Extending the Cascaded Gaussian Mixture Regression Framework for Cross-Speaker Acoustic-Articulatory Mapping

Laurent Girin, Thomas Hueber, and Xavier Alameda-Pineda, *Member, IEEE*

***Abstract*—This paper addresses the adaptation of an acoustic-articulatory inversion model of a reference speaker to the voice of another source speaker, using a limited amount of audio-only data. In this study, the articulatory-acoustic relationship of the reference speaker is modeled by a Gaussian mixture model and inference of articulatory data from acoustic data is made by the associated Gaussian mixture regression (GMR). To address speaker adaptation, we previously proposed a general framework called Cascaded-GMR (C-GMR) which decomposes the adaptation process into two consecutive steps: spectral conversion between source and reference speaker and acoustic-articulatory inversion of converted spectral trajectories. In particular, we proposed the integrated C-GMR technique (IC-GMR) in which both steps are tied together in the same probabilistic model. In this paper, we extend the C-GMR framework with another model called Joint-GMR (J-GMR). Contrary to the IC-GMR, this model aims at exploiting all potential acoustic-articulatory relationships, including those between the source speaker's acoustics and the reference speaker's articulation. We present the full derivation of the exact expectation–maximization (EM) training algorithm for the J-GMR. It exploits the *missing data* methodology of machine learning to deal with limited adaptation data. We provide an extensive evaluation of the J-GMR on both synthetic acoustic-articulatory data and on the multispeaker MOCHA EMA database. We compare the J-GMR performance to other models of the C-GMR framework, notably the IC-GMR, and discuss their respective merits.**

***Index Terms*—Acoustic-articulatory inversion, EM, Gaussian mixture regression, GMM, missing data, speaker adaptation.**

I. Introduction

THE Gaussian Mixture Regression (GMR) is an efficient regression technique derived from the well-known Gaussian Mixture Model (GMM) [1]. The GMR is widely used in different areas of speech processing, e.g. voice conversion [2], [3], in image processing, e.g. head pose estimation from depth data [4], generation of hand writing [5], and in robotics [6], [7]. In the present paper, we consider the application of GMR to the *speech acoustic-articulatory inversion problem*, i.e. estimating trajectories of speech articulators (jaws, lips, tongue, palate) from speech acoustic data [8]–[10]. Such model can be used in the context of pronunciation training to automatically animate a virtual talking head displaying the internal speech articulators, using only the speaker's voice. Such acoustic-articulatory GMR is generally trained on a large dataset of input-output joint observations recorded on a single speaker, later on referred to as the *reference* speaker. Using this reference GMR with the speech signal produced by a new speaker (hereafter referred to as the *source* speaker) can lead to poorly estimated articulatory trajectories. Indeed, because of the differences in the voice characteristics and in the speech production strategies across speakers, the new input data does not follow the statistical distribution of the reference acoustic data. Therefore, in this paper we address the problem of GMR *speaker adaptation*: We consider a GMR adaptation process that can be used to easily adapt a virtual talking head to any new speaker. Moreover, the adaptation process must be designed to work with a tiny set of *input-only*, i.e. acoustic, observations from the source speaker (in practice using a few sentences), in order to guarantee a user-friendly non-invasive system. Indeed, in real-world applications collecting data from a new user comes at high cost, especially for articulatory data.

The general speaker adaptation and normalization problem has been considered in, e.g., [11], [12]. In a recent study [13], the articulatory data from a new source speaker are obtained as a weighted sum of articulatory data from a set of different reference speakers, which is reminiscent of the eigenvoice decomposition of [14]. One limitation of this method is that it necessitates several parallel sets of acoustic-articulatory data from different speakers. The present study avoids this constraint by using acoustic and articulatory data from a single reference speaker. Regarding previous GMR-based approaches, [15] proposed to adapt the model parameters related to input observations using two state-of-the-art adaptation techniques for GMM, namely: maximum a posteriori (MAP) [16] and maximum likelihood linear regression (MLLR) [17]. Then, we proposed a general framework called cascaded GMR (C-GMR) and derived two implementations [18]. The first one, referred to as Split-C-GMR (SC-GMR), is a simple chaining of two separate GMRs: a first

Manuscript received October 26, 2016; revised December 21, 2016; accepted December 23, 2016. Date of publication January 11, 2017; date of current version February 9, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sabato Marco Siniscalchi.

L. Girin is with the University of Grenoble Alpes, GIPSA-lab, Grenoble 38040, France, and also with INRIA Grenoble Rhône-Alpes, Montbonnot 56521, France (e-mail: laurent.girin@gipsa-lab.grenoble-inp.fr).

T. Hueber is with the CNRS, University of Grenoble Alpes, GIPSA-lab, Grenoble 38040, France (e-mail: thomas.hueber@gipsa-lab.grenoble-inp.fr).

X. Alameda-Pineda is with INRIA Grenoble Rhône-Alpes, Montbonnot 56521, France (e-mail: xavier.alameda-pineda@inria.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2017.2651398

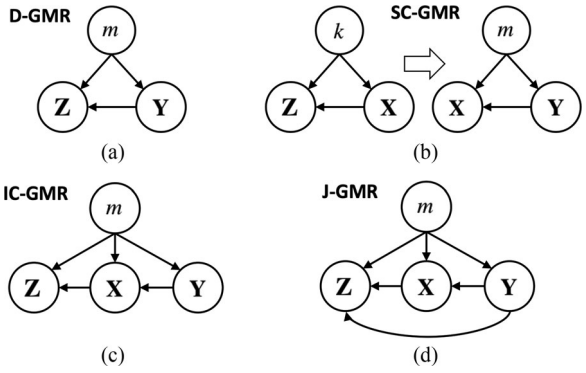

Fig. 1. Graphical representation of the generative models associated to the D-, SC-, IC-, and J-GMR. In the present applicative framework, $\mathbf{Y}$ is a reference articulatory feature vector, $\mathbf{X}$ is a reference acoustic feature vector, and $\mathbf{Z}$ is a source acoustic feature vector.

GMR maps the source acoustic feature vector, denoted $\mathbf{Z}$, into a reference acoustic feature vector, denoted $\mathbf{X}$, and then a second GMR maps $\mathbf{X}$ into the output articulatory feature vector, denoted $\mathbf{Y}$, lying in the reference speaker articulatory space (see Fig. 1(b)). The second implementation, referred to as Integrated-C-GMR (IC-GMR) combines the two successive mappings in a single probabilistic model (see Fig. 1(c)). Indeed, the $\mathbf{Z}$-to-$\mathbf{X}$ and $\mathbf{X}$-to-$\mathbf{Y}$ mappings are integrated at the mixture component level, sharing the $\mathbf{X}$ space. Importantly, the EM algorithm associated to the IC-GMR model [18] uses the general methodology of *missing data* [19], [20], explicitly taking into account the tiny amount of adaptation data from the source speaker. Specifically, the source data consisted in a small subset of the sentences of the reference training set, and the complement of this subset was considered missing. The IC-GMR was shown to provide superior performance to the SC-GMR and also to a direct GMR between $\mathbf{Z}$ and $\mathbf{Y}$ that disregards the $\mathbf{X}$ data (D-GMR, see Fig. 1(a)).

As seen in Fig. 1(c), the IC-GMR does not explicitly model any direct statistical dependency between $\mathbf{Z}$ and $\mathbf{Y}$ (i.e. in the graphical model, there is no arrow between $\mathbf{Z}$ and $\mathbf{Y}$). In other words, the cascade is "forced" to pass through $\mathbf{X}$, the reference speaker's acoustic space. In a general manner, adding such link would enable the output $\mathbf{Y}$ to be jointly inferred from $\mathbf{Z}$ and $\mathbf{X}$. In the above-mentioned limited parallel dataset strategy [18] (the source data consist in a small subset of the sentences of the reference training set) the acoustics of the source and the reference speaker are not physically linked, but they share the same phonetic content. Therefore, adding the $\mathbf{Z}$-$\mathbf{Y}$ link to the IC-GMR model enables to exploit the correlation associated to the shared phonetic content. Even if the direct $\mathbf{ZY}$ correlation happened to be weaker than the other cross-correlations ($\mathbf{ZX}$ and $\mathbf{XY}$), the impact of exploiting this direct link and thus estimating $\mathbf{Y}$ *jointly from* $\mathbf{X}$ *and* $\mathbf{Z}$, remains to be explored and properly evaluated. The resulting generative probabilistic model is equivalent to a joint multi-variate GMM on $\{\mathbf{Z}, \mathbf{X}, \mathbf{Y}\}$, and we can thus refer to this model as Joint GMM (J-GMM), and to the associated regressor as J-GMR.

The research question addressed in this paper is: "Is there any benefit of explicitly modeling a direct link between the source speaker's acoustics ($\mathbf{Z}$) and the reference speaker's articulation ($\mathbf{Y}$), with special emphasis on the case of very limited amount of adaptation data?" To this aim:

1) We present the J-GMR model and the corresponding exact EM training algorithm with missing data;
2) We provide an extensive evaluation of the J-GMR model on both synthetic acoustic-articulatory data and on the multi-speaker MOCHA EMA database, and compare its performance to other models of the C-GMR framework (i.e. SC-GMR, and IC-GMR).

At this point, it must be mentioned that an alike J-GMM-based regressor was proposed in [21], but with different assumptions regarding the nature of the adaptation data. Besides, in [21] the training algorithm for the J-GMM was not fully detailed: Only parameters update rules were given, and importantly, these update rules are different from the ones we derive in the present paper. Also importantly, the inference equation used in [21] is actually the IC-GMR inference equation, and thus it does not correspond to the structure of the J-GMM.[1] In other words, the inference equation in [21] does not exploit the direct link between $\mathbf{Z}$ and $\mathbf{Y}$, leading to a potential inconsistency between training and regression stages. Therefore, one additional goal of the present article is to wrap those very recent studies in GMR-based acoustic-articulatory inversion (i.e. [21], [15], [18]) and provide a complete and consistent methodological framework for training and using the J-GMR model.

The remaining of this paper is organized as follows. Section II gives a brief technical presentation of the GMR speaker adaptation problem in the present acoustic-articulatory inversion context, and of the D-GMR, SC-GMR and IC-GMR solutions. Section III presents and discusses the J-GMM and the associated J-GMR inference equation. The complete derivation of the corresponding EM algorithm under the missing data configuration of [18] is presented in Section IV. The theoretical differences with the solution proposed in [21] are further discussed. In Section V, we evaluate the practical performance of the proposed J-GMR under the task of speech acoustic-to-articulatory inversion with two different datasets, one synthetic and one of real data, and compare it to the performance of the D-GMR, SC-GMR, and IC-GMR. We discuss the research question risen above in the light of the experimental results. Section VI concludes the paper.

II. Cascaded GMR

A. Definitions, Notations and Working Hypothesis

Let us consider a GMM and the associated GMR between realizations of input $\mathbf{X}$ and output $\mathbf{Y}$ random (column) vectors, of arbitrary finite dimension. In the present speech acoustic-to-articulatory inversion framework, $\mathbf{Y}$ is an articulatory feature vector and $\mathbf{X}$ is a corresponding acoustic feature vector,

[1] Note that this IC-GMR inference equation was initially proposed in [15], without the overall generative framework.

both from the reference speaker. Let us define $\mathbf{V} = [\mathbf{X}^\top, \mathbf{Y}^\top]^\top$ where $^\top$ denotes the transpose operator. Let $p(\mathbf{X} = \mathbf{x}; \mathbf{\Theta}_\mathbf{X})$ denote the probability density function (PDF) of $\mathbf{X}$.[2] Let $\mathcal{N}(\mathbf{x}; \mu_\mathbf{X}, \mathbf{\Sigma}_{\mathbf{XX}})$ denote the Gaussian distribution evaluated at $\mathbf{x}$ with mean vector $\mu_\mathbf{X}$ and covariance matrix $\mathbf{\Sigma}_{\mathbf{XX}}$. Let $\mathbf{\Sigma}_{\mathbf{XY}}$ denote the cross-covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$ and $\mathbf{\Lambda}_{\mathbf{XX}}$ the precision matrix of $\mathbf{X}$ (similarly for cross-terms). With these notations, the PDF of a GMM on $\mathbf{V}$ writes:

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

where M is the number of components, $p(m) = \pi_m \geq 0$ and $\sum_{m=1}^{M} \pi_m = 1$. Let $\mathcal{D}_\mathbf{xy} = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N}$ denote a large training dataset of N i.i.d. vector pairs drawn from the $(\mathbf{X}, \mathbf{Y})$ distribution. In practice, these data are feature vectors extracted from synchronized articulatory and acoustic recordings of the reference speaker. The parameters of the above GMM reference model are estimated from $\mathcal{D}_\mathbf{xy}$, using an EM algorithm. Then, inference of $\mathbf{y}$ given a new observed value $\mathbf{x}$ can be performed by the minimum mean squared error (MMSE) estimator, which is the posterior mean $\hat{\mathbf{y}} = \mathrm{E}[\mathbf{Y}|\mathbf{x}]$:

$$\hat{\mathbf{y}} = \sum_{m=1}^{M} p(m|\mathbf{x}) \left(\mu_{\mathbf{Y},m} + \mathbf{\Sigma}_{\mathbf{YX},m}\mathbf{\Sigma}_{\mathbf{XX},m}^{-1}(\mathbf{x} - \mu_{\mathbf{X},m})\right), \quad (2)$$

with $p(m|\mathbf{x}) = \frac{\pi_m \mathcal{N}(\mathbf{x}; \mu_{\mathbf{X},m}, \mathbf{\Sigma}_{\mathbf{XX},m})}{\sum_{k=1}^{M} \pi_k \mathcal{N}(\mathbf{x}; \mu_{\mathbf{X},k}, \mathbf{\Sigma}_{\mathbf{XX},k})}$. Alternatively, one may consider maximum a posteriori (MAP) inference using $\hat{\mathbf{y}} = \arg\max_\mathbf{y} p(\mathbf{y}|\mathbf{x})$.

Let us now consider a new input vector $\mathbf{Z}$ following a different statistical distribution that the one of $\mathbf{X}$. Here, $\mathbf{Z}$ is an acoustic feature vector from the source speaker, to which the reference GMR has to be adapted. We assume that a tiny dataset $\mathcal{D}_\mathbf{z}$ of new input vectors $\mathbf{z}$ is available for the adaptation. As in [18], we assume that $\mathcal{D}_\mathbf{z}$ can be aligned with a subset of the reference input dataset: This requires that the new speaker pronounces a subset of the sentences contained in $\mathcal{D}_\mathbf{xy}$ and that these new recordings are time-aligned with the corresponding recordings of the reference speaker (e.g. using dynamic time warping (DTW) techniques). Since the working hypothesis is that the data tuples are i.i.d., we can reorder the dataset and write without loss of generality $\mathcal{D}_\mathbf{z} = \{\mathbf{z}_n\}_{n=1}^{N_0}$, with $N_0 \ll N$.

B. D-GMR, SC-GMR and IC-GMR

In this section, we briefly recall the three approaches for GMR adaptation considered in [18], which will be used here as baseline. Their graphical representation in Fig. 1 has already been discussed.

The first one is a direct $\mathbf{Z}$-to-$\mathbf{Y}$ GMR (D-GMR). Inference of $\mathbf{y}$ given an observed value $\mathbf{z}$ is done using (2), replacing $\mathbf{x}$ and $\mathbf{X}$ with $\mathbf{z}$ and $\mathbf{Z}$:

$$\hat{\mathbf{y}} = \sum_{m=1}^{M} p(m|\mathbf{z}) \left(\mu_{\mathbf{Y},m} + \mathbf{\Sigma}_{\mathbf{YZ},m}\mathbf{\Sigma}_{\mathbf{ZZ},m}^{-1}(\mathbf{z} - \mu_{\mathbf{Z},m})\right), \quad (3)$$

with $p(m|\mathbf{z}) = \frac{\pi_m \mathcal{N}(\mathbf{z}; \mu_{\mathbf{Z},m}, \mathbf{\Sigma}_{\mathbf{ZZ},m})}{\sum_{k=1}^{M} \pi_k \mathcal{N}(\mathbf{z}; \mu_{\mathbf{Z},k}, \mathbf{\Sigma}_{\mathbf{ZZ},k})}$. The parameters are trained with $\mathcal{D}_\mathbf{zy} = \{\mathbf{z}_n, \mathbf{y}_n\}_{n=1}^{N_0}$.

The second and third models are instances of cascaded GMR. As mentioned in the introduction, the Split-Cascaded GMR (SC-GMR) consists of chaining two distinct GMRs: a $\mathbf{Z}$-to-$\mathbf{X}$ GMR followed by the reference $\mathbf{X}$-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. Note also that the first GMR is trained with the N_0 samples of $\mathcal{D}_\mathbf{zx} = \{\mathbf{z}_n, \mathbf{x}_n\}_{n=1}^{N_0}$, while the second GMR is the reference GMR trained with the N samples of $\mathcal{D}_\mathbf{xy}$.

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. 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 generative mixture model is defined as:

$$p(\mathbf{o}) = \sum_{m=1}^{M} p(m)p(\mathbf{y}|m)p(\mathbf{x}|\mathbf{y}, m)p(\mathbf{z}|\mathbf{x}, m), \quad (4)$$

where all PDFs are Gaussian, and we have defined $\mathbf{O} = [\mathbf{X}^\top, \mathbf{Y}^\top, \mathbf{Z}^\top]^\top$ for concision. We show in Section III that (4) is a particular case of GMM and we thus refer to this generative model as the IC-GMM. The corresponding inference equation, i.e. the IC-GMR, is given by [18]:

$$\hat{\mathbf{y}} = \sum_{m=1}^{M} p(m|\mathbf{z}) \times \left(\mu_{\mathbf{Y},m} + \mathbf{\Sigma}_{\mathbf{YX},m}\mathbf{\Sigma}_{\mathbf{XX},m}^{-1}\mathbf{\Sigma}_{\mathbf{XZ},m}\mathbf{\Sigma}_{\mathbf{ZZ},m}^{-1}(\mathbf{z} - \mu_{\mathbf{Z},m})\right). \quad (5)$$

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

$$\mathbf{\Sigma}_{\mathbf{YZ},m} = \mathbf{\Sigma}_{\mathbf{YX},m}\mathbf{\Sigma}_{\mathbf{XX},m}^{-1}\mathbf{\Sigma}_{\mathbf{XZ},m}. \quad (6)$$

The IC-GMR is trained with the complete set of available $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ data, i.e. $\mathcal{D}_\mathbf{z} \cup \mathcal{D}_\mathbf{xy}$, using the EM algorithm presented in [18].

III. JOINT GMR

In this section we present the proposed *Joint* GMM generative model (J-GMM) and the associated inference equation. We also discuss the relationship with previous works, i.e. [21] and [18]. The Joint GMM on $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ is defined as:

$$p(\mathbf{o}) = \sum_{m=1}^{M} p(m)p(\mathbf{o}|m; \mathbf{\Theta}_m) = \sum_{m=1}^{M} \pi_m \mathcal{N}(\mathbf{o}; \mu_m, \mathbf{\Sigma}_m), \quad (7)$$

[2] In the following, for concision we omit $\mathbf{X}$ and we may omit $\mathbf{\Theta}_\mathbf{X}$, depending on the context.

where $\mathbf{\Theta}_m = \{\mu_m, \mathbf{\Sigma}_m\}$ are the parameters of the m-th Gaussian component, and thus $\mathbf{\Theta} = \cup_{m=1}^{M}\{\pi_m, \mathbf{\Theta}_m\}$. In order to derive the associated inference equation we first compute:

$$p(\mathbf{y}|\mathbf{z}) = \int_{\mathbf{X}} \sum_{m=1}^{M} p(\mathbf{x}, \mathbf{y}, m|\mathbf{z}) d\mathbf{x} = \sum_{m=1}^{M} p(m|\mathbf{z}) p(\mathbf{y}|\mathbf{z}, m). \quad (8)$$

Since the conditional and marginal distributions of a Gaussian are Gaussian as well, (8) is a GMM. Therefore, the J-GMR inference equation under the MMSE criterion turns out to be identical to the usual expression for a direct $\mathbf{Z}$-to-$\mathbf{Y}$ GMR, i.e. (3).[3] At first sight, this may look a bit strange since this gives the impression of by-passing the information contained in $\mathbf{X}$. However, this is not the case: although its inference equation is identical, the complete "joint" process for GMR adaptation is not equivalent to a GMR build directly from $(\mathbf{z}, \mathbf{y})$ training data, i.e. the D-GMR. Indeed, as shown in the next section, the estimation of the J-GMR parameters with the EM algorithm uses all the available data, i.e. $\mathcal{D}_{\mathbf{xy}}$ and $\mathcal{D}_{\mathbf{z}}$, hence including all $\mathbf{x}$ data. In summary, the D-GMR and the J-GMR inference equations are identical but these two models differ by their underlying generative model and associated training procedure, leading to different parameter values (even when using the same adaptation dataset).

As mentioned in the introduction, a J-GMM-based model has already been considered in [21] as the underlying generative model of $\mathbf{O}$ in the present speaker adaptation problem. However, [21] performs MAP inference instead of MMSE inference. More importantly, even if the underlying generative model is a J-GMM, the inference equation in [21] corresponds to the IC-GMR. In details, $p(\mathbf{o})$ in (1-2) of [21] corresponds to (7), while $p(\mathbf{y}|\mathbf{z})$ in (6) of [21] corresponds to (5). Indeed, this posterior PDF $p(\mathbf{y}|\mathbf{z})$ assumes no direct link between $\mathbf{Z}$ and $\mathbf{Y}$, which is correct for the IC-GMM but incorrect for the J-GMM. Thus the inference in [21] is not consistent with the J-GMM.[4]

Remarkably, (6) characterizes the IC-GMR developed in [18] as a particular case of the J-GMR. In Appendix A, we show that this is also true at the mixture model level, i.e. the IC-GMM (4) is a particular case of the J-GMM (7) with (6). The matrix product $\mathbf{\Sigma}_{\mathbf{XZ},m}\mathbf{\Sigma}_{\mathbf{ZZ},m}^{-1}$ in (5) 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 is shown in the next section.

[3]Alternately a MAP estimator can be used by taking the argmax of (8).

[4]Note that in [21], the details of the derivation of the intermediate form (6) into the GMR form (7) are not provided. In contrast, we provided detailed derivation in [18], where (19) is shown to result into two equivalent forms of a GMR expression (25) and (26). Also, to be fully precise, (7) in [21] corresponds to (26) in [18] up to two differences that we interpret as typos: First, the term $\mathbf{\Sigma}_m^{(x,x)}$ in (9) of [21] should be $\mathbf{\Sigma}_m^{(x,x)^{-1}}$, see e.g., (5) in [18]; and second, a right-sided term $\mathbf{\Sigma}_m^{(x,x)^{-1}}$ is missing in (10) in [21], i.e. $\mathbf{\Sigma}_m^{(a,s)}\mathbf{\Sigma}_m^{(y,y)^{-1}}\mathbf{\Sigma}_m^{(y,x)}$ should be $\mathbf{\Sigma}_m^{(a,s)}\mathbf{\Sigma}_m^{(y,y)^{-1}}\mathbf{\Sigma}_m^{(y,x)}\mathbf{\Sigma}_m^{(x,x)^{-1}}$ (see (26) in [18]). Without this matrix, "unnormalized" input data are propagated into the mapping process.

IV. EM Algorithm for J-GMM with Missing Data

This section introduces the exact EM algorithm associated to the J-GMM, explicitly handling an incomplete adaptation dataset using the general methodology of missing data (with the same data configuration as in [18], which was reminded in Section II-A). The EM iteratively maximizes the expected complete-data log-likelihood. At iteration $i+1$, the E-step computes the auxiliary 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 first describe the E and M steps, then we detail the initialization process. Finally we comment the link between the EM algorithms of the IC-GMM and J-GMM, and the differences between the proposed EM and the EM for J-GMM given in [21]. The associated source code is available at: `https://git.gipsa-lab.grenoble-inp.fr/cgmr.git`.

A. E-step

In order to derive the auxiliary function $Q(\mathbf{\Theta}, \mathbf{\Theta}^{(i)})$, we follow the general methodology given in [20]–(Section 9.4) and [22]. In [18], we have shown that this leads to the general expression:

$$\begin{aligned} Q(\mathbf{\Theta}, \mathbf{\Theta}^{(i)}) = &\sum_{n=1}^{N_0}\sum_{m=1}^{M} \gamma_{nm}^{(i+1)} \log p(\mathbf{o}_n, m; \mathbf{\Theta}_m) \\ &+ \sum_{n=N_0+1}^{N}\sum_{m=1}^{M} \frac{1}{p(\mathbf{v}_n; \mathbf{\Theta}_{\mathbf{V}}^{(i)})} \int p\left(\mathbf{o}_n, m; \mathbf{\Theta}_m^{(i)}\right) \\ &\times \log p(\mathbf{o}_n, m; \mathbf{\Theta}_m) d\mathbf{z}_n. \end{aligned} \quad (9)$$

where

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

are the so-called *responsibilities* (of component m explaining observation $\mathbf{o}_n$) [20]. Eq. (9) is valid for any mixture model on i.i.d. vectors $(\mathbf{V}, \mathbf{Z})$ with partly missing $\mathbf{z}$ data. Here we study the particular case of the J-GMM. For this aim, we denote $\mu_{\mathbf{Z}|\mathbf{v}_n,m}^{(i)}$ the posterior mean of $\mathbf{Z}$ given $\mathbf{v}_n$ and given that the data were generated by the m-th Gaussian component with parameters $\mathbf{\Theta}_m^{(i)}$:

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

Let us define $\mathbf{o}'_{nm} = [\mathbf{v}_n^\top, \mu_{\mathbf{Z}|\mathbf{v}_n,m}^{(i)\top}]^\top$ 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 (11). Let us arbitrarily extend $\mathbf{o}'_{nm}$ with $\mathbf{o}'_{nm} = \mathbf{o}_n$ for $n \in [1, N_0]$, and let us extend the definition of the responsibilities to the incomplete data vectors $\mathbf{v}_n$:

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

Then, $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(i)})$ is given by:

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(i)}) = \sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_{nm}^{(i+1)}\left(\log\pi_m - \frac{\log|\boldsymbol{\Sigma}_m| + (\mathbf{o}'_{nm}-\mu_m)^\top\boldsymbol{\Sigma}_m^{-1}(\mathbf{o}'_{nm}-\mu_m)}{2}\right) - \frac{1}{2}\sum_{m=1}^{M}\left(\sum_{n=N_0+1}^{N}\gamma_{nm}^{(i+1)}\right)\mathrm{Tr}\left[\boldsymbol{\Lambda}_{\mathbf{ZZ},m}(\boldsymbol{\Lambda}_{\mathbf{ZZ},m}^{(i)})^{-1}\right]. \tag{13}$$

The proof is given in Appendix B. The first double sum in (13) 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 the corresponding $\mathbf{x}$ and $\mathbf{y}$ data and the 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 the missing data.

B. 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 [20]. For $m \in [1, M]$, we have:

$$\pi_m^{(i+1)} = \frac{1}{N}\sum_{n=1}^{N}\gamma_{nm}^{(i+1)}. \tag{14}$$

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:

$$\mu_m^{(i+1)} = \frac{\sum_{n=1}^{N}\gamma_{nm}^{(i+1)}\mathbf{o}'_{nm}}{\sum_{n=1}^{N}\gamma_{nm}^{(i+1)}}. \tag{15}$$

This expression is the 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 (13) as a function of $\boldsymbol{\Sigma}_m^{-1}$ by completing $(\boldsymbol{\Lambda}_{\mathbf{ZZ},m}^{(i)})^{-1}$ with zeros to obtain the matrix $(\boldsymbol{\Lambda}_{\mathbf{ZZ},m}^{0\,(i)})^{-1}$:

$$(\boldsymbol{\Lambda}_{\mathbf{ZZ},m}^{0\,(i)})^{-1} = \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \left(\boldsymbol{\Lambda}_{\mathbf{ZZ},m}^{(i)}\right)^{-1}\end{bmatrix}. \tag{16}$$

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:

$$\boldsymbol{\Sigma}_m^{(i+1)} = \frac{1}{\sum_{n=1}^{N}\gamma_{nm}^{(i+1)}}\left[\sum_{n=1}^{N}\gamma_{nm}^{(i+1)}(\mathbf{o}'_{nm}-\mu_m)(\mathbf{o}'_{nm}-\mu_m)^\top + \left(\sum_{n=N_0+1}^{N}\gamma_{nm}^{(i+1)}\right)(\boldsymbol{\Lambda}_{\mathbf{ZZ},m}^{0(i)})^{-1}\right]. \tag{17}$$

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 $\boldsymbol{\Lambda}_{\mathbf{ZZ},m}^{(i)} = \left[(\boldsymbol{\Sigma}_m^{(i)})^{-1}\right]_{\mathbf{ZZ}} \neq (\boldsymbol{\Sigma}_{\mathbf{ZZ},m}^{(i)})^{-1}$.

C. EM Initialization

The initialization of the proposed EM algorithm takes a very peculiar aspect. Indeed, as a result of the nature of the adaptation process, the reference $\mathbf{X}$-$\mathbf{Y}$ GMM model is used to initialize the marginal parameters in $(\mathbf{X}, \mathbf{Y})$. As for the $\mathbf{Z}$ parameters, we adopt the two possible following strategies. Strategy 1 is data-driven: The marginal parameters in $\mathbf{Z}$ are initialized using the adaptation data $\mathcal{D}_\mathbf{z} = \{\mathbf{z}_n\}_{n=1}^{N_0}$. The cross-term parameters in $(\mathbf{Z}, \mathbf{X})$ and $(\mathbf{Z}, \mathbf{Y})$ are initialized by constructing the sufficient statistics using $\{\mathbf{z}_n, \mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N_0}$. Since the number of adaptation data is limited, and the related statistics may be poorly reliable, we also propose Strategy 2 which is a "blind" strategy: The $\mathbf{ZX}$ cross-covariance matrix is set to the identity matrix, the $\mathbf{ZY}$ cross-covariance matrix is set to zero and the covariance of $\mathbf{Z}$ is set to the covariance of $\mathbf{X}$. As shown in Section V, this simple blind initialization exhibited significantly better performance than the one exploiting the statistics of the adaptation set in our experiments, for small adaptation sets. Finally, remember that, whatever the initialization, *all model parameters are jointly updated* by alternating the E and M steps, using both reference data $\mathcal{D}_\mathbf{xy}$ and aligned adaptation data $\mathcal{D}_\mathbf{z}$.

D. Link between J-GMM and IC-GMM Revisited

We have seen in Section III that the IC-GMM is a particular (constrained) version of the J-GMM. However, the EM for the IC-GMM presented in [18] (which exploits the linear-Gaussian form of the IC-GMM) is *not* derivable as a particular case of the EM for J-GMM provided in the present section. More precisely, if one attempts to estimate the IC-GMR parameters with the algorithm we introduce in this section, the M-step should be constrained by (6). Naturally, the complexity of the resulting constrained algorithm would be much higher than the complexity of the (unconstrained) EM of [18]. 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 IC-GMM deals with constrained covariance matrices, whereas the J-GMM uses fully free covariance matrices.

E. Analysis of the Differences with the EM for J-GMM of [21]

As mentioned before, the data configuration is different than the one used in [21]. The EM of [21] exhibits symmetric terms relative to missing $\mathbf{z}$ data *and* missing $\mathbf{y}$ data. In the present study, we only considered missing $\mathbf{z}$ data. However it is straightforward to extend the proposed framework to the case of additional missing $\mathbf{y}$ data, and this would also lead to "symmetric

terms". One prominent feature of the current approach is that we consider the availability of a tiny set of fully complete data samples $(\mathbf{z}, \mathbf{x}, \mathbf{y})$, which comes at no cost. Indeed, in any case the source speaker is asked to pronounce a given set of sentences and we simply choose this set to be a subset of $\mathcal{D}_{\mathbf{xy}}$.

For such mixture models as the J-GMM, the E-step basically boils down to compute the responsibilities, i.e. the posterior weights that represent the contribution of each mixture component to explain the data. In the present paper, as well as in [18], the responsibilities are expressed directly as a function of the observed data only (and of the current estimate of the parameters). This is usual in EM with missing data. Surprisingly, in [21]–(11,12) the responsibilities are expressed as a function of the complete data vectors, where the missing data is replaced with an arbitrary estimate computed from observed data.

In the M-step of Section IV-B, the update of the parameters of the m-th component depends only on the missing data estimated using this component. Again this is what one naturally obtains with the principled formulation of EM algorithms with missing data [19], [20] (recall the definition of $\mathbf{o}'_{nm}$ after (11)).[5] In contrast, [21] uses the per-component estimates of the missing data (their equations (15) and (16) that are consistent with (11)) to compute estimates of the missing data that are averaged along components (their equations (13) and (14)). Consequently, their final estimates result in a dubious form (responsibilities can be grouped) and do not match the ones in (15) and (17), obtained following the proposed principled derivation.

Given that the derivation of the EM is not detailed in [21], it is difficult to know where these contradictions arise from. Due to the similarity in terminology and in the formulation, we believe that, at the very least, this discussion is required.

V. Experiments

The performance of the J-GMR was evaluated on the speech acoustic-to-articulatory inversion task (i.e. recovering movement of the tongue, lips, jaw and velum from speech acoustics), and compared to the D-GMR, SC-GMR and IC-GMR. Two series of experiments were conducted: the first one on synthetic data, the second one on real data.

A. Synthetic Data – Set Up

Experiments on synthetic data were conducted using a so-called articulatory synthesizer. This allowed us to carry out a first investigation of the J-GMR behavior by controlling finely the structure of the adaptation dataset (as opposed to the real data of Section V-B). A synthetic dataset of vowels was thus generated using the Variable Linear Articulatory Model (VLAM) [23]. 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 calculates the corresponding area function using 29 tubes of variable length and then deduces the corresponding spectrum using acoustic simulation [24]. 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. The difference in vocal tract length induces a shift of the formants along the frequency axis as illustrated in Fig. 2. Moreover, this shift is non-linear, justifying the use of a non-linear (or locally linear) mapping model such as the GMR.

[5]The same principle was observed in the EM for the IC-GMM in [18].

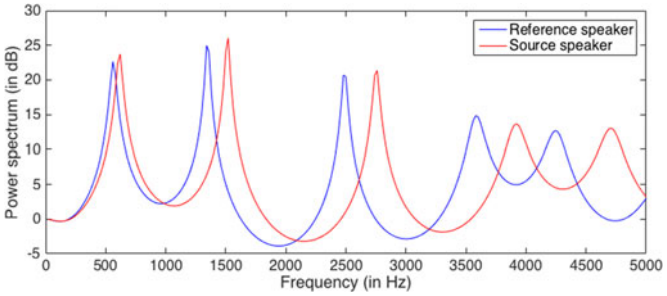

Fig. 2. Power spectra generated by VLAM for the same articulatory configuration but for two different vocal tract lengths.

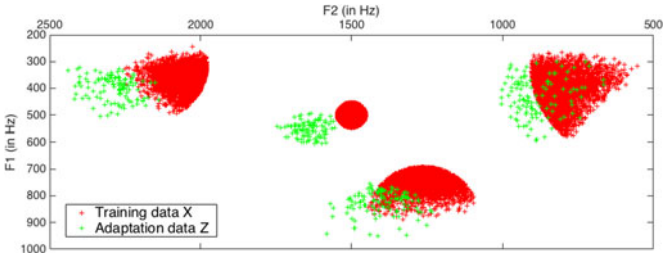

Fig. 3. Synthetic data generated using VLAM in the F2-F1 acoustic space.

We generated a dataset of $(\mathbf{z}, \mathbf{x}, \mathbf{y})$ triplets structured into four clusters simulating the 4 following vowels: /a/, /i/, /u/, /ə/. In these experiments, the spectrum is described by the position and the amplitude of the 4 first formants, which are easily captured on such synthetic data, hence 8-dimensional $\mathbf{x}$ and $\mathbf{z}$ observations. We generated 20,000 triplets (5,000 for each of the 4 vowels). These data are displayed in Fig. 3 (red points) along with a selection of 467 adaptation vectors (green points), in the two first formant frequencies (F1-F2) plane.

B. MOCHA EMA – Set Up

Experiments on real data were conducted using electromagnetic articulatory (EMA) recordings. We used the publicly available Multichannel Articulatory Database (MOCHA) [25] provided by the Center for Speech Technology Research (University of Edinburgh). It includes acoustic-articulatory data of two speakers: fsew0 (female) and msak0 (male). Both speakers uttered 460 sentences extracted from the British TIMIT corpus, representing 20.6 min of speech for fsew0, and 17.4 min of speech for msak0.

Mel-frequency cepstral coefficients (MFCC) were used here to represent the acoustic content of the speech signal. Each audio

observation ($\mathbf{x}$ and $\mathbf{z}$) was a 26-dimensional vector composed of 13 MFCC coefficients and their first derivatives. These vectors were extracted from the 16-kHz speech waveform every 10 ms, leading to a total of about 123,800 vectors for fsew0 and of about 104,600 vectors for msak0.

Regarding the articulatory data, each observation $\mathbf{y}$ was a 14-dimensional vector gathering the 2D coordinates of 7 electromagnetic actuation coils describing the lips, tongue, jaw and velum positions in the midsagittal plane of the reference speaker's vocal tract, every 10 ms. These articulatory data were normalized following the procedure described in [26]. This normalization consists in centering and whitening the data (i.e. subtracting the mean value of each feature and dividing by its standard deviation) on a per-file basis. The mean (resp. standard deviation) of each feature was then low-pass filtered to alleviate the DC drift observed in the raw MOCHA database (see Fig. 3.6, p. 71 in [26]). Note that this has become a de-facto standard procedure, see [8] and [9].

We conducted two series of experiments: adaptation of reference speaker fsew0 to source speaker msak0 (denoted msak0→ fsew0) and adaptation of reference speaker mask0 to source speaker fsew0 (denoted fsew0→ msak0).

C. Experimental Protocol

For the synthetic data, the complete set of $(\mathbf{z}, \mathbf{x}, \mathbf{y})$ triplets, are naturally aligned. For the MOCHA data, dynamic time warping (DTW) was used to time-align each of the sentences pronounced by the source speaker with the corresponding sentence pronounced by the reference speaker. The source speaker's acoustics was warped onto the reference speaker's acoustics (and by synchronicity onto the reference speaker's articulatory data).

For the experiments on the synthetic dataset, the EM algorithm for training the reference $\mathbf{X}$-$\mathbf{Y}$ model (and also the $\mathbf{Z}$-$\mathbf{X}$ model for the SC-GMR) was initialized using the k-means algorithm, repeated 5 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 30-fold cross-validation protocol: The data was divided in 30 subsets of approximate equal size, 29 subsets were used 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 during the training of the reference $\mathbf{X}$-$\mathbf{Y}$ model.[6] In the majority of these experiments, the optimal value of M was found to be 16. Similarly, the number K of components of the $\mathbf{Z}$-$\mathbf{X}$ model of the SC-GMR was set by cross-validation within the set $\{2, 4, 8, 12, 16\}$.

For the experiments with MOCHA, a similar procedure was used, though with different settings to adapt to the difference in dataset size and dimension. Here, all methods were evaluated under a 5-fold cross-validation protocol (four subsets for training and one subset for test, all of approximate equal size). In each of the five folds, $k/20$ of the size of the training set was used for adaptation, with $k \in [1, 10]$. This results in 50 experiments for each of the two aforementioned configurations (msak0→ fsew0 and fsew0→ msak0). As for M, the number of mixture components, cross-validation on the reference model for the MOCHA dataset let to an optimal value $M = 128$. However, the results for $M = 128, 64$, and 32 were found to be quite close, which is consistent with the results reported in previous literature [8]. Given that the J-GMR and IC-GMR models have more parameters than the reference model, and are thus more prone to overfitting, we chose to set $M = 32$. As for K, it was set using the same cross-validation procedure as for the synthetic data case.

[6]Remember that, in nature, the number of mixture components M of the J-GMR and IC-GMR is imposed by the reference model.

TABLE I
DATA-TO-PARAMETERS RATIO FOR THE SYNTHETIC DATASET AND FOR MOCHA (FOR BOTH SPEAKERS), FOR ALL MODELS AND FOR THE TWO EXTREME VALUES OF N_0 REPORTED IN FIGS. 4 AND 5

Data	Synthetic		fsew0→ msak0		msak0→ fsew0	
N_0	Low	High	Low	High	Low	High
Reference	137	137	121	121	144	144
D-GMR	3	34	6	61	7	71
SC-GMR	89	100	69	88	81	104
IC-GMR	77	87	56	72	67	86
J-GMR	63	70	47	61	56	72

For both synthetic and real data experiments, the performance was assessed by calculating the average Root Mean Squared Error (RMSE) between the articulatory trajectories estimated from the source speaker's acoustics, and the ones generated by the reference speaker (for the real data experiments, the reference speaker's acoustics and articulatory test data were aligned on the source speaker's acoustics using DTW). 95% confidence interval of RMSE measures were obtained using paired t-tests. For the synthetic data, the vectors were generated independently (i.e. with no temporal structure), hence each vector provides an independent RMSE measure. As for the MOCHA dataset, independence between samples was assumed by considering the average RMSE on 5 consecutive frames, i.e. 50 ms. Note that the RMSE values for the synthetic data are unitless, since the VLAM articulatory data are arbitrary control parameters.

In order to discuss complexity and accuracy issues for the different models, we define the data-to-parameters ratio (DPR) as the total number of (scalar) data divided by the total number of (scalar) parameters to estimate. This simple measure provides prior information on how much the model is prone to overfit: The lower the DPR is (meaning less training data or more complex models) the more the model is prone to overfitting. Table I presents the DPR values for the synthetic dataset and for MOCHA, for each model, and for the two extreme values of N_0 in the reported figures (see below). Note that the DPR is not a performance measure per se; it rather provides a potential explanation for the behavior of the models under evaluation. Indeed, in practice we observed that training models with DPR below 20 is risky, since the overfitting phenomenon may be predominant, impairing the generalization capabilities of the trained model.

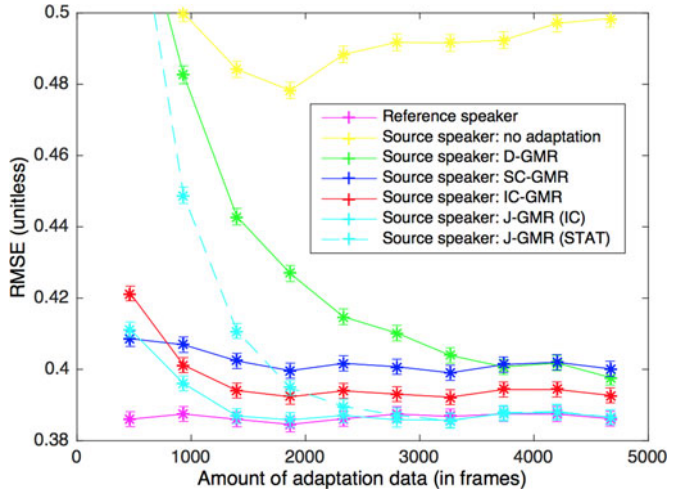

Fig. 4. RMSE (unitless) of the Z-to-Y mapping as a function of the size of the adaptation data (in number of vectors), for D-GMR, SC-GMR, IC-GMR and J-GMR (lower and upper bounds are given by the X-Y mapping in magenta and the Z-to-Y mapping with no adaptation in yellow; error bars represent 95% confidence intervals).

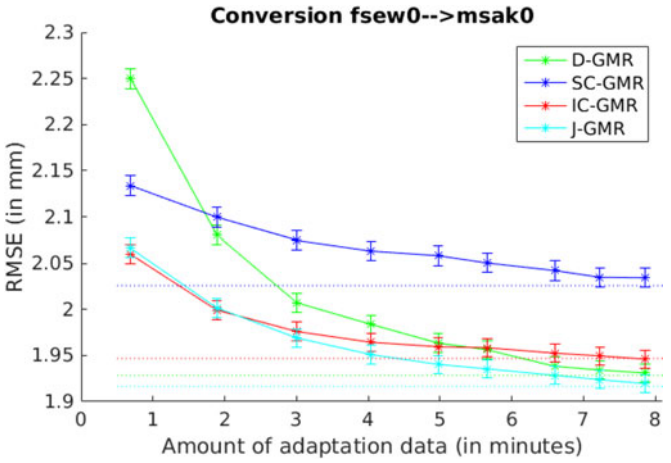

Fig. 5. RMSE (in mm) with 95% confidence intervals for source speaker fsew0 as a function of the amount of adaptation data, for D-, SC-, IC- and J-GMR, and their respective *oracle* in dotted lines.

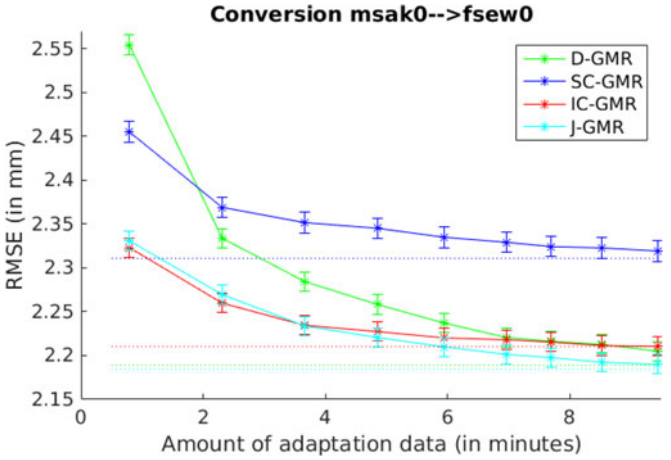

Fig. 6. RMSE (in mm) with 95% confidence intervals for source speaker msak0 as a function of the amount of adaptation data, for D-, SC-, IC- and J-GMR, and their respective *oracle* in dotted lines.

We can see in Table I that all values are significantly larger than 20, except for the D-GMR with small N_0, as will be discussed later.

D. Synthetic Data – Results

The RMSE for the J-GMR, as well as for the D-GMR, SC-GMR and IC-GMR are plotted in Fig. 4, as a function of N_0, the size of the adaptation set. The performance of the J-GMR, SC-GMR and IC-GMR are relatively 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, i.e. $\mathcal{D}_{\mathbf{zy}} = \{\mathbf{z}_n, \mathbf{y}_n\}_{n=1}^{N_0}$. As illustrated by the DPR values in Table I, this is a quite limited dataset compared to the dataset exploited by the C-GMR family, i.e. $\mathcal{D}_{\mathbf{z}} \cup \mathcal{D}_{\mathbf{xy}} = \{\mathbf{z}_n\}_{n=1}^{N_0} \cup \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N}$. For low N_0 values, this results in poor performance of the D-GMR, with possible severe overfitting. 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 [18], the IC-GMR performs better than the SC-GMR, except for the lower N_0 value. Importantly, we observe a systematic and statistically significant improvement of the proposed J-GMR over the IC-GMR, for all N_0 values. The gain of J-GMR over IC-GMR is within the approximate range 1.5%–2.5% of RMSE depending on N_0. Subsequently, the J-GMR also clearly outperforms the SC-GMR, except for the lower N_0 for which the difference between J-GMR and SC-GMR is not significant. These results illustrate that the J-GMR is able to better exploit the statistical relations between $\mathbf{z}$, $\mathbf{x}$ and $\mathbf{y}$ data compared to the other C-GMR models. Indeed, while the $\mathbf{Z}$-$\mathbf{X}$ and $\mathbf{X}$-$\mathbf{Y}$ statistical relationships are exploited by the SC-, IC- and J-GMR, only the latter directly exploits the $\mathbf{Z}$-$\mathbf{Y}$ statistical relationship. Therefore, only in the J-GMR the mapping is not forced to pass through $\mathbf{X}$, which is shown to be beneficial in this set of experiments.

Regarding the initialization strategy of the J-GMR, we notice that for the lower range of N_0 values the blind initialization strategy clearly outperforms the one based on the statistics of the adaptation set (denoted with the suffix "(STAT)" in Fig. 4). This shows that in that case, the amount of adaptation data is not sufficient to calculate reliable statistics to be exploited in model parameter estimation. When the adaptation set grows in size (over approx. 3,000 vectors), the difference in performance between the two initialization strategies becomes not significant, if any. Therefore, in the following, we will favor the blind initialization strategy.

E. MOCHA EMA – Results

The results of the experiments fsew0→ msak0 and msak0→ fsew0 on the MOCHA EMA dataset are shown in Figs. 5 and 6

respectively. Here also, the curves plot the RMSE against the amount of adaptation data. Similarly to [18] and similarly to the synthetic data experiments, for small adaptation sets, the IC-GMR clearly outperforms the D-GMR model. This is observed for the two source speakers msak0 and fsew0. The same tendency is observed with the proposed J-GMR model since the J-GMR performance is close to the IC-GMR performance (see below). SC-GMR also outperforms D-GMR, but only for the lowest N_0 value, since the difference between SC-GMR and IC-GMR is higher than in the synthetic data case. Altogether, these first general results confirm the results obtained on synthetic data and, again, they can be explained by the fact that the D-GMR exploits only the reference speaker's articulatory data that can be associated with the source speaker's audio data (see the small corresponding DPR values in Table I). This corroborates the benefit of (i) relying on an intermediate representation space, for instance the reference acoustic space $\mathbf{X}$, and (ii) exploiting all available $(\mathbf{x}, \mathbf{y})$ observations during the adaptation process. The fact that both J-GMR and IC-GMR clearly outperform SC-GMR everywhere seems to support the interest of a model structure where $\mathbf{X}$ is a single common representation space tied to both input $\mathbf{Z}$ and output $\mathbf{Y}$ at the mixture component level (as already observed for the IC-GMR in [18]).

As for the comparison between J-GMR and IC-GMR, these results also confirm the potential interest of using the J-GMR method over the IC-GMR. Indeed, while the two methods perform closely for tiny amounts of adaptation data, the J-GMR exhibits better results than the IC-GMR for larger amounts of adaptation data. More precisely, we can identify three different zones in the RMSE plots of both source speakers. First the data scarcity zone (below 3 min of adaptation data), where the IC-GMR shows equivalent performance than the J-GMR (for fsew0→ mska0 conversion) or slightly better performance but not in a statistically significant manner (for the msak0→ fsew0).

Second, the data abundance zone (above 7 min and more than 9 min of adaptation data for fsew0→ msak0 and msak0→ fsew0 respectively), where the D-GMR has enough data to show competitive performance compared to the J-GMR (see the correct DPR values for the D-GMR for high N_0 in Table I). At the same time, the RMSE of the IC-GMR is here higher than the RMSE of D-GMR and J-GMR in a statistically significant manner. Therefore, it would appear that the constraint associated to the IC-GMR model penalizes its performance when enough adaptation data is available. This would suggest that more data implies more complex underlying links, some of which cannot be captured well by the IC-GMR model. This explanation is reinforced by the results under the so-called "oracle" settings, when all data is used at adaptation time, i.e. $N_0 = N$, which can be seen as the right limit of the plots. The result of the oracle settings for the four models are represented with dotted lines in Figs. 5 and 6. We can see that the J-GMR is able to better exploit the overall statistical correlations than the IC-GMR. Interestingly, the J-GMR oracle RMSE is below the D-GMR oracle RMSE, whereas the IC-GMR oracle RMSE is above. Hence, even for large adaptation data, it appears to be a good thing to exploit $\mathbf{x}$ at the mixture component level, but it is not such a good thing to do it in a too constrained manner.

This behavior is also observed, in a somewhat less intense manner, in the third zone (between 3 min and 7/9 min of adaptation data). Here the IC-GMR starts exhibiting worse performance than J-GMR (the difference is statistically significant from 5 min and 7 min of adaptation data for fsew0→ msak0 and msak0→ fsew0, respectively). At the same time, the D-GMR does not have enough data yet to approach the performance of the J-GMR. Our understanding is that, within this range, the complexity of the adaptation data overwhelms the IC-GMR, while not yet containing enough information to optimally exploit the $\mathbf{Z}$-$\mathbf{Y}$ link.

Overall, the privileged choice for cross-speaker acoustic-articulatory inversion appears to be the J-GMR. Indeed, if not enough adaptation data is available, the J-GMR has equivalent or close performance to the IC-GMR. In case a large amount of adaptation data is available, the J-GMR and the D-GMR perform closely, with a small advantage for the J-GMR, and this is further confirmed by the oracle results. Finally, the J-GMR has proven to be the most effective model in half-way situations between adaptation data scarcity and abundance.

VI. Conclusion

In this paper, we extended the general framework of Cascaded-GMR introduced in [18] with a new model called J-GMR. Similarly to the IC-GMR of [18], the J-GMR relies on a common intermediate $\mathbf{X}$ space tying the input and output spaces at the mixture component level. But in contrast to the IC-GMR, the J-GMR enables a direct link between input and output in addition to the $\mathbf{Z}$-$\mathbf{X}$-$\mathbf{Y}$ cascaded path. We provided the exact EM training algorithm for the J-GMR explicitly considering missing input data, and applied this model to the cross-speaker acoustic-articulatory inversion task.

The reported experiments on both synthetic and real data show that the J-GMR outperforms the D-GMR, especially for small adaptation datasets, as was already observed for the IC-GMR in [18]. Moreover, we can provide an answer to the question stated in the introduction: Including an explicit link to the probabilistic model between the reference speaker's articulatory space and the source speaker's auditory space is beneficial for the present adaptation task. On the synthetic dataset, the J-GMR outperforms systematically the IC-GMR. On the real data, the J-GMR performs similarly to the IC-GMR for limited adaptation datasets but outperforms the IC-GMR for larger ones. The data-to-parameters ratio of the J-GMR is slightly inferior to the one of the IC-GMR, reflecting a slightly higher complexity of the J-GMR over the IC-GMR. However, in our experimental set-up this difference did not have a negative effect on the performance of the J-GMR.

More generally, this article extends the Cascaded-GMR framework both from the theoretical and experimental perspectives. We believe that all models from this library of cascaded GMR adaptation techniques can be of potential interest for other applications, depending on the amount of adaptation data and their latent structure.

This research line could be enriched in, at least, two directions. First, by investigating the use of other methodological

frameworks, such as deep neural networks architectures or robust low-rank techniques [27], [28]. Second, by running extensive experiments in real-world scenarios, beyond laboratory conditions.

ACKNOWLEDGMENT

The authors would like to thank L.-J. Boë for his help with the VLAM model.

APPENDIX A
LINK BETWEEN THE J-GMM AND IC-GMM

We show here that the IC-GMM (4) more largely presented in [18] is a particular case of the J-GMM (7) with (6). Without loss of generality, the density components of the J-GMM can be rewritten as:

$$p(\mathbf{o}|m) = \pi_m p(\mathbf{y}|m)p(\mathbf{x}|\mathbf{y},m)p(\mathbf{z}|\mathbf{x},\mathbf{y},m), \quad (18)$$

where all pdfs are Gaussian. Here the conditional pdf of $\mathbf{Z}$ depends on both $\mathbf{x}$ and $\mathbf{y}$, whereas in the IC-GMM it depends only on $\mathbf{x}$ (see Fig. 1). Setting $p(\mathbf{z}|\mathbf{x},\mathbf{y},m) = p(\mathbf{z}|\mathbf{x},m)$ is equivalent to say that $\mathbf{Z}$ and $\mathbf{Y}$ are *conditionally independent* given $\mathbf{x}$, which can be expressed equivalently as $p(\mathbf{y},\mathbf{z}|\mathbf{x},m) = p(\mathbf{y}|\mathbf{x},m)p(\mathbf{z}|\mathbf{x},m)$ [20]–Section 8.2. Let us denote $\mathbf{U} = [\mathbf{Y}^\top, \mathbf{Z}^\top]^\top$. $p(\mathbf{u}|\mathbf{x},m)$ is a Gaussian pdf with covariance matrix $\mathbf{\Sigma}_{\mathbf{UU}|\mathbf{x},m} = \mathbf{\Sigma}_{\mathbf{UU},m} - \mathbf{\Sigma}_{\mathbf{UX},m}\mathbf{\Sigma}^{-1}_{\mathbf{XX},m}\mathbf{\Sigma}_{\mathbf{XU},m}$ [20, Sec. 2.3]. It is easy to show that the block diagonal term of this matrix is $\mathbf{\Sigma}_{\mathbf{YZ},m} - \mathbf{\Sigma}_{\mathbf{YX},m}\mathbf{\Sigma}^{-1}_{\mathbf{XX},m}\mathbf{\Sigma}_{\mathbf{XZ},m}$. Therefore, the conditional independence holds if and only if this block-diagonal term is null, i.e. (6). Alternately, we can write $p(\mathbf{o}|m)$ as a multivariate Gaussian and decompose the argument of the exponential function: it is then easy to show that $p(\mathbf{z}|\mathbf{x},\mathbf{y},m) = p(\mathbf{z}|\mathbf{x},m)$ for all values of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ (and m), if and only if all entries of $\mathbf{\Lambda}_{\mathbf{ZY},m} (= \mathbf{\Lambda}^\top_{\mathbf{YZ},m})$ are zero, for all $m \in [1, M]$. Of course the two conditions are equivalent: Since $\mathbf{\Lambda}_{\mathbf{UU},m} = \mathbf{\Sigma}^{-1}_{\mathbf{UU}|\mathbf{x},m}$ [20]–(2.79) and (2.82), $\mathbf{\Lambda}_{\mathbf{UU},m}$ is block-diagonal if and only if $\mathbf{\Sigma}_{\mathbf{UU}|\mathbf{x},m}$ is block-diagonal.

APPENDIX B
CALCULATION OF Q FOR THE JOINT GMM MODEL

In [18], we provided the general expression (9) of Q which is valid for any mixture model of the form $p(\mathbf{o};\mathbf{\Theta}) = \sum_{m=1}^{M} p(m)p(\mathbf{o}|m;\mathbf{\Theta}_m)$ parameterized by $\mathbf{\Theta}$, and applied on a i.i.d. random vector $\mathbf{O} = [\mathbf{V}^\top, \mathbf{Z}^\top]^\top$ with missing $\mathbf{z}$ data for $n \in [N_0+1, N]$. We now further calculate this expression for the J-GMM model defined in (7). Injecting (7) into (9) leads to:

$$\begin{aligned} Q(\mathbf{\Theta},\mathbf{\Theta}^{(i)}) = &\sum_{n=1}^{N_0}\sum_{m=1}^{M}\gamma_{nm}^{(i+1)}\left(\log\frac{\pi_m}{|\mathbf{\Sigma}_m|^{1/2}} - \frac{1}{2}\|\mathbf{o}_n-\mu_m\|_{\mathbf{\Sigma}}\right) \\ &+ \sum_{n=N_0+1}^{N}\sum_{m=1}^{M}\frac{1}{p(\mathbf{v}_n;\mathbf{\Theta}^{(i)}_{\mathbf{V}})}\left(\left(\log\frac{\pi_m}{|\mathbf{\Sigma}_m|^{1/2}}\right)\right. \\ &\times\int p(\mathbf{o}_n,m;\mathbf{\Theta}^{(i)})d\mathbf{z}_n - \int\frac{1}{2} \\ &\left.\times\|\mathbf{o}_n-\mu_m\|_{\mathbf{\Sigma}_m}p(\mathbf{o}_n,m;\mathbf{\Theta}^{(i)})\mathrm{d}\mathbf{z}_n\right), \end{aligned} \quad (19)$$

where $\|\mathbf{x}\|_{\mathbf{\Sigma}} = \mathbf{x}^\top\mathbf{\Sigma}^{-1}\mathbf{x}$ stands for the Mahalanobis distance, $^{(i)}$ denotes the i-th iteration, and $\gamma_{nm}^{(i+1)}$ are defined in (10).

To further develop (19), we first notice that for Gaussian vectors we have:

$$\begin{aligned}\int_{\mathbf{Z}} p\left(\mathbf{o}_n,m;\mathbf{\Theta}^{(i)}\right)d\mathbf{z}_n &= p\left(\mathbf{v}_n,m;\mathbf{\Theta}^{(i)}_{\mathbf{V}}\right) \\ &= \pi_m\mathcal{N}\left(\mathbf{v};\mu^{(i)}_{\mathbf{V},m},\mathbf{\Sigma}^{(i)}_{\mathbf{VV},m}\right). \end{aligned} \quad (20)$$

More importantly, we need to calculate:

$$\begin{aligned} f(\mathbf{v}_n) &= \int_{\mathbf{Z}} -\frac{1}{2}\|\mathbf{o}_n-\mu_m\|_{\mathbf{\Sigma}_m}p\left(\mathbf{o}_n,m;\mathbf{\Theta}^{(i)}\right)d\mathbf{z}_n \\ &= \int_{\mathbf{Z}} -\frac{1}{2}\|\mathbf{o}_n-\mu_m\|_{\mathbf{\Sigma}_m}\frac{\pi_m e^{-\frac{1}{2}\|\mathbf{o}_n-\mu_m\|_{\mathbf{\Sigma}_m}}}{\sqrt{(2\pi)^D|\mathbf{\Sigma}^{(i)}_m|}}\mathrm{d}\mathbf{z}_n. \end{aligned} \quad (21)$$

The literature on matrix calculus provides a formula to integrate a quadratic term multiplied by another exponential quadratic term over the complete vector, but not over a subvector. Therefore, we need to separate the terms in $\mathbf{v}_n$ and the terms in $\mathbf{z}_n$. Using the precision matrix $\mathbf{\Lambda}_m = \mathbf{\Sigma}^{-1}_m$, we can first develop the quadratic term as:

$$\begin{aligned}\|\mathbf{o}_n-\mu_m\|_{\mathbf{\Sigma}_m} &= -(\mathbf{o}_n-\mu_m)^\top\mathbf{\Lambda}_m(\mathbf{o}_n-\mu_m) \\ &= -(\mathbf{v}_n-\mu_{\mathbf{V},m})^\top\mathbf{\Lambda}_{\mathbf{VV},m}(\mathbf{v}_n-\mu_{\mathbf{V},m}) \\ &\quad -2(\mathbf{v}_n-\mu_{\mathbf{V},m})^\top\mathbf{\Lambda}_{\mathbf{VZ},m}(\mathbf{z}_n-\mu_{\mathbf{Z},m}) \\ &\quad -(\mathbf{z}_n-\mu_{\mathbf{Z},m})^\top\mathbf{\Lambda}_{\mathbf{ZZ},m}(\mathbf{z}_n-\mu_{\mathbf{Z},m}), \end{aligned} \quad (22)$$

then reorganize it into (see [29, Sec. 8.1.6]):

$$\begin{aligned}\|\mathbf{o}_n-\mu_m\|_{\mathbf{\Sigma}_m} &= -(\mathbf{o}_n-\mu_m)^\top\mathbf{\Lambda}_m(\mathbf{o}_n-\mu_m) \\ &= -\left\|\mathbf{z}_n-\mu_{\mathbf{Z},m}+\mathbf{\Lambda}^{-1}_{\mathbf{ZZ},m}\mathbf{\Lambda}_{\mathbf{ZV},m}(\mathbf{v}_n-\mu_{\mathbf{V},m})\right\|_{\mathbf{\Lambda}^{-1}_{\mathbf{ZZ},m}} \\ &\quad +(\mathbf{v}_n-\mu_{\mathbf{V},m})^\top\mathbf{\Lambda}_{\mathbf{VZ},m}\mathbf{\Lambda}^{-1}_{\mathbf{ZZ},m}\mathbf{\Lambda}_{\mathbf{ZV},m}(\mathbf{v}_n-\mu_{\mathbf{V},m}) \\ &\quad -(\mathbf{v}_n-\mu_{\mathbf{V},m})^\top\mathbf{\Lambda}_{\mathbf{VV},m}(\mathbf{v}_n-\mu_{\mathbf{V},m}). \end{aligned} \quad (23)$$

In the first term on the right hand side, we can recognize the posterior mean vector of $\mathbf{Z}$ given $\mathbf{v}_n$, i.e.:

$$\begin{aligned}\mu_{\mathbf{Z}|\mathbf{v}_n,m} &= \mu_{\mathbf{Z},m} - \mathbf{\Lambda}^{-1}_{\mathbf{ZZ},m}\mathbf{\Lambda}_{\mathbf{ZV},m}(\mathbf{v}_n-\mu_{\mathbf{V},m}) \quad (24)\\ &= \mu_{\mathbf{Z},m} + \mathbf{\Sigma}_{\mathbf{ZV},m}\mathbf{\Sigma}^{-1}_{\mathbf{VV},m}(\mathbf{v}_n-\mu_{\mathbf{V},m}). \quad (25)\end{aligned}$$

Besides, the two last terms of (23) can be factorized. We can recognize the inverse covariance matrix of $\mathbf{V}$ for component m, $\mathbf{\Sigma}^{-1}_{\mathbf{VV},m} = \mathbf{\Lambda}_{\mathbf{VV},m} - \mathbf{\Lambda}_{\mathbf{VZ},m}\mathbf{\Lambda}^{-1}_{\mathbf{ZZ},m}\mathbf{\Lambda}_{\mathbf{ZV},m}$ [20]–(2.91), and thus we have:

$$\begin{aligned}\|\mathbf{o}_n-\mu_m\|_{\mathbf{\Sigma}_m} &= \|\mathbf{z}_n-\mu_{\mathbf{Z}|\mathbf{v}_n,m}\|_{\mathbf{\Lambda}^{-1}_{\mathbf{ZZ},m}} \\ &\quad +\|\mathbf{v}_n-\mu_{\mathbf{V},m}\|_{\mathbf{\Sigma}_{\mathbf{VV},m}}. \end{aligned} \quad (26)$$

Of course, the same result holds at iteration $i+1$:

$$\begin{aligned}\|\mathbf{o}_n-\mu^{(i)}_m\|_{\mathbf{\Sigma}^{(i)}_m} &= \|\mathbf{z}_n-\mu^{(i)}_{\mathbf{Z}|\mathbf{v}_n,m}\|_{\mathbf{\Lambda}^{(i)-1}_{\mathbf{ZZ},m}} \\ &\quad +\|\mathbf{v}_n-\mu^{(i)}_{\mathbf{V},m}\|_{\mathbf{\Sigma}^{(i)}_{\mathbf{VV},m}}. \end{aligned} \quad (27)$$

Reinjecting (22) and (27) into (21) leads to:

$$f(\mathbf{v}_n) = \int_{\mathbf{Z}} \Big(-\|\mathbf{v}_n - \mu_{\mathbf{V},m}\|_{\mathbf{\Lambda}^{-1}_{\mathbf{VV},m}} - \|\mathbf{z}_n - \mu_{\mathbf{Z},m}\|_{\mathbf{\Lambda}^{-1}_{\mathbf{ZZ},m}} - 2(\mathbf{v}_n - \mu_{\mathbf{V},m})^\top \mathbf{\Lambda}_{\mathbf{VZ},m}(\mathbf{z}_n - \mu_{\mathbf{Z},m}) \Big) \times \frac{\pi_m e^{-\frac{1}{2}\left\|\mathbf{z}_n - \mu^{(i)}_{\mathbf{Z}|\mathbf{v}_n,m}\right\|_{\mathbf{\Lambda}^{(i)-1}_{\mathbf{ZZ},m}}} e^{-\frac{1}{2}\left\|\mathbf{v}_n - \mu^{(i)}_{\mathbf{V},m}\right\|_{\mathbf{\Sigma}^{(i)}_{\mathbf{VV},m}}}}{2\sqrt{(2\pi)^D |\mathbf{\Sigma}^{(i)}_m|}} \mathrm{d}\mathbf{z}_n . \tag{28}$$

Separating $\mathbf{v}_n$ and $\mathbf{z}_n$, the calculation of $|\mathbf{\Sigma}^{(i)}_m|$ can be done by noting that:

$$\left|\mathbf{\Sigma}^{(i)}_m\right| = \left|\mathbf{\Sigma}^{(i)}_{\mathbf{VV},m}\right| \left|\mathbf{\Sigma}^{(i)}_{\mathbf{ZZ},m} - \mathbf{\Sigma}^{(i)}_{\mathbf{ZV},m} \mathbf{\Sigma}^{(i)-1}_{\mathbf{VV},m} \mathbf{\Sigma}^{(i)}_{\mathbf{VZ},m}\right| = \left|\mathbf{\Sigma}^{(i)}_{\mathbf{VV},m}\right| \left|\mathbf{\Lambda}^{(i)-1}_{\mathbf{ZZ},m}\right|, \tag{29}$$

and thus:

$$\begin{aligned} &\frac{\pi_m e^{-\frac{1}{2}\left\|\mathbf{z}_n - \mu^{(i)}_{\mathbf{Z}|\mathbf{v}_n,m}\right\|_{\mathbf{\Lambda}^{(i)-1}_{\mathbf{ZZ},m}}} e^{-\frac{1}{2}\|\mathbf{v}_n - \mu^{(i)}_{\mathbf{V},m})\|_{\mathbf{\Sigma}^{(i)}_{\mathbf{VV},m}}}}{\sqrt{(2\pi)^D |\mathbf{\Sigma}^{(i)}_m|}} \\ &= \frac{e^{-\frac{1}{2}\|\mathbf{z}_n - \mu^{(i)}_{\mathbf{Z}|\mathbf{v}_n,m}\|_{\mathbf{\Lambda}^{(i)-1}_{\mathbf{ZZ},m}}}}{\sqrt{(2\pi)^{D_\mathbf{Z}} |\mathbf{\Lambda}^{(i)-1}_{\mathbf{ZZ},m}|}} \times \frac{\pi_m e^{-\frac{1}{2}\|\mathbf{v}_n - \mu^{(i)}_{\mathbf{V},m})\|_{\mathbf{\Sigma}^{(i)}_{\mathbf{VV},m}}}}{\sqrt{(2\pi)^{D_\mathbf{V}} |\mathbf{\Sigma}^{(i)}_{\mathbf{VV},m}|}} \\ &= \frac{e^{-\frac{1}{2}\|\mathbf{z}_n - \mu^{(i)}_{\mathbf{Z}|\mathbf{v}_n,m}\|_{\mathbf{\Lambda}^{(i)-1}_{\mathbf{ZZ},m}}}}{\sqrt{(2\pi)^{D_\mathbf{Z}} |\mathbf{\Lambda}^{(i)-1}_{\mathbf{ZZ},m}|}} \times p\left(\mathbf{v}_n, m; \mathbf{\Theta}^{(i)}_{\mathbf{V},m}\right). \end{aligned} \tag{30}$$

Therefore, we have:

$$\begin{aligned} f(\mathbf{v}_n) = p\left(\mathbf{v}_n, m; \mathbf{\Theta}^{(i)}_{\mathbf{V}}\right) \int_{\mathbf{Z}} \Big(&- \|\mathbf{v}_n - \mu_{\mathbf{V},m}\|_{\mathbf{\Lambda}^{-1}_{\mathbf{VV},m}} - \|\mathbf{z}_n - \mu_{\mathbf{Z},m}\|_{\mathbf{\Lambda}^{-1}_{\mathbf{ZZ},m}} - 2(\mathbf{v}_n - \mu_{\mathbf{V},m})^\top \\ &\times \mathbf{\Lambda}_{\mathbf{VZ},m}(\mathbf{z}_n - \mu_{\mathbf{Z},m}) \Big) \\ &\times \frac{e^{-\frac{1}{2}\|\mathbf{z}_n - \mu^{(i)}_{\mathbf{Z}|\mathbf{v}_n,m}\|_{\mathbf{\Lambda}^{(i)-1}_{\mathbf{ZZ},m}}}}{2\sqrt{(2\pi)^{D_\mathbf{Z}} |\mathbf{\Lambda}^{(i)-1}_{\mathbf{ZZ},m}|}} \mathrm{d}\mathbf{z}_n . \end{aligned} \tag{31}$$

After [29]–(351) and (357), we obtain:

$$\begin{aligned} f(\mathbf{v}_n) = p\left(\mathbf{v}_n, m|\mathbf{\Theta}^{(i)}_{\mathbf{V}}\right) \Big(&- \frac{1}{2}\|\mathbf{v}_n - \mu_{\mathbf{V},m}\|_{\mathbf{\Lambda}^{-1}_{\mathbf{VV},m}} \\ &- (\mathbf{v}_n - \mu_{\mathbf{V},m})^\top \mathbf{\Lambda}_{\mathbf{VZ},m} \left(\mu^{(i)}_{\mathbf{Z}|\mathbf{v}_n,m} - \mu_{\mathbf{Z},m}\right) \\ &- \frac{1}{2}\left\|\mu^{(i)}_{\mathbf{Z}|\mathbf{v}_n,m} - \mu_{\mathbf{Z},m}\right\|_{\mathbf{\Lambda}^{-1}_{\mathbf{ZZ},m}} - \frac{1}{2}\mathrm{Tr}\left[\mathbf{\Lambda}_{\mathbf{ZZ},m} \mathbf{\Lambda}^{(i)-1}_{\mathbf{ZZ},m}\right] \Big), \end{aligned}$$

which can be reorganized into:

$$\begin{aligned} f(\mathbf{v}_n) = &-\frac{1}{2} p(\mathbf{v}_n, m; \mathbf{\Theta}^{(i)}_{\mathbf{V}}) \\ &\times \left(\|\mathbf{o}'_{nm} - \mu_m\|_{\mathbf{\Lambda}^{-1}_m} + \mathrm{Tr}\left[\mathbf{\Lambda}_{\mathbf{ZZ},m} \mathbf{\Lambda}^{(i)-1}_{\mathbf{ZZ},m}\right] \right). \end{aligned} \tag{32}$$

Using (20), (32) and the extended definition of responsibilities for $n \in [N_0 + 1, N]$, (19) can be rewritten into (13). ■

REFERENCES

[1] G. McLachlan and D. Peel, *Finite Mixture Models*. New York, NY, USA: Wiley, 2000.

[2] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

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

[4] Y. Tian, L. Sigal, H. Badino, F. de la Torre, and Y. Liu, "Latent Gaussian mixture regression for human pose estimation," in *Proc. Asian Conf. Comp. Vis.*, Queenstown, New Zealand, 2010, pp. 679–690.

[5] A. Chowriappa, R. Rodrigues, T. Kesavadas, V. Govindaraju, and A. Bisantz, "Generation of handwriting by active shape modeling and global local approximation (GLA) adaptation," in *Proc. Int. Conf. Frontiers Handwriting Recognit.*, Kolkata, India, 2010, pp. 206–211.

[6] 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 Robot. Autom. Mag.*, vol. 17, no. 2, pp. 44–54, Jun. 2010.

[7] X. Alameda-Pineda and R. Horaud, "Vision-guided robot hearing," *Int. J. Rob. Res.*, vol. 34, no. 4/5, pp. 437–456, 2015.

[8] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Commun.*, vol. 50, no. 3, pp. 215–227, 2008.

[9] H. Zen, Y. Nankaku, and K. Tokuda, "Continuous stochastic feature mapping based on trajectory HMMs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 417–430, Feb. 2011.

[10] G. Ananthakrishnan and O. Engwall, "Mapping between acoustic and articulatory gestures," *Speech Commun.*, vol. 53, no. 4, pp. 567–589, 2011.

[11] S. Dusan and L. Deng, "Vocal-tract length normalization for acoustic-to-articulatory mapping using neural networks," *J. Acoust. Soc. Amer.*, vol. 106, no. 4, p. 2181, 1999.

[12] G. Sivaraman, V. Mitra, H. Nam, M. Tiede, and C. Espy-Wilson , "Vocal tract length normalization for speaker independent acoustic-to-articulatory speech inversion," in *Proc. Interspeech*, 2016, pp. 455–459.

[13] A. Ji, M. Johnson, and J. Berry, "Parallel reference speaker weighting for kinematic-independent acoustic-to-articulatory inversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 10, pp. 1865–1875, Oct. 2016.

[14] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 695–707, Nov. 2000.

[15] T. Hueber, G. Bailly, P. Badin, and F. Elisei, "Speaker adaptation of an acoustic-articulatory inversion model using cascaded Gaussian mixture regressions," in *Proc. Interspeech*, Lyon, France, 2013, pp. 2753–2757.

[16] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech, Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[17] M. J. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Comput. Speech Lang.*, vol. 10, no. 4, pp. 249–264, 1996.

[18] T. Hueber, L. Girin, X. Alameda-Pineda, and G. Bailly, "Speaker-adaptive acoustic-articulatory inversion using cascaded Gaussian mixture regression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2246–2259, Dec. 2015.

[19] G. McLachlan and K. Thriyambakam, *The EM Algorithm and Extensions*. New York, NY, USA: Wiley, 1997.

[20] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag, 2006.

[21] H. Uchida, D. Saito, N. Minematsu, and K. Hirose, "Statistical acoustic-to-articulatory mapping unified with speaker normalization based on voice conversion," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 588–592.

[22] Z. Ghahramani and M. I. Jordan, "Learning from incomplete data," Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep., 1994.

[23] 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," *J. Phonet.*, vol. 35, no. 1, pp. 1–19, 2007.
[24] P. Badin and G. Fant, "Notes on vocal tract computation," Dept. Speech, Music and Hearing, KTH, Stockholm, Sweden, *Quarterly Progress and Status Report*, 1984, pp. 53–108.
[25] A. A. Wrench, "A multi-channel/multi-speaker articulatory database for continuous speech recognition research," *Phonus.*, vol. 5, pp. 1–13, 2000.
[26] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal," Ph.D. dissertation, Univ. Edinburgh, Edinburgh, U.K., 2002.
[27] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe, "Analyzing free-standing conversational groups: a multimodal approach," in *Proc. 23rd ACM Int. Conf. Multimedia.*, 2015, pp. 5–14.
[28] X. Alameda-Pineda, E. Ricci, Y. Yan, and N. Sebe, "Recognizing emotions from abstract paintings using non-linear matrix completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5240–5248.
[29] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," 2012.

Laurent Girin received the M.Sc. and Ph.D. degrees in signal processing from the Institut National Polytechnique de Grenoble (INPG), Grenoble, France, in 1994 and 1997, respectively. In 1999, he joined the Ecole Nationale Supérieure d'Electronique et de Radioélectricité de Grenoble (ENSERG), as an Associate Professor. He is currently a Professor at Phelma (Physics, Electronics, and Materials Department of Grenoble-INP), where he lectures signal processing theory and applications to audio. His research interests include at GIPSA-Lab (Grenoble Laboratory of Image, Speech, Signal, and Automation). It deals with different aspects of speech and audio processing (analysis, modeling, coding, transformation, synthesis), with a special interest in joint audio/visual speech processing and source separation. He is also a regular collaborator at INRIA (French Computer Science Research Institute), Grenoble, as an associate member of the Perception Team.

Thomas Hueber received the Engineering degree in electronics, telecommunication, and computer science from CPE Lyon, Villeurbanne, France, and the M.Sc. degree in image processing from the University of Lyon, Lyon, France, in 2006 and the Ph.D. degree in computer science from Pierre and Marie Curie University, Paris, France, in 2009 and working on *silent speech interfaces*. In 2010, he joined GIPSA-lab (Grenoble, France) as a Postdoctoral researcher and became a tenured CNRS researcher in 2011. His research interests include multimodal speech processing (recognition, synthesis, conversion), with a special interest in speech biosignals (such as the articulatory movements, muscle and brain activities), their modeling using machine learning techniques, and their use in assistive technologies. In 2011, he received the 6th Christian Benoit Award (ACB/ISCA/AVISA), and in 2015, the best paper award in Speech Communication (Eurasip-ISCA).

Xavier Alameda-Pineda received the M.Sc. degree in mathematics and telecommunications from BarcelonaTech, Barcelona, Spain, and in computer science from Grenoble-INP, Grenoble, France, and received the Ph.D. degree in mathematics and computer science from Université Joseph Fourier, Saint-Martin-d'Hères, France, in 2013 and working on the Perception Team at INRIA, France. He moved to Treno, Italy, for a Postdoctoral fellowship at the MHUG team, University of Trento. Since late 2016, he is a Tenured Research Scientist in the Perception team at INRIA. His research interests include multimodal machine learning and signal processing for scene analysis and robotics. He is the winner of the Best Paper Award at ACM Multimedia'15, the Best Student Paper Award at IEEE WASPAA'15, and the Best Scientific Paper Award of Image, Speech, Signal, and Video Processing at IEEE International Conference on Pattern Recognition'16.