F0 Transformation within the Voice Conversion Framework

Zdeněk Hanzlíček, Jindřich Matoušek

Department of Cybernetics, University of West Bohemia, Pilsen, Czech Republic

zhanzlic@kky.zcu.cz, jmatouse@kky.zcu.cz

10.21437/Interspeech.2007-549

Abstract

In this paper, several experiments on F_0 transformation within the voice conversion framework are presented. The conversion system is based on a probabilistic transformation of line spectral frequencies and residual prediction. Three probabilistic methods of instantaneous F_0 transformation are described and compared. Moreover, a new modification of inter-speaker residual prediction is proposed which utilizes the information on target F_0 directly during the determination of suitable residuum. Preference listening tests confirmed that this modification outperformed the standard version of residual prediction.

Index Terms: voice conversion, f0 transformation, residual prediction

1. Introduction

The aim of voice conversion is to transform an utterance pronounced by a source speaker so that it sounds as if it is spoken by a target speaker.

This paper is focused on the problem of F_0 transformation. The task could be divided into two virtually independent parts. Firstly, the target F_0 trajectory is to be obtained. Secondly, the reconstructed speech has to follow this F_0 trajectory.

Two basic approaches to F_0 transformation exist. The first converts the instantaneous F_0 frame by frame (e.g [1]), the other describes and converts the whole F_0 trajectory (e.g. [2] or [3]). An ample comparison of both approaches is presented in [4]. This study concerns only a specific sort of F_0 transformation functions – frame by frame F_0 conversion based on a probabilistic description.

Our baseline voice conversion system (see [5] and [6]) utilizes the pitch-synchronous linear prediction (LP) analysis, each speech frame is two pitch long with one pitch overlap. LP parameters are represented by their line spectral frequencies (LSFs) which are converted by employing a probabilistic function (e.g. [7] or [8]). Residual signal is represented by its amplitude and phase FFT-spectra which are transformed by using residual prediction (e.g. [8] or [9]). The reconstruction of speech is performed by a simple OLA method.

In this study, a new modification of inter-speaker residual prediction is proposed which utilizes the information on target F_0 directly during the determination of a suitable residuum. Thus during the reconstruction of speech, only slight modification is necessary. This version of residual prediction outperforms the standard one; this is confirmed by preference listening tests.

The conversion functions are estimated from parallel training data: LSF sequences extracted from equal utterances from source and target speaker are time-aligned by using DTW algorithm. In the following text all training and testing data are supposed to be time-aligned.

This paper is organized as follows. In Section 2, a simple method for LSF transformation using GMM is described. Section 3 deals with two simple methods for converting fundamental frequency. Section 4 gives account of combined LSF and F_0 conversion. In Section 5, a new modification of residual prediction is proposed. In Section 6, all the described methods are evaluated and compared. Finally, Section 7 concludes this paper.

2. LSF transformation

The interrelation between source and target speaker's LSFs (x and y, respectively) is described by a joint GMM with Q mixtures

$$p(x,y) = \sum_{q=1}^{Q} \alpha_q \mathcal{N}\Big\{ \begin{bmatrix} x \\ y \end{bmatrix}; \mu_q, \Sigma_q \Big\}. \quad (1)$$

All unknown parameters are estimated by employing the expectation-maximization (EM) algorithm; for initialization the binary split k-means algorithm is used. The mean vectors μ_q and covariance matrices Σ_q can be decomposed into blocks corresponding to source and target speaker's components

$$\mu_q = \begin{bmatrix} \mu_q^x \\ \mu_q^y \end{bmatrix} \qquad \Sigma_q = \begin{bmatrix} \Sigma_q^{xx} & \Sigma_q^{xy} \\ \Sigma_q^{yx} & \Sigma_q^{yy} \end{bmatrix}. \quad (2)$$

The transformation function is defined as conditional expectation of target y given source x

$$\tilde{y} = E\{y|x\} = \sum_{q=1}^{Q} p(q|x) \Big[\mu_q^y + \Sigma_q^{yx} \big(\Sigma_q^{xx}\big)^{-1} (x - \mu_q^x) \Big]. \quad (3)$$

where $p(q|x)$ is the conditional probability of mixture q given source x

$$p(q|x) = \frac{\alpha_q \mathcal{N}\{x; \mu_q^x, \Sigma_q^{xx}\}}{\sum_{i=1}^{Q} \alpha_i \mathcal{N}\{x; \mu_i^x, \Sigma_i^{xx}\}}. \quad (4)$$

The conversion function is determined for voiced and unvoiced data separately. Although unvoiced speech is unimportant for speaker identity perception, the conversion of unvoiced speech proved good on transitions between voiced and unvoiced speech.

3. F0 transformation

3.1. F0 normalization

This method is also known as Gaussian normalization or mean/variance transformation. It is is usually used as a reference method because of its simplicity and good performance. It is based on the assumption that the instantaneous F_0 from

source and target speakers (f_x and f_y, respectively) have Gaussian distribution

$$p(f_x) = \mathcal{N}\{f_x; \mu_x, \sigma_x\} \qquad p(f_y) = \mathcal{N}\{f_y; \mu_y, \sigma_y\}. \quad (5)$$

The transformation function that converts the mean and variance from values μ_x, σ_x to values μ_y, σ_y is given by

$$\tilde{f_y} = \mu_y + \frac{\sigma_y}{\sigma_x}(f_x - \mu_x). \quad (6)$$

The advantage of this method is that no parallel data are needed. The transformation function can be obtained from arbitrary (representative) speech data from both speakers.

3.2. Simple F0 expectation

The disadvantage of the previous method is that there is no means to exploit the information included in data parallelism.

Similarly as in the case of LSF conversion, time-aligned source and target instantaneous F_0 values are described with a joint GMM

$$p(f_x, f_y) = \sum_{q=1}^{Q} \alpha_q \mathcal{N}\left\{ \begin{bmatrix} f_x \\ f_y \end{bmatrix}; \begin{bmatrix} \mu_q^x \\ \mu_q^y \end{bmatrix}, \begin{bmatrix} \sigma_q^{xx} & \sigma_q^{xy} \\ \sigma_q^{yx} & \sigma_q^{yy} \end{bmatrix} \right\}. \quad (7)$$

The converted fundamental frequency $\tilde{f_y}$ is given as the conditional expectation of target f_y given source f_x

$$\tilde{f_y} = E\{f_y | f_x\} = \sum_{q=1}^{Q} p(q|f_x) \Big[\mu_q^y + \frac{\sigma_q^{yx}}{\sigma_q^{xx}}(f_x - \mu_q^x) \Big]. \quad (8)$$

4. Combined F0 & LSFs transformation

This method was already introduced in [1]; however, the implemented system employed the Harmonic plus Noise Model of speech production.

A possible interdependency between LSF and F_0 is exploited here – they are converted together using one transformation function. Formally, new variables can be introduced

$$\chi = \begin{bmatrix} x \\ f_x \end{bmatrix} \qquad \psi = \begin{bmatrix} y \\ f_y \end{bmatrix}. \quad (9)$$

Again, the joint distribution of χ and ψ is estimated using EM algorithm

$$p(\chi, \psi) = \sum_{q=1}^{Q} \alpha_q \mathcal{N}\left\{ \begin{bmatrix} \chi \\ \psi \end{bmatrix}; \mu_q, \Sigma_q \right\} \quad (10)$$

and the conversion function is defined as the conditional expectation

$$\tilde{\psi} = E\{\psi | \chi\}. \quad (11)$$

However, the simple composition of LSFs and fundamental frequency in Eq. (9) is unsuitable because the importance of particular components is not well-balanced. It leads to bad initialization. In [1], the fundamental frequency was normalized and logarithmized. In our experiments we found out that a good solution to this problem can be obtained in the following way

$$\chi = \begin{bmatrix} 100 \cdot x \\ f_x \end{bmatrix} \qquad \psi = \begin{bmatrix} 100 \cdot y \\ f_y \end{bmatrix} \quad (12)$$

Of course, to obtain the proper results, all transformation formulas have to be consistently modified.

5. Residual prediction

Residual prediction is a technique which allows the estimation of the suitable residuum for given LPC or similar parameters. Traditionally, the residual prediction is based on probabilistic description of cepstral parameter space – with a GMM. For each mixture of this model, a typical amplitude and phase residual spectrum is calculated. For more details see e.g. [9] or [8].

In [5] and [6] a new approach to residual prediction – so-called inter-speaker residual prediction was introduced. It is briefly described in the following subsection. In the second subsection, a new modification of this approach is proposed which utilizes the information on the desired (target) F_0. Due to the better determination of the residual signal, a smaller signal modification is necessary during the speech reconstruction stage.

5.1. Simple inter-speaker residual prediction

Within inter-speaker residual prediction, the residuum of the target speaker is estimated from the source speaker's parameters. Consistently recorded utterances and correct time-alignment are presupposed.

In our experiments on residual prediction, LSFs slightly outperformed cepstral parameters, thus they are employed in our experiments.

A non-probabilistic description of source LSF space is used, the LSFs are clustered into Q ($Q \approx 20$) classes by employing the bisective k-means algorithm; each class q is represented by its LSF centroid $\bar{x}_q$. The pertinence of parameter vector x_n to class q is expressed by

$$w(q|x_n) = \frac{\left[(\bar{x}_q - x_n)^\top (\bar{x}_q - x_n)\right]^{-1}}{\sum_{i=1}^{Q} \left[(\bar{x}_i - x_n)^\top (\bar{x}_i - x_n)\right]^{-1}} \quad (13)$$

For each parameter class q, the typical residual amplitude spectrum $\hat{r}_q$ is calculated as a weighted average over all training data

$$\hat{r}_q = \frac{\sum_{n=1}^{N} r_n w(q|x_n)}{\sum_{n=1}^{N} w(q|x_n)} \quad (14)$$

and the typical residual phase spectrum $\hat{\varphi}_q$ is selected

$$\hat{\varphi}_q = \varphi_{n^*} \qquad n^* = \arg\max_{n=1\ldots N} w(q|x_n). \quad (15)$$

In order to calculate the amplitude spectrum, all spectra have to be resampled (interpolated) to have the same length. Cubic spline interpolation is used and the target length is given as the average of all residua lengths. Due to consistency, phase spectra have to be interpolated to the same length; nearest neighbour interpolation is employed.

In the transformation stage, the residual amplitude spectrum $\tilde{r}_n$ is calculated as the weighted average over all classes

$$\tilde{r}_n = \sum_{q=1}^{Q} \hat{r}_q w(q|x_n) \quad (16)$$

and the residual phase spectrum is selected from parameter class q^* with the highest weight $w(q|x_n)$

$$\tilde{\varphi}_n = \hat{\varphi}_{q^*} \qquad q^* = \arg\max_{q=1\ldots Q} w(q|x_n). \quad (17)$$

The determined amplitude and phase FFT-spectra have to be resampled (interpolated) to the length which corresponds to the desired instantaneous F_0. Then, by using the inverse FFT

algorithm and by filtering with converted LPC filter, a new two-pitch speech segment is obtained whose instantaneous F_0 corresponds to the desired value. Resulting speech is reconstructed by employing a simple OLA method.

5.2. Extended inter-speaker residual prediction

By employing the simple residual prediction, converted speech can suffer from artifacts which are largely caused by phase spectrum interpolation. Especially in cases of larger interpolation.

A possible solution to this problem is to store more residua which correspond to different instantaneous F_0 values. As previously, the parameter space of the source speaker is divided into Q classes. All training data are uniquely classified into these classes. For each class q, a set R_q of pertaining data indices is established

$$R_q = \big\{k; 1 \leq k \leq N \ \wedge \ d(\bar{x}_q, x_k) = \min_{i=1\ldots Q} d(\bar{x}_i, x_k)\big\}. \quad (18)$$

Within each parameter class, the data are divided into L_q subclasses according to their instantaneous F_0, the number of subclasses L_q can differ for particular parameter classes q. Each F_0 subclass is described by its centroid $\bar{f}_q^\ell$ (q-th parameter class, ℓ-th F_0 subclass) and the set of data belonging into this subclass is defined as a set R_q^ℓ of corresponding indices

$$R_q^\ell = \big\{k; \ k \in R_q \ \wedge \ d(\bar{f}_q^\ell, f_k) = \min_{i=1\ldots L_q} d(\bar{f}_q^i, f_k)\big\}. \quad (19)$$

For each F_0 subclass, a typical residual amplitude spectrum $\hat{r}_q^\ell$ is determined as the weighted average of amplitude spectra belonging into this subclass

$$\hat{r}_q^\ell = \frac{\sum_{n \in R_q^\ell} r_n w(q|x_n)}{\sum_{n \in R_q^\ell} w(q|x_n)} \quad (20)$$

and the typical residual phase spectrum $\hat{\varphi}_q^\ell$ is selected

$$\hat{\varphi}_q^\ell = \varphi_{n^*} \qquad n^* = \arg\max_{n \in R_q^\ell} w(q|x_n) \quad (21)$$

In comparison with the previous method, amplitude and phase spectra are interpolated to the average lengths within particular F_0 subclasses. Thus the amount of signal modification is significantly smaller.

During prediction, the residual amplitude spectrum $\tilde{r}_q$ is calculated as the weighted average over all classes. However, from each class q only one subclass ℓ_q is selected whose centroid $\bar{f}_q^\ell$ is the nearest to the desired fundamental frequency $\tilde{f}_n$

$$\tilde{r}_n = \sum_{q=1}^{Q} \hat{r}_q^{\ell_q} w(q|x_n) \qquad \ell_q = \arg\min_{\ell=1\ldots L_q} d(\bar{f}_q^\ell, \tilde{f}_n) \quad (22)$$

The residual phase spectrum is selected from the parameter class q^* with the highest weight $w(q|x_n)$ from the F_0 subclass ℓ^* with the nearest central frequency $\bar{f}_q^\ell$

$$\tilde{\varphi}_n = \hat{\varphi}_{q^*}^{\ell^*} \qquad \begin{aligned} q^* &= \arg\max_{q=1\ldots Q} w(q|x_n) \\ \ell^* &= \arg\min_{\ell=1\ldots L_{q^*}} d(\bar{f}_{q^*}^\ell, \tilde{f}_n) \end{aligned} \quad (23)$$

Again, the resulting amplitude and phase FFT-spectra have to be interpolated to the length given by the desired F_0 $\tilde{f}_n$. The impact of this interpolation should be less significant, because the original and target lengths are very close to each other.

6. Experiments and results

In this section, assessment and comparison of all aforementioned methods is presented. In the first subsection, mathematical evaluation of F_0 and LSF transformations are presented. The second and third subsections deal with subjective evaluation by listening tests.

6.1. Speech data

Speech data for our experiments were recorded in an anechoic chamber. Firstly, one female speaker recorded the reference utterances. It was a set of 55 quite short sentences (about 6 words long); all sentences were in the Czech language. Subsequently, 4 other speakers (2 males and 2 females) listened to these reference utterances and repeated them. In this way, better pronunciation and prosodic consistency is guaranteed. Along with the speech signal, the EGG signal was recorded to ensure more robust pitch-mark detection and F_0 contour estimation.

6.2. Objective evaluation – LSF and F_0 transformation

In all experiments, conversion from reference speaker to all other speakers was performed. 40 utterances were used for training and 15 different utterances for assessment.

The performance of F_0 transformation can be expressed using average error (Euclidean distance) between transformed and target (time-aligned) trajectories, $\tilde{f}_y$ and f_y, respectively.

$$E(\tilde{f}_y, f_y) = \frac{1}{N} \sum_{n=1}^{N} d(\tilde{f}_y(n), f_y(n)) \quad (24)$$

The results are stated in Table 1. To expose the consistency of results for all speakers, the results are presented separately. It is interesting that the error between transformed and target F_0 trajectories does not probably depend on primary distance between source and target F_0 trajectories.

Table 1: *Comparison of F0 transformation methods – average F0 errors [Hz].*

Target speaker	Male 1	Male 2	Fem. 1	Fem. 2
Default F0 distance (source – target)	50.64	68.12	30.38	21.76
F_0 normalization	13.14	12.32	16.71	15.66
Simple F_0 expect. (1 mixture)	13.08	11.77	16.75	15.75
Simple F_0 expect. (10 mixtures)	12.89	12.32	16.35	15.57
Combined expect. (10 mixtures)	11.97	11.15	14.89	14.45

The effectiveness of the conversion can be also evaluated by so-called performance index which is defined by

$$I_F = 1 - \frac{E(\tilde{f}_y, f_y)}{E(f_x, f_y)} \quad (25)$$

The higher value of performance index signifies the better conversion performance (max. value is 1). Comparison of F_0 transformation methods is presented in Table 2.

Similarly, for evaluation of LSF conversion, performance index I_{LSF} is used

$$I_{LSF} = 1 - \frac{E(\tilde{y}, y)}{E(x, y)} \quad (26)$$

Table 2: *Comparison of F0 transformation methods – performance indices.*

Target speaker	Male 1	Male 2	Fem. 1	Fem. 2
F_0 normalization	0.740	0.819	0.450	0.280
Simple F_0 expect. (1 mixture)	0.742	0.827	0.449	0.276
Simple F_0 expect. (10 mixtures)	0.745	0.830	0.462	0.284
Combined expect. (10 mixtures)	0.764	0.836	0.510	0.336

The contribution of combined LSF & F_0 transformation compared to simple LSF transformation is stated in Table 3.

Table 3: *Comparison of LSF transformation methods – performance indices.*

Target speaker	Male 1	Male 2	Fem. 1	Fem. 2
Simple expect. (10 mixtures)	0.406	0.323	0.314	0.337
Combined expect. (10 mixtures)	0.412	0.335	0.317	0.344

6.3. Simple vs. extended residual prediction

For comparison of the proposed two versions of residual prediction, the standard preference test was employed. In both cases, combined LSF & F_0 conversion was utilized. Ten participants of the test listened to 10 pairs of utterances, one was transformed by using simple residual prediction and the other by using extended residual prediction. The listeners selected utterances which sounded better. Results are presented in Table 4.

Table 4: *Simple vs. extended residual prediction*

Target speaker	Male	Female	Both
Simple RP preferred	1.9 %	0.0 %	1.0 %
Extended RP preferred	74.0 %	66.7 %	70.3 %
Cannot decide	24.1 %	33.3 %	28.7 %

6.4. Speaker discrimination test

Within this test, the combined LSF & F_0 conversion and extended residual prediction were employed. An extension of ABX test was used. Ten participants listened to triplets of utterances: original source and target (A and B in a random order) and transformed (X); they had to decide whether X sounds like A or B and rate their decision according to the following scale

1. X sounds like A
2. X sounds rather like A
3. cannot make a decision
4. X sounds rather like B
5. X sounds like B

Moreover, the listeners were allowed to use real numbers (e.g. 1.5 or 2.5) in indecisive cases.

Cases when A was from target and B from source speaker was reversed (including ratings). Thus all results correspond to the case when A is source and B target utterance. Then the higher rating, the more effective conversion. Average results are presented in Table 5.

Table 5: *Speaker discrimination test – average rating.*

Target sp.	Male	Female	Both
Average rat.	4.36	3.54	3.94

7. Conclusion

In this paper, three probabilistic conversion functions for F_0 transformation were compared. The transformation based on combined LSF & F_0 conditional expectation outperforms all other methods. Moreover, the conversion of LSF is also improved in this way. Furthermore, a new modification of inter-speaker residual prediction was proposed and compared to the traditional version by listening tests. All listeners definitely preferred the new method.

8. Acknowledgements

Support for this work was provided by the Ministry of Education of the Czech Republic, project No. 2C06020, and the EU 6th Framework Programme IST-034434.

9. References

[1] En-Najjary, T., Rosec, O. and Chonavel, T., "A Voice Conversion Method Based on Joint Pitch and Spectral Envelope Transformation", Proceedings of Interspeech 2004 - ICSLP, pp. 1225–1228

[2] Gillet, B. and King, S., "Transforming F0 Contours", Proceedings of Eurospeech 2003, pp. 101–104

[3] Chappell, D. T. and Hansen, J. H. L., "Speaker-Specific Pitch Contour Modeling and Modification", Proceedings of ICASSP 1998, pp. 885–888

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

[5] Hanzlíček, Z. and Matoušek J., "First Steps towards New Czech Voice Conversion System", TSD 2006, Lecture Notes in Artificial Intelligence 4188, Springer-Verlag, Berlin, Heidelberg, 2006, pp. 383–390

[6] Hanzlíček, Z., "On Residual Prediction in Voice Conversion Task", Proceedings of the 16th Czech-German Workshop on Speech Processing, ÚRE AVČR, Prague, Czech Republic, 2006, pp. 90–97

[7] Stylianou, Y., Cappé, O. and Moulines, E., "Continuos Probabilistic Transform for Voice Conversion", IEEE Transactions on Speech and Audio Processing, Vol. 6, 1998, pp. 131–142

[8] Kain, A., "High Resolution Voice Transformation", Ph.D. thesis, OGI School of Science & Engineering, Oregon Health & Science University, 2001

[9] Sündermann, D., Bonafonte, A., Ney, H. and Höge, H., "A Study on Residual Prediction Techniques for Voice Conversion", Proceedings of ICASSP 2005 pp. 13–16