCCs without energy + Δ) by their left and right acoustic contexts. We dub this method as *tri-frame* alignment. The 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 [14] with excellent results and, recently, in F0 transformation as well [17]. 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 [14] 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 [11]. 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 [14] and

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

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 averaging in GMM-based conversion, the converted F0 contours tend to be *oversmoothed* [15]. In this study, we apply HEQ to reduce oversmoothing effect. Specifically, we apply HEQ as a postprocessing method after the GMM-based conversion 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 into L bins uniformly: $x_{\min} = a_1 < a_2 < ... < a_{L+1} = x_{\max}$ 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 bin. Note that HEQ is both nonparametric and nonlinear transformation.

4. Experiments

4.1. Experimental Setup

We conduct voice conversion experiments on the *CMU Arctic corpus* [8]. 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) conversions. 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. We utilize the robust pitch tracking algorithm (RAPT) [13] and STRAIGHT [7] for speech analysis and synthesis.

Both objective and subjective evaluation are conducted to assess the performance of the proposed approach. For the objective evaluation, Pearson's correlation coefficient $(-1 \leq r \leq 1)$ is used to measure similarity of the target and the converted F0 contours over all voiced frames. High r indicates similarity of the two contours, ideal value being the maximum $r = 1$. Since the converted F0 contour is not time-aligned with the target F0 contour, dynamic time warping (DTW) alignment using MFCCs is performed 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 experiments. In comparison, $K = 8$ Gaussians were used for 90 training utterances in [4].

4.2. Objective Evaluation Results

We first compare the proposed tri-frame alignment to mono-frame alignment using only one frame context. The results in Table 2 indicate that tri-frame slightly increases the correlations for both conversions. We next study the effects of adding F0 deltas and the HEQ-based postprocessing by using the tri-frame based alignment. The results in Table 3 indicate the importance of delta coefficients. The HEQ post-processing also increases the correlation for both conversions.

Table 2: Results for mono- and tri-frame based alignment in GMM conversion (no F0 deltas).

	RMS→AWB	AWB→SLT
Alignment	correlation	correlation
Mono-frame	0.623	0.576
Tri-frame	0.626	0.594

As a summary, Table 4 contrasts the full proposed system to the two baseline methods (mean/var conversion and mono-

Table 3: 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

Table 4: Comparison of the baseline and the proposed methods.

	RMS→AWB	AWB→SLT
Method	correlation	correlation
Baseline 1: Mean/var	0.638	0.584
Baseline 2: Mono-frame GMM	0.623	0.576
Tri-frame GMM+ΔF0 + HEQ	0.655	0.618

frame GMM). The proposed system gives highest correlations. Overall, the correlations are not perfect but can be considered high, given the requirement of text-independence.

4.3. Subjective Evaluation

For the subjective evaluation part we conducted a number of ABX tests. We first presented the target utterance as a reference (X), then the subject listened to two versions of speech, A and B, which had been converted using two alternative methods. Subjects were asked to choose whether A or B sounded more similar to the target, or choose "equal" in the case (s)he could not hear any 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 compare the proposed method with the same baseline methods as in Table 4. In each test, 10 listeners participated and 10 sentence pairs were used. Tables 6 and 5 indicate and confirm that the proposed method yields better F0 transformation compared to both of the baseline methods.

Table 5: ABX results comparing tri-frame GMM+ΔF0+HEQ with mean/var conversion method.

	Tri-frame GMM +ΔF0+HEQ	Mean/var	Equal
RMS→AWB	52%	17%	31%
AWB→SLT	47%	21%	32%

Table 6: ABX results comparing tri-frame GMM+ΔF0+HEQ with mono-frame GMM.

	Tri-frame GMM +ΔF0+HEQ	Mono-frame GMM	Equal
RMS→AWB	55%	18%	27%
AWB→SLT	53%	17%	30%

5. Conclusions

We proposed a text-independent F0 transformation system which does not require parallel training data. Our objective evaluation indicated that F0 deltas helps to create better mimics of the target F0 contours. The proposed system improved both the mean/var and the baseline GMM conversion methods in both objective and subjective evaluations.

6. References

[1] D.T. Chappel and J.H.L. Hansen. Speaker-specific pitch contour modeling and modification. In *ICASSP*, volume 2, pages 885–888, Seattle, Washington, USA, May 1998.

[2] D. Erro and A. Moreno. Frame alignment method for cross-lingual voice conversion. In *Interspeech*, pages 1969–1972, Antwerp, Belgium, August 2007.

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

[4] E. Helander and J. Nurminen. A novel method for prosody prediction in voice conversion. In *ICASSP*, volume 4, pages 509–512, Honolulu, Hawaii, Apr. 2007.

[5] Z. Inanoglu. *Transforming Pitch in a Voice Conversion Framework*. Master's thesis, St. Edmund's College, University of Cambridge, Cambridge, July 2003.

[6] A. Kain and M.W. Macon. Spectral voice conversion for text-to-speech synthesis. In *ICASSP*, volume 1, 1998.

[7] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné. 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 Comm.*, 27(3):187–208, 1999.

[8] J. Kominek and A.W. Black. The CMU Arctic speech databases. In *5th ISCA Workshop on Speech Synth.*, 2004.

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

[10] Y. Stylianou. Voice transformation: A survey. In *ICASSP*, pages 3585–3588, Taipei, Taiwan, April 2009.

[11] Y. Stylianou, O. Cappé, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE T. Speech, Audio & Lang. Proc.*, 6(2):131–142, Mar. 1998.

[12] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, and J. Hirschberg. Text-independent cross-language voice conversion. In *Interspeech*, pages 2262–2265, 2006.

[13] D. Talkin. A robust algorithm for pitch tracking (RAPT). *Speech coding and synthesis*, 495:518, 1995.

[14] T. Toda, A.W. Black, and K. Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE T. Audio, Speech & Lang. Proc.*, 15(8):2222–2235, Nov. 2007.

[15] T. Toda, H. Saruwatari, and K. Shikano. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. In *ICASSP*, volume 2, pages 841–844, 2001.

[16] O. Turk and L.M. Arslan. Voice conversion methods for vocal tract and pitch contour modification. In *Eurospeech*, pages 2845–2848, Geneva, Sept. 2003.

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