Dynamic Model Selection 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

Statistical methods for voice conversion are usually based on a single model selected in order to represent a tradeoff between goodness of fit and complexity. In this paper we assume that the best model may change over time, depending on the source acoustic features. We present a new method for spectral voice conversion[1] called Dynamic Model Selection (DMS), in which a set of potential best models with increasing complexity - including a mixture of Gaussian and probabilistic principal component analyzers - are considered during the conversion of a source speech signal into a target speech signal. This set is built during the learning phase, according to the Bayes information criterion (BIC). During the conversion, the best model is dynamically selected among the models in the set, according to the acoustical features of each source frame. Subjective tests show that the method improves the conversion in terms of proximity to the target and quality.

Index Terms: Voice conversion, model selection.

1. Introduction

A typical application of *voice conversion* technique (VC) is *speaker conversion*: the speech signal of a source speaker is modified to be perceived as if it had been uttered by a target speaker [1]. The overall methodology for speaker conversion is to first learn a mapping function of acoustic features of a source speaker to those of a target speaker. To learn this mapping function, 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 unique model is used for the conversion. This model is selected among others during the learning phase according to the spectral distortion obtained from the conversion of a test corpus or to an informational criterion such as the Bayesian information criterion (BIC[9]). In this paper, assuming that the best model may change over time according to the source acoustic features, we propose a new conversion method 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 datas, low complexity general model is selected for the conversion. However, if the source datas are close to training datas, a more complex and precise model is selected leading to a more accurate conversion.

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

2. Proposed Approach

In subsection 2.1, we introduce the Gaussian mixture modeling framework for spectral conversion. Then, in subsection 2.2, we introduce mixtures of probabilistic principal component analyzers which provide an entire range of covariance structure that incrementally includes more covariance information. This will allow us to define a whole range of models with increasing number of components from which the best model will be selected in order to perform the conversion at each source frame. The DMS method is presented in subsection 2.3.

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[10]. 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 random 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 features vector. We will denote $z = (x, y) = \{(x_n, y_n)\}_{n\in\mathcal{N}}$ a realization of this random process, in which x_n and y_n are the acoustic features vectors

[1]This study was supported by FEDER Angelstudio: Générateur d'avatars personnalisés ; 2009-2011

at frame n for the source and that for the target, respectively. We assume that Z is an independent and identically distributed random process (i.i.d.) such as $p(z) = \prod_{n\in\mathcal{N}} p(z_n)$. We introduce the auxiliary i.i.d random 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 automatically aligned by Dynamic Time Warping (DTW).

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

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

with

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

In each mixture component $k \in \mathcal{U}$, the conditional target mean vector $E_{k,n}^y$ for the given source acoustic features vector is calculated by a simple linear conversion based on the correlation between the source and target acoustic features 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 features vector belonging to each one of the mixture components:

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

2.2. Mixtures of Probabilistic Principal Component Analyzers

In the following of this paper, a set of potential best models with increasing number of parameters will be built during the training from which the best model will be chosen at each frame during the conversion step. The increase of the number of components of a Gaussian mixture is limited by the increasing complexity of the model due to the large number of parameters associated with the covariance matrices. One way to solve this problem is to use diagonal structures, but the performances are then sacrified because the latter are unable to model the underlying second order statistics. Mixture of *Probabilistic Principal Component Analyzers* (PPCAs) is a method proposed by Tipping and Bishop [11] to solve the inflexibility of GMMs by performing a pseudo-local *Principal Component Analysis* (PCA) on each mixture component. It has been applied to VC in [12]. Modeling covariance structure with a mixture of PPCAs provides an entire range of covariance structures that incrementally includes more covariance information.

PPCA's statistical model assumes that a set of q latent variables is responsible for generating the d-dimensional data set z_n. Under several assumptions given in [11], the observations z_n are Gaussian with mean μ and model covariance

$$\Sigma = WW^t + \sigma^2 I \tag{6}$$

in which the $d \times q$ matrix W contains the factor loadings which account for the statistical dependencies between the individual variables of z_n (e.g. correlations between LSF) and the specific factors σ explain small disturbances in each individual random variable of z_n (e.g. sensor noise about each individual LSF). It gives the flexibility of removing dimensions from W which in turn adds to σ^2, the average variance not captured in the projection. Several of these models can be combine into a mixture of PPCAs as described in [11]. The only difference with a GMM is that each of the K component densities are represented with a single PPCA model rather than with a multivariate normal distribution. An equivalent expression for the conditional expectation in equation (4) can be found by posing: $\Sigma_k^{xx} = W_k^x(W_k^x)^t + \sigma_k^2 I$ and $\Sigma_k^{yx} = W_k^y(W_k^x)^t$. Mixture of PPCAs can be seen as a more general case of the GMMs for spectral conversion as shown in [12]. In the following it will be use in order to define models with increasing number of mixtures while keeping a reasonable model complexity.

2.3. Dynamic model selection

In classical speaker conversion methods, a unique 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 test corpus or by using methods from the model selection research field. Information criteria 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 a very 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 propose 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 and mixtures of PPCAs is built according to

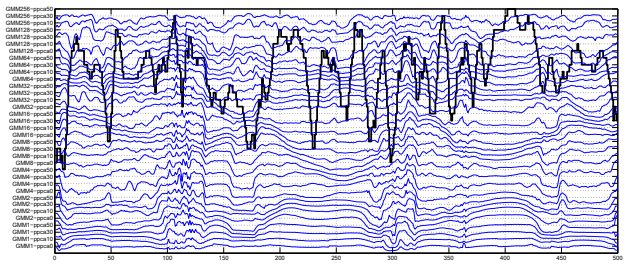

Figure 1: *Example of Dynamic Model Selection along a segment of a speech utterance. On the left the set of potential models. In bold line: selected models at each frame n, in light lines: LSF representation of the source spectral envelope.*

the BIC criterion 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 the case where M is a mixture of PPCA, Σ_k^{xx} is replaced by $W_k^x (W_k^x)^t + \sigma_k^2 I$. 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. An example of model selection along a segment of a speech utterance is given on Figure 1: complex models are used on stable spectrum parts while simpler and general models are used in transition parts.

3. Voice Conversion System

3.1. Corpora description

3 corpora have been recorded by 3 speakers with different French accent:

- the speaker S has a standard accent;
- the speaker A has a Hispanic accent;
- the speaker B has a French Canadian accent.

The voices of the speakers S,A and B, will be denoted as voice S, voice A and voice B respectively. The voice S will be use as the source voice and the voices A and B will be used as target voices during conversion. Each corpora includes 200 phonetically balanced utterances, which represents between 9 and 10 minutes of speech depending on speakers speech rate (<9 minutes for voices S and A and >10 minutes for voice B). The voices S and A were recorded in an anechoic chamber (32 bits, Fe=48kHz). The recording quality of the speaker B is lower (16 bits, 44.1khz compressed MP3 format) and slightly reverberated. The recordings were downsampled to Fe=24kHz for the experiments.

The corpora of voices S and A on the one hand and voice S and B on the other hand were aligned by DTW. Each of the two obtained parallel corpora (S, A) and (S, B) was then splitted into a training part (190 parallel utterances) and a test part (10 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-components GMM 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 datas was found to improve quality in an informal listening test.

3.2. 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 has been set to 30 in agreement with [13]. Linear Spectral Frequencies (LSF) parametrization was chosen due to their good linear interpolation properties. Analysis and synthesis were done using phase vocoder [15].

3.3. Model pre-selection for the DMS

Several joint models including GMMs with full covariance matrices and mixture of PPCAs with different values of q (3, 5, 10, 30 and 50) were estimated on both corpora (S, A) and (S, B) by considering different numbers of mixture components K (2, 4, 8, 16, 32, 64, 128, 256, 384 and 512). Then, for each corpus and for each K, the best model among the estimated ones was selected according to the BIC criterion (the best model being the one with the lowest BIC value). The selected model topologies in the following were valid for both corpora but with differents estimates values:

- Up to K=32 mixture components: GMM with full covariance;
- for K=64: mixture of PPCAs with q=50;
- for K=128: mixture of PPCAs with q=30;
- for K=256 and K=384: mixture of PPCAs with q=10;
- for K=512: mixture of PPCAs with q=5.

This set of best potential models - denoted $\mathcal{M}$ - will be used during the evaluation of the DMS method in the following section. The best model of $\mathcal{M}$ - which was the GMM with full covariance matrices and 16 components - will be used as the reference model for the classical conversion method and denoted as GMM16 in the following.

4. Subjective evaluation

In this section, we aim to evaluate the DMS method according to the perceived conversion effect and to the quality of the resulting converted voice. Two subjective tests[2] were designed aiming to evaluate these 2 aspects of VC. Both tests were performed on 27 listeners including 14 speech processing experts and 13 non-experts, 21 French native speakers, 6 non-native French speakers.

[2]The tests can be found online at `http://recherche.ircam.fr/equipes/analyse-synthese/lanchant/index.php/Main/TestDMS`

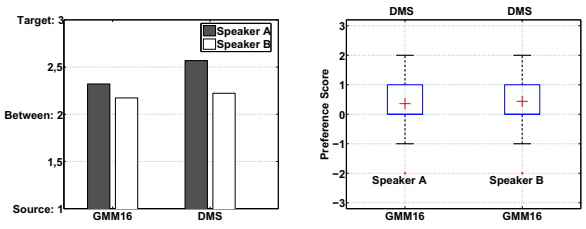

Figure 2: *Left: conversion effect test considering both method (the vote scale has 5 values from 0 to 4). Right: quality preference scores for the DMS method for speaker A (left) with mean=0.31, and for speaker B (right) with mean=0.36*

4.1. Conversion Effect

The objective of this first subjective test was to evaluate the *perceived conversion effect* after modification of the source speech S. We aimed to qualify if the converted voice was perceived closer to the target, to the source or between them. For each target voice (A and B), 3 speech utterances of the source converted considering both methods (GMM16 vs DMS) have been presented to listeners. Also, three utterances of both original source and target voice speakers have been introduced in the evaluation data to assess the ability of discrimination of listeners between the source and target voice speakers. At each trial, an utterance was randomly chosen among the 12 utterances and presented to the listener which had to vote on the perceived position relative to the source and target on a scale of 5 values, from 0 if perceived as the source to 4 if perceived as the target voice.

The results of this first test are presented on the left part of Figure 2. Note that the vote values axis as been zoomed between 1 and 3 for convenience but that the votes have been given on a scale from 0 to 4 during the test. For both target voice, the french accent is clearly identified. However, most listeners couldn't recognize neither the source nor the target, which is a well-known effect (*third-speaker effect*) reported in the bibliography, especially in intra-gender voice conversion task. For the target voice A, the mean values of the votes are 2.32 for GMM16 model and 2.56 for the DMS method. For the target voice B the mean values are lower with 2.17 for the GMM16 model and 2.22 for the DMS method which could be explain by the presence of reverberation in the recordings of the target which are not present in the converted speech signal. However, for both speaker the DMS method outperform the GMM16 based method in term of perceived conversion effect.

4.2. Converted Speech quality

The second test was focused on the *quality* of the converted signals. A comparison category rating test (CCR[16]) was used to assess the quality of the speech converted by the DMS method in comparison to the speech converted by the GMM16 model. 6 utterances were chosen to generate the test samples for each method. Utterances were presented in random order to the listeners for evaluation. They were asked to attribute a score to the quality of the second sample of a pair compared to the quality of the first one on the comparison mean opinion score (CMOS) scale. The ranking of the 2 methods was evaluated by averaging the scores of the CCR test for each method. The results are presented on the right part of Figure 2. Only the preference score for the DMS method is presented on the figure (the score for the classical GMM16 method being the negative value of the latter).

The results show that for each target speaker the new DMS method provides a better conversion quality than the one based on the single GMM16 model, with a preference mean equal to 0.31 for the target speaker A and 0.36 for the speaker target B. A *multi-class one-way analysis of variance* (ANOVA[17]) for speaker A ($F(1,322)=29.6$;p-value<0.001) and speaker B ($F(1,322)=36.9$;p-value<0.001) confirms that the difference in quality between both methods is significant, for both target speakers cases.

5. Conclusion

A new approach for spectral voice conversion involving several models has been proposed. In this approach, a set of potential best models is chosen during the learning phase. During the conversion, the model selection is achieved dynamically on this set, at each source frame, according to its acoustical features. Subjective tests 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 further work, we will focus on the choice of more appropriate criteria for the selection of the best model during the conversion. We will also apply this framework to the conversion method described in [8] to improve the quality of synthesis by taking into account the dynamic features and the global variance.

6. 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*, vol. 16, no. 2, 1995, pp. 207–216.

[4] H. Valbret, E. Moulines, and J. Tubach, "Voice transformation using psola technique," in *Speech Communication*, no. 11, 1992, 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*, vol. 6, no. 2, 1998, 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] 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.

[11] M. E. Tipping and B. C. M., "Mixtures of probabilistic principal component analysers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[12] A. Wilde, M. Martinez, "Probabilistic principal component analysis applied to voice conversion," in *Asilomar Conference on Signals Systems and Computers*, vol. 2, 2004, pp. 2255–2259.

[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] ——, "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] R. I.-U. P.800, "Methods for subjective determination of transmission quality," in *International Telecommunication Union*, Aug. 1996.

[17] J. Hair, R. Anderson, M. Tatham, and W. Black, *Multivariate data analysis*. Prentice-Hall, 1995.