OBJECTIVE EVALUATION OF THE DYNAMIC MODEL SELECTION METHOD FOR SPECTRAL VOICE CONVERSION

Pierre Lanchantin, Xavier Rodet

IRCAM - CNRS-UMR9912-STMS,
Analysis-Synthesis Team,
1, place Igor-Stravinsky,
75004 Paris, France
lanchantin@ircam.fr, rod@ircam.fr

ABSTRACT

Spectral voice conversion is usually performed using a single model selected in order to represent a tradeoff between goodness of fit and complexity. Recently, we proposed a new method for spectral voice conversion, called *Dynamic Model Selection* (DMS), in which we assumed that the model topology may change over time, depending on the source acoustic features. In this method a set of models with increasing complexity is considered during the conversion of a source speech signal into a target speech signal. During the conversion, the best model is dynamically selected among the models in the set, according to the acoustical features of each source frame. In this paper, we present an objective evaluation demonstrating that this new method improves the conversion by reducing the transformation error compared to methods based on a single model.

Index Terms— Voice conversion, Gaussian Mixture Regression, model selection.

1. INTRODUCTION

The aim of *speaker conversion* - a typical application of *voice conversion* technique (VC) - is to modify the speech signal of a source speaker to be perceived as if it had been uttered by a target speaker [1]. The overall methodology for speaker conversion is to define and learn a mapping function of acoustic features of a source speaker to those of a target speaker. Several approaches have been proposed such as vector quantization [2], neural networks [3] or multivariate linear regression [4] among others statistical methods [5]. One of the most popular statistical method, proposed by Stylianou and al. [6], is based on a *Gaussian mixture model* (GMM) that defines a continuous mapping between the features of source and target voices. The comparative study [7] suggests a better performance of this method compared to vector quantization, neural networks and multiple linear regression. Although this type of method is relatively efficient, conversion performance are still insufficient regarding speech quality: the frame by frame conversion process induces inappropriate spectral parameter trajectories and the converted spectrum can be excessively smoothed. Toda and al. have recently proposed in [8] a method based on maximum likelihood estimation of a parameter trajectory, which greatly improves the quality of synthesis by taking into account the dynamic features and the global variance.

In most statistical methods of speaker conversion, a single model is used for the conversion. This model is selected, among others during the training phase, according to the spectral distortion obtained from the conversion of a development corpus or to an informational criterion such as the Bayesian information criterion (BIC [9]). Recently, assuming that the best model may change over time according to the source acoustic features, we proposed in [10] a new method for spectral conversion called *Dynamic Model Selection* (DMS) based on the use of several models in parallel. At each frame of the source vector, the most appropriate model is selected according to the values of the acoustic features: if the values are far from training data, low complexity general model is selected for the conversion. However, if the source data are close to training data, a more complex and precise model is selected leading to a more accurate conversion. Subjective tests were performed and showed that the method is promising as it can improve the conversion in terms of proximity to the target and quality compared to the method based on a single model. In this paper, we continue this work by presenting an objective evaluation of the method according to the performance index defined by Kain in [16].

The paper is organized as follows: Section 2 presents the proposed approach and the voice conversion system is described in Section 3; finally objective evaluation of the method is presented and discussed in Sections 4 and 5

This study was supported by FEDER Angelstudio : Générateur d'Avatars personnalisés ; 2009-2011

2. PROPOSED APPROACH

In the next subsection, we introduce the Gaussian mixture modeling framework for spectral conversion. We define a set of models with increasing number of components from which we can select the best model to perform the conversion of each source frame. This DMS method is finally presented in subsection 2.2.

2.1. Spectral conversion with Gaussian mixture models

Stylianou and al. [6] proposed to model the source speaker acoustic probability space with a GMM. The cross-covariance of the target speaker with source speaker and the mean of the target speaker were then estimated using least squares optimization of an overdetermined set of linear equations. Kain extended Stylianou's work by modeling directly the joint probability density of the source and target speaker's acoustic space [11]. This method allows the system to capture all the existing correlations between the source and target speaker's acoustic features. We briefly describe the method in the following.

Let $Z = (X, Y)$ be a joint random process in which $X = \{X_n\}_{n\in\mathcal{N}}$ and $Y = \{Y_n\}_{n\in\mathcal{N}}$ are the source and target acoustic features processes respectively, and $\mathcal{N}$ the set of frame indexes. Each X_n and Y_n takes its values in $\mathbb{R}^d$ where d is the dimension of the acoustic feature vector. We will denote $z = (x, y) = \{(x_n, y_n)\}_{n\in\mathcal{N}}$ a realization of this process, in which x_n and y_n are the acoustic features vector at frame n for the source and that for the target, respectively. We assume that Z is an independent and identically distributed process (i.i.d.) such as $p(z) = \prod_{n\in\mathcal{N}} p(z_n)$. We introduce the auxiliary i.i.d process of mixture components $U = \{U_n\}_{n\in\mathcal{N}}$, each U_n taking its values in $\mathcal{U}$ with cardinal K. The joint probability density of the source and target feature vectors is then modeled by a GMM as follows

$$p(z_n|\phi) = \sum_{k=1}^{K} p(u_n = k)p(z_n|u_n = k, \phi_k) \tag{1}$$

with $p(z_n|u_n = k, \phi_k) = \mathcal{N}(z_n; \bar{\mu}_k^z, \Sigma_k^z)$. ϕ is the GMM parameters set which consists of the weight $p(u_n = k)$, the mean vector $\bar{\mu}_k^z = \begin{bmatrix} \bar{\mu}_k^x \\ \bar{\mu}_k^y \end{bmatrix}$ and the covariance matrix $\Sigma_k^z = \begin{bmatrix} \Sigma_k^{xx} & \Sigma_k^{xy} \\ \Sigma_k^{yx} & \Sigma_k^{yy} \end{bmatrix}$ for all mixture components $k \in \mathcal{U}$. $\bar{\mu}_k^x$ and $\bar{\mu}_k^y$ are the mean vector of the k-th mixture component for the source and that for the target, respectively. Σ_k^{xx} and Σ_k^{yy} are the covariance matrix of the k-th mixture component for the source and that for the target, respectively. Σ_k^{xy} and Σ_k^{yx} are the cross-covariance matrix of the k-th mixture component for the source and that for the target, respectively. ϕ is estimated by Expectation-Maximization on a parallel corpus $z = (x, y)$ in which x and y have been temporally aligned.

The conditional probability density of y_n given x_n is also a GMM as follows :

$$p(y_n|x_n; \phi) = \sum_{k=1}^{K} p(u_n = k|x_n; \phi_k)p(y_n|x_n, u_n = k; \phi_k) \tag{2}$$

where

$$\begin{cases} p(u_n = k|x_n; \phi_k) \propto p(u_n = k)\mathcal{N}(x_n; \bar{\mu}_k^x, \Sigma_k^{xx}) \\ p(y_n|x_n, u_n = k; \phi_k) = \mathcal{N}(y_n; E_{k,n}^y, C_k^y) \end{cases} \tag{3}$$

with

$$\begin{cases} E_{k,n}^y = \bar{\mu}_k^y + \Sigma_k^{yx} \left(\Sigma_k^{xx}\right)^{-1} (x_n - \bar{\mu}_k^x) \\ C_k^y = \Sigma_k^{yy} - \Sigma_k^{yx} \left(\Sigma_k^{xx}\right)^{-1} \Sigma_k^{xy} \end{cases} \tag{4}$$

In each mixture component $k \in \mathcal{U}$, the conditional target mean vector $E_{k,n}^y$ for the given source acoustic feature vector is calculated by a simple linear conversion based on the correlation between the source and target acoustic feature vector as shown in equation (4).

The conversion is finally performed on the basis of the minimum mean-square error (MMSE) : the converted feature vector is the weighted sum of the conditional mean vectors in which the weights are the posterior probabilities of the source acoustic feature vector belonging to each one of the mixture components:

$$\hat{y}_n = E[y_n|x_n] = \sum_{k=1}^{K} p(u_n = k|x_n; \phi_k)E_{k,n}^y \tag{5}$$

The conditional covariance matrix can also be evaluated, giving a kind of confidence measure for the conditional mean vector for each $n \in \mathcal{N}$

$$C[y_n|x_n] = \sum_{k=1}^{K} p(u_n = k|x_n; \phi_k)^2 C_k^y \tag{6}$$

2.2. Dynamic model selection

In classical speaker conversion methods, a single model is selected during the training step and used for the conversion. This model is selected among others according to the spectral distortion obtained from the conversion of a development corpus or by using methods from the models selection research field. Information Criteria such as BIC [9] have been designed for this purpose. A good model will balance goodness of fit and complexity, so it should have neither a very low bias nor avery low variance. A model with too large a variance due to overparametrization will give poor performance on datas different or far from the training datas because of the high variance of the local estimators resulting in overfitting. The model undergoes oscillations that are both very large and whose features strongly depend on the exact positions of the points leading to a model with a huge variance and very large

response errors. However, the same model will give excellent conversion performances on datas similar or close to the training ones

The *Dynamic Model Selection* (DMS) method that we initially proposed in [10] consists of using several models in parallel assuming that the best model may change over time according to the source acoustic features. To do so, a set of potential best models $\mathcal{M}$ including GMMs with increasing number of components is built during the training step. During the conversion step, at each frame $n \in \mathcal{N}$, the most appropriate model is chosen according to the likelihood of the source datas given each model as

$$\hat{M}_n = \arg\max_{M \in \mathcal{M}} p(x_n|M) \tag{7}$$

with

$$p(x_n|M) = \sum_{k=1}^{K} p(u_n = k)\mathcal{N}(x_n; \bar{\mu}_k^x, \Sigma_k^{xx}) \tag{8}$$

the values of $p(u_n)$ $\bar{\mu}_k^x$, Σ_k^{xx} and K, depending on the model M. In this way, we aim to use a general model with low complexity if the values are far from training datas and a more complex and precise model if the source datas are closer to training datas, leading to a better conversion.

3. VOICE CONVERSION SYSTEM

3.1. Corpuses description

Two corpuses have been recorded by French speakers with different accents (Standard and Hispanic). Each corpus includes 200 phonetically balanced utterances, which represent between 9 and 10 minutes of speech depending on speakers speech rate. The corpuses were recorded in an anechoic chamber (32 bits, Fe=48kHz). The recordings were downsampled to Fe=24kHz for the experiments. The corpuses were then aligned by a *Dynamic Time Warping* (DTW) algorithm constrained by phone boundaries extracted using *ircamAlign* [12], an HMM-based alignment system based on the HTK toolbox. The obtained parallel corpus was then splitted into a training part (180 parallel utterances) and a test part (20 parallel utterances).

To avoid alignment errors, such as matching between voiced and unvoiced segment - which can lead to a bad estimation of the joint probability density - a pre-rejecting method of mismatched source-target frames was applied: a two-class GMM model was first estimated for each corpus and couples $z_n = (x_n, y_n)$ for which $\arg\max_{u_n} p(u_n|x_n) \neq \arg\max_{u_n} p(u_n|y_n)$ were rejected from the training corpus as they were considered as mismatched frames. These frames represent a nearly 9% of both parallel training corpora. During our experiments, this rejecting method of poorly matched source-target data was found to improve quality in an informal listening test.

3.1.1. Spectral features

Line spectral frequencies (LSF) were used as spectral features vector for the source x_n and target y_n for each $n \in \mathcal{N}$. To do so, the spectral envelope was estimated each 2.5 ms by True Envelope method on a Mel scale (MTE) and coded by Linear Predictive Coding (LPC). The optimal order considering the MTELPC [13, 14] estimator is 30 according to $\hat{O}_{opt} = 0.15 * F_s/F_0$ [13]. Linear Spectral Frequencies (LSF) parametrization was chosen due to its good linear interpolation properties. Analysis and synthesis were done using phase vocoder [15].

4. OBJECTIVE EVALUATION

In this section, we aim to evaluate the DMS according to transformation error. To do so, several joint densities modeled by GMM with full covariance matrix with increasing number of components - from 1 to 32 - were estimated on the parallel corpus. We used the performance index proposed by Kain [16] defined as a ratio of two measures:

- the *trans-speaker* error which is the spectral distance between the converted speech and the target speech determining the proximity of the converted speech to the target speaker's one;
- the *inter-speaker* error which measures the spectral distance between the source and target speaker.

The performance index, denoted P in the following, is computed according to the following equation:

$$P = 1 - \sum_{n=1}^{N} \frac{d(y_n, \hat{y}_n)}{d(y_n, x_n)} \tag{9}$$

where $\hat{y}_n$ is the converted target vector and the distance d is Euclidean. P equal zero if the transformation error equals the inter-speaker error, and less than zero if the transformation error is even larger. In the opposite way, P approaches one as the transformation error approaches zero. Finally, the *intra-speaker* error, defined in [16] as a measure of how much variablity is present from one rendition to the next of the same sentence, approximates the lower bound of an achievable transformation and, then, the optimal value of P will be less that one. In our evaluation, source vectors of the test parallel corpus are first converted according to each GMM model and the performance indexes are computed according to equation (9). Then source vectors are converted using the DMS method with maximum likelihood criteria as defined in equation (7). The results of the evaluation are presented in the next section.

5. RESULTS & DISCUSSION

Performance indexes corresponding to each GMM model with different number of components are presented in plain

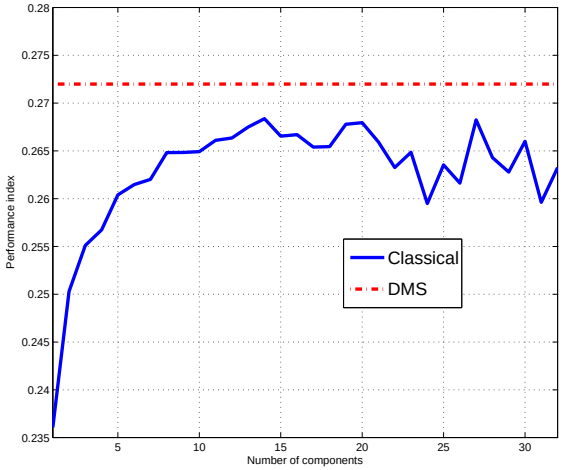

Fig. 1. Performance index for classical (plain line) and DMS (dash-dotted line) methods.

	Performance Index ($P(10^{-1})$)						
DMS_{opt}	**3.951**						
DMS_{ML}	**2.720**						
GMM_m	14	27	20	...	3	2	1
$P(10^{-1})$	**2.684**	2.682	2.680		2.551	2.503	2.361

Table 1. Performance indexes comparison between the different models and methods presented in this paper.

line in the Fig. 1 (referred as Classical method). They are also given in decreasing order in the lowest part of table 1. The best performance is obtained using the GMM with 14 components with $P = 0.2684$.

To evaluate the optimal performance index which could be reach using DMS with the given set of models - denoted as DMS_{opt} in Table 1 - we first selected models maximizing the performance index P at each frame. We obtain an average performance index $P = 0.3951$. Using the ML criterion defined in equation (7) - denoted as DMS_{ML} in table 1 and presented in dotted line in Fig. 1 - we obtain a performance index $P = 0.2720$ which represents an error reduction of 7.5% relative to the optimal performance, which can be considered as significant. However, the maximum likelihood criterion seems far from being optimal and better performance could be obtain using a more optimal criterion as suggested by the optimal performance value.

6. CONCLUSION & FURTHER WORK

We presented an objective evaluation of the recently proposed DMS method for spectral voice conversion. In this method, several models with increasing complexity are used in parallel and the model selection is achieved dynamically for each source frame according to its acoustical features. Objective tests show that the method can improve the conversion in terms of transformation error. Further work includes the definition of a more optimal selection criterion which could lead to better performance using the DMS method.

7. REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H Kuwabara, "Voice conversion through vector quantization," *J. Acoust. Soc. Jpn. (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[2] M. Abe, S. Nakamura, and H. Kawabara, "Voice conversion through vector quantization," in *Proc. International Conference on Acoustics Speech and Signal Processing (ICASSP'88)*, 1988, pp. 655–658.

[3] M. Narendranath, H. Murthy, S. Rajendran, and B. Yegnanarayan, "Transformation of formants for voice conversion using artificial neural networks," in *Speech Communication*, 1995, vol. 16, pp. 207–216.

[4] H. Valbret, E. Moulines, and J. Tubach, "Voice transformation using psola technique," in *Speech Communication*, 1992, number 11, pp. 175–187.

[5] H. Kuwabara and Y. Sagisaka, "Acoustic characteristics of speaker individuality: control and conversion," *Speech Communication*, vol. 16, no. 2, pp. 165–173, 1995.

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

[7] G. Baudoin and Y. Stylianou, "On the transformation of the speech spectrum for voice conversion," in *Proc. ICSLP '96*, Philadelpia USA, 1996, pp. 1404–1408.

[8] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[9] Y. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete Multivariate Analysis*, MIT Press Cambridge, 1975.

[10] P. Lanchantin and X. Rodet, "Dynamic Model Selection for Spectral Voice Conversion," in *Interspeech 2010*, Makuhari, Japan, September 2010.

[11] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. International Conference on Acoustics Speech and Signal Processing (ICASSP'88)*, 1998, pp. 285–288.

[12] P. Lanchantin, A.C. Morris, X. Rodet, and C. Veaux, "Automatic phoneme segmentation with relaxed textual constraints," in *LREC'08 Proceedings*, Marrakech, Morocco, 2008.

[13] F. Villavicencio, A. Röbel, and X Rodet, "Extending efficient spectral envelope modeling to mel-frequency based representation," in *IEEE 2008 International Conference on Acoustics, Speech, and Signal processing (ICASSP'08)*, 2008, pp. 1625–1628.

[14] F. Villavicencio, Röbel A., and X Rodet, "Applying improved spectral envelope modeling for high-quality voice conversion," in *IEEE's International Conference on Acoustics, Speech, and Signal processing (ICASSP'09)*, 2009.

[15] P. Depalle and G. Poirrot, "Svp: A modular system for analysis, processing and synthesis of sound signals," in *Proceedings of the International Computer Music Conference*, 1991.

[16] A. Kain, *High resolution voice transformation*, Ph.D. dissertation, OGI School of Science and Engineering at Oregon Health and Science University, 2001.