Pitch and Duration Transformation with Non-parallel Data

Damien Lolive, Nelly Barbot, Olivier Boeffard

IRISA / University of Rennes 1 - ENSSAT
6 rue de Kerampont, B.P. 80518, F-22305 Lannion Cedex
France
{damien.lolive,nelly.barbot,olivier.boeffard}@irisa.fr

Abstract

In a voice transformation context, prosody transformation using parallel corpora is quite unrealistic as such corpora are difficult and also expensive to build. Based on this observation, we propose an approach for transforming prosody using non-parallel corpora thanks to the MLLR adaptation strategy. This methodology is applied to the joint transformation of duration and F_0 at the syllable level. The source data are modelled by a GMM which is adapted to the target by applying a linear transformation to the mean vectors of the gaussian mixture. This methodology is applied to the conversion of duration and F_0 between two french speakers and is evaluated by cross validation between the models and the test datasets. Taking the target model as reference, the adaptation enables to make 80% of the path between source and target data.

1. Introduction

The goal of a voice transformation system is to modify utterances of a source speaker to be perceived as if they were spoken by a target speaker. During the last years, some technical domains as biometric identification or Text-To-Speech synthesis, have used the voice transformation methodology. Concerning biometric identification, prosody transformation and voice transformation in general may be used to test speaker verification and speaker identification systems. In the field of speech synthesis, voice transformation may have an high impact insofar as a speech unit corpus, describing a TTS voice, paired with a set of transformation functions substitutes the classical approach for which each voice needs a complete acoustic unit inventory.

A voice transformation system has to satisfy two main requirements: a transformation of the segmental acoustic features and the one of supra-segmental features. In this paper. We focus here on prosody transformation and more particularly on the duration and the fundamental frequency, F_0. Usually, such a transformation system can be decomposed into three stages: stylization, classification and then transformation. In the literature, a great amount of recent works deals with prosody transformation and more particularly F_0 [1, 2]. A standard approach consists in modifying the F_0 by applying a linear or polynomial transformation which is based on global parameters of the source and the target voices [3, 4]. Some other approaches decompose that complex transformation problem into subproblems doing a partition of the feature space, as done for example by the *codebook* solution [3, 5].

It is necessary, in classical voice conversion systems, to have, for each sentence, one example from the source speaker and another one from the target speaker. As a consequence, two parallel corpora have to be used which is a restrictive hypothesis, not always applicable according to the desired application. Relaxing this constraint may soften the design of applications. A possible answer to this problem, in the case of parametric models, can be found in model adaptation via a speaker adaptation methodology as MLLR, Maximum Likelihood Linear Regression, [6]. In speech synthesis, *Tamura & al.*, [7], jointly model mel-cesptral coefficients, F_0 and duration using multi-stream MSD-HMM. Their goal is to obtain speaker dependent phoneme models by MLLR adaptation of speaker independant models.

In this paper, we propose a methodology for prosody transformation using non-parallel corpora. We consider prosodic information at the syllable level. The underlying idea is that a melodic sentence can be decomposed into smaller elements which can be put together to build a complete melodic sentence [8]. Considering this hypothesis, prosody conversion can be done based on transformed syllable sequences. For a syllable, duration and F_0 are represented by vector of constant size. We have chosen to represent a speaker melodic space by a GMM. Next, the adaptation of source GMM parameters with the target data is done by applying a MLLR methodology. A transformation function of prosodic feature vectors based on the adapted GMM parameters is also proposed. A transformed vector is computed as a weighted sum of the adapted GMM centroids of the source speaker. The weighting coefficients are the *a posteriori* probabilities of a GMM component given the observed source vector. This approach has been already proposed for spectral conversion and has shown a higher transformation efficiency than the mapping codebook approach [9].

The paper is organized as follows. First, the duration and F_0 models are presented in section 2. Section 3 details the GMM modelling as well as the adaptation methodology to the target speaker voice. The transformation function is also detailed in this section. Then, the experimental methodology is introduced in section 4. Finally, the results are given and discussed in section 5.

2. Data pre-processing

2.1. Interpolation and smoothing

Sentence level F_0 contours are pre-processed in a similar way to the one proposed in [10]. First, an interpolation is done to eliminate unvoiced parts of each F_0 contour. This interpolation follows the hypothesis according to which a continuous melodic gesture exists, the fundamental frequency value is then masked during unvoiced parts. Moreover, the F_0 contours obtained after interpolation are smoothed using a cubic spline in order to suppress micro-melodic variations.

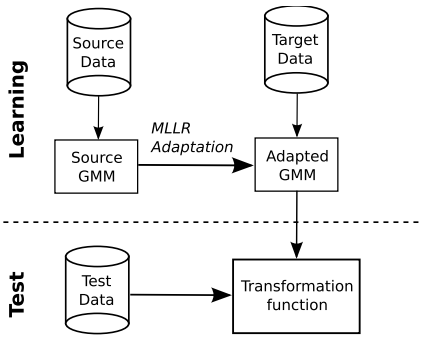

Figure 1: Architecture of the conversion system with source GMM adaptation to target data.

2.2. Duration representation

To represent duration, we used the syllable structure of the utterance. A syllable can be split into three parts: onset, nucleus and coda. Onset and coda parts may be empty. From this structure, a vector $\mathbf{D} = (d_{\text{onset}}, d_{\text{nucleus}}, d_{\text{coda}})$ is built to characterize the repartition of duration at a syllabic level. The duration of each part is computed as a multiple of 10 ms.

2.3. F_0 stylization

A F_0 contour on a syllable basis is represented by the triplet $\mathbf{F} = (F_0^{10\%}, F_0^{50\%}, F_0^{90\%})$. Each coordinate of that vector is the F_0 value located respectively at 10%, 50% and 90% of the time support. This process is just like normalizing the F_0 contour duration with respect to a single time support for all contours, as it is done in [11]. This normalization is a simple method to eliminate the F_0 contour length variations while and allowing comparison of the different shapes.

2.4. Syllable prosody

The prosody of a syllable (F_0 and duration) is a dimension 6 vector: $\mathbf{x} = (d_{\text{onset}}, d_{\text{nucleus}}, d_{\text{coda}}, F_0^{10\%}, F_0^{50\%}, F_0^{90\%})$. This vector enables the joint transformation of F_0 and duration. It is based on the structure of a syllable (for duration), while F_0 is represented in an arbitrary way. Consequently, other stylizations of syllable prosody are conceivable within this framework.

3. Prosody transformation

The goal of our approach is to transform prosody between source and target speakers without parallel corpora. GMM are used to model source and target vectors and the conversion function is based on the source/target adapted GMM parameters. Figure 1 illustrates this methodology. The first step consists in learning a GMM on data from the source speaker and then adapting the parameters of the source model using data from the target speaker. The second step of the prosody conversion system deals with the use of the models to transform F_0 and duration.

3.1. GMM modelling

For a speaker, we consider the set $\mathbf{X}$ of the vectors $\mathbf{x}$ representing the prosody of each syllable. A GMM $\mathcal{M}_{\mathbf{X}}$ with M gaussians is chosen to model the dataset $\mathbf{X}$: its probability distribution is given by

$$P(\mathbf{x}|\mathbf{\Theta}) = \sum_{m=1}^{M} \alpha_m P(\mathbf{x}|\theta_m)$$

where $\mathbf{\Theta} = (\alpha_1, \ldots, \alpha_m, \theta_1, \ldots, \theta_M)$ is the set of parameters. α_m is the mixing coefficient associated to the mth gaussian with parameters $\theta_m = (\mu_m, \Sigma_m)$ and distribution $P(\mathbf{x}|\theta_m)$.

The EM algorithm, implemented to learn the GMM parameters, is an iterative algorithm whose goal is to maximize the loglikelihood of the data and the model, [12].

3.2. GMM Adaptation

The MLLR adaptation (Maximum Likelihood Linear Regression) method is proposed in [6, 13]. Let us consider a GMM $\mathcal{M}_{\mathbf{X}}$ with parameters $\mathbf{\Theta} = (\alpha, \mu, \mathbf{\Sigma})$ learnt on a dataset $\mathbf{X}$. The goal is to adapt the parameters of the GMM $\mathcal{M}_{\mathbf{X}}$ thanks to a new dataset $\mathbf{Y}$ by computing a linear transformation on the parameters μ_m and $\mathbf{\Sigma}_m$ for each gaussian m and maximizing the likelihood of the adapted GMM on the new dataset:

$$\begin{aligned} \hat{\mu}_m &= \mathbf{A}_m \mu_m + \mathbf{b}_m \\ \hat{\mathbf{\Sigma}}_m &= \mathbf{\Sigma}_m \end{aligned}$$

where $\mathbf{A}_m$ is a full transformation matrix. In this paper, we use MLLR adaptation to adapt only the gaussian means.

The MLLR approach consists in finding a set of transformation matrices which, when applied to the gaussian means, maximize the likelihood of the adaptation data. Thus, a new estimate of the mean, $\hat{\mu}_m$, of the gaussian m is found by:

$$\hat{\mu}_m = \widehat{\mathbf{W}}_m \xi_m$$

where $\widehat{\mathbf{W}}_m$ is the adaptation matrix of size $n \times (n+1)$ (n is the size of the dataset) and $\xi_m = [1\, \mu_1 \ldots \mu_n]$ is the extended mean vector. We thus have:

$$\widehat{\mathbf{W}}_m = [\hat{\mathbf{b}}_m\ \hat{\mathbf{A}}_m]$$

Estimating $\widehat{\mathbf{W}}_m$ is achieved by applying the EM algorithm with the adaptation dataset.

3.3. Pitch and duration transformation

The source voice prosody has to be modified in order to make it similar to the target voice prosody. We can define a transformation function in the following way:

$$\mathbf{x}' = \mathcal{F}(\mathbf{x}, \mathcal{M}_{\mathbf{X}}) = \sum_m P(m|\mathbf{x}) \mu_m \qquad (1)$$

where $P(m|\mathbf{x})$ is the probability that $\mathbf{x}$ belongs to the class m and μ_m is the mean of the gaussian m from GMM $\mathcal{M}_{\mathbf{X}}$.

In particular, to transform the source vector to the target vector space, we use the adapted GMM $\widehat{\mathcal{M}_{\mathbf{X}}}$. The transformation function can be written using the adaptation matrix computed with the MLLR:

$$\mathcal{F}(x, \widehat{\mathcal{M}_{\mathbf{X}}}) = \sum_m P(m|\mathbf{x}) \hat{\mu}_m = \sum_m P(m|\mathbf{x}) (\widehat{\mathbf{W}}_m \xi_m) \quad (2)$$

where ξ_m is the extended mean vector of class m from the source GMM $\mathcal{M}_{\mathbf{X}}$, $\widehat{\mathbf{W}}_m$ is the adaptation matrix for this gaussian.

Table 1: Loglikelihood for $\mathcal{M}_{\mathbf{X}}$ and $\mathcal{M}_{\mathbf{Y}}$ and results for the adaptation of $\mathcal{M}_{\mathbf{X}}$ using $\mathbf{Y}$ with 95% confidence intervals on learning sets.

	Training	Validation
$\mathcal{M}_{\mathbf{X}}$	-18.00 ± 0.09	-18.05 ± 0.19
$\mathcal{M}_{\mathbf{Y}}$	-17.84 ± 0.10	-17.97 ± 0.21
$\widehat{\mathcal{M}_{\mathbf{X}}}$	-18.14 ± 0.10	-18.28 ± 0.21

Table 2: Loglikelihood values for the source $\mathcal{M}_{\mathbf{X}}$, target $\mathcal{M}_{\mathbf{Y}}$ and adapted $\widehat{\mathcal{M}_{\mathbf{X}}}$ GMM on the source $\mathbf{X}$ and target $\mathbf{Y}$ validation data.

	X	Y
$\mathcal{M}_{\mathbf{X}}$	-18.05 ± 0.19	-19.73 ± 0.23
$\mathcal{M}_{\mathbf{Y}}$	-19.07 ± 0.19	-17.97 ± 0.21
$\widehat{\mathcal{M}_{\mathbf{X}}}$	-18.88 ± 0.22	-18.28 ± 0.21

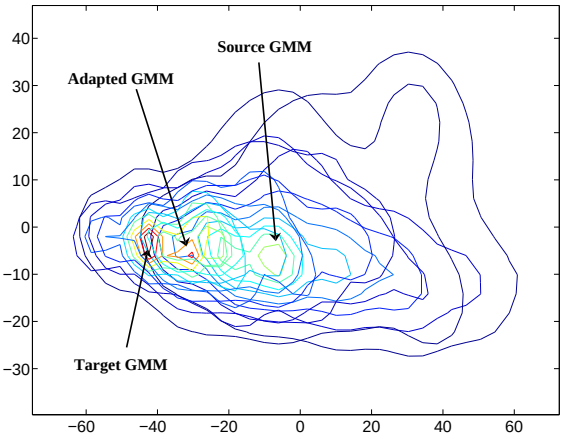

Figure 2: Projection of the source, target and adapted GMM probability densities. The projection is realized on the source and target data mean space.

4. Experimental protocol

4.1. Data

For the experiments, two corpora are used. The source voice is a read speech corpus used in a TTS system. The target voice comes from the french Ester corpus and corresponds to broadcast news. We have chosen the speaker named "Simon Tivolle". For both corpora, a segmentation process into phones is carried out in an automatic manner. The mean fundamental frequency, F_0, has been analyzed automatically with the help of the autocorrelation function. Each phonetic sequence is segmented into syllables. The F_0 contours are interpolated and then smoothed before representing the prosody as detailed in section 2.

For each voice, 5000 syllables are used for training the GMM and 1500 syllables for the validation. The prosody transformation between the two selected voices is not easy. Indeed, the source voice is read speech with a relatively smooth prosody whereas the second voice contains a more diversified prosody which is particular to the journalistic style.

4.2. Experiments

The GMM used for the experiments have 32 gaussians with diagonal covariance matrices. The adaptation and the F_0 and duration transformation are difficult to evaluate. Indeed, in the framework of prosody transformation using non parallel corpora, a perceptual test is clearly difficult unless having parallel sub-corpus dedicated to the evaluation. As we do not have parallel corpora, we propose to evaluate the methodology by cross validation on the loglikelihood between the datasets and the models. We consider the source $\mathcal{M}_{\mathbf{X}}$ and target $\mathcal{M}_{\mathbf{Y}}$ GMM respectively trained on source and target data. We also consider the adapted GMM $\widehat{\mathcal{M}_{\mathbf{X}}}$ trained from $\mathcal{M}_{\mathbf{X}}$ and adapted thanks to the target data. Moreover, we consider the transformation of the source data using the adapted model, described in (2). In order to evaluate the quality of these new data with respect to the target data, it is necessary to evaluate first the impact of the transformation function (1) on the initial data.

5. Results and discussion

In table 1, we can observe the loglikelihood values for the GMM training and the adaptation of the source GMM from the source data to the target data. When the source GMM is learnt on the source data, we can notice that the loglikelihood values on the training and the validation data are close together. The same comment is valid for the target GMM and the adapted GMM considering the target data. In particular, these results mean that the models are representative of the data and there is no overfitting effect.

Table 2 shows a comparison of the loglikelihood values for the source $\mathcal{M}_{\mathbf{X}}$, target $\mathcal{M}_{\mathbf{Y}}$ and adapted $\widehat{\mathcal{M}_{\mathbf{X}}}$ GMM respectively on the source $\mathbf{X}$ and target $\mathbf{Y}$ datasets. By analyzing this table line by line, we can notice that, for $\mathcal{M}_{\mathbf{X}}$, the loglikelihood is better on the $\mathbf{X}$ dataset than on the $\mathbf{Y}$ dataset. For $\mathcal{M}_{\mathbf{Y}}$, the inverse phenomenon happens. It points out the fact that $\mathbf{X}$ and $\mathbf{Y}$ have different distributions of values. Moreover, the GMM $\widehat{\mathcal{M}_{\mathbf{X}}}$ has a higher loglikelihood value on the target dataset, $\mathbf{Y}$, than on the source dataset, $\mathbf{X}$. These results show that the adapted GMM $\widehat{\mathcal{M}_{\mathbf{X}}}$ is closer to the distribution of $\mathbf{Y}$ than the one of $\mathbf{X}$. The MLLR adaptation process moved the distributions of the gaussians of GMM $\mathcal{M}_{\mathbf{X}}$ towards those of GMM $\mathcal{M}_{\mathbf{Y}}$. This movement can be observed on figure 2. This figure shows clearly that the adapted GMM is closer to the target GMM than the source GMM.

Let us consider $\mathbf{X}'$, $\mathbf{Y}'$, $\mathbf{Z}'$ three new datasets obtained by the transformation of respectively:

- source data with the source GMM $\mathcal{M}_{\mathbf{X}}$, by applying (1),
- target data with the target GMM $\mathcal{M}_{\mathbf{Y}}$, by applying (1),
- source data with the adapted GMM $\widehat{\mathcal{M}_{\mathbf{X}}}$, by applying (2).

From these new transformed datasets, three more GMM can be trained: $\mathcal{M}_{\mathbf{X}'}$, $\mathcal{M}_{\mathbf{Y}'}$, $\mathcal{M}_{\mathbf{Z}'}$, respectively using $\mathbf{X}'$, $\mathbf{Y}'$ and $\mathbf{Z}'$.

We can now take a look at the transformation function behavior thanks to the results summarized in table 3. The results concerning the three GMM trained from $\mathbf{X}'$, $\mathbf{Y}'$, $\mathbf{Z}'$ are presented in this table. In the first part of it, we can notice that the GMM $\mathcal{M}_{\mathbf{X}'}$, $\mathcal{M}_{\mathbf{Y}'}$, $\mathcal{M}_{\mathbf{Z}'}$ give bad results on $\mathbf{X}$ et $\mathbf{Y}$. One can explain this phenomenon by the fact that the transformation function shows a tendency to project the data near the gaussian means. Indeed, equation (1) indicates that the transformed

Table 3: Comparison of loglikelihood values for the GMM $\mathcal{M}_{\mathbf{X}'}$, $\mathcal{M}_{\mathbf{Y}'}$ and $\mathcal{M}_{\mathbf{Z}'}$ on original and transformed validation data.

	$\mathbf{X}$	$\mathbf{Y}$	$\mathbf{X}'$	$\mathbf{Y}'$	$\mathbf{Z}'$
$\mathcal{M}_{\mathbf{X}'}$	-28.04 ± 0.89	-31.58 ± 0.99	-5.12 ± 0.55	-23.92 ± 0.38	-23.18 ± 0.39
$\mathcal{M}_{\mathbf{Y}'}$	-30.21 ± 1.04	-29.02 ± 0.73	-20.70 ± 0.59	-8.89 ± 0.43	-14.77 ± 0.33
$\mathcal{M}_{\mathbf{Z}'}$	-27.43 ± 0.65	-27.34 ± 0.38	-20.05 ± 0.44	-17.96 ± 0.31	-4.93 ± 0.59

Table 4: Comparison of loglikelihood values for the GMM $\mathcal{M}_{\mathbf{X}}$, $\mathcal{M}_{\mathbf{Y}}$ and $\widehat{\mathcal{M}_{\mathbf{X}}}$ on transformed validation data.

	$\mathbf{X}'$	$\mathbf{Y}'$	$\mathbf{Z}'$
$\mathcal{M}_{\mathbf{X}}$	-15.52 ± 0.15	-16.87 ± 0.18	-16.47 ± 0.17
$\mathcal{M}_{\mathbf{Y}}$	-17.41 ± 0.19	-15.46 ± 0.17	-15.84 ± 0.16
$\widehat{\mathcal{M}_{\mathbf{X}}}$	-16.85 ± 0.20	-15.88 ± 0.16	-15.49 ± 0.14

value is equal to the sum of the gaussian means weighted by a class membership probability. More precisely, the transformed data are located within the convex envelop formed by the GMM gaussian means. Then, the transformed data variance is uniquely related to this class membership probability. From that, the transformed data have a lower variance than the original data. The right part of table 3 confirms this comment by showing high differences between the three datasets $\mathbf{X}'$, $\mathbf{Y}'$, $\mathbf{Z}'$ for a fixed GMM. For each line, we find the same ranking for the models as the one established in table 4.

In order to cope with the lack of variability of the transformed data, it would be interesting to adapt not only the gaussian means but also their variance and integrate it into the transformation function.

Thanks to these three transformed datasets (transformed by functions of the same kind), we will be able to specify the behavior of the adaptation process. In table 4, we can notice that the results obtained on the three transformed datasets are of the same order for GMM $\mathcal{M}_{\mathbf{Y}}$ and $\widehat{\mathcal{M}_{\mathbf{X}}}$. These two GMM give better results on $\mathbf{Y}'$ and $\mathbf{Z}'$ than on $\mathbf{X}'$. GMM $\mathcal{M}_{\mathbf{Y}}$ is better on the adapted data $\mathbf{Z}'$ than on the source data $\mathbf{X}'$. This result shows that the adapted data (source data transformed by the adapted GMM) have their distribution closer to $\mathbf{Y}'$ than $\mathbf{X}'$. The same observation holds for $\widehat{\mathcal{M}_{\mathbf{X}}}$. Therefore, we can conclude that the adaptation of the source GMM to the target data effectively enables the estimation of a transformation function between source and target.

6. Conclusions

A methodology to answer the problem of prosody transformation using non parallel corpora is presented. Duration and F_0, at a syllabic level, are represented by a vector of constant size. A GMM is estimated on the duration and F_0 observations for a source speaker. We propose to apply first a MLLR approach to adapt the source parameters with respect to the target voice data. Next, we present a model to linearly transform a prosodic source vector. Using this model, the adapted centroids are weighted by the a posteriori distribution of the source vectors.

The experimental protocol is essentially based on cross validation between the models and the test datasets. An exhaustive comparison between data and models shows that: on the one hand, the adapted GMM efficiently models the target data and on the other hand, the transformation function produces data as likely for the target data as the target data themselves. Taking the target model as a reference, the adaptation enables to make 80% of the path between source and target data.

The evaluation of a non parallel method is difficult and the use of two parallel corpora for evaluation purposes is helpful. The proposed methodology is entirely non parallel as we do not have yet two parallel evaluation corpora. Further experiments are planned to confront this method to well-tried methods.

7. References

[1] Z. Inanoglu, "Transforming pitch in a voice conversion framework," St. Edmond's College, University of Cambridge, Tech. Rep., July 2003.

[2] B. Gillett and S. King, "Transforming f0 contours," in *Proc. of Eurospeech Conference*, 2003, pp. 1713–1716.

[3] D. T. Chappell and J. H. L. Hansen, "Speaker-specific pitch contour modeling and modification," in *Proc. of ICASSP*, vol. 2, 1998, pp. 885–888.

[4] T. Ceyssens, W. Verhelst, and P. Wambacq, "On the construction of a pitch conversion system," in *Proc. of EUSIPCO*, 2002, pp. 1301–1304.

[5] E. Helander and J. Nurminen, "A novel method for prosody prediction in voice conversion," in *Proc. of ICASSP*, vol. 4, 2007, pp. 509–512.

[6] C. J. Leggetter and P. C. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression," in *Proc. of Eurospeech Conference*, 1995, pp. 1155–1158.

[7] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for hmm-based speech synthesis using mllr," in *Proc. of ICASSP*, vol. 2, 2001, pp. 805–808.

[8] P. Mertens, "Automatic recognition of intonation in french and dutch," in *Proc. of Eurospeech conference*, 1989, pp. 46–50.

[9] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 131–142, 1998.

[10] Y. Yamashita, T. Ishida, and K. Shimadera, "A Stochastic F0 Contour Model Based on Clustering and a Probabilistic Measure," *IEICE Transactions on Information and Systems*, vol. E86-D, no. 3, pp. 543–549, 2003.

[11] U. D. Reichel, "Data-driven extraction of intonation contour classes," in *Proc. of the 6th ISCA Workshop on Speech Synthesis*, 2007, pp. 240–245.

[12] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," International Computer Science Institute, Tech. Rep., 1998.

[13] M. Gales and P. Woodland, "Mean and variance adaptation within the mllr framework," *Computer Speech & Language*, vol. 10, pp. 249–264, 1996.