Adaptation of a Gaussian Mixture Regressor to a New Input Distribution: Extending the C-GMR Framework

Laurent Girin[1,2], Thomas Hueber[1] and Xavier Alameda-Pineda[2,3]

[1] CNRS / Grenoble Alpes Univ. / GIPSA-lab, Grenoble, France,
firstname.lastname@gipsa-lab.grenoble-inp.fr,
[2] INRIA Rhône-Alpes, France,
[3] University of Trento, Italy,
xavier.alamedapineda@unitn.it

Abstract. This paper addresses the problem of the adaptation of a Gaussian Mixture Regression (GMR) to a new input distribution, using a limited amount of input-only examples. We propose a new model for GMR adaptation, called Joint GMR (J-GMR), that extends the previously published framework of Cascaded GMR (C-GMR). We provide an exact EM training algorithm for the J-GMR. We discuss the merits of the J-GMR with respect to the C-GMR and illustrate its performance with experiments on speech acoustic-to-articulatory inversion.

Keywords: GMM, Gaussian Mixture Regression, Adaptation, EM algorithm.

1 Introduction

The Gaussian Mixture Regression (GMR) is an efficient regression technique derived from the Gaussian Mixture Model (GMM) [1]. The GMR is widely used in different areas of speech processing, e.g. voice conversion [2, 3], acoustic-articulatory mapping [4, 5], in image processing, e.g. head pose estimation from depth data [6], and in robotics [7].

Let us consider a GMR that has been trained on a large dataset of *input-output* joint observations. The problem addressed in this paper is the *adaptation* of this GMR to a (moderate) change in the distribution of input data using a limited set of new *input-only* samples. In a practical context, this aims at using a well-estimated GMR with input observations that no more faithfully follow the distribution observed during training. For example, in speech processing, this happens when considering a speaker different from the one used for training. To address this problem, we first proposed in [8] to adapt the model parameters related to input observations using two state-of-the-art adaptation techniques for GMM which are maximum a posteriori (MAP) [9] and maximum likelihood linear regression (MLLR) [10]. Then, we proposed in [11] a general framework called Cascaded GMR (C-GMR) and derived two implementations (see Fig. 1).

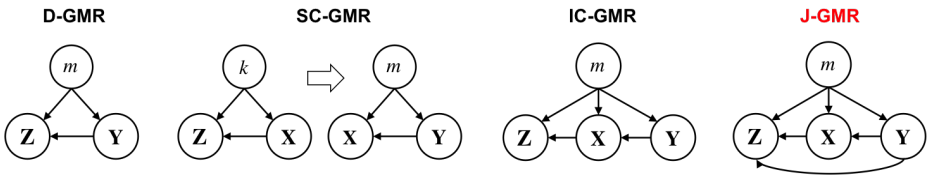

Fig. 1. Graphical representation of the D-, SC-, IC-, and J-GMR.

The first one, referred to as Split-C-GMR (SC-GMR), is a simple chaining of two consecutive GMRs. The second one, referred to as Integrated-C-GMR (IC-GMR) combines the two successive GMRs in a single probabilistic model. The IC-GMR put into practice the general *missing data* methodology of machine learning [12] to exploit both the small and potentially sparse adaptation dataset and the large and dense training dataset. In [11], this model was shown to provide superior performance to the SC-GMR and to a s a direct GMR (D-GMR) trained only with new input data and corresponding output data. The D-GMR, SC-GMR and IC-GMR are briefly presented in Section 2.

In the present paper, we extend the general framework of C-GMR to a new model, called Joint GMR (J-GMR), presented in Section 3. Compared to the IC-GMR, the J-GMR can be considered as more general, in the sense that it aims at modeling the statistical dependencies between all the considered variables. In Section 4, we provide the exact associated EM algorithm [13] used to perform the adaptation to new input data. As for the IC-GMR, the J-GMR and associated EM algorithm consider explicitly the incomplete adaptation dataset jointly with the training dataset, using the missing data methodology. In Section 5 we illustrate the interest of the new model: The performance of the J-GMR is favorably compared to the D-JMR, SC-GMR and IC-GMR in a speech acoustic-to-articulatory inversion task on simulated data. Section 6 concludes the paper.

2 Cascaded GMR

2.1 Definitions, notations and working hypothesis

Let us consider a GMR between realizations of input $\mathbf{X}$ and output $\mathbf{Y}$ (column) random vectors, of arbitrary finite dimension. Let us define a new input vector $\mathbf{Z}$ to which the GMR is to be adapted. Let us define $\mathbf{V} = [\mathbf{X}^\top, \mathbf{Y}^\top]^\top$ and $\mathbf{O} = [\mathbf{X}^\top, \mathbf{Y}^\top, \mathbf{Z}^\top]^\top$, where $^\top$ denotes the transpose operator. Let $p(\mathbf{X} = \mathbf{x}; \boldsymbol{\Theta}_\mathbf{X})$ denote the probability density function (PDF) of $\mathbf{X}$ (for simplicity, we omit $\mathbf{X}$ and may omit $\boldsymbol{\Theta}_\mathbf{X}$). Let $\mathcal{N}(\mathbf{x}; \mu_\mathbf{X}, \boldsymbol{\Sigma}_{\mathbf{XX}})$ denote the Gaussian distribution of $\mathbf{X}$ with mean vector $\mu_\mathbf{X}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{XX}}$. Let $\boldsymbol{\Sigma}_{\mathbf{XY}}$ denote the cross-covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$, and $\boldsymbol{\Lambda}_{\mathbf{XX}}$ the precision matrix of $\mathbf{X}$ (similarly for cross-terms). With these notations, the PDF of a GMM on $\mathbf{V}$ writes:

$$p(\mathbf{v}) = \sum_{m=1}^{M} \pi_m \mathcal{N}(\mathbf{v}; \mu_{\mathbf{V},m}, \boldsymbol{\Sigma}_{\mathbf{VV},m}), \tag{1}$$

where M is the number of components, $\pi_m \geq 0$, and $\sum_{m=1}^{M} \pi_m = 1$.

Let us denote $\mathcal{D}_{\mathbf{xy}} = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N}$ the large dataset of N i.i.d. vector pairs drawn from the ($\mathbf{X}$,$\mathbf{Y}$) distribution, that is used for the training of the $\mathbf{X}$-to-$\mathbf{Y}$ GMR. We assume that a limited dataset $\mathcal{D}_{\mathbf{z}}$ of new input vectors $\mathbf{z}$ is available for the adaptation of the GMR. Moreover, we assume that $\mathcal{D}_{\mathbf{z}}$ can be aligned with a subset of the reference input dataset (e.g., in voice conversion, by time-aligning the same sentence pronounced by two speakers). Since reordering of the dataset is arbitrary, we denote $\mathcal{D}_{\mathbf{z}} = \{\mathbf{z}_n\}_{n=1}^{N_0}$, with $N_0 \ll N$.

2.2 D-GMR, SC-GMR and IC-GMR

In this section, we briefly recall three approaches for GMR adaptation considered in [11], which will be used here as a baseline. Their graphical representation is illustrated in Fig. 1. The first one is a direct $\mathbf{Z}$-to-$\mathbf{Y}$ GMR trained using $\mathcal{D}_{\mathbf{zy}} = \{\mathbf{z}_n, \mathbf{y}_n\}_{n=1}^{N_0}$. Inference of $\mathbf{y}$ given an observed value $\mathbf{z}$ is performed by the Minimum Mean Squared Error (MMSE) estimator, which is the posterior mean $\hat{\mathbf{y}} = \mathrm{E}[\mathbf{Y}|\mathbf{z}]$:

$$\hat{\mathbf{y}} = \sum_{m=1}^{M} p(m|\mathbf{z}) \left(\mu_{\mathbf{Y},m} + \mathbf{\Sigma}_{\mathbf{YZ},m}\mathbf{\Sigma}_{\mathbf{ZZ},m}^{-1}(\mathbf{z} - \mu_{\mathbf{Z},m})\right), \tag{2}$$

with $p(m|\mathbf{z}) = \frac{\pi_m \mathcal{N}(\mathbf{z}|\mu_{\mathbf{Z},m}, \mathbf{\Sigma}_{\mathbf{ZZ},m})}{\sum_{m=1}^{M} \pi_m \mathcal{N}(\mathbf{z}|\mu_{\mathbf{Z},m}, \mathbf{\Sigma}_{\mathbf{ZZ},m})}$ and M is the number of mixture components. This model is referred to as D-GMR. The second and third models are instances of cascaded GMR. The split-cascaded GMR (SC-GMR) consists of chaining two distinct GMRs: a $\mathbf{Z}$-to-$\mathbf{X}$ GMR followed by a $\mathbf{X}$-to-$\mathbf{Y}$ GMR. The inference equation thus consists in chaining $\hat{\mathbf{x}} = \mathrm{E}[\mathbf{X}|\mathbf{z}]$ and $\hat{\mathbf{y}} = \mathrm{E}[\mathbf{Y}|\hat{\mathbf{x}}]$, where both expectations follow (2) with their respective parameters. Note that the two GMRs may have a different number of mixture components. The integrated-cascaded GMR (IC-GMR) combines the $\mathbf{Z}$-to-$\mathbf{X}$ mapping and the $\mathbf{X}$-to-$\mathbf{Y}$ mapping into a single GMR-based mapping process. Importantly, this is made at the component level of the GMR, i.e. *within the mixture*, as opposed to the SC-GMR (see Fig. 1). The corresponding IC mixture model is defined by:

$$p(\mathbf{o}) = \sum_{m=1}^{M} \pi_m p(\mathbf{y}|m) p(\mathbf{x}|\mathbf{y}, m) p(\mathbf{z}|\mathbf{x}, m), \tag{3}$$

where all PDFs are Gaussian. The IC-GMR is given by:

$$\hat{\mathbf{y}} = \sum_{m=1}^{M} p(m|\mathbf{z})[\mu_{\mathbf{Y},m} + \mathbf{\Sigma}_{\mathbf{YX},m}\mathbf{\Sigma}_{\mathbf{XX},m}^{-1}\mathbf{\Sigma}_{\mathbf{XZ},m}\mathbf{\Sigma}_{\mathbf{ZZ},m}^{-1}(\mathbf{z} - \mu_{\mathbf{Z},m})]. \tag{4}$$

The above equation is a $\mathbf{Z}$-to-$\mathbf{Y}$ GMR with a specific form of the covariance matrix, i.e. $\mathbf{\Sigma}_{\mathbf{YZ},m}$ is not a free parameter:

$$\mathbf{\Sigma}_{\mathbf{YZ},m} = \mathbf{\Sigma}_{\mathbf{YX},m}\mathbf{\Sigma}_{\mathbf{XX},m}^{-1}\mathbf{\Sigma}_{\mathbf{XZ},m}. \tag{5}$$

3 Joint GMR

Let us define the *Joint* GMM on ($\mathbf{X}$,$\mathbf{Y}$,$\mathbf{Z}$) as:

$$p(\mathbf{o}) = \sum_{m=1}^{M} \pi_m \mathcal{N}(\mathbf{o}; \mu_m, \mathbf{\Sigma}_m). \tag{6}$$

The associated J-GMR inference equation is again given by the posterior mean $\hat{\mathbf{y}} = \mathrm{E}[\mathbf{Y}|\mathbf{z}]$. Here, we have:

$$p(\mathbf{y}|\mathbf{z}) = \int_{\mathbf{X}} \sum_{m=1}^{M} p(\mathbf{x}, \mathbf{y}, m|\mathbf{z}) d\mathbf{x} = \sum_{m=1}^{M} p(m|\mathbf{z}) p(\mathbf{y}|\mathbf{z}, m). \tag{7}$$

Since the conditional and marginal distributions of a Gaussian are Gaussian as well, (7) is a GMM. Therefore, the J-GMR inference equation turns out to be identical to the usual expression for a direct $\mathbf{Z}$-to-$\mathbf{Y}$ GMR, i.e. (2). This gives the impression of by-passing the information contained in $\mathbf{X}$. However, this is not the case: the complete proposed process for GMR adaptation is not equivalent to a GMR build directly from $(\mathbf{z}, \mathbf{y})$ training data. Indeed, as shown in the next section, the estimation of the J-GMR parameters with the EM algorithm exploits all the available data, i.e. $\mathcal{D}_{\mathbf{xy}}$ and $\mathcal{D}_{\mathbf{z}}$, hence including all $\mathbf{x}$ data.

Remarkably, (5) characterizes the IC-GMR as a particular case of the J-GMR. This is also true at the mixture model level, i.e. (3) is a particular case of (6) with (5). The matrix product $\mathbf{\Sigma}_{\mathbf{XZ},m}\mathbf{\Sigma}_{\mathbf{ZZ},m}^{-1}$ in (4) enables to go from $\mathbf{z}$ to $\mathbf{x}$, and then $\mathbf{\Sigma}_{\mathbf{YX},m}\mathbf{\Sigma}_{\mathbf{XX},m}^{-1}$ enables to go from $\mathbf{x}$ to $\mathbf{y}$, so that the IC-GMR goes from $\mathbf{z}$ to $\mathbf{y}$ "passing through $\mathbf{x}$". In contrast, the J-GMR enables to go directly from $\mathbf{z}$ to $\mathbf{y}$, though again, it is not equivalent to the $\mathbf{Z}$-$\mathbf{Y}$ D-GMR since $\mathbf{x}$ data are used at training time, as shown in the next section.

4 EM algorithm for J-GMR

This section introduces the exact EM algorithm associated to the J-GMR, explicitly handling incomplete adaptation datasets using the general methodology of missing data. The EM iteratively maximizes the expected complete-data log-likelihood, denoted by Q. At iteration $i+1$, the E-step computes the function $Q(\mathbf{\Theta}, \mathbf{\Theta}^{(i)})$, where $\mathbf{\Theta}^{(i)}$ are the parameters computed at iteration i. The M-step maximizes Q with respect to $\mathbf{\Theta}$, obtaining $\mathbf{\Theta}^{(i+1)}$. In the following we describe the E-step, the M step, the initialization process, and finally we comment the link between the EM algorithms of the IC-GMR and J-GMR.

4.1 E-step

In order to derive the expected complete-data log-likelihood $Q(\mathbf{\Theta}, \mathbf{\Theta}^{(i)})$, we follow the general methodology given in [14]–(Section 9.4) and [12]. In [11], we have

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(i)}) = \sum_{n=1}^{N_0} \sum_{m=1}^{M} \gamma_{nm}^{(i+1)} \log p(\mathbf{o}_n, m; \boldsymbol{\Theta}_m) + \sum_{n=N_0+1}^{N} \sum_{m=1}^{M} \frac{1}{p(\mathbf{v}_n; \boldsymbol{\Theta}_{\mathbf{V}}^{(i)})} \int p(\mathbf{o}_n, m; \boldsymbol{\Theta}_m^{(i)}) \log p(\mathbf{o}_n, m; \boldsymbol{\Theta}_m) d\mathbf{z}_n. \quad (8)$$

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(i)}) = \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{nm}^{(i+1)} \left(\log \pi_m - \frac{\log |\boldsymbol{\Sigma}_m| + (\mathbf{o}'_{nm} - \mu_m)^\top \boldsymbol{\Sigma}_m^{-1} (\mathbf{o}'_{nm} - \mu_m)}{2} \right) - \frac{1}{2} \sum_{m=1}^{M} \left(\sum_{n=N_0+1}^{N} \gamma_{nm}^{(i+1)} \right) \mathrm{Tr} \left[\boldsymbol{\Lambda}_{\mathbf{ZZ},m} (\boldsymbol{\Lambda}_{\mathbf{ZZ},m}^{(i)})^{-1} \right]. \quad (9)$$

shown that this leads to the general expression (8), where

$$\gamma_{nm}^{(i+1)} = \frac{p(\mathbf{o}_n, m; \boldsymbol{\Theta}_m^{(i)})}{p(\mathbf{o}_n; \boldsymbol{\Theta}^{(i)})}, \quad n \in [1, N_0], \quad (10)$$

are the so-called *responsibilities* (of component m explaining observation $\mathbf{o}_n$) [14]. Eq. (8) is valid for any mixture model on i.i.d. vectors $(\mathbf{Z}, \mathbf{V})$ with partly missing $\mathbf{z}$ data. Here we study the particular case of the J-GMR. For this aim, we denote $\mu_{\mathbf{Z}_n|\mathbf{v}_n,m}^{(i)}$ the posterior mean of $\mathbf{Z}_n$ given $\mathbf{v}_n$ for the m-th Gaussian component with parameters $\boldsymbol{\Theta}_m^{(i)}$, i.e.:

$$\mu_{\mathbf{Z}_n|\mathbf{v}_n,m}^{(i)} = \mu_{\mathbf{Z},m}^{(i)} + \boldsymbol{\Sigma}_{\mathbf{ZV},m}^{(i)} \left(\boldsymbol{\Sigma}_{\mathbf{VV},m}^{(i)} \right)^{-1} (\mathbf{v}_n - \mu_{\mathbf{V},m}^{(i)}). \quad (11)$$

Let us define $\mathbf{o}'_{nm} = [\mathbf{v}_n^\top, \mu_{\mathbf{Z}_n|\mathbf{v}_n,m}^{(i)\top}]^\top$ if $n \in [N_0 + 1, N]$, i.e. $\mathbf{o}'_{nm}$ is an "augmented" observation vector in which for $n \in [N_0 + 1, N]$ the missing data vector $\mathbf{z}_n$ is replaced with its estimate $\mu_{\mathbf{Z}_n|\mathbf{v}_n,m}^{(i)}$. Let us arbitrarily extend $\mathbf{o}'_{nm}$ with $\mathbf{o}'_{nm} = \mathbf{o}_n$ for $n \in [1, N_0]$, and the definition of the responsibilities to the incomplete data vectors $\mathbf{v}_n$:

$$\gamma_{nm}^{(i+1)} = \frac{p(\mathbf{v}_n, m; \boldsymbol{\Theta}_{\mathbf{V},m}^{(i)})}{p(\mathbf{v}_n; \boldsymbol{\Theta}_{\mathbf{V}}^{(i)})}, \quad n \in [N_0 + 1, N]. \quad (12)$$

Then, $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(i)})$ is given by (9). The proof is provided in [15]. The first double sum in (9) is similar to the one found in the usual EM for GMM (without missing data), except that for $n \in [N_0 + 1, N]$ missing $\mathbf{z}$ data are replaced with their estimate using corresponding $\mathbf{x}$ and $\mathbf{y}$ data and current parameter values, and responsibilities are calculated using available data only. The second term is a correction term that, as seen below, modifies the estimation of the covariance matrices $\boldsymbol{\Sigma}_m$ in the M-step to take into account missing data.

4.2 M-step

Priors: Maximization of $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(i)})$ with respect to the priors π_m is identical to the classical case of GMM without missing data [14]. For $m \in [1, M]$, we have:

$$\pi_m^{(i+1)} = \frac{1}{N} \sum_{n=1}^{N} \gamma_{nm}^{(i+1)}. \tag{13}$$

Mean vectors: For $m \in [1, M]$, derivating $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(i)})$ with respect to μ_m and setting the result to zero leads to:

$$\mu_m^{(i+1)} = \frac{\sum_{n=1}^{N} \gamma_{nm}^{(i+1)} \mathbf{o}'_{nm}}{\sum_{n=1}^{N} \gamma_{nm}^{(i+1)}}. \tag{14}$$

This expression is an empirical mean, similar to the classical GMM case, except for the specific definition of observation vectors and responsibilities for $n \in [N_0 + 1, N]$.

Covariance matrices: Let us first express the trace in (9) as a function of $\boldsymbol{\Sigma}_m^{-1}$ by completing $(\boldsymbol{\Lambda}_{\mathbf{ZZ},m}^{(i)})^{-1}$ with zeros to obtain the matrix $(\boldsymbol{\Lambda}_{\mathbf{ZZ},m}^{0\,(i)})^{-1}$:

$$(\boldsymbol{\Lambda}_{\mathbf{ZZ},m}^{0\,(i)})^{-1} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{\Lambda}_{\mathbf{ZZ},m}^{(i)})^{-1} \end{bmatrix}. \tag{15}$$

Thus, $\mathrm{Tr}\left[\boldsymbol{\Lambda}_{\mathbf{ZZ},m}(\boldsymbol{\Lambda}_{\mathbf{ZZ},m}^{(i)})^{-1}\right] = \mathrm{Tr}\left[\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\Lambda}_{\mathbf{ZZ},m}^{0\,(i)})^{-1}\right]$, and by canceling the derivative of $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(i)})$ with respect to $\boldsymbol{\Sigma}_m^{-1}$ we get:

$$\boldsymbol{\Sigma}_m^{(i+1)} = \frac{1}{\sum_{n=1}^{N} \gamma_{nm}^{(i+1)}} \left[\sum_{n=1}^{N} \gamma_{nm}^{(i+1)} (\mathbf{o}'_{nm} - \mu_m)(\mathbf{o}'_{nm} - \mu_m)^\top + \left(\sum_{n=N_0+1}^{N} \gamma_{nm}^{(i+1)} \right) (\boldsymbol{\Lambda}_{\mathbf{ZZ},m}^{0\,(i)})^{-1} \right]. \tag{16}$$

The first term is the empirical covariance matrix and is similar to the classical GMM without missing data, except again for the specific definition of observation vectors and responsibilities for $n \in [N_0 + 1, N]$. The second term can be seen as an additional correction term that deals with the absence of observed $\mathbf{z}$ data vectors for $n \in [N_0 + 1, N]$. We remark that $\boldsymbol{\Sigma}_m^{(i+1)}$ depends on all the terms of $\boldsymbol{\Sigma}_m^{(i)}$ obtained at previous iteration, since $\boldsymbol{\Lambda}_{\mathbf{ZZ},m}^{(i)} = [\boldsymbol{\Sigma}_m^{(i)\,-1}]_{\mathbf{ZZ}} \neq (\boldsymbol{\Sigma}_{\mathbf{ZZ},m}^{(i)})^{-1}$.

4.3 EM Initialization

Similarly to [11], the initialization of the proposed EM algorithm takes a very peculiar aspect. Indeed, the reference $(\mathbf{X}, \mathbf{Y})$ GMR model is used to initialize the marginal parameters in $(\mathbf{X}, \mathbf{Y})$ of the Joint GMM. The marginal parameters in $\mathbf{Z}$ are initialized using the aligned adaptation data $\mathcal{D}_{\mathbf{z}} = \{\mathbf{z}_n\}_{n=1}^{N_0}$. The cross-term parameters are initialized by constructing the sufficient statistics using $\{\mathbf{z}_n, \mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N_0}$.

4.4 A remark on the link between the J-GMR EM and the IC-GMR EM

We already noticed that the IC mixture model (3) can be seen as a constrained version of the Joint GMM (6). However, the EM for the IC-GMR presented in [11] (which exploits the linear-Gaussian form of the IC-GMR) is *not* derivable as a particular case of the EM of Section 4. More precisely, if one attempts to estimate the IC-GMR parameters with the algorithm we present in this paper, the M-step should be constrained by (5). Naturally, the complexity of the resulting constrained algorithm would be much higher. Consequently, even if the IC-GMR and the J-GMR models are closely related, the two learning algorithms are intrinsically different. This difference arises from the fact that the mixture model underlying the IC-GMR deals with constrained covariance matrices, whereas the Joint GMM uses fully free covariance matrices.

5 Experiments

The performance of the J-GMR was evaluated on a speech acoustic-to-articulatory inversion task which consists in recovering the movements of the tongue, lips, jaw and velum from the speech's acoustics. The goal is to adapt an acoustic-to-articulatory (i.e. $\mathbf{X}$-to-$\mathbf{Y}$) GMR trained on a large dataset $\mathcal{D}_{\mathbf{xy}}$ from a reference speaker given a small amount of audio observations only $\mathcal{D}_{\mathbf{z}}$ from another speaker (referred here to as the *source speaker*). In this study, experiments were conducted on synthetic data obtained using a so-called articulatory synthesizer. This allows us to better understand the behavior of the J-GMR model by controlling finely the structure of the adaptation dataset (as opposed to in-vivo data recorded using motion capture techniques on real human speakers). A synthetic dataset of vowels was thus generated using the Variable Linear Articulatory Model (VLAM) [16]. VLAM consists of a vocal tract model driven by seven control parameters (lips aperture and protrusion; jaw; tongue body, dorsum and apex; velum). For a given articulatory configuration, VLAM deduces the corresponding spectrum using acoustic simulation [17]. Among other articulatory synthesizers, VLAM is of particular interest in our study. Indeed, it integrates a model of the vocal tract growth and enables to generate two different spectra from the same articulatory configuration but different vocal tract length. We used this feature to simulate a parallel acoustic-articulatory dataset for two speakers (reference and source) with different vocal tract length corresponding to speaker age of 25 years and 17 years respectively. We generated $20,000$ triplets $(\mathbf{z}, \mathbf{x}, \mathbf{y})$ structured into four clusters simulating the 4 following vowels: /a/, /i/, /u/, /ə/. In our experiments, the spectrum is described by the position and the amplitude of the 4 first formants (i.e. local maxima in the power spectrum), hence 8-dimensional $\mathbf{x}$ and $\mathbf{z}$ observations. Fig. 2 displays the $20,000$ training acoustic data $\mathbf{x}$ (in the two first formant frequencies plane F1-F2; red points) and a selection of 467 adaptation vectors $\mathbf{z}$ (green points).

The EM algorithms for training the reference $\mathbf{X}$-$\mathbf{Y}$ model (and also the $\mathbf{Z}$-$\mathbf{X}$ model for the SC-GMR) were initialized using a k-means algorithm, repeated 5

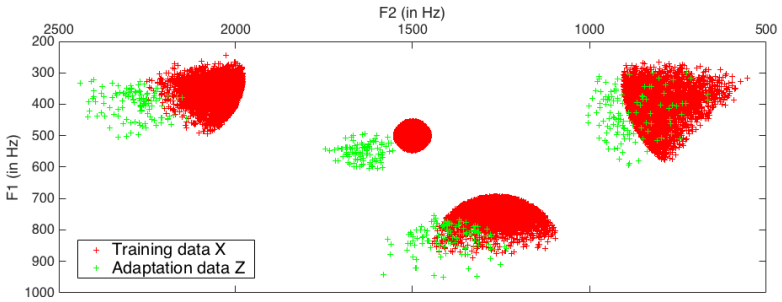

Fig. 2. Synthetic data generated using VLAM, displayed in the F2-F1 acoustic space.

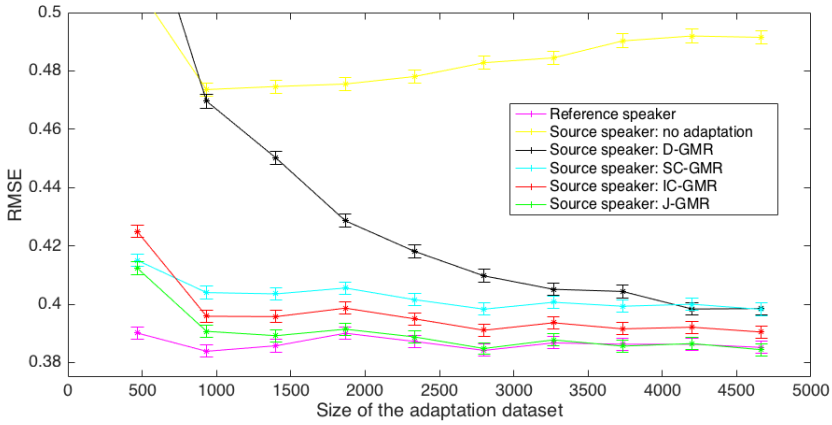

Fig. 3. RMSE (in cm) of the Z-to-Y mapping as a function of the size of the adaptation data (in number of vectors), for D-GMR, SC-GMR, IC-GMR and J-GMR (lower and upper bounds are given by the X-Y mapping in magenta and the Z-to-Y mapping with no adaption in yellow; error bars represent 95% confidence intervals.)

times (only the best initial model was kept for training). For all EMs, the number of iterations was empirically set to 50. All methods were evaluated under a K-fold cross-validation protocol (with $K = 30$). The data was divided in 30 subsets of approximate equal size: 29 subsets for training and 1 subset for test, considering all permutations. In each of the 30 folds, $k/30$ of the size of the training set was used for adaptation, with $k \in [1, 10]$. For a given value of k, we conducted 10 experiments with a different adaptation dataset. For each experiment, the optimal number of mixture components (within $M = 2, 4, 8, 12, 16, 20$) was determined using cross-validation. The performance was assessed by calculating the average Root Mean Squared Error (RMSE) between the articulatory trajectories $\mathbf{y}$ estimated from the source speaker's acoustics, and the ones recorded on the reference speaker. For each RMSE measure, 95% confidence interval (from which statistical significance between two regression techniques can be assessed) was obtained using paired t-test.

The RMSE for the J-GMR, as well as for the D-GMR, SC-GMR and IC-GMR are plotted in Fig. 3, as a function of N_0, the size of the adaptation set. The performance of the J-GMR, SC-GMR and IC-GMR are quite close, and are clearly better than without adaptation and than the D-GMR, especially for low values of N_0. This latter result comes from the fact that the D-GMR exploits only the limited amount of reference speaker's articulatory data that can be associated with the source speaker's audio data. This tends to validate the benefit of exploiting all available $(\mathbf{x}, \mathbf{y})$ observations during the adaptation process, as done in the C-GMR framework. As in [11], the IC-GMR performs better than the SC-GMR, except for the lower N_0. Importantly, we observe a systematic and statistically significant improvement (within the approximate range 1.5%–2.5% of RMSE depending on N_0) of the proposed J-GMR over both the IC-GMR and the SC-GMR (except for the lower N_0 for which the difference between J-GMR and SC-GMR is not significant). This illustrates that the J-GMR is able to better exploit the statistical relations between $\mathbf{z}$, $\mathbf{x}$ and $\mathbf{y}$ data. The link between $\mathbf{x}$ and $\mathbf{y}$ is exploited in the J-GMR, as in the IC-GMR, though the new direct link between $\mathbf{z}$ and $\mathbf{y}$ is also exploited. In short the mapping is not exclusively forced to pass through $\mathbf{x}$, which is shown to be beneficial in the present set of experiments.

6 Conclusion

In this paper, we have extended the general framework of Cascaded-GMR with a new model called J-GMR, for which we provided the exact EM training algorithm explicitly considering missing data. The J-GMR has been shown to perform better than the D-GMR, SC-GMR and IC-GMR in our acoustic-to-articulatory inversion experiments. Altogether, the results show the benefit of considering an intermediate $\mathbf{Z}$-to-$\mathbf{X}$ mapping in the general $\mathbf{Z}$-to-$\mathbf{Y}$ mapping process. This is done explicitly in the SC-GMR and implicitly in both IC-GMR and J-GMR. The relative performance of all these C-GMR models may depend on the latent structure of the data. Hence we believe that all models from this library of GMR adaptation techniques can be of potential interest for other applications. The MATLAB source code of the IC-GMR training and mapping algorithms is available at http://www.gipsa-lab.fr/∼thomas.hueber/cgmr/, and the code for the J-GMR will be added if the present paper is accepted.

As a moderation note, it has to be remembered that, so far, the whole proposed C-GMR framework relies on the assumption that the adaptation data can be aligned with a subset of the training data (see Section 2.1), which is a strong hypothesis. In our future work, we will work on relaxing this assumption, e.g. only considering identification of the class of each adaptation data, which will extend the potential applications of the proposed C-GMR models.

Acknowledgements

The authors warmly thank Louis-Jean Boë for his help with the VLAM model.

References

1. G. McLachlan and D. Peel, *Finite Mixture Models*. New-York, USA: John Wiley and sons, 2000.
2. Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.
3. T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.
4. ——, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.
5. H. Zen, Y. Nankaku, and K. Tokuda, "Continuous stochastic feature mapping based on trajectory HMMs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 417–430, 2011.
6. Y. Tian, L. Sigal, H. Badino, F. de la Torre, and Y. Liu, "Latent gaussian mixture regression for human pose estimation," in *ACCV*, 2010, pp. 679–690.
7. S. Calinon, F. D'halluin, E. L. Sauser, D. G. Caldwell, and A. G. Billard, "Learning and reproduction of gestures by imitation: An approach based on hidden Markov model and Gaussian mixture regression," *IEEE Robotics and Automation Magazine*, vol. 17, no. 2, pp. 44–54, 2010.
8. T. Hueber, G. Bailly, P. Badin, and F. Elisei, "Speaker adaptation of an acoustic-articulatory inversion model using cascaded Gaussian mixture regressions," in *Proceedings of Interspeech*, Lyon, France, 2013, pp. 2753–2757.
9. J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.
10. M. J. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech & Language*, vol. 10, no. 4, pp. 249–264, 1996.
11. T. Hueber, L. Girin, X. Alameda-Pineda, and G. Bailly, "Speaker-adaptive acoustic-articulatory inversion using cascaded Gaussian mixture regression," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 12, pp. 2246–2259, 2015.
12. Z. Ghahramani and M. I. Jordan, "Learning from incomplete data," MIT, Cambridge, MA, USA, Tech. Rep., 1994.
13. G. McLachlan and K. Thriyambakam, *The EM algorithm and extensions*. New-York, USA: John Wiley and sons, 1997.
14. C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.
15. L. Girin, T. Hueber, and X. Alameda-Pineda, "Appendix to: Adaptation of a Gaussian mixture regressor to a new input distribution: Extending the C-GMR framework," Tech. Rep., 2016, Available at http://www.gipsa-lab.grenoble-inp.fr/~laurent.girin/demo/JGMRappendix.pdf.
16. L. Ménard, J.-L. Schwartz, L.-J. Boë, and J. Aubin, "Articulatory–acoustic relationships during vocal tract growth for french vowels: Analysis of real data and simulations with an articulatory model," *Journal of Phonetics*, vol. 35, no. 1, pp. 1–19, 2007.
17. P. Badin and G. Fant, "Notes on vocal tract computation," *Quarterly Progress and Status Report, Dept for Speech, Music and Hearing, KTH, Stockholm*, pp. 53–108, 1984.