

# A voice conversion method based on joint pitch and spectral envelope transformation

*Taoufik En-Najjary\**, *Olivier Rosec\** and *Thierry Chonavel\*\**

\* France Telecom R&D, DIH/IPS, 2 avenue Pierre Marzin, 22307 Lannion Cedex, France  
{taoufik.ennajjary, olivier.rosec}@francetelecom.com

\*\* ENST Bretagne, Département SC, BP 832, 29285 Brest Cedex, France  
Thierry.Chonavel@enst-bretagne.fr

## Abstract

Most of the research in Voice Conversion (VC) is devoted to spectral transformation while the conversion of prosodic features is essentially obtained through a simple linear transformation of pitch. These separate transformations lead to an unsatisfactory speech conversion quality, especially when the speaking styles of the source and target speakers are different. In this paper, we propose a method capable of jointly converting pitch and spectral envelope information. The parameters to be transformed are obtained by combining scaled pitch values with the spectral envelope parameters for the voiced frames and only spectral envelope parameters for the unvoiced ones. These parameters are clustered using a Gaussian Mixture Model (GMM). Then the transformation functions are determined using a conditional expectation estimator. Tests carried out show that, this process leads to a satisfactory pitch transformation. Moreover, it makes the spectral envelope transformation more robust.

## 1. Introduction

The purpose of VC is to modify a source speaker's speech so that it is perceived as if a target speaker had uttered it. A possible application of such a technology is the personalization of Text-To-Speech (TTS) systems. Indeed, voice conversion offers a quick and automatic way to create additional voices for a speech synthesizer, while avoiding the actual recording and processing of large speech databases for each new voice, which is known to be a very tiresome and expensive task.

In order to achieve VC, the only necessary speech material is a small recording of both speakers uttering the same message. Given these limited databases a learning procedure aims to estimate a transformation function, which will then be applied to the source speaker's signal in order to mimic the speech of the target speaker. Due to the small amount of data used, VC essentially addresses the problem of transforming segmental properties, such as spectral envelope and pitch.

Until now, many methods have been developed for spectral envelope modification [1, 2, 3, 4, 5]. These are of course useful, insofar as they make it possible to approach the target spectral envelope. Unfortunately, these techniques remain insufficient for correctly mimicking the target speaker's voice; other parameters related to prosody are judged crucial for the speech perception and must therefore be

taken into account. Amongst these prosodic parameters, pitch information is of particular importance.

As yet, relatively little research has tackled the delicate problem of pitch conversion. The general approach for pitch modifications in the framework of VC boils down to respecting the global speech scale of the target speaker. These modifications were improved in [6] in order to take the pitch slope within a sentence into account. In another method [7] a piecewise linear transformation was proposed in order to map pitch contours from one speaker to another. However such modifications remain global; only pitch characteristics defined on the whole database are taken into account.

Syrdal and Steele [8] provide evidence that the first formant and pitch are dependent, suggesting that in order to preserve a high quality of the speech signal, any change to one of these parameters must be accompanied by a suitable modification of the other. Such an approach is adopted in [9], where the spectral envelope is modified according to pitch modification by a vector quantization (VQ) codebook mapping technique, using three codebooks for respectively low, medium and high-pitched speech frames. Another interesting method [10] uses a GMM in order to predict the average evolution of the spectral envelope over all occurring pitch values within a speech database. In TTS synthesis, this technique allows coarse spectral transformations to be done in the case of extreme pitch modifications.

In a previous work [11], we proposed a method in order to successively modify the spectral envelope and the pitch. Initially, a spectral envelope transformation function is applied to the source speaker's spectral parameters so as to approach those of the target speaker. Then, a pitch prediction function is used to estimate the pitch of the target speaker from the transformed spectral envelope. The process of pitch prediction provides convincing results insofar as the pitch prediction error is rather weak (about 4 Hz). However, any error of the transformed spectral envelope parameters is reflected automatically on the predicted pitch, which makes this method not very robust and thus limits its interest for VC applications.

In this work, a new method for the joint transformation of the pitch and the spectral envelope based on GMM is proposed. The paper is organized as follows. The next section gives the outline of the method, while sections 3 and 4 respectively describe the experimental setup and present the evaluation results.

## 2. Voice conversion algorithm based on GMM

### 2.1. Gaussian Mixture Model (GMM)

The learning procedure described in this paper aims to fit a GMM model to the data. Formally, a GMM allows the probability distribution of a random variable  $z$  to be modeled as the sum of  $Q$  Gaussian components, also referred to as classes. Its probability density function can be written as

$$p(z) = \sum_{i=1}^Q \alpha_i N(z; \mu_i, \Sigma_i), \text{ with } \sum_{i=1}^Q \alpha_i = 1, \alpha_i \geq 0.$$

$N(z; \mu, \Sigma)$  denotes the Gaussian distribution with mean vector  $\mu$  and covariance matrix  $\Sigma$  and  $\alpha_i$  denotes the prior probability that vector  $z$  belongs to the  $i^{\text{th}}$  class. The model parameters  $(\alpha, \mu, \Sigma)$  are estimated using the Expectation Maximization (EM) algorithm [12] which is an iterative method for computing maximum likelihood parameter estimates.

### 2.2. Estimation of the mapping functions

#### 2.2.1. Transformation function for voiced frames

The GMM defined above will be applied to parameter vectors including at the same time the spectral parameters and pitch both for the source and target speakers. As these input vectors combine heterogeneous information, particular care must be taken so as to control the relative weights of the spectral and pitch components.

Finding the relative importance of spectral and pitch information is a rather difficult problem. A simple scheme is adopted here, also used in [13] in a speaker identification context, where the pitch values are scaled according to:

$$F_{\log} = \log(F_0 / \bar{F}_0), \quad (1)$$

where  $F_0$  is the fundamental frequency in Hz and where  $\bar{F}_0$  is the average pitch value determined on all the voiced frames contained in the training database.

The GMM is used here to jointly model the parameters of the source and target speakers. Let  $x = [c_x^T, g_x]^T$  be a vector of parameters obtained by incorporating the spectral coefficients  $c_x^T$  and the normalized pitch  $g_x$  of the source speaker (where T denotes the transposition operator). In the same way, we obtain  $y$  the vector of parameters for the target speaker. Let  $X = [x_1, x_2, \dots, x_N]$  be the sequence of  $N$  feature vectors describing a succession of speech frames produced by the source speaker, and  $Y = [y_1, y_2, \dots, y_N]$  be the corresponding sequence as produced by the target speaker. Then the aim of the learning stage is to estimate a function  $F$  such that the transformed vector  $F(x)$  best matches the target vector  $y$ .

In the same approach as in [5], the combination of source and target vectors  $z = [x^T, y^T]^T$  is used to estimate the GMM parameters  $(\alpha, \mu, \Sigma)$ . Then, the mapping function is chosen to be the regression, which can be written as

$$\hat{y}_{CE} = F(x) = E[y|x] \\ = \sum_{i=1}^Q h_i(x) [\mu_i^y + \Sigma_i^{xy} (\Sigma_i^{xx})^{-1} (x - \mu_i^x)], \quad (2)$$

$$\text{where } h_i(x) = \frac{\alpha_i N(x; \mu_i^x, \Sigma_i^{xx})}{\sum_{j=1}^Q \alpha_j N(x; \mu_j^x, \Sigma_j^{xx})} \quad (3)$$

is the *a posteriori* probability that a given input vector  $x$  belongs to the  $i^{\text{th}}$  class.  $\mu_i^x$  and  $\mu_i^y$  denote mean vectors of class  $i$  for the source and target speakers respectively.  $\Sigma_i^{xx}$  is the covariance matrix of class  $i$  for the source speaker.  $\Sigma_i^{yx}$  denotes the cross-covariance matrix of class  $i$  for the source and target speakers.

#### 2.2.2. Transformation function for unvoiced frames

For unvoiced frames, only spectral envelope features are transformed. The learning procedure remains the same, as GMM is used to model the joint density of the source and target speaker's spectral parameters. This leads to a transformation function similar to (2).

## 3. Experiments

### 3.1. Datasets

The database used in order to train and test the conversion function consists of 962 short-paired utterances from one male and one female speakers, which corresponds for each speaker to about 15 minutes of speech, sampled at 16kHz. The phonetic transcription and segmentation have been checked and manually corrected. The training set was obtained by holding out 300 paired utterances of the total dataset. The 662 remaining utterances are used for tests.

### 3.2. Analysis and Training

This conversion system is evaluated using the Harmonic plus Noise Model (HNM) [4], which allows high quality prosodic as well as spectral modifications.

As in [14], the present work uses a simplified version of the HNM. The major difference is in the use of a constant maximum voicing frequency, which is set to half the sampling frequency. This enables the whole spectrum to be represented by the regularized discrete cepstrum method presented in [15]. The analysis is performed asynchronously with a constant frame rate of 10 msec. Note that the asynchronous mode is used only for the data needed to train the conversion function. For the subsequent voice conversion, the analysis is carried out pitch-synchronously because it allows higher quality prosodic modifications. The order of the cepstral representation is set to 20 and the  $c(0)$  coefficient is not included in the parameters to be transformed.

In order to train the speech conversion function, the first step consists in aligning the source and the target speech feature streams in such a way that they describe the same phonetic content frame by frame. This is achieved by a Dynamic Time Warping (DTW) algorithm. The phonetic boundaries have been considered as anchor points for the DTW procedure. The pairs of aligned speech frames where

one speaker's speech was voiced and the other speaker's was unvoiced were rejected from the training sets. The remaining time aligned data comprise 30,000 spectral vectors, which corresponds to about 5 minutes of speech. Then the paired vectors are split in two datasets according to whether the frames are voiced or not. For each frame, the pitch is normalized according to (1) and combined with the cepstral coefficients.

Then the estimation of the GMM parameters is achieved by the EM algorithm, initialized by a classical VQ procedure, and run until either the likelihood increase is below a given threshold, or 20 iterations are exceeded. To prevent singularities, a small value was added to the diagonal elements of the covariance matrices after each iteration. This training procedure was applied to several GMM, with a number of mixture components varying as a power of 2 between 8 and up to 64.

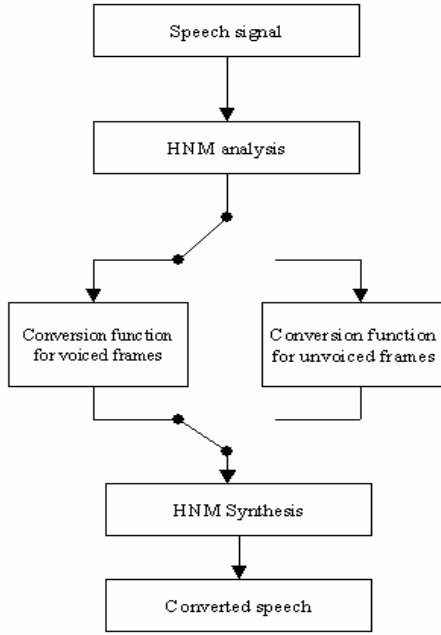


Figure 1: Bloc diagram of the voice conversion system.

### 3.3. Conversion

Once the conversion functions have been estimated, the voice transformation is performed, as indicated in Figure 1. Pitch and spectral features are extracted in a pitch-synchronous way. If the frame is voiced, the pitch is scaled according to (1) and combined with cepstral vector as described in subsection 2.2.1, before the transformation function for voiced frames is applied. Then the converted pitch is inverse-scaled using the average pitch of the target speaker. If the frame is unvoiced we apply directly the transformation function for unvoiced frame. Once the pitch and the spectral envelope are transformed, the converted speech is generated by an HNM synthesizer.

## 4. Evaluation

In these experiments we compare the difference between our approach and a classical spectral transformation [5] combined

with a simple linear modification of pitch ( $F_0$ ) used in the literature which can be written as:

$$\hat{F}_0 = ((F_0 - \mu_{src}) / \sigma_{src}) * \sigma_{tar} + \mu_{tar}, \quad (4)$$

where  $\mu_{src}$ ,  $\sigma_{src}$  are the mean and standard deviation for the source speaker respectively, and  $\mu_{tar}$ ,  $\sigma_{tar}$  are those for the target speaker.

### 4.1. Objective Evaluation

Figure 2 shows a typical example of pitch conversion obtained by application of our method. The joint conversion of the spectral envelope and the pitch offers a rather satisfactory pitch transformation. By comparison, a simple scaling of the pitch by a linear transformation according to (4) does not make it possible to reflect differences between the source pitch contour shapes and the target ones. Thus, the process is particularly interesting because it allows a suitable pitch conversion, even if the speaking styles of the two speakers are different. To quantify this improvement on the whole test dataset, we used the Normalized Pitch Distortion (NPD) defined as:

$$NPD = \sqrt{\frac{\sum_{n=1}^N (F_{0,n}^y - \hat{F}_{0,n}^y)^2}{\sum_{n=1}^N (F_{0,n}^y - F_{0,n}^x)^2}}. \quad (5)$$

This distortion decreases from 0.19 in the case of SPT to 0.15 with the proposed method for male-to-female conversion, and from 0.17 to 0.12 for female-to-male conversion.

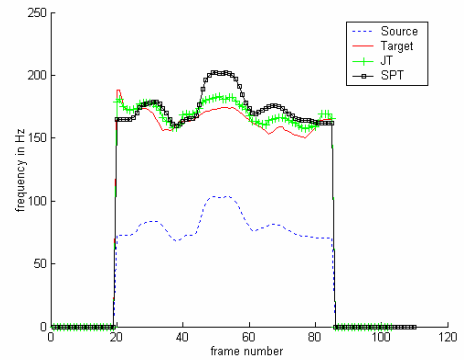


Figure 2: Example of pitch transformation: Joint Transformation (JT), Simple Pitch Transformation (SPT).

For spectral envelope conversion, objective tests are carried out using the Normalized Cepstral Distortion (NCD) defined by:

$$NCD = \frac{\sum_{n=1}^N \|c_{y,n} - \hat{c}_{y,n}\|_2}{\sum_{n=1}^N \|c_{y,n} - c_{x,n}\|_2}, \quad (6)$$

where  $c_{y,n}$ ,  $\hat{c}_{y,n}$ , and  $c_{x,n}$  respectively denote the target, converted and source spectral parameters. As shown on figure 3, this NCD is smaller in the case of joint conversion than in the case when spectral envelope is transformed alone. These results can be explained by the fact that the pitch and spectral envelope information are correlated as observed in [10, 11]. Thus, modifying both quantities globally makes the conversion process more robust. For unvoiced data, spectral envelope conversion results are also satisfactory as depicted on figure 4.

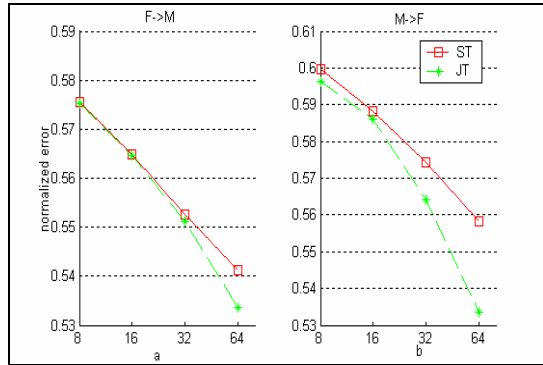


Figure 3: NCD for voiced data produced by joint transformation (JT) and spectral transformation only (ST) as a function of the number of components for female-to-male (a) and male-to-female (b) conversions.

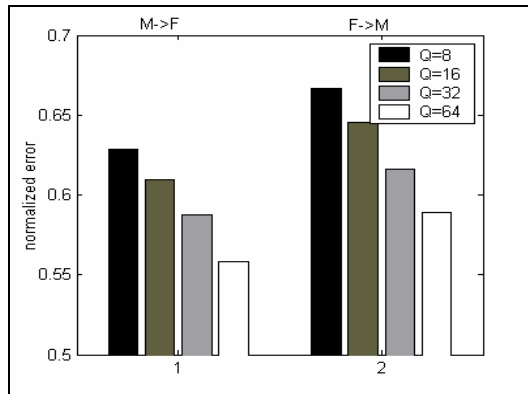


Figure 4: Normalized conversion errors for unvoiced data as a function of the number of GMM component, the figures represent the male-to-female (1) and female-to-male (2) conversion errors.

#### 4.2. Subjective Evaluation

In order to subjectively evaluate our method, 10 utterances were extracted from the test dataset. For each utterance, the converted speech signals obtained by both methods were randomly presented to 10 listeners. Each listener was then asked to select the converted signal which seemed to him closer to the target utterance.

The results are satisfactory as the speech converted with the proposed method was preferred in 97% of the cases.

## 5. Conclusion

This paper proposed a new GMM-based method enabling the joint conversion of pitch and spectral envelope. Objective as well as subjective tests have shown that this process increases the conversion performance.

Further experiments will be done in the near future on larger and richer databases in order to attest to which extent the proposed method can convert speech with different speaking styles.

## 6. References

- [1] M. Abe, S. Nakamura, K. Shikano and H. Kuwabara, "Voice conversion through vector quantization", Proceedings of IEEE ICASSP, pp 655-658, 1988.
- [2] H. Valbret, E. Moulines and J. P. Tubach, "Voice transformation using PSOLA technique", Speech Communications, vol. 11, pp. 175-187, 1995.
- [3] N. Iwahashi and Y. Sagisaka, "Speech spectrum transformation by speaker interpolation", Proceedings of IEEE ICASSP, vol. 1, pp. 461-465, 1994.
- [4] Y. Stylianou, "Harmonic plus Noise Model for speech, combined with statistical methods, for speech and speaker modification", PhD thesis, Ecole Nationale Supérieure des Telecommunications, Paris, France, 1996.
- [5] A. Kain and Mr. Macon, "Text-to-speech voice adaptation from sparse training dated", Proceedings of ICSLP 1998.
- [6] T. Ceysens, W. Verhelst and P. Wambacq, "One the construction of a pitch conversion system", Proceedings of EUSIPCO, 2002.
- [7] B. Gillet and S. King, "Transforming F0 Contours", Proceeding of EUROSPEECH 2003.
- [8] A.K. Syrdal and S.A. Steele, "Vowel F1 have has function of announcer fundamental frequency", 110<sup>th</sup> Meeting of JASA, flight. 78, Fall 1985.
- [9] K. Tanaka and M. Abe, "A new fundamental frequency modification algorithm with transformation of spectrum envelope according to  $F_0$ ", Proceedings of IEEE ICASSP, vol.2, pp. 951-954, 1997.
- [10] A. Kain and Y. Stylianou, "Stochastic modeling of Spectral adjustment for high quality pitch modification", Proceedings of IEEE ICASSP, vol. 2, pp. 949-952, 2000.
- [11] T. En-Najjary, O. Rosec and T. Chonavel, "A new method for pitch prediction from spectral envelope and its application in voice conversion", Proceeding of EUROSPEECH 2003.
- [12] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum likelihood from incomplete dated via the EM algorithm", Journal of the Royal Statistical Society Serie B, flight 39, pp. 1-38, 1977.
- [13] C. Miyajima, Y. Hattori, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Text-independent speaker identification using gaussian mixture models based on multi-space probability distribution", IEICE Transactions on Information ans Systems, vol. E84, 2001.
- [14] G. Baudoin and Y. Stylianou, "On the transformation of the speech spectrum for voice conversion", Proceedings of ICSLP, 1996.
- [15] O. Cappé, J. Laroche and E. Moulines, "Regularized estimate of cepstrum envelope from discrete frequency points", IEEE ASSP Applications of Signal Processing to Audio and Acoustics, pp. 213-216, 1995.