Voice Conversion based on Probabilistic Parameter Transformation and Extended Inter-Speaker Residual Prediction *

Zdeněk Hanzlíček and Jindřich Matoušek

University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
zhanzlic@kky.zcu.cz, jmatouse@kky.zcu.cz

Abstract. Voice conversion is a process which modifies speech produced by one speaker so that it sounds as if it is uttered by another speaker. In this paper a new voice conversion system is presented. The system requires parallel training data. By using linear prediction analysis, speech is described with line spectral frequencies and the corresponding residua. LSFs are converted together with instantaneous F_0 by joint probabilistic function. The residua are transformed by employing residual prediction. In this paper, a new modification of residual prediction is introduced which uses information on the desired target F_0 to determine a proper residuum and it also allows an efficient control of F_0 in resulting speech.

Key words: voice conversion, 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.

In [1] and [2], some initial experiments on voice conversion were presented. In this paper, a new voice conversion system is described. This system requires parallel training data which is analysed by using pitch-synchronous linear prediction. Thus, speech frames are represented by line spectral frequencies (LSFs), corresponding residua and instantaneous fundamental frequency. LSFs are converted together with instantaneous F_0 by using a joint probabilistic function. The residua are transformed by employing residual prediction – a method which estimates a suitable residuum for a given parameter vector. In this paper, a new extension of this method is introduced which uses information on the desired target F_0 to determine a proper residuum and it also allows an efficient control of F_0 in resulting speech.

This paper is organized as follows. Section 2 describes speech data used in our experiments and methods used for its analysis, synthesis and time-alignment. Section 3 deals with parameter transformation for voiced and unvoiced speech.

* 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.

Section 4 describes the extended inter-speaker residual prediction. In Section 5, the performance of our conversion system is evaluated. Finally, Section 6 concludes the paper and outlines our future work.

2 Speech Data

Speech data for our experiments were recorded under special conditions in an anechoic chamber. Along with the speech signal, the glottal signal (EGG) was recorded to ensure more robust pitch-mark detection and F_0 contour estimation.

Firstly, one female speaker recorded the reference utterances – a set of 55 short sentences. All sentences were in the Czech language. Subsequently, four other speakers (two males and two females) listened to these reference utterances and tried to repeat them in the reference speaker's style. This should guarantee better pronunciation and prosodic consistency among all speakers.

2.1 Speech Analysis and Synthesis

Our voice conversion system employs pitch-synchronous linear prediction (LP) analysis. Each voiced frame is two pitch long with one pitch overlap. Pitch-marks are extracted from the EGG signal. Unvoiced frames are 10 msec long with 5 msec overlap. LP parameters are represented by their line spectral frequencies (LSFs), which are converted by employing a probabilistic function (e.g. [3] or [5]). Residual signal is represented by its amplitude and phase FFT-spectra, which are transformed by using residual prediction (e.g. [5] or [6]).

The reconstruction of speech is performed by a simple OLA method. For analysis and synthesis we employ special weight windows; in both cases a square root of Hann window is used. This is a trade-off between efficacious speech description and smooth frame composition on condition of correct speech reconstruction.

2.2 Speech Data Alignment and Selection

To find the conversion function properly, the training data has to be correctly time-aligned. This is performed by the dynamic time warping (DTW) algorithm. For each frame, the feature vector consists of delta-LSFs and V/U flag whose value is 1 for voiced and 0 for unvoiced frame.

After time-alignment, some suspicious data have to be excluded from the training set, because they probably correspond to incorrect time-alignment caused e.g. by prosodic or pronunciation mismatch:

- pairs composed of one voiced and one unvoiced frame
- long constant sections (horizontal or vertical) of warping function
- frame pairs with a very low energy or with too different energy values
- frame pairs with too different values of normalized F_0

3 Parameter Transformation

Parameters (LSFs) are transformed using a probabilistic conversion function based on the description of training data with a Gaussian mixture model (GMM). The conversion function is determined for voiced and unvoiced speech separately. Although unvoiced speech is supposed to be unimportant for speaker identity perception, the conversion of unvoiced speech proved good on transitions between voiced and unvoiced speech. Without unvoiced speech transformation, some unusual source speaker's glimmer was noticed in the converted utterances.

3.1 Simple LSF Transformation

This approach to parameter transformation was proposed by Stylianou et al. [3] and later improved by Kain et al. [5]. However, we used it only for the conversion of unvoiced speech. The interrelation between source and target speaker's LSFs (x and y, respectively) is described by a joint GMM with Q mixtures

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

All unknown parameters (mixture weights α_q, mean vectors μ_q and covariance matrices Σ_q) are estimated by employing the expectation-maximization (EM) algorithm. The mean vectors μ_q and covariance matrices Σ_q consist of blocks which correspond to source and target speaker's components

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

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

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

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

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

3.2 Combined LSF & $\mathbf{F_0}$ Transformation

This extension of the aforementioned simple LSF transformation was introduced by En-Najjary at al. [7]; however, the implemented system employed the Harmonic plus Noise Model of speech production.

This method exploits the interdependency between LSFs and instantaneous F_0; they are converted together by using one transformation function. Formally, new variables are introduced

$$\chi = \begin{bmatrix} 10^2 \cdot x \\ f_x \end{bmatrix} \qquad \psi = \begin{bmatrix} 10^2 \cdot y \\ f_y \end{bmatrix}. \tag{5}$$

A simple composition of LSFs and instantaneous F_0 would be unsuitable because the importance of particular components would not be well-balanced. This is the reason for introducing the weighting factor 10^2; this value was experimentally selected and performs well for all speaker combinations. In [7], the balancing of components is solved by the normalization of fundamental frequency.

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

$$p(\chi,\psi) = \sum_{q=1}^{Q} \alpha_q \mathcal{N}\left\{ \begin{bmatrix} \chi \\ \psi \end{bmatrix}; \mu_q = \begin{bmatrix} \mu_q^{\chi} \\ \mu_q^{\psi} \end{bmatrix}, \Sigma_q = \begin{bmatrix} \Sigma_q^{\chi\chi} & \Sigma_q^{\chi\psi} \\ \Sigma_q^{\psi\chi} & \Sigma_q^{\psi\psi} \end{bmatrix} \right\} \tag{6}$$

and the conversion function is defined as the conditional expectation of target ψ given source χ

$$\tilde{\psi} = E\{\psi|\chi\} = \sum_{q=1}^{Q} p(q|\chi)\Big[\mu_q^{\psi} + \Sigma_q^{\psi\chi}\big(\Sigma_q^{\chi\chi}\big)^{-1}(\chi - \mu_q^{\chi})\Big]. \tag{7}$$

The resulting vector $\tilde{\psi}$ is decomposed into LSFs $\tilde{y}$ and instantaneous F_0 $\tilde{f}_y$ which is further used in the extended residual prediction method (see Section 4).

4 Residual Prediction

Residual prediction is a method which allows the estimation of a suitable residuum for a given parameter vector. It would be unsatisfactory to use the original source speaker's residua because the residual signal still contains significant information on speaker identity, mainly in voiced speech.

In voice conversion framework (see e.g. [6] or [5]), the residual prediction is traditionally based on probabilistic description of source speaker's cepstral parameter space – with a GMM. For each mixture of this model, a typical residual signal is determined; it is represented by its amplitude and phase residual spectrum. Naturally, this method is only used for voiced frames. In unvoiced speech, residua are adopted from source speech without any modification.

In [1] and [2], a new approach to residual prediction – so-called inter-speaker residual prediction – was introduced. In comparison with the traditional residual prediction, the cardinal difference is that the target speaker's residua are estimated directly by using the source speaker's parameter vectors. Moreover, the source speaker's parameter space is described in a non-probabilistic manner.

In this paper, a new extension of this method is proposed which uses information on the desired instantaneous F_0 during the selection of a suitable residuum and facilitates a simple and efficient control of F_0 in the transformed speech.

4.1 Training Stage

A non-probabilistic description of source LSF space is used. Source LSFs are clustered into Q classes by employing the binary split k-means algorithm; a reasonable value of Q is about 20. Each class q is represented by its LSF centroid $\bar{x}_q$. The pertinence of parameter vector x_n ($n = 1, 2, \ldots N$) to class q ($q = 1, 2, \ldots Q$) can be expressed by the following weight

$$w(q|x_n) = \frac{\left[d(\bar{x}_q, x_n)\right]^{-1}}{\sum_{i=1}^{Q}\left[d(\bar{x}_i, x_n)\right]^{-1}} \tag{8}$$

All training data are uniquely classified into these classes. For each class q, a set R_q of pertaining data indices is established

$$R_q = \left\{k; 1 \leq k \leq N \ \wedge \ w(q|x_k) = \max_{i=1\ldots Q} w(i|x_k)\right\}. \tag{9}$$

Thus all data x_r for $r \in R_q$ belongs into class q. Within each parameter class q, the data is divided into L_q subclasses according to their instantaneous F_0. The number of subclasses L_q differs for particular parameter classes q. Each F_0 subclass is described by its central frequency $\bar{f}_q^{\ell}$ (q-th LSF 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

$$R_q^{\ell} = \left\{k; \ k \in R_q \ \wedge \ d(\bar{f}_q^{\ell}, f_k) = \min_{i=1\ldots L_q} d(\bar{f}_q^{i}, f_k)\right\}. \tag{10}$$

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

$$\hat{r}_q^{\ell} = \frac{\sum_{n \in R_q^{\ell}} r_n w(q|x_n)}{\sum_{n \in R_q^{\ell}} w(q|x_n)}. \tag{11}$$

Although all FFT-spectra cover the same frequency range given by the sampling frequency f_s, their lengths in samples are different because they correspond to the lengths of pitch-synchronously segmented frames. Thus all spectra have to be interpolated to the same length; cubic spline interpolation is used and the target length equals to the average length of all spectra within particular subclasses.

Similarly, the typical residual phase spectrum $\hat{\varphi}_q^{\ell}$ is determined. However because of phase warping problem, it is not calculated but it is only simply selected

$$\hat{\varphi}_q^{\ell} = \varphi_{n^*} \qquad n^* = \arg\max_{n \in R_q^{\ell}} w(q|x_n). \tag{12}$$

The selected residual phase spectrum should be interpolated to the same length as amplitude spectrum. To avoid the phase warping problem, nearest neighbour interpolation is used.

4.2 Transformation Stage

In the transformation stage, the desired target instantaneous fundamental frequency $\tilde{f}_n$ has to be known for each voiced frame; it is obtained by combined LSF & F_0 transformation (see Section 3.2).

The target 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 $\bar{f}_q^{\ell_q}$ is the nearest to the desired fundamental frequency $\tilde{f}_n$

$$\tilde{r}_n = \sum_{q=1}^{Q} \hat{r}_q^{\ell_q} w(q|x_n) \qquad \ell_q = \underset{\ell=1\ldots L_q}{\arg\min}\, d(\bar{f}_q^{\ell}, \tilde{f}_n) \tag{13}$$

The target 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 $\bar{f}_{q^*}^{\ell^*}$

$$\begin{aligned} \tilde{\varphi}_n = \hat{\varphi}_{q^*}^{\ell^*} \qquad & q^* = \underset{q=1\ldots Q}{\arg\max}\, w(q|x_n) \\ & \ell^* = \underset{\ell=1\ldots L_{q^*}}{\arg\min}\, d(\bar{f}_{q^*}^{\ell}, \tilde{f}_n) \end{aligned} \tag{14}$$

The resulting amplitude and phase FFT-spectra have to be interpolated to the length given by the desired F_0 $\tilde{f}_n$. The speech quality deterioration caused by this interpolation should not be significant, because the length of predicted residuum is very close to the target length.

5 Experiments and Results

In this section, the assessment of the described conversion system is presented. In the first subsection, mathematical evaluation of LSF and F_0 transformation is presented. The second subsection deals with subjective evaluation by listening tests.

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

5.1 Objective Evaluation – LSF and F_0 Transformation

The performance of LSF transformation can be expressed by using the performance index I_{LSF}

$$I_{LSF} = 1 - \frac{E(\tilde{y}, y)}{E(x, y)}, \tag{15}$$

where $E(x, y)$ is the average Euclidean distance between LSFs of 2 time-aligned utterances $x = \{x_1, x_2, \ldots x_N\}$ and $y = \{y_1, y_2, \ldots y_N\}$

$$E(x, y) = \frac{1}{N} \sum_{n=1}^{N} (x_n - y_n)^\top (x_n - y_n). \tag{16}$$

The higher value of performance index signifies the better conversion performance (maximum value is 1).

Similarly, the F_0 transformation can be evaluated by performance index I_{F_0}

$$I_{F_0} = 1 - \frac{E(\tilde{f}_y, f_y)}{E(f_x, f_y)}, \tag{17}$$

or it can be also simply assessed by using average Euclidean distance $E(\tilde{f}_y, f_y)$ between transformed $\tilde{f}_y$ and target f_y (the result is in Hz).

Results are stated in Table 1. They are presented separately for each speaker to expose that the outcomes are speaker dependent.

Table 1. *Mathematical evaluation of LSF and F_0 transformation performance.*

Target speaker	Male 1	Male 2	Female 1	Female 2
LSF performance index (voiced speech)	0.412	0.335	0.317	0.344
LSF performance index (unvoiced speech)	0.316	0.254	0.237	0.217
F_0 performance index	0.764	0.836	0.510	0.336
Default F_0 distance [Hz] (source – target)	50.64	68.12	30.38	21.76
Final F_0 distance [Hz] (transformed – target)	11.97	11.15	14.89	14.49

Though some performance indices were proposed which should facilitate more complex transformation assessment (e.g. spectral performance index), they do not often correspond to the real speech quality and resulting speaker identity as it is perceived by people. Thus, the best way of evaluating a voice conversion system in a complex way is listening tests.

5.2 Subjective Evaluation – Speaker Discrimination Test

An extension of standard ABX test was used. 10 participants listened to triplets of utterances: original source and target (A and B in a random order) and transformed (X). They made decisions 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. X is halfway between A and B
4. X sounds rather like B
5. X sounds like B

For unified result interpretation, cases when A was from target and B from source speaker were reversed. Thus all results correspond to the case when A is source and B target utterance. Then the higher rating signifies the more effective conversion. Average rating for female-to-male conversion was 4.36 (i.e. listeners were sure of the target speaker identity) and for female-to-female conversion was 3.54 (i.e. the identity was closer to the target speaker, but it was not so persuasive).

6 Conclusion and Future Work

In this paper a new voice conversion system was introduced which is based on probabilistic transformation of LSF and fundamental frequency and which utilizes the extended inter-speaker residual prediction for determination of proper residual signals.

Speaker discrimination tests revealed that the identity of converted speech is closer to the target speaker. However in cases of similar source and target voices (female-to-female conversion), the decision was not definite. This was probably caused by insufficient speaking style consistency and a small amount of training data. Generally, all speakers had some difficulty reproducing the reference utterances; they focused on mimicking but their speech lost its fluency and naturalness. Thus, in our future work we will concentrate on approaches which do not require parallel training data.

References

1. Hanzlíček, Z. and Matoušek, J.: First Steps towards New Czech Voice Conversion System. In Proceedings of TSD 2006, Lecture Notes in Artificial Intelligence 4188, Springer-Verlag, Berlin, Heidelberg (2006) 383–390
2. Hanzlíček, Z.: On Residual Prediction in Voice Conversion Task. In Proceedings of the 16th Czech-German Workshop on Speech Processing, ÚRE AVČR, Prague, Czech Republic (2006) 90–97
3. Stylianou, Y., Cappé, O., Moulines, E.: Continuous Probabilistic Transform for Voice Conversion. IEEE Transactions on Speech and Audio Processing, Vol.6, No.2 (1998) 131–142
4. Kain, A., Macon, M. W.: Design and Evaluation of Voice Conversion Algorithm Based on Spectral Envelope Mapping and Residual Prediction. Proceedings of ICASSP'01 (2001)
5. Kain, A.: High Resolution Voice Transformation. Ph.D. thesis, Oregon Health & Science University, Portland, USA (2001)
6. Sündermann, D., Bonafonte, A., Ney, H., Höge, H.: A Study on Residual Prediction Techniques for Voice Conversion. In Proceedings of ICASSP'05 (2005) 13–16
7. En-Najjary, T., Rosec, O. and Chonavel, T.: A Voice Conversion Method Based on Joint Pitch and Spectral Envelope Transformation. In Proceedings of Interspeech 2004 - ICSLP (2004) 1225–1228