A new method for pitch prediction from spectral envelope and its application in voice conversion

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

This paper deals with the estimation of pitch from only spectral envelope information. The proposed method uses a Gaussian Mixture Model (GMM) to characterize the joint distribution of the spectral envelope parameters and pitch-normalized values. During the learning stage, the model parameters are estimated by means of the EM algorithm. Then, a regression is made which enables the determination of a pitch prediction function from spectral envelope coefficients. Some results are presented which show the accuracy of the proposed method in terms of pitch prediction. Finally, the application of this method in a voice conversion system is described.

1.   Introduction

A voice conversion (VC) system aims to transform an utterance spoken by a given source speaker in such a way that it is perceived as if another target speaker had spoken it. A possible application of such a technology is the personalization of Text-To-Speech (TTS) synthesis systems. Indeed, VC 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.

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 voice; other parameters related to prosody are judged crucial for the speech perception and must therefore be taken into account. Amongst the 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. However such modifications remain global; only pitch characteristics defined on the whole database are taken into account.

Syrdal and Steele [7] 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 [8], where the spectral envelope is modified according to pitch modification by a vector quantization codebook mapping technique, using three codebooks for respectively low, medium and high-pitched speech frames. Another interesting method [9] uses a Gaussian mixture model (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 this paper, a method based on a GMM is described, which enables the estimation of a pitch prediction function from spectral envelope information. The paper is organized as follows. The analysis procedure as well as the learning stage of the proposed method are presented in section 2. Section 3 presents prediction results obtained by this new technique, and section 4 describes its application in a VC system.

2.   Pitch prediction using GMM

2.1. Analysis

The method described below aims to predict pitch values from spectral envelope information. In order to estimate this so-called pitch prediction function, an analysis has to be done so as to obtain a set of spectral coefficient vectors as well as the associated estimated pitch values.

In the scope of VC, an interesting model is the Harmonic plus Noise Model (HNM) [4] as it allows high quality prosodic as well as spectral modifications. Given a voiced frame, this model splits the speech spectrum into two parts delimited by a so-called maximum voicing frequency. The lower part of the spectrum is approximated by a sum of harmonically related sine waves while the upper part is modelled by an AR-filtered white Gaussian noise.

The fundamental frequency is computed by a temporal method and the estimated values are manually checked to assure that there is no pitch doubling or halving error. The remaining of the HNM analysis used in this paper is similar to the one developed in [4]. 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 [10].

2.2. Learning procedure

2.2.1. GMM modelling

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 modelled as the sum of Q multivariate Gaussian components, also referred to as classes. Its probability density function can be written as

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

where $N(z;\mu,\Sigma)$ denotes the n-dimensional normal distribution with mean vector μ and covariance matrix Σ, and α_i denotes the prior probability that the ith class generated z. The model parameters (α,μ,Σ) are estimated using the expectation maximization (EM) algorithm [11], an iterative method for computing the maximum likelihood parameter estimates.

2.2.2. Pitch prediction function

Let $X = [x_1 x_2 ... x_N]$ be a sequence of pitch values and $Y = [y_1 y_2 ... y_N]$ be the corresponding set of cepstral coefficient vectors. The goal of the learning procedure is to estimate a function F such that the predicted pitch values $\hat{x} = F(y)$ best match in some sense the actual observed pitch values y.

The approach adopted here is to combine the pitch information with the spectral envelope and to model their joint density by means of a GMM. More precisely, let $z = [y^T x]^T$ denote the obtained joint vector. In a first stage, the training step aims at estimating the GMM parameters (α,μ,Σ) of the joint probability density function $p(z)$. Then the pitch prediction is chosen to be the regression which can be written as

$$F(y) = E[x|y] = \sum_{i=1}^{Q} h_i(y) \left[\mu_i^x + \Sigma_i^{xy} (\Sigma_i^{yy})^{-1} (y - \mu_i^y) \right]$$

where

$$h_i(y) = \frac{\alpha_i N(y;\mu_i^y,\Sigma_i^{yy})}{\sum_{j=1}^{Q} \alpha_i N(y;\mu_i^y,\Sigma_i^{yy})} \quad (1)$$

is the posterior probability that a given input vector y belongs to the ith class, with

$$\Sigma_i = \begin{bmatrix} \Sigma_i^{yy} & \Sigma_i^{yx} \\ \Sigma_i^{xy} & \Sigma_i^{xx} \end{bmatrix} \text{ and } \mu_i = \begin{bmatrix} \mu_i^y \\ \mu_i^x \end{bmatrix}.$$

2.2.3. Pitch normalization

The method described here above operates a clustering on vectors containing cepstral coefficients and pitch values. 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 [12], where the pitch values are normalized according to:

$$F_{\log} = \log(F_0 / \overline{F_0}), \quad (2)$$

where F_0 is the fundamental frequency in Hz and where $\overline{F_0}$ is the average pitch value determined on all the voiced frames contained in the training database. The interest of this normalization procedure will be clarified in the following section.

3. Simulation results

This section presents some results of the application of the method in the case of a French female speaker. F_0 values and discrete cepstrum parameters are extracted from a training database containing 25 minutes of speech sampled at 16 kHz. Of course, only the frames declared as voiced during the analysis are retained for the learning stage. The order of cepstral vectors is set to 20 and the first cepstral coefficient, which is related to the frame energy, is ignored. The GMM model is implemented with 64 components. During the learning step, the EM algorithm is initialized with a classical vector quantization technique.
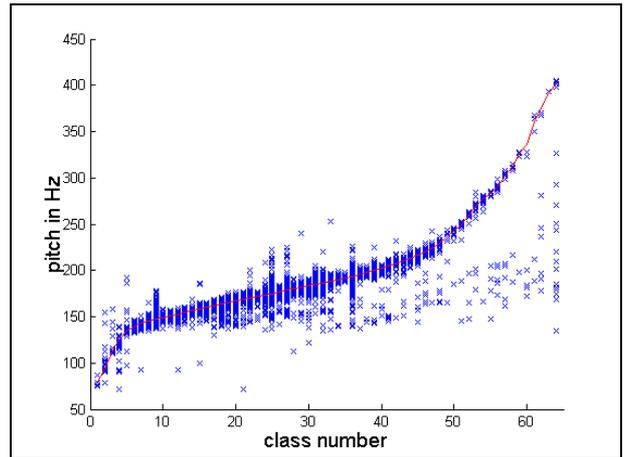
Figure 1: *Distribution of the observed pitch values together with the mean for each estimated class: training without normalization.*

Initially, the normalization of the pitch presented in the previous section is omitted. The mixture component maximizing the posterior probability (1) is assigned to each input frame. Given this classification, the mean pitch value is calculated for each class. Figure 1 shows the distribution of the observed pitch values together with its mean as a function of the class label (class labels are determined so that the mean pitch values of each class appears in ascending order). It reveals a certain correlation between pitch and spectral envelope, but also exhibits a high variance of the pitch within each class, as well as a strong overlapping of pitch values between distinct classes. After normalizing the pitch according to (2), this distribution becomes much more narrow within each class, as underlined by figure 2.
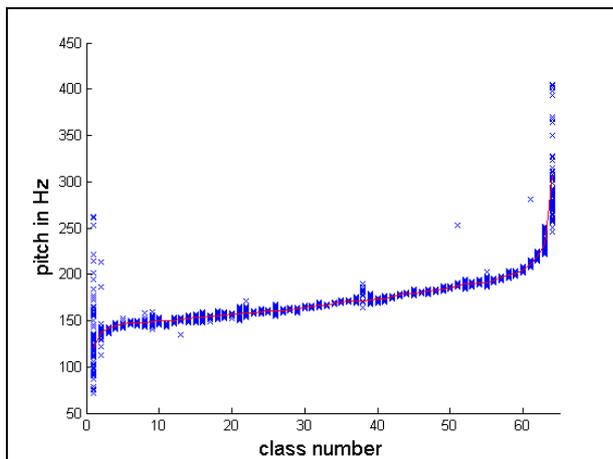
F0 values	<150 Hz	150-250 Hz	>250 Hz	Total
Mean (Hz)	0.6	-0.1	0.6	-0.02
Standard deviation (Hz)	4.7	2.5	28.5	4.2
Relative occurence (%)	11.4	87.4	1.2	100

Table 1: Mean and standard deviation of the pitch prediction error obtained by the proposed method.

Figure 2: Distribution of the observed pitch values together with the mean for each estimated class: training with normalization.

In order to illustrate the prediction capacities of the method presented here, it is tested on a set of utterances, which are excluded from the learning database. In the experiment, the prediction algorithm is only applied to voiced frames. The result obtained on an utterance of the test database is shown in figure 3. It appears that the observed and estimated pitch contours are quite similar, although a noticeable error occurs in the neighbourhood of frame 40. To get a more precise idea of the algorithm behaviour, the error between the observed and predicted pitch values is also calculated. Table 1 shows the results obtained globally, as well as on three pitch intervals. Globally, the prediction error has a mean of 0.02 Hz and a standard deviation of 4.2 Hz. Moreover, considering the 150-250 Hz pitch region, which contains 87.4% of the analyzed data, this standard deviation is even reduced to 2.5 Hz. On the contrary, for higher pitch values it increases up to 28.5 Hz. This degraded performance can be explained by the lack of speech frames having such extreme pitch values. Indeed, few data are assigned to classes whose mean pitch is high and thus, the parameters of their mixture components might be badly estimated.
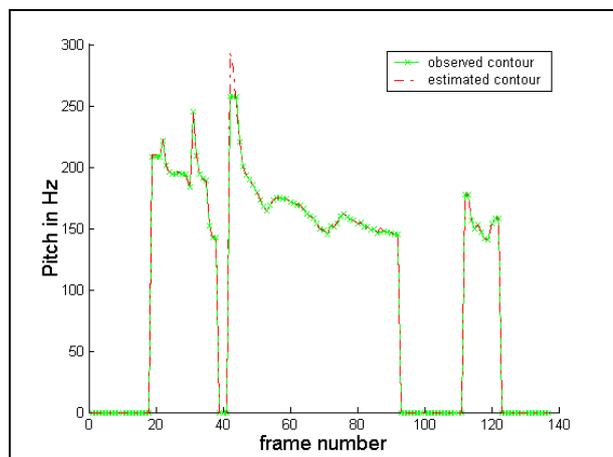

**Figure 3:** *Observed and estimated pitch contours for the utterance: "Quatre-vingt-dix".*

4. Application in voice conversion

This section presents the application of the proposed prediction algorithm in the framework of VC. Figure 4 presents the overall architecture of a VC system enabling spectral transformation as well as fundamental frequency modification according to the converted spectral envelope.

First, the source speech signal is analyzed, in order to get a set of parameters that can be directly used by the VC system. However, it must be noted that very often this analysis module is not part of the conversion system itself. For example, in the case of TTS synthesis systems, the analysis is done off-line and so, the parameters that will be subject to transformation are stored with all the information needed by the speech synthesizer.

The VC itself starts with the transformation of the spectral envelope. In the literature, the conversion function is implemented using a variety of techniques, such as vector quantization with mapping codebooks [1], dynamic frequency warping [2], speaker interpolation [3], or GMM [4, 5]. Comparative tests carried out in [13] show that conversion by GMM is that which offers the best results. Thus, GMM based

spectral conversion is used in the current VC system here with an implementation similar to [5].
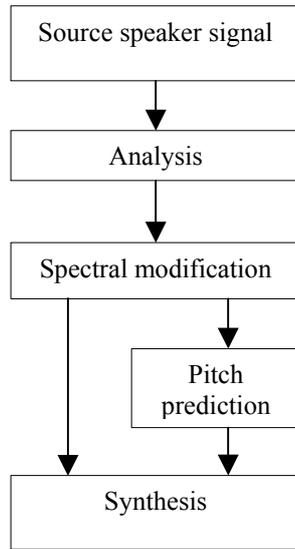
Figure 4: Voice conversion system associating spectral envelope and pitch modification

The next step of the VC process concerns the application of the proposed algorithm in order to predict a target fundamental frequency from the converted spectral coefficients. As already mentioned in the description of the learning stage, the pitch predictor is only applied to voiced speech frames. These estimated pitch values can then be used as prosodic contours for the VC system.

Finally, a speech synthesis algorithm receives this transformed spectral and pitch information and generates an output converted speech signal. Amongst existing techniques, HNM synthesis is particularly well suited to such a conversion process as it is based on a parametric modelling and coding scheme of the speech signal, where pitch and spectral envelope information is stored and is thus directly available for synthesis.

5. Conclusion

This paper proposes a new GMM-based method enabling an accurate prediction of the pitch from spectral information. Preliminary results are promising as the pitch prediction error has a standard deviation of 4.2 Hz. The integration of the above method in a VC application is also described. Further experiments and more specifically subjective tests will be done in the near future so as to attest the effectiveness of the proposed pitch conversion function when combined in a complete VC system.

6. References

[1] Mr. 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, 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 Ceyssens, W Verhelst and P. Wambacq, "One the construction of a pitch conversion system", Proceedings of EUSIPCO, 2002.

[7] A.K. Syrdal and S.A. Steele,"Vowel F1 have has function of announcer fundamental frequency ", 110th Meeting of JASA, flight. 78, Fall 1985.

[8] K Tanaka and Mr. Abe, " A new fundamental frequency modification algorithm with transformation of spectrum envelope according to F_0", Proceedings of IEEE ICASSP, vol.2, 1997.

[9] A. Kain and Y. Stylianou, "Spectral Stochastic modeling of adjustment for high quality pitch modification", Proceedings of IEEE ICASSP, 2000.

[10] O Cappé, J Laroche and E Moulines, "Regularized estimate of cepstrum envelope from discrete frequency points", IEEE ASSP Workshop one application of signal processing to audio and acoustics, Mohong, 1995.

[11] 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.

[12] 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.

[13] G. Baudoin and Y. Stylianou, "On the transformation of the speech spectrum for voice conversion", Proceedings of ICSLP, 1996.