Weighted Frequency Warping for Voice Conversion

Daniel Erro, Asunción Moreno

Department of Signal Theory and Communications
Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
derro@gps.tsc.upc.edu, asuncion@gps.tsc.upc.edu

Abstract

This paper presents a new voice conversion method called Weighted Frequency Warping (WFW), which combines the well known GMM approach and the frequency warping approach. The harmonic plus stochastic model has been used to analyze, modify and synthesize the speech signal. Special phase manipulation procedures have been designed to allow the system to work in pitch-asynchronous mode. The experiments show that the proposed technique reaches a high degree of similarity between the converted and target speakers, and the naturalness and quality of the resynthesized speech is much higher than those of classical GMM-based systems.

Index Terms: voice conversion, speech synthesis, harmonic model, GMM, weighted frequency warping

1. Introduction

The goal of voice conversion systems is to modify the voice of a source speaker for it to be perceived as if it was uttered by another speaker (target speaker). For this purpose, relevant characteristics of the source speaker are identified and replaced by the characteristics of the target speaker without losing any information or modifying the message. In speech synthesis, voice conversion techniques have important applications. Indeed, text-to-speech synthesis systems (TTS) usually generate their output by selecting and concatenating speech units taken from a database, which has been previously built by recording the voice of a professional or skilled speaker. A voice conversion block could be included in the TTS system to transform the recorded voice, so that it would not be necessary to record a database for each of the potential users of the system.

Several solutions for the voice conversion problem have been proposed since the first codebook-based transformation method was developed by Abe et al. [1]. Arslan et al. tried to avoid the spectral discontinuities caused by the hard partition of the acoustic space by means of a fuzzy classification [2]. Other techniques tried to represent the correspondence between the frequency axis of the source and target speakers by a warping function [3, 4]. Due to the low degree of modification, the reached quality was high, but the conversion was not successful because the amplitude of the formants could not be manipulated. One of the most important advances was the use of gaussian mixture models (GMM) to implement a continuous probabilistic spectral transformation based on the partition of the acoustic space into overlapping classes [5, 6]. The spectral envelopes were successfully converted without discontinuities, but the problem of over-smoothing appeared. Other works based on GMM transformations attempted to solve it [7, 8, 9]. At the same time, other types of acoustic classification such as hidden Markov models or decision trees were investigated [10], and the efforts of many authors focused on the residuals of the vocal tract parameterization [6, 11]. Nevertheless, the problem of high-quality voice conversion for real applications is not completely solved. There is still a compromise between the degree of transformation of voices and the quality reached by the different conversion methods.

The main goal of this work is to design a voice conversion method that successfully converts voices without significant quality degradation. We propose a new technique named Weighted Frequency Warping (WFW) where the voice is transformed via frequency warping, that is reported to have a high quality, combined with GMM that provides good conversion results. A different frequency warping function is calculated for each frame by means of a linear combination of basis functions. The weights of the combination and the shapes of the basis functions are obtained from a trained GMM, which is also used to increase the similarity between the warped source speaker and the target speaker. The model assumed for the speech signal is the harmonic plus stochastic model (HSM) [14], which provides maximum flexibility and capacity of manipulation. The implemented voice conversion system is expected to operate not only integrated into a TTS system, but also as an independent device which analyzes, converts and re-synthesizes speech. For this purpose, the classical HSM implementation has been modified to be pitch-asynchronous. Thus, the problems of pitch marking and accurate separation of signal periods are avoided, and the analysis rate can be adjusted depending on the applications. In exchange, the phase coherence is a crucial point, so new procedures for high-quality prosodic modification are also proposed in this paper.

The paper is structured as follows. In section 2, the new WFW voice conversion method is described. In section 3, some aspects about the implementation of the system are explained in detail, including the speech model and the prosodic modification procedures. Section 4 contains the results and the discussion of the experiments that compare the proposed approach with other reference systems, drawing the conclusions that are listed in section 5.

2. The New Method: WFW

The spectral conversion method based on GMMs has a good performance in terms of similarity between the converted and target voices, but the converted speech has a lower quality because of several factors: over-smoothing and broadening of the formants, effects of the conversion in the analysis/synthesis system (residual, phase spectrum…), etc. Although different solutions have been proposed for each of these problems, there is still a need of preserving the quality and naturalness of the signal. On the other hand, it is known that the degradation caused by frequency-warping-based transformations is minimal, although the conversion scores are not as high as in GMM-based systems. The Weighted Frequency Warping method (WFW) is a combination between these two approaches.

As described in [6], GMM-based voice conversion systems use a set of time-aligned LSF vectors of the source and target speakers, $\{[x_k^T\ y_k^T]^T\}$ to estimate the parameters $\{\alpha_i, \mu_i, \Sigma_i\}$ of a joint model of m gaussian mixtures. Once the model has been trained, the transformation function $F(x)$ is given by the following equations:

$$F(x) = \sum_{i=1}^{m} p_i(x) \cdot \left[\mu_i^y + \Sigma_i^{yx} \left(\Sigma_i^{xx}\right)^{-1} \left(x - \mu_i^x\right)\right] \quad (1a)$$

$$p_i(x) = \frac{\alpha_i N\left(x, \mu_i^x, \Sigma_i^{xx}\right)}{\sum_{j=1}^{m} \alpha_j N\left(x, \mu_j^x, \Sigma_j^{xx}\right)} \quad \mu_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix} \quad \Sigma_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix} \quad (1b, c, d)$$

where $p_i(x)$ is the probability that a LSF vector x belongs to the i^{th} gaussian component of the GMM. Observing the spectral envelopes given by the mean LSF vectors of each of the m acoustic classes of the GMM, μ_i^x and μ_i^y, it can be seen that their formant structure is quite similar. In this paper we propose to use the position of these formants to establish a piecewise linear frequency-warping function $W_i(f)$ for each of the m acoustic classes, as it is shown in figure 1. We assume that phonemes with similar formant structures, which are linked to the same gaussian component of the GMM, should be associated with similar frequency warping trajectories. Thus, the probabilities $p_i(x)$ can be used as weights for a linear combination of the m different warping trajectories.

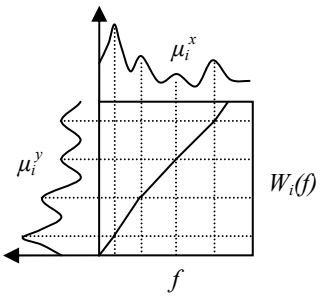

Figure 1 – *Frequency warping function for the i^{th} acoustic class.*

Once the GMM has been trained, the Weighted Frequency Warping method consists of the following steps:

1. Given a source frame to be converted, the associated LSF vector x is calculated. The probabilities $p_i(x)$ are calculated from the trained GMM (1b).
2. The frequency-warping function for the frame represented by x is obtained by a weighted combination of W_i's.

   $$W(f) = \sum_{i=1}^{m} p_i(x) \cdot W_i(f) \quad (2)$$

   Thus, a different warping trajectory is assigned to each frame. The soft classification provided by the GMM makes the warping function evolve slowly in time, avoiding the noise caused by the discontinuities.
3. Let $A(f)$ and $\theta(f)$ be the magnitude and phase estimators of the spectrum at the current frame. The converted amplitude and phase are obtained by warping the source envelopes.

   $$A'(f) = A\left(W^{-1}(f)\right), \quad \theta'(f) = \theta\left(W^{-1}(f)\right) \quad (3a, b)$$

This step does not completely transform the source voice into the target speaker's voice because the formants are only reallocated while their amplitude remains unmodified. The information provided by the GMM transformation allows a simple solution to this problem:

4. A converted $F(x)$ is obtained by means of the transformation function (1a) and the corresponding all-pole envelope is calculated. The energy is measured at the bands 100-300Hz, 300-800Hz, 800-2500Hz, 2500-3500Hz and 3500-5000Hz, which are likely to contain different formants. Constant multiplicative factors are used inside each band to correct the energy of the frequency-warped speech frame.

3. Implementation of the WFW system

3.1. Analysis and Reconstruction of Signals

In this paper, a modified version of the HSM model [14] is used for the analysis and reconstruction of the speech signal. The speech signal is modeled by a harmonic component and a stochastic component. The harmonic component is a sum of sinusoids whose amplitudes, frequencies and phases are determined for each speech frame. The stochastic component is modeled by a LPC filter driven by white noise. The signal parameters are measured at a constant frame rate of f_s/N frame/sec. f_s is the sampling frequency and N corresponds typically to a time interval of 8 or 10 ms. From now on, the center of the k^{th} analysis frame will be called *point k*. Pitch and voiced/unvoiced decision are taken at every frame k. In voiced frames, the harmonic component is characterized by the amplitudes and phases of the harmonics below 5 KHz. It is not adequate to apply a commonly used time-varying cut-off frequency, because the voice conversion method uses a parameterization of the spectral envelope based on the amplitudes of the harmonics. Afterwards, the harmonic waveform is reconstructed and subtracted from the original signal in order to isolate the stochastic component, which is analyzed in N-length frames centered at the points k using linear predictive coding (LPC).

The signal is reconstructed by overlapping and adding $2N$-length frames centered at each point k. Each frame contains the sum of the measured harmonics with constant amplitudes, frequencies and phases, and the stochastic contribution, generated by filtering white gaussian noise with the measured LPC-filters. A triangular window is used to overlap-add the frames in order to obtain the time-varying synthetic signal.

$$s^{(k)}[n] = \sum_{j=1}^{J^{(k)}} A_j^{(k)} \cos\left(2\pi j f_0^{(k)} n / f_s + \varphi_j^{(k)}\right) + \sigma[n] * h_{LPC}^{(k)}[n] \quad (4a)$$

$$s[kN + m] = \left(\tfrac{N-m}{N}\right) \cdot s^{(k)}[m] + \left(\tfrac{m}{N}\right) \cdot s^{(k+1)}[m - N] \quad (4b)$$

m is in the range [0, N–1]. The speech signal resynthesized from the measured parameters is almost indistinguishable from the original. More details about the analysis-synthesis process can be found in [12].

3.2. Prosodic Modifications

As a pitch-asynchronous scheme is being used, the prosodic modification of the signal implies the challenge of modifying the phases of the harmonics without altering the phase coherence between frames or causing artifacts. For this purpose, we developed new strategies to manipulate the phases. We consider that the phases $\varphi_j^{(k)}$ measured at a certain point k are the sum of two components: a linear-in-frequency term given by the parameter $\alpha^{(k)}$, and the phase contribution of the time-varying vocal tract, $\theta_j^{(k)}$.

$$\varphi_j^{(k)} = j\alpha^{(k)} + \theta_j^{(k)} \quad (5)$$

The **duration modification** can be carried out by increasing or decreasing the distance N between the synthesis points in equation (4b), so that the amplitude and fundamental frequency variations get adapted to the new time scale. On the other hand, if the phases were kept unmodified, fixed at the

center of the frames, the waveform coherence between consecutive points would be lost, causing artifacts and noisy pitch variations. Therefore, the change in N needs to be compensated with a phase manipulation in a way that the waveform and pitch of the duration-modified signal are similar to the original. This manipulation should affect only to the linear-in-frequency phase term. Assuming that the fundamental frequency varies linearly from point k–1 to k, we define the function Ψ which represents the expected phase increment of the first harmonic between those points, affecting only the linear-in-frequency term:

$$\alpha^{(k)} - \alpha^{(k-1)} \cong \Psi\left(f_0^{(k-1)}, f_0^{(k)}, N\right) = \pi N \left(f_0^{(k-1)} + f_0^{(k)}\right) / f_s \quad (6)$$

If N is substituted by N', the following phase correction is applied:

$$\Delta\varphi_1^{(k)} = \Psi\left(f_0^{(k-1)}, f_0^{(k)}, N'\right) - \Psi\left(f_0^{(k-1)}, f_0^{(k)}, N\right) \quad (7a)$$

$$\varphi_j'^{(k)} = \varphi_j^{(k)} + j\sum_{\kappa=2}^{k} \Delta\varphi_1^{(\kappa)} \quad j = 1 \ldots J^{(k)} \quad \forall k > 1 \quad (7b)$$

$J^{(k)}$ is the number of harmonics in frame k. This correction compensates the modification of N without affecting the small local variations in the vocal tract phase response. The stochastic coefficients are not modified.

For the **pitch modifications**, the amplitudes of the new harmonics $A'^{(k)}_j$ are obtained by a simple linear interpolation between the measured log-amplitudes in order to maintain the formant structure unaltered. The vocal tract contribution to the phases of the new harmonics, $\theta'^{(k)}_j$, can be obtained by means of a linear interpolation of the real and imaginary parts of the complex amplitudes $A_j^{(k)}\exp(i\cdot\theta_j^{(k)})$. The values of $\theta_j^{(k)}$ are calculated from the original phases $\varphi_j^{(k)}$ by subtracting the linear-in-frequency phase term given by $\alpha^{(k)}$. We estimate $\alpha^{(k)}$ using the following formula.

$$\alpha^{(k)} = \arg\max_{\alpha} \sum_{j=1}^{J^{(k)}} A_j^{(k)} \cos\left(\varphi_j^{(k)} - j\alpha\right) \quad (8a)$$

$$\theta_j^{(k)} = \varphi_j^{(k)} - j\alpha^{(k)} \quad (8b)$$

Finally, the relative position of the synthesis point within the new pitch period is now different and the linear term has to be corrected to compensate the modification of the periodicity. The phase correction to be performed is given by (7b) with

$$\Delta\varphi_1^{(k)} = \Psi\left(f_0'^{(k-1)}, f_0'^{(k)}, N\right) - \Psi\left(f_0^{(k-1)}, f_0^{(k)}, N\right) \quad (9)$$

The stochastic coefficients are not modified.

3.3. Voice Conversion

Although this work focuses on the spectral characteristics of the voice, a basic prosodic transformation is also applied. The fundamental frequency is characterized by a log-normal distribution. During the training phase, an estimate of the average value μ and variance σ of $\log f_0$ is calculated for each speaker. Our basic prosodic modification consists of replacing the source speaker's μ and σ by the values of the target speaker. The frequencies of the harmonic sinusoids are then scaled according to the new pitch values.

$$\log f_0^{(\text{converted})} = \mu^{(\text{target})} + \frac{\sigma^{(\text{target})}}{\sigma^{(\text{source})}}\left(\log f_0^{(\text{source})} - \mu^{(\text{source})}\right) \quad (10)$$

The WFW method is applied to the HSM model as follows. In the training phase, the harmonic amplitudes $\{A_j^{(k)}\}$ of the source and target training data are measured and the discrete all-pole modeling technique [13] is applied to obtain the optimal all-pole filters that better fit $\{A_j^{(k)}\}$. The all-pole filter coefficients are translated into LSF vectors. The LSF vector pairs extracted from the training data are used to train the GMM (1). Finally, the warping function $W_i(f)$ associated to each gaussian component is calculated from the GMM parameters. In the conversion phase, the current LSF vector x of the source speaker is determined. The probabilistic weights $p_i(x)$ (1b) are calculated to obtain the warping function $W(f)$ (2) of the current frame. The magnitude envelope $A(f)$ of the current frame is estimated by means of a linear interpolation between the measured log-amplitudes. The phase envelope $\theta(f)$ is estimated by linearly interpolating the real and imaginary parts of the complex amplitudes $A_j^{(k)}\exp(i\cdot\theta_j^{(k)})$, as in section 3.2. Warped envelopes $A'(f)$ and $\theta'(f)$ are calculated (3). Target amplitudes $\{A'^{(k)}_j\}$ and phases $\{\theta'^{(k)}_j\}$ are calculated by resampling the warped envelopes $A'(f)$ and $\theta'(f)$ at the positions of the new harmonics defined by the transformed f_0 (10). The linear-in-frequency phase term is adjusted according to the new f_0, as explained in section 3.2. In some cases, especially when the source speaker is a woman and the target speaker is a man, there is not enough information available from the source envelopes to fill all the harmonics below 5 KHz. In this situation, the envelope $F(x)$ obtained by means of the classical GMM transformation (1a) is used to get the missing data. Finally, the energy is corrected inside the bands defined in section 2 using the harmonic amplitudes obtained from the envelope $F(x)$.

It is known that the conversion of the stochastic component is not as relevant as the harmonic conversion [5, 11]. Nevertheless, a high correlation is observed between the vocal tract shape and the LPC envelope of the stochastic component. A reason is that the separation between harmonic and stochastic components in the voiced frames is not perfect. Furthermore, the assumption that there are no harmonics beyond the cut-off frequency of 5 KHz is not completely realistic, even though the quality reached in resynthesis is very high. For these reasons, it seems adequate to predict the stochastic component of the target speaker from his vocal tract LSF parameters at voiced frames. Using the GMM previously trained, the prediction is carried out using the following expression:

$$y_{stoch} = \sum_{i=1}^{m} p_i(y) \cdot \left[\eta_i + \Gamma_i\left(\Sigma_i^{yy}\right)^{-1}\left(y - \mu_i^{y}\right)\right] \quad (11)$$

where y_{stoch} is the LSF representation of the stochastic component that corresponds to the target speaker's LSF envelope given by y. The optimal vectors η_i and matrices Γ_i are calculated from the training data of the target speaker. During the conversion phase, the transformed LSF vector $F(x)$ of equation (1a) is used in (11) instead of y. The stochastic component of the unvoiced frames is left unmodified, because its conversion does not lead to any important improvement and it can cause a small loss of quality.

4. Experimental Results

The audio database used for the VC evaluation contains more than 150 sentences in Spanish, uttered by two male and two female speakers. The sampling frequency is 16 KHz and the average duration of the sentences is 5 seconds. All the sentences were analyzed and parameterized according to the model described in section 3, and 80% of them were used for the training of the conversion functions. The recorded parallel sentences were aligned for each pair of speakers using HMM-based forced recognition. An 8th order GMM was estimated with 14th order LSF vectors. Three methods were compared:

- **TTS**: it is a TD-PSOLA TTS system based on concatenation of units extracted from the training sentences of the target speaker. Obviously, it does not convert voices, but it is useful as a reference.
- **GMM**: it is a GMM-based voice conversion system using the HSM model. The converted amplitudes and phases are calculated by resampling the envelope of the all-pole filter given by the converted LSF vector. The pitch and the stochastic component are transformed using expressions (10) and (11), respectively.
- **WFW**: the new voice conversion system described above.

One male and one female speaker were chosen as source, and the other two speakers were taken as target, so four different conversion directions were considered: male to male (m2m), female to female (f2f), male to female (m2f) and female to male (f2m). Five sentences were converted and resynthesized for all the combinations of methods and conversion directions, and 15 volunteers were asked to listen to the converted sentences in random order. Only three of them were skilled listeners. For each pair of voices, listeners were asked to judge if the two voices belonged to the same person using a 5-point scale, from 1 (completely different) to 5 (identical). The final conversion score was obtained by averaging all the individual scores. On the other hand, the listeners were asked to rate the quality of the sentences from 1 point (bad) to 5 points (excellent). The table 1 shows the results of the perceptual test.

Conversion

	f2f	f2m	m2f	m2m	All
TTS	3.67	3.93	3.93	3.87	**3.85**
GMM	3.13	3.27	2.47	3.07	**2.98**
WFW	3.00	2.53	3.27	2.93	**2.93**

Quality

	f2f	f2m	m2f	m2m	All
TTS	2.53	2.87	2.47	2.67	**2.63**
GMM	3.13	3.33	2.53	2.73	**2.93**
WFW	4.20	3.60	3.00	3.27	**3.52**

Table 1 – Results of the perceptual test

The conversion score obtained by the TTS system can be considered the maximum score reachable with the training data. Due to the small size of the database for a concatenative TTS, there are concatenation artifacts. The opinion of the listeners seems to be strongly influenced by the concatenation artifacts and gives an idea of the difficulty of reaching a score higher than 4. There is a small loss of conversion accuracy from GMM to WFW. This can be a consequence of the fact that details from the source speaker persist when the frequency warping function is applied. Looking at the different conversion directions it can be seen that the main significant differences are located in the cross-gender conversion cases. In particular, WFW fails when converting from female to male. The reason is the strong f_0 contrast between those speakers, because the source spectral envelopes are defined by few harmonics, while a high number of target harmonics have to be extracted from them.

Looking at the quality scores, it can be seen that the increment of quality from GMM to WFW is important. Furthermore, the improvements are visible and consistent in every conversion direction.

Some other informal tests have been carried out to evaluate the WFW system using less training data, and the results seem to be very similar to those displayed in table 1.

5. Conclusions

This paper shows that the voice conversion techniques based on gaussian mixture models and frequency warping can be effectively combined to compensate the drawbacks of both methods. A good balance is obtained between the conversion and quality scores reached by means of the proposed method, WFW. The improvements in the quality of the converted synthetic speech are very important with respect to GMM methods (around 0.5 points in a MOS test).

In future works, non-parallel training corpora will be used to evaluate the conversion system.

6. Acknowledgements

This work was partially supported by TC-STAR (Technology and Corpora for Speech-to-Speech Translation, FP6-506738) and AVIVAVOZ (TEC2006-13694-C03).

7. References

[1] M.Abe, S.Nakamura, K.Shikano, H.Kuwabara, "Voice Conversion through Vector Quantization", ICASSP, 1988.

[2] L.M.Arslan, "Speaker Transformation Algorithm using Segmental Codebooks (STASC)", Speech Communication, 1999.

[3] H.Valbret, E.Moulines, J.P.Tubach, "Voice Transformation using PSOLA Technique", Speech Communication, 1992.

[4] D.Sündermann, H.Ney, "VTLN-based voice conversion", ISSPIT, 2003.

[5] Y.Stylianou, O.Cappé, E.Moulines, "Continuous Probabilistic Transform for Voice Conversion", IEEE Trans. On Speech and Audio Proc., 1998.

[6] A.Kain, "High-Resolution Voice Transformation", PhD Thesis, OGI School of Science and Engineering, 2001.

[7] T.Toda, H.Saruwatari, K.Shikano, "Voice Conversion Algorithm based on Gaussian Mixture Model with Dynamic Frequency Warping of Straight Spectrum", ICASSP, 2001.

[8] Y.Chen, M.Chu, E.Chang, J.Liu, R.Liu, "Voice Conversion with Smoothed GMM and MAP Adaptation", Europ. Conf. on Speech Communication and Tech., 2003.

[9] H.Ye, S.Young, "Quality-enhanced Voice Morphing using Maximum Likelihood Transformations", IEEE Trans. On Audio, Speech and Lang. Proc., 2006.

[10] H.Duxans, A.Bonafonte, A.Kain, J.Van Santen, "Including dynamic and phonetic information in voice conversion systems", ICSLP, 2004.

[11] D.Sündermann, A.Bonafonte, H.Ney, "A Study on Residual Prediction Techniques for Voice Conversion", ICASSP, 2005.

[12] D.Erro, A.Moreno, "A Pitch-Asynchronous Simple Method for Speech Synthesis by Diphone Concatenation using the Deterministic plus Stochastic Model", SPECOM, 2005.

[13] A.El-Jaroudi, J.Makhoul, "Discrete All-Pole Modeling", IEEE Trans. on Signal Proc., 1991.

[14] Y.Stylianou, "Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification", PhD thesis, École Nationale Supérieure des Télécommunications, 1996.