Automatic Animation of an Articulatory Tongue Model from Ultrasound Images of the Vocal Tract

Diandra Fabre[1], Thomas Hueber[1], Laurent Girin[1,2], Xavier Alameda-Pineda[2,3], Pierre Badin[1]

1 Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
2 INRIA, Perception Team, 38334, Montbonnot, France
3 Univ. of Trento, DISI, 38123, Trento, Italy

Abstract

Visual biofeedback is the process of gaining awareness of physiological functions through the display of visual information. As speech is concerned, visual biofeedback usually consists in showing a speaker his/her own articulatory movements, which has proven useful in applications such as speech therapy or second language learning. This article presents a novel method for automatically animating an articulatory tongue model from ultrasound images. Integrating this model into a virtual talking head enables to overcome the limitations of displaying raw ultrasound images, and provides a more complete and user-friendly feedback by showing not only the tongue, but also the palate, teeth, pharynx, etc. Altogether, these cues are expected to lead to an easier understanding of the tongue movements. Our approach is based on a probabilistic model which converts raw ultrasound images of the vocal tract into control parameters of the articulatory tongue model. We investigated several mapping techniques such as the Gaussian Mixture Regression (GMR), and in particular the Cascaded Gaussian Mixture Regression (C-GMR) techniques, recently proposed in the context of acoustic-articulatory inversion. Both techniques are evaluated on a multispeaker database. The C-GMR consists in the adaptation of a GMR reference model, trained with a large dataset of multimodal articulatory data from a reference speaker, to a new source speaker using a small set of adaptation data recorded during a preliminary enrollment session (system calibration). By using prior information from the reference model, the C-GMR approach is able (i) to maintain good mapping performance while minimizing the amount of adaptation data (and thus limiting the duration of the enrollment session), and (ii) to generalize to articulatory configurations not seen during enrollment better than the GMR approach. As a result, the C-GMR appears to be a good mapping technique for a practical system of visual biofeedback.

Preprint submitted to Speech Communication *August 29, 2017*

Keywords: speech production, ultrasound imaging, articulatory talking head, adaptation, Gaussian mixture regression, biofeedback

1. Introduction

Several studies have shown that providing a speaker with visual information about his/her tongue can be useful in the context of speech therapy and second language (L2) *pronunciation training* (see [1] for a short overview). This paradigm can be referred to as *visual biofeedback* (see [2, 3, 4]). The combination of visual information with acoustic and tactile feedback is expected to increase the articulatory awareness of the speaker (i.e. the patient or the L2 learner, see [5, 6]).

Several techniques can be used to capture the tongue movements during speech production. For instance, electropalatography (EPG) provides an accurate description of the tongue-palate contacts occurring when speaking, in terms of both timing and localization. This technique has been used in the context of speech therapy [7, 8] and of L2 pronunciation training [9]. EPG is more relevant for consonants, but can be also used to a certain extent for vowels [10, 11]. Note that EPG requires the speaker to wear an artificial palate with embedded contact sensors that must be fitted to his/her palate, which makes its deployment in the clinical field relatively limited.

More recently, the use of medical ultrasound imaging for biofeedback has also been investigated. Ultrasound imaging is a non-invasive and clinically safe technique able to monitor tongue movements during speech [12], for both vowels and consonants. Moreover, affordable and portable devices are now available. In a typical setup, the probe is placed beneath the speaker's chin and the vocal tract is imaged in the midsagittal plane. This configuration provides a partial view of the upper surface of the tongue as illustrated in Fig. 1. Promising results have been obtained in various contexts, such as the speech rehabilitation of the English /r/ [13, 14] and persisting speech sound disorders [15, 4].

However, as mentioned in [3, 17], the raw ultrasound images may sometimes be difficult to interpret by a non-specialist. We conjecture several explanations: (i) Raw ultrasound images are quite noisy due to the presence of speckle, see for instance Fig. 1(right); (ii) Some parts of the tongue contour might be poorly imaged when the tongue surface is oriented nearly parallel to the ultrasound beam (e.g. the back of the tongue when pronouncing the phoneme /u/); (iii) For some phonemes and some subjects, the apex (tongue tip) and parts of the tongue root can be hidden

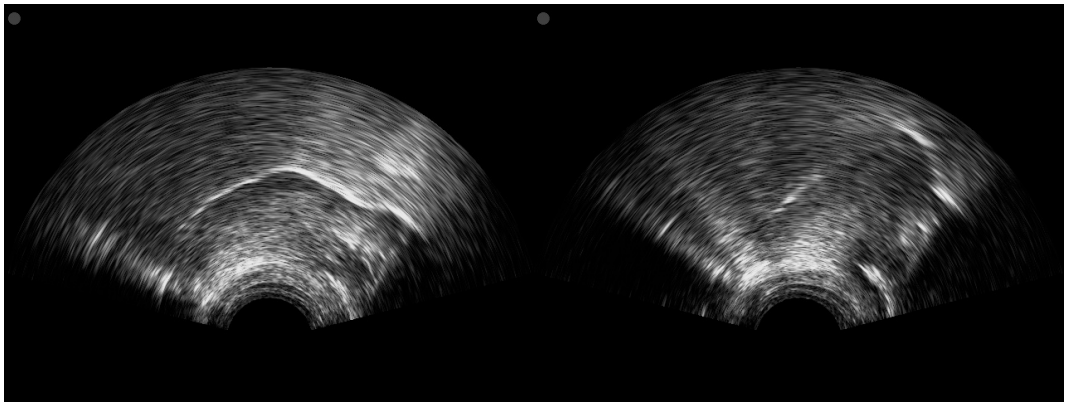

Figure 1: Raw ultrasound images in the midsagittal plane recorded during the realization of phonemes /a/ (left) and /i/ (right). The brightest white line is the reflection from the air just above the tongue surface, and the large conical black regions anterior to the tongue tip and posterior to the tongue back are the acoustic shadows of the jaw and hyoid bone respectively [16].

.

by ultrasonic shadows created by the jaw and hyoid bones, respectively; and (iv) Importantly, the tongue is displayed *out of any spatial context* since the palate, teeth and pharynx are not visible in an ultrasound image of the vocal tract.

Finally, visual biofeedback may be provided by estimating articulatory movements from speech acoustics (rather than directly capturing them). This problem is known as the acoustic-articulatory inversion problem. It has been extensively addressed in the speech processing literature as a topic on its own [18, 19, 20, 21, 22, 23, 24, 25, 26]. Estimated articulatory movements can be visualized in an intuitive way by means of a computer-animated talking head displaying the internal speech production apparatus, i.e. not only the tongue but also the palate, jaw, teeth and pharynx. Such a tool is here referred to as an *articulatory talking head* (ATH) [27, 28, 29, 30]. The combination of an acoustic-articulatory inversion system with an ATH has been proposed in our previous work [31, 32]. Such an approach is interesting since it does not require expensive and sophisticated sensors during practical usage of the system (since only a microphone is needed to capture the user's voice). However, this approach suffers from the following two limitations: i) the user should be able to vocalize, which may not be the case for some pathologies or when training a specific articulatory motor skill, and ii) the performance of the mapping is often phoneme-dependent (and can be limited for occlusives).

Based on this overall analysis of existing techniques, we focus in this work on the direct capture

of the articulatory movements using ultrasound imaging. We aim to design a visual biofeedback system based on ultrasound imaging and able to deliver a clear and easy-to-understand view of the tongue movements in real-time. For that purpose, we combine ultrasound imaging with a computer-animated ATH. Such combination has somehow already been proposed in [29], in the context of L2 pronunciation training. In that study, an ATH was used in a Wizard-of-Oz paradigm, where a hidden expert controls the feedback provided to the participant. The ATH, based on a 3D tongue model built from MRI data, was used to help French speakers pronounce unfamiliar Swedish phonemes (the alveolar trill /r/ and the velar fricative /ʃ/). The expert phonetician evaluated the acoustic production of the learner. Then, he selected the most similar movement from a database of pre-calculated talking head animations and provided it as a feedback to the learner. During this experiment, the tongue movements of the learner were monitored using ultrasound imaging.

The automatic animation of an articulatory tongue model from ultrasound images has been addressed in [33]. In this preliminary study, the authors propose to control a 3D tongue model from tongue ultrasound images, using either a contour tracking method as in [34], or by tracking speckle patterns.

In the present article, we introduce a method for animating *automatically* and in real-time the ATH's tongue model from the ultrasound images obtained from any arbitrary speaker, referred here to as the *source speaker*. We use the ATH developed at GIPSA-lab [28], illustrated in Fig. 2(c) and (d-right). Similarly to [29], this ATH embeds a geometrical model of the tongue built from static 3D MRI data recorded on a so-called *reference speaker* (Fig. 2(b)). In [35], it was shown that this tongue model can be controlled by a set of 2D coordinates representing the positions of three flesh points of the surface of the tongue in the midsagittal plane (see Fig. 2(d)). These 2D coordinates were obtained from articulatory data recorded by electromagnetic articulography (EMA) on the same reference speaker. In the following, we refer to this set of 2D coordinates as the *EMA control parameters*, which are combined in an EMA control parameters vector (or simply EMA vector).

Given this framework, the core issue addressed in the present study is thus "how to convert a raw ultrasound image of the source speaker's tongue into a set of EMA parameters controlling the articulatory tongue model of the reference speaker?" At first sight, this problem is a regression problem between data lying in different spaces, that can be addressed with standard supervised machine learning techniques: A statistical conversion model between input and output features has

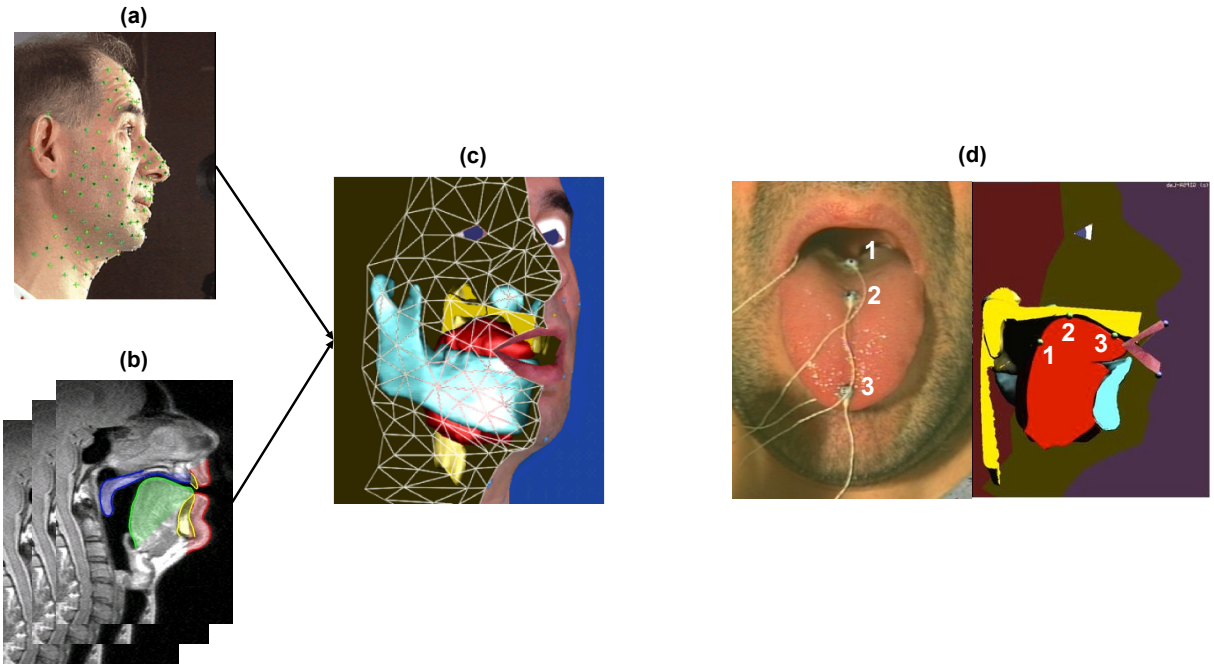

Figure 2: Articulatory talking head [28] built from video data (a) and static 3D MRI data (b) of the reference speaker. (c) 3D display of the articulatory talking head revealing the tongue, jaw and velum models built from the manual segmentation of MRI images. (d) EMA coils attached to the reference speaker's tongue; these coils can be used to animate the tongue model of the talking head (and thus referred to as the EMA control parameters). The EMA coils and the corresponding numbers are displayed here on the talking head (d,right) for clarification.

to be designed and trained from data. Popular choices for such a model include artificial neural networks (ANN) and Gaussian mixture regression (GMR). The GMR, which is central in this study, is the regressor derived from the well-known Gaussian mixture model (GMM), both being widely used in voice conversion [36, 37].

This data-driven machine learning approach requires a joint dataset of inputs (ultrasound images from the source speaker, i.e. the system user) and outputs (EMA control parameters from the reference speaker) for the supervised training of the conversion model. We propose the following usage scenario: in a preliminary enrollment session, the source speaker is invited to pronounce a set of training utterances, for which the EMA control parameters (from the reference speaker) are available. This scenario is illustrated in Fig. 3. Once the training is done, the system can be used for biofeedback: new ultrasound images are converted into EMA parameters which are finally used to animate the tongue model of the talking head.

However, two particular constraints must be carefully taken into account in the derivation of an

appropriate machine learning methodology to solve the present problem:

- Reducing the amount of training data required from the source speaker. For practical reasons, the enrollment session should be as short as possible, i.e. the input dataset should be as limited as possible, while keeping acceptable mapping performance.

- Optimizing the generalization capability of the conversion model. Indeed, this would be useful to deal with pathological speech or speech produced by a L2 learner. The model should be able to generalize to articulations not seen during training. For instance, the source speaker may not be able to pronounce a phoneme during the enrollment session, but (s)he is expected to pronounce it correctly after the therapy/learning, and thus (s)he should be given the appropriate visual feedback.

Besides, it is important to note that we do not aim at fitting the geometry of the ATH to the geometry of the source speaker's vocal tract. Our goal is to represent the tongue movements of the source speaker in the reference speaker's vocal tract (i.e. in the ATH). In a clinical scenario, this would allow to use the same ATH to display the patient's pathological articulation and the correct articulation. This correct articulation could be provided by pre-recorded animations of the ATH, or by the speech therapist with a system adapted to his/her articulation patterns. [1]

In this study, we first considered the above mapping problem using a conventional GMR. This starting point is reminiscent of our previous work [38]. Then, in order to better cope with the two particular constraints of a clinical or a L2 pronunciation training scenario, we investigated the use of the Cascaded Gaussian Mixture Regression (C-GMR) techniques, recently proposed in the context of acoustic-articulatory inversion [32].[2] Indeed, the core idea of C-GMR is to benefit from prior information coming from a large dataset of both input and output data from the reference speaker. In the present study, this dataset is composed of ultrasound images of the reference speaker's tongue

[1] Here, the therapist/teacher will have first to record a set of training utterances used to estimate a conversion model. Then, he/she will be able to animate automatically the tongue model of the ATH from ultrasound images of his/her own tongue.

[2] More specifically, in [32] we mapped acoustic features, (Mel-Frequency Cepstral coefficients, MFCC) extracted from audio speech signals to Electro-Magnetic Articulatory (EMA) parameters. The C-GMR was used to adapt the mapping model trained on a reference speaker to a new speaker using audio-only adaptation data .

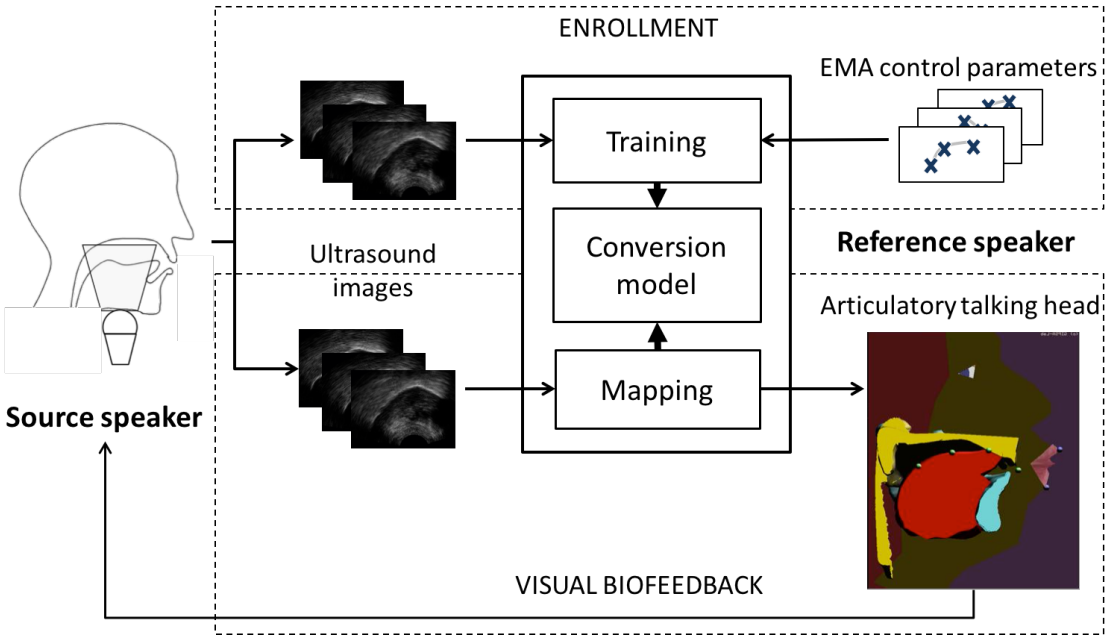

Figure 3: Usage scenario for the proposed visual biofeedback system: During the enrollment session (top), the source speaker (i.e the system user) records a set of utterances for which the corresponding EMA control parameters are available. The recorded ultrasound images and the reference's speaker EMA control parameters are used to train the conversion model. During the mapping/biofeedback stage (bottom), new EMA control parameters, used to animate the ATH, are estimated directly from the raw ultrasound images.

and corresponding EMA control parameters, all recorded in advance. These data are used to build a robust conversion model between the input and output spaces. The model is then adapted using a limited amount of ultrasound data of a new source speaker and is then used to convert inputs from this speaker. This prior knowledge provided by this reference model is expected to reduce the amount of data required from the source speaker while increasing the generalization capability of the system, which are precisely the two critical aspects mentioned above.

The experimental evaluation was conducted on a multispeaker dataset (one reference speaker with ultrasound, EMA and audio, and two source speakers with ultrasound and audio), recorded for this study. We compared two different C-GMR approaches with the conventional GMR. We showed that one particular C-GMR model called the Integrated C-GMR copes well with the two constraints mentioned above and is thus a good candidate for a practical system of visual biofeedback.

The rest of the paper is organized as follows. Section 2 presents the different techniques proposed to address the mapping problem considered in this study, including the conventional GMR and the

Cascaded-GMR. Then we describe the experiments conducted to validate the proposed system. The experimental set-up is detailed in Section 3. The results are reported and discussed in Section 4. Limitations and perspectives are presented in Section 5. Section 6 concludes the paper.

2. Mapping Methodology

2.1. Gaussian mixture model and Gaussian mixture regression

First, we briefly recall the principle of Gaussian mixture regression (GMR) which is central in this study. Let $\mathbf{X}$ and $\mathbf{Y}$ be two random column vectors of dimension D_X and D_Y respectively. Let $\mathbf{J}$ denote the concatenation of $\mathbf{X}$ and $\mathbf{Y}$ such as $\mathbf{J} = [\mathbf{X}^\top, \mathbf{Y}^\top]^\top$ (where $^\top$ denotes the transpose operator). Let $p(\mathbf{x}; \mathbf{\Theta_X})$ denote the probability density function (PDF) of $\mathbf{X}$, parametrized by the set of parameters $\mathbf{\Theta_X}$.[3] Let $\mathcal{N}(\mathbf{x}; \mu_\mathbf{X}, \mathbf{\Sigma_{XX}})$ denote the Gaussian distribution on $\mathbf{X}$ with mean vector $\mu_\mathbf{X}$ and covariance matrix $\mathbf{\Sigma_{XX}}$. Let $\mathbf{\Sigma_{XY}}$ denote the covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$. A Gaussian mixture model (GMM) on ($\mathbf{X}$, $\mathbf{Y}$) consists of a weighted sum of Gaussian PDFs:

$$p(\mathbf{j}; \mathbf{\Theta_J}) = \sum_{m=1}^{M} \pi_m \mathcal{N}(\mathbf{j}; \mu_{\mathbf{J},m}, \mathbf{\Sigma_{JJ},}_m), \tag{1}$$

where M is the number of components of the mixture. For each component m, $\pi_m = p(m)$ is the prior probability satisfying $\sum_{m=1}^{M} \pi_m = 1$, $\mu_{\mathbf{J},m} = [\mu_{\mathbf{X},m}^\top, \mu_{\mathbf{Y},m}^\top]^\top$ is the mean vector and $\mathbf{\Sigma_{JJ},}_m$ is the covariance matrix given by:

$$\mathbf{\Sigma_{JJ},}_m = \begin{bmatrix} \mathbf{\Sigma_{XX},}_m & \mathbf{\Sigma_{XY},}_m \\ \mathbf{\Sigma_{YX},}_m & \mathbf{\Sigma_{YY},}_m \end{bmatrix}. \tag{2}$$

The classical Expectation-Maximization (EM) algorithm (see [39] (ch. 9)) for GMM can be used to estimate these parameters given a training set of joint observations $(\mathbf{x}, \mathbf{y})$. The GMR used to map $\mathbf{x}$ into an estimated value $\hat{\mathbf{y}}$ of $\mathbf{y}$ is defined as in [40]:

$$\hat{\mathbf{y}} = \mathrm{E}[\mathbf{Y}|\mathbf{x}; \mathbf{\Theta_J}] = \sum_{m=1}^{M} p(m|\mathbf{x}; \mathbf{\Theta_X}) \mu_{\mathbf{Y}|\mathbf{x},m} \tag{3}$$

[3] $p(\mathbf{x}; \mathbf{\Theta_X})$ is a shortcut for $p(\mathbf{X} = \mathbf{x}; \mathbf{\Theta_X})$.

Bold upper-case letters denote random vectors, and corresponding bold lower-case letters denote their realizations.

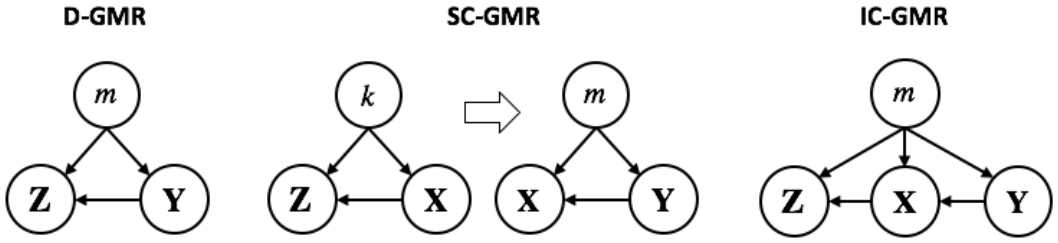

Figure 4: Graphical representation of Direct-GMR (D-GMR), Split Cascaded GMR (SC-GMR) and Integrated Cascaded GMR (IC-GMR). In this study, $\mathbf{Z}$ and $\mathbf{X}$ refer to ultrasound features for the source and reference speaker respectively. $\mathbf{Y}$ refers to EMA control parameters of the ATH. m and k refer to the mixture component index. The large blank arrow in the SC-GMR indicates the direction of cascading (i.e. the output $\mathbf{X}$ of the first model is sent as input of the second model). All horizontal thin arrows are right-to-left because these graphical models represent the *generative* models associated to the presented equations. At inference time, right-sided variables are inferred from left-sided variables.

with

$$\mu_{\mathbf{Y}|\mathbf{x},m} = \mu_{\mathbf{Y},m} + \mathbf{\Sigma}_{\mathbf{YX},m}\mathbf{\Sigma}^{-1}_{\mathbf{XX},m}(\mathbf{x} - \mu_{\mathbf{X},m}), \tag{4}$$

$$p(m|\mathbf{x};\mathbf{\Theta}_{\mathbf{X}}) = \frac{\pi_m \mathcal{N}\left(\mathbf{x};\mu_{\mathbf{X},m},\mathbf{\Sigma}_{\mathbf{XX},m}\right)}{\sum_{i=1}^{M}\pi_i \mathcal{N}\left(\mathbf{x};\mu_{\mathbf{X},i},\mathbf{\Sigma}_{\mathbf{XX},i}\right)}. \tag{5}$$

This mapping amounts to minimizing the mean squared error (MSE) between $\mathbf{y}$ and $\hat{\mathbf{y}}$ assuming statistical independence and identical distribution of the observations.

2.2. Direct mapping from source speaker's ultrasound to reference speaker's EMA data (D-GMR)

Let $\mathbf{Z}$ denote a random vector of ultrasound features derived from the raw ultrasound images of the source speaker (the extraction of such features will be detailed in Section 3.2). Let $\mathbf{Y}$ be a corresponding vector of EMA control parameters, describing the tongue position of the reference speaker while uttering the same phoneme (the alignment procedure between source and reference speakers' data is described later). Let N_0 denote the number of ultrasound images recorded by the source speaker during the enrollment session, and let $\{\mathbf{z}_n\}_{n=1}^{N_0} = \mathbf{z}_{1:N_0}$ denote the corresponding set of ultrasound feature vectors derived from these images.

A first straightforward way to address the regression problem considered in this study is to directly model the statistical relationship between the ultrasound feature vector of the source speaker $\mathbf{Z}$ and the EMA parameter vector of the reference speaker $\mathbf{Y}$. This can be done with a $\mathbf{Z}$-$\mathbf{Y}$ GMM, from which we can directly derive the corresponding $\mathbf{Z}$-to-$\mathbf{Y}$ GMR (as described in Section 2.1).

This approach addresses simultaneously the cross-speaker and cross-modality mapping issues and is here referred to as the "direct" GMR (D-GMR; see Fig. 4 (left)). Importantly, training this model requires to associate the $\mathbf{z}_{1:N_0}$ enrollment data with "corresponding" $\mathbf{y}_{1:N_0}$ EMA data from the reference speaker. This is done by time-aligning each recorded sentence pronounced by the source speaker with the same sentence pronounced by the reference speaker, using a dynamic time warping (DTW) algorithm. However, since ultrasound features lie in a totally different space than the EMA features, the DTW cannot be applied directly to these data. Therefore, we perform this alignment in the acoustic space, and extend it to the ultrasound images and the EMA for the source speaker and the reference speaker respectively. Complementary technical details on the alignment procedure are given in Section 3.3. After this procedure, the source speaker's dataset $\mathbf{z}_{1:N_0}$ is assumed to be aligned with the reference speaker's dataset $\mathbf{y}_{1:N_0}$. The EM algorithm for GMM can then be applied to the set $\{\mathbf{z}_{1:N_0}, \mathbf{y}_{1:N_0}\}$. Note that the need for recording the audio signal in addition to the ultrasound data during the enrollment session is one drawback of the D-GMR approach.

2.3. Cascaded Gaussian Mixture Regression (C-GMR)

As briefly stated in the introduction, the core motivation of the C-GMR framework [32] is to benefit from prior information on the reference speaker's articulatory space. Such prior information is given by a GMM trained on a set of joint observations $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{N} = \{\mathbf{x}_{1:N}, \mathbf{y}_{1:N}\}$, where $\mathbf{X}$ is a feature vector derived from an ultrasound image of the reference speaker. This dataset can be obtained by recording the tongue movements of the reference speaker simultaneously using ultrasound imaging and EMA.[4] Therefore, the $\mathbf{X}$-$\mathbf{Y}$ GMM models the statistical relationships between two different "articulatory" vectors representing the same tongue movement, for the same reference speaker, but captured using two different devices (i.e. ultrasound imaging and EMA). Since this reference model does not involve the source speaker, it can be trained "in the laboratory" on a very large dataset. In practice, this means that the number N of $(\mathbf{x}, \mathbf{y})$ vector pairs in this dataset can be chosen to be significantly larger than the number N_0 of $\mathbf{z}$ vectors in the dataset recorded by the source speaker in the enrollment session (in short, $N_0 \leq N$ and potentially $N_0 \ll N$). The

[4]However, this was not directly achievable due to practical experimental issues, and the two modalities had to be recorded separately, and then aligned. This is detailed in Section 3.1.

reference model is thus expected to be well-estimated and to finely describe the articulatory space of the reference speaker, in both the ultrasound and EMA modalities.

The C-GMR approach exploits the reference **X**-**Y** GMM by splitting the **Z**-to-**Y** regression in two steps:

1. a **Z**-to-**X** mapping step which models the statistical relationships between source and reference speaker's ultrasound data (i.e. a cross-speaker monomodal mapping)
2. a **X**-to-**Y** mapping step derived from the reference **X**-**Y** GMM, which models the statistical relationships between the ultrasound data and the EMA data of the reference speaker (i.e. a single-speaker cross-modal mapping).

In our previous work [32], we proposed two versions of the C-GMR framework. As shown in Fig. 4(middle), the first one, referred to as the *Split C-GMR* (SC-GMR), is a straightforward chaining of two GMRs. The output of the first GMR is calculated before being injected as input of the second GMR. In other words, we have $\hat{\mathbf{y}} = \mathrm{E}[\mathbf{Y}|\hat{\mathbf{x}}; \mathbf{\Theta_J}]$ with $\hat{\mathbf{x}} = \mathrm{E}[\mathbf{X}|\mathbf{z}; \mathbf{\Theta_I}]$ (with $\mathbf{I} = [\mathbf{Z}^\top, \mathbf{X}^\top]^\top$), where both expectations follow (3) with their respective parameters. Importantly, in the SC-GMR, the two GMRs may have a different number of mixture components. As stated above, the reference **X**-**Y** GMR is trained from the N available ($\mathbf{x}$,$\mathbf{y}$) joint observations of the reference speaker, whereas the **Z**-**X** GMR is trained on a (much) smaller dataset that consists of $\mathbf{z}_{1:N_0}$ and a corresponding subset $\mathbf{x}_{1:N_0}$ of ultrasound feature vectors extracted from the reference speaker dataset.[5] The number of components of the reference **X**-**Y** GMR is thus expected to be larger than the number of components of the **Z**-**X** GMR.

The second version of the C-GMR framework is referred to as the *Integrated C-GMR* (IC-GMR) since it *integrates* the two GMRs listed above into a single GMR-based mapping process, as shown in Fig. 4(right). Very importantly, this combination is made at the component level of the GMR, as opposed to the SC-GMR. In other words, the plugged **Z**-to-**X** and **X**-to-**Y** regressors share the same component assignment variable m. The goal is here to make the source input vector **Z** benefit from the partitioning of the articulatory space of the reference speaker (i.e. **X**-**Y**) which is assumed to be well estimated on a large dataset. Mathematically, this principle is implemented as follows (see [32] for a complete description). In the IC-GMR model, the statistical dependencies between

[5] Again, this requires an alignment procedure, which is detailed in Section 3.3.

$\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are modeled as:

$$p(\mathbf{x},\mathbf{y},\mathbf{z};\boldsymbol{\Theta}) = \sum_{m=1}^{M} p(m)p(\mathbf{y}|m;\boldsymbol{\Theta}_{\mathbf{Y},m})p(\mathbf{x}|\mathbf{y},m;\boldsymbol{\Theta}_{\mathbf{X}|\mathbf{Y},m}) \times p(\mathbf{z}|\mathbf{x},m;\boldsymbol{\Theta}_{\mathbf{Z}|\mathbf{X},m}), \quad (6)$$

where for each component m, $\pi_m = p(m)$ is the prior component weight, $p(\mathbf{y}|m;\boldsymbol{\Theta}_{\mathbf{Y},m})$ is a Gaussian distribution, and the conditional PDFs $p(\mathbf{x}|\mathbf{y},m;\boldsymbol{\Theta}_{\mathbf{X}|\mathbf{Y},m})$ and $p(\mathbf{z}|\mathbf{x},m;\boldsymbol{\Theta}_{\mathbf{Z}|\mathbf{X},m})$ are Linear-Gaussian distributions (i.e. a Gaussian distribution with the mean being an affine function of the conditional variable). All Gaussian distributions have full-covariance matrices. The minimum MSE estimation $\hat{\mathbf{y}}$ of $\mathbf{y}$ given $\mathbf{z}$ is given by its posterior mean (see [32] for the complete derivation):

$$\hat{\mathbf{y}} = \mathrm{E}[\mathbf{Y}|\mathbf{z}] = \sum_{m=1}^{M} p(m|\mathbf{z};\boldsymbol{\Theta}_{\mathbf{Z}})(\mu_{\mathbf{Y},m} + \boldsymbol{\Sigma}_{\mathbf{YX},m}\boldsymbol{\Sigma}_{\mathbf{XX},m}^{-1}\boldsymbol{\Sigma}_{\mathbf{XZ},m}\boldsymbol{\Sigma}_{\mathbf{ZZ},m}^{-1}(\mathbf{z}-\mu_{\mathbf{Z},m})). \quad (7)$$

The component weights $p(m|\mathbf{z};\boldsymbol{\Theta}_{\mathbf{Z}})$ are obtained by applying the classical formula (5) with the marginal distributions $p(\mathbf{z}|m;\boldsymbol{\Theta}_{\mathbf{Z},m})$, which can be obtained from the distributions defined above. Similarly to the GMR, the above equation enables the mapping to be performed in real-time.

The full derivation of the exact EM training algorithm of the IC-GMR was presented in [32]. This EM algorithm jointly exploits the large dataset ($\mathbf{x}_{1:N}$,$\mathbf{y}_{1:N}$) from the reference speaker and the small (aligned) enrollment dataset $\mathbf{z}_{1:N_0}$. Remember that the amount of source speaker's data may be relatively small since the enrollment stage is generally limited in time. It can also be sparse because the source speaker may not be able to pronounce one or several phonemes. Interestingly, the proposed training algorithm explicitly applies the *missing data* methodology of machine learning [41, 40] to deal with this limited and/or sparse aspect of the enrollment dataset. The core idea of this training approach is to infer the missing information of the source speaker from the reference speaker data, which is assumed to be (much) larger (i.e. $N > N_0$) and denser (i.e. the reference speaker is able to pronounce correctly all the phonemes of a given language).[6]

[6]Technically, this probabilistic inference occurs during each iteration of the E-step: The observations $\mathbf{z}_{N_0+1:N}$ which are considered as *missing* (i.e. they lack in the "ideal" joint set of $\mathbf{z}_{1:N},\mathbf{x}_{1:N},\mathbf{y}_{1:N}$ of N training observations) are replaced with their conditional mean $\mu_{\mathbf{Z}|\mathbf{x}_n,m}$ for $n \in [N_0+1,N]$. This inference can be seen as a $\mathbf{X}$-to-$\mathbf{Z}$ GMR which can be calculated using (4), for each component of the IC-GMR. Then the responsibility of each component m is calculated following (5), given the completed dataset $\mathbf{z}_{1:N},\mathbf{x}_{1:N},\mathbf{y}_{1:N}$ (which includes the true observations $\mathbf{z}_{1:N_0}$ and the inferred observations $\mathbf{z}_{N_0+1:N}$) and the current values of model parameters. See [32] for details.

3. Experimental set-up and protocol

The D-GMR and C-GMR approaches (i.e. SC-GMR and IC-GMR) were evaluated on a multispeaker database, with special focus on the impact of the enrollment stage, in terms of size and content, on the mapping performance. In this section, we describe the experimental set-up and protocol. The results will be reported and discussed in the next section.

3.1. Data acquisition

As mentioned earlier, the training of the reference **X**-**Y** GMM requires to simultaneously record a large amount of ultrasound and EMA tongue data on the reference speaker. Such simultaneous recordings have recently been showed to be feasible by Aron et al. [42]. In their study, the ultrasound probe was held manually by the speaker and could move up and down with the jaw. EMA was used to track the displacement of the probe (with sensors attached to it). Calibration and registration techniques were then used to combine the two modalities. This approach is probably the most suitable for recording the data for the reference speaker in our case, and should be investigated in future studies. In the setup available in our laboratory, the probe is kept fixed with respect to the speaker's head by means of a stabilization helmet, manufactured by the Articulate Instruments company. Since we found that the ultrasound system, and mainly the helmet, interfered with the magnetic field of the EMA device, resulting in measurement distortion, we built the parallel "ultrasound-EMA" dataset in two steps, using a different approach.

We started from an existing EMA+audio dataset recorded on the reference speaker [31]. These data were recorded using the Carstens AG200 EMA system with a sampling rate of 100Hz. They consist of approximately 17 minutes of speech (after removing the long pauses), for a total of $1,109$ sequences (i.e. all French vowels, 224 VCVs, 109 isolated words, 88 sentences, all items being repeated twice). Then, in another session, we recorded an ultrasound+audio database with the same reference speaker. The speaker was asked to pronounce the same speech material as for the EMA+audio dataset.

Ultrasound images (640×480 grayscale images, 60 fps) were acquired using the *Terason T3000* medical ultrasound system, with a 128-element microconvex transducer. Ultrasound frequency range was set to 3 MHz-5 MHz, scanning angle to 140°, and penetration depth to 7 cm. The acoustic speech signal (44.1 kHz, 32 bits) was recorded synchronously with ultrasound thanks to

the *Ultraspeech 1.3* software [43]. The ultrasound images were processed with the feature extraction procedure described in Section 3.2 and aligned with the EMA vectors as detailed in Section 3.3.

Two additional ultrasound databases were recorded for this study for evaluating the cross-speaker (source-to-reference) and cross-modality (ultrasound-to-EMA) articulatory mapping using D-GMR, SC-GMR and IC-GMR. For that purpose, one male subject (distinct from the reference speaker) and one female subject, neither having articulation disorders, were asked to pronounce the same database while being recorded using ultrasound (and audio), with the same experimental setup as for the reference speaker. In the following, these two source speakers will be referred to as M1 and F1. As for the reference dataset, each of these two enrollment ultrasound datasets was processed with feature extraction and alignment, as described in the next subsections.

3.2. Extraction of ultrasound feature vectors from raw ultrasound images

Visual features were extracted from raw ultrasound images with the Principal Component Analysis or PCA-based technique proposed by Turk and Pentland for face recognition (also known as the EigenFace technique) [44]. At training stage, each image was down-sampled to 64×64 pixels, normalized by its mean value, and transformed into a 4096×1 vector. A PCA was then performed on the whole training dataset of N_i images (i.e. on a $N_i \times 4096$ matrix). The basis vectors that best explain the variation of the pixel intensities are here called *EigenTongue* [45]. At feature extraction stage, each new ultrasound image was pre-processed in the same way and projected onto the set of *EigenTongue*. An ultrasound feature vector was finally defined as the set of D first coordinates in that space. The number of coordinates was determined by selecting the EigenTongue that carry 80% of the variance of the pixels, which was found to be $D = 30$ in this study, for both the reference and source speakers. Ultrasound feature vector sequences were finally re-sampled from 60 Hz to 100 Hz in order to fit the sampling rate of EMA.

3.3. Data alignment

The alignment procedure for the reference speaker dataset $\{\mathbf{x}_{1:N}, \mathbf{y}_{1:N}\}$ followed the line already described in Section 2.2 for the alignment of the $\{\mathbf{z}_{1:N_0}, \mathbf{y}_{1:N_0}\}$ dataset, i.e. the alignment was done using the audio signals: each audio sequence of the ultrasound+audio dataset was aligned with the corresponding audio sequence of the EMA+audio dataset, using a DTW-based procedure. The resulting alignment was then applied to the ultrasound feature vector sequences. Note that in all

our experiments, the spectral content of the acoustic speech signal was parametrized using Mel-frequency cepstral coefficients (MFCC) decomposition, performed with the HTK toolkit [46] and a standard configuration: The audio signal was downsampled to 16 kHz, 16 bit quantization, and analyzed with a 20 ms-length window with 5 ms shift. The MFCC analysis results in vectors of 26 coefficients including static coefficients and their first derivative. MFCC feature sequences were downsampled from 200 Hz to 100 Hz in order to fit the sampling rate of EMA.

For the SC-GMR and IC-GMR, two approaches were investigated to align the ultrasound data from the source and reference speakers, in order to build the datasets $\{\mathbf{z}_{1:N_0}, \mathbf{x}_{1:N_0}\}$ (for the SC-GMR) and $\{\mathbf{z}_{1:N_0}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}\}$ (for the IC-GMR). The first approach is the same as the one used for the datasets of the D-GMR and of the reference GMR: It exploits the audio signal recorded simultaneously with the ultrasound images. With the clinical application in mind, the second approach aims at simplifying the experimental set-up by getting rid of the audio recording. To that purpose, the DTW was applied directly to ultrasound feature vectors (rather than to audio vectors (MFCC)). Common EigenTongue basis vectors were thus estimated on a training set containing an equal number of ultrasound images from both source and reference speakers. This allows to represent the images of both speakers in the same joint image feature space and thus to align the sequences of the two speakers using DTW with these features. This approach will be referred to as "SC-GMR / IC-GMR (no audio)".

3.4. Training reference, D-GMR and C-GMR models

The **X**-**Y** reference GMM was trained using the (large and aligned) parallel ultrasound-EMA dataset from the reference speaker $\{\mathbf{x}_{1:N}, \mathbf{y}_{1:N}\}$. As stated in Section 2.2, the D-GMR was trained using the aligned dataset $\{\mathbf{z}_{1:N_0}, \mathbf{y}_{1:N_0}\}$. The **Z**-**X** GMM of the SC-GMR was trained using the aligned dataset $\{\mathbf{z}_{1:N_0}, \mathbf{x}_{1:N_0}\}$. For these three models, the EM training algorithm for GMM was used, initialized with k-means clustering (the whole training procedure was repeated 5 times).

A cross-validation procedure was used to determine the optimal number of mixture components M (from 2 to 16), using 20% of the training set as a validation set. For the **X**-**Y** reference GMM, the optimal value was found to be $M = 16$. For the D-GMR and the SC-GMR, this number depended on the amount of enrollment data, i.e. N_0 (the more data are available, the phonetically denser the articulatory space is likely to be, the more components are needed to model it).

For the reference speaker, and for each step of the cross-validation procedure, the *EigenTongue* decomposition basis was built by selecting randomly $N_i = 2,000$ images among the available N training images. The number $2,000$ was chosen empirically. As in [45], we found that it was sufficient to cover the tongue shape variability and perform the *EigenTongue* decomposition. For the source speaker, as well as for the reference speaker for the SC-GMR/IC-GMR (no audio) experiments (for which a common *EigenTongue* basis vectors are estimated), all of the N_0 frames of the enrollment dataset were used if $N_0 \leq 2,000$, and a subset of $2,000$ images was randomly selected from the enrollment dataset otherwise.

Finally, the IC-GMR was trained using the aligned dataset $\{\mathbf{z}_{1:N_0}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}\}$. By design, it inherits from the structure of the reference **X**-**Y** GMM, with the optimal value $M = 16$ found in our experiments. As described in [32], the IC-GMR parameters related to the **X**-**Y** mapping were initialized by the reference GMR.

4. Results

In this section, we first report the performance of the reference model. Then we present and discuss the performance of the complete cross-speaker cross-modality mapping. Finally, we explore two issues that are relevant to the target applications (speech therapy and L2 learning): the influence of the amount of enrollment data on the performance, and the generalization capabilities of the GMR and C-GMR.

4.1. Evaluation metric

In all the following experiments, the performance of the different models was assessed by calculating the (average) root mean squared error (RMSE) between original and estimated EMA control parameters, as:

$$RMSE(i) = \sqrt{\frac{1}{N_v(i)} \frac{1}{D_Y} \sum_{n=1}^{N_v(i)} \sum_{d=1}^{D_Y} (y_{dn} - \hat{y}_{dn})^2}, \tag{8}$$

$$RMSE_{\text{Average}} = \frac{1}{N_s} \sum_{i=1}^{N_s} RMSE(i), \tag{9}$$

where i is the utterance index, $N_v(i)$ is the number of feature vectors in utterance i, D_Y and y_{dn} are the dimension and the entries of vectors $\mathbf{y}$, respectively, and N_s is the total number of utterances in the test dataset.

4.2. Performance of the **X***-to-*$\mathbf{Y}$ *reference model*

First, the performance of the reference speaker's model (i.e. ultrasound-to-EMA mapping) was evaluated using a 5-fold cross-validation technique. The ultrasound-EMA database of the reference speaker was divided into 5 subsets of approximately equal size. For each trial, 4 subsets were used for training the model, while the remaining subset was used for testing.

For the **X**-to-$\mathbf{Y}$ reference GMR, the $RMSE_{\text{Average}}$ (RMSE averaged over all test sequences and all EMA control parameters) was found to be 2.2 mm. This performance is globally satisfactory. Interestingly, it is better than the performance reported in the literature on GMM-based acoustic-to-articulatory mapping (as in [47]). This may be explained by the fact that the ultrasound-to-EMA mapping may be less complex than the acoustic-to-EMA mapping since the tongue information is directly available in the ultrasound image, and not mixed with other articulatory information as it is in the audio signal (such as the movement of the lips, the laryngeal activity, etc.). This performance is considered as the baseline of the other mapping methods explored in the following.

The performance of the reference GMR model for each EMA control parameter coordinate is presented in Table 1. Here, the RMSE was calculated independently for each entry of $\mathbf{y}$ (hence no summation on d was applied in (8); instead the RMSE was calculated for each d value independently; still the per-entry RMSE is averaged over the complete test dataset). We report also the value of the Pearson correlation coefficient for each EMA control parameter. With a minimal and maximal correlation of 0.79 and 0.94 respectively, the performance is satisfactory and relatively homogeneous. However, it appears that the EMA control parameters related to the middle tongue (mid) are more accurately estimated than the ones related to the tongue tip and the tongue back. These differences could be partly explained by the fact that these tongue parts are sometimes hidden in the ultrasound images by the acoustic shadows of the jaw and hyoid bones.

4.3. Performance of the cross-speaker cross-modality mapping methods

We discuss here the performance of the D-GMR, SC-GMR and IC-GMR techniques as a function of the amount of enrollment data (i.e. N_0). The experimental protocol was the following: i) The database recorded by the reference and the two source speakers was partitioned in 5 subsets of approximately equal size; 4 subsets were used to build the training and enrollment dataset while

EMA control parameter	RMSE (in mm) ± CI	Pearson corr. coef.
$\mathbf{tip_h}$	2.3 ± 1.1	0.89
$\mathbf{tip_v}$	2.0 ± 0.9	0.85
$\mathbf{mid_h}$	1.9 ± 0.9	0.89
$\mathbf{mid_v}$	1.7 ± 0.9	0.90
$\mathbf{back_h}$	1.6 ± 0.9	0.94
$\mathbf{back_v}$	2.2 ± 1.1	0.79

Table 1: Performance of the **X**-to-**Y** GMR (ultrasound-to-EMA mapping for the reference speaker) in terms of (average) RMSE with 95% confidence interval (CI) and Pearson correlation coefficient, for each individual EMA control parameter (tip, mid, back; $_h$ and $_v$ denote the horizontal and vertical coordinates in the 2D midsagittal plane, respectively).

the remaining subset was used for test; ii) A set of enrollment sentences was randomly selected from the sentences available in the 4 subsets used for training; The size of the enrollment dataset varied from 1/20 to 1/2 of the size of the training set with 8 intermediary sizes; iii) Step (ii) was repeated for all the five permutations of Step (i) (5-fold validation). This resulted in 50 experiments for each source speaker F1 and M1.

Results are presented in Fig. 5 for F1 and in Fig. 6 for M1. First, we observe that the RMSE of all models is significantly larger than the RMSE observed when processing ultrasound images of the reference speaker (which is 2.2 mm, see Section 4.2 and "Ref" in Fig. 5 and 6). For example, the IC-GMR, which is the best method for small N_0 values (see extended discussion below), provides nearly 1 mm worse RMSE for both speaker M1 and F1, for about 1 min of adaptation data (long silences at the beginning and end of each sentence being excluded). This remains true even for a relatively large enrollment dataset (though the RMSE decreases with N_0, see below). This is an expected result which can be partially explained by articulatory idiosyncrasies, i.e. differences between two speakers in terms of morphology and articulatory strategies when pronouncing the same phoneme. Here, this phenomenon is not taken into account in the evaluation which considers the articulatory strategy of the reference speaker as the ground truth. Therefore, a vector of EMA parameters estimated from a source speaker's ultrasound image can be slightly different from the one observed when the reference speaker is pronouncing the same phoneme, without leading to an

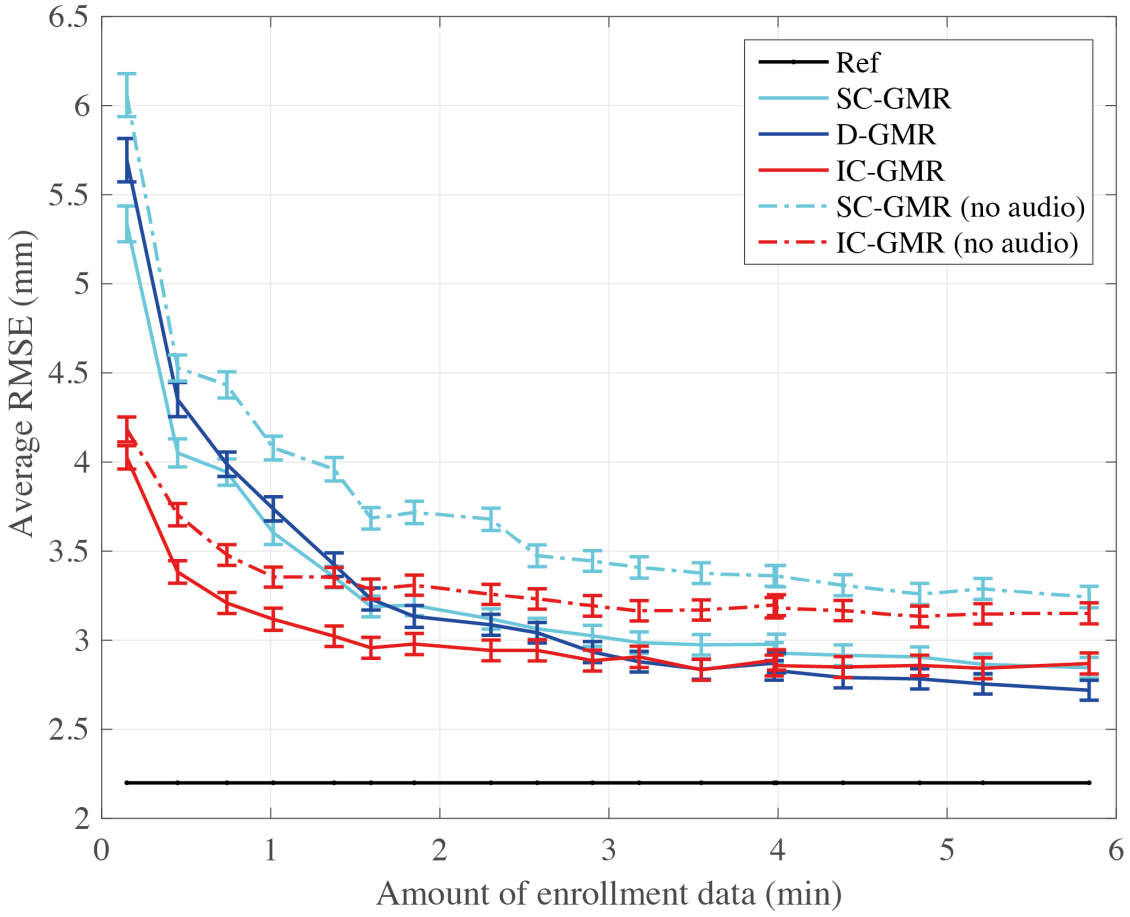

Figure 5: Performance of Z-to-Y mapping in RMSE (in mm) with 95% confidence interval as a function of the amount of enrollment data, for the D-GMR, SC-GMR, and IC-GMR, for source speaker F1. For SC-GMR (no audio) and IC-GMR (no audio), the alignment of source and reference speakers' ultrasound data is performed without exploiting the audio recordings. The "Ref" line stands for the performance of the reference **X**-**Y** GMR

.

incorrect visual feedback.

Then, we observe also that the performances of all mapping techniques considered in this study (D-GMR, SC-GMR and IC-GMR) increase with N_0, as expected. However they do not have the same starting level (i.e. RMSE value at lower N_0 value) and they do not have the same decrease rate. In a general way, the D-GMR starts with the higher RMSE value and has the fastest decrease, whereas the IC-GMR starts with the lower RMSE value and has the lower decrease rate. The IC-GMR curve and the D-GMR curve cross each other at approximately 3.5 min of adaptation data

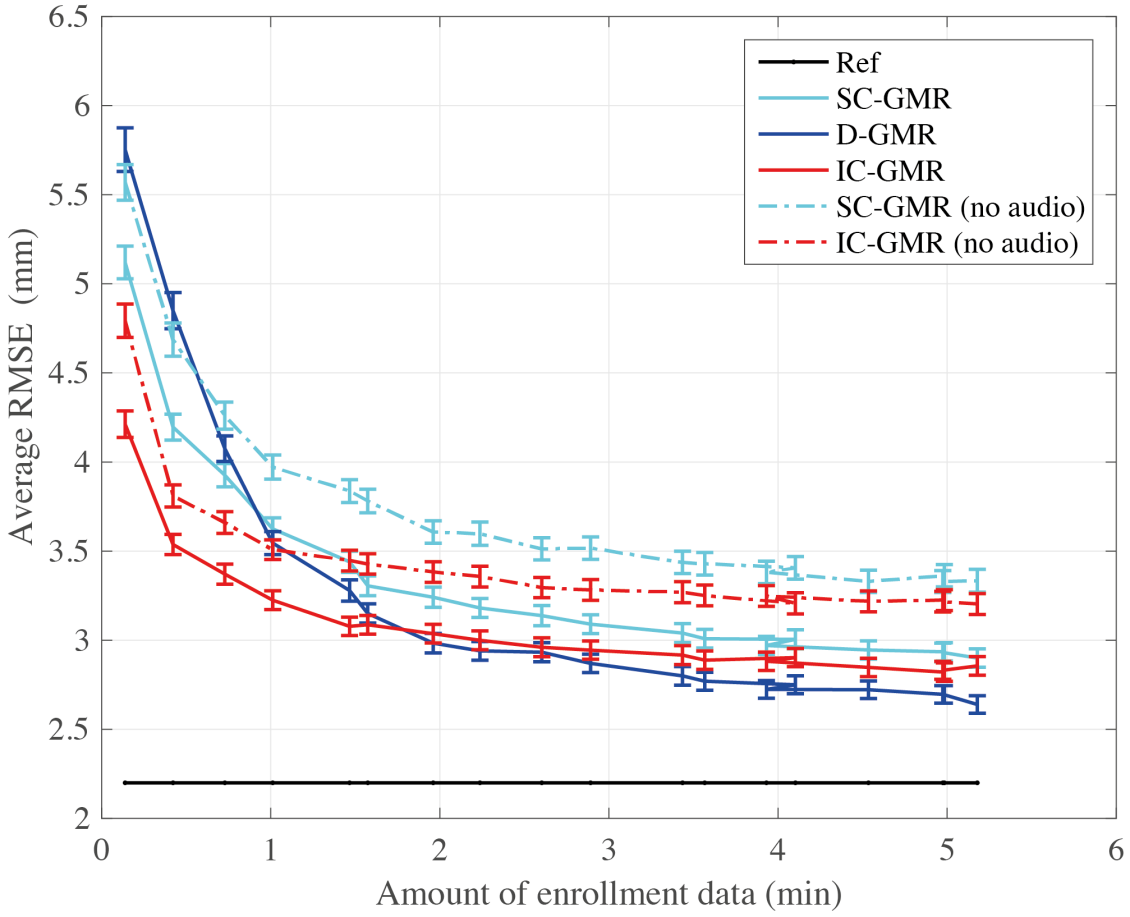

Figure 6: Performance of Z-to-Y mapping in RMSE (in mm) with 95% confidence interval as a function of the amount of enrollment data, for source speaker M1

for speaker F1 and at approximately 1.75 min of adaptation data for speaker M1. The SC-GMR lies between the D-GMR and the IC-GMR for small N_0 values, and performs worse than both D-GMR and IC-GMR for higher N_0 values. For an enrollment dataset larger than about 4 min for speaker F1 (with statistical significance starting at about 5 min) and 2 min for speaker M1 (with statistical significance starting at about 3 min), the best performance is achieved with the D-GMR model. As could be expected, this shows that when enough data from the source speaker is available, there is no need to exploit prior information on the reference speaker. Conversely, for a small size of the enrollment dataset (which is one of the main goals of this study), the best performance is clearly obtained with the IC-GMR technique. This tendency is observed for both source speakers. The IC-GMR outperforms all other methods below approximately 3 min for F1,

and below approximately 1.5 min for M1. As an example, at 1 min of adaptation data, the gain of IC-GMR over D-GMR is 0.6 mm for F1, and 0.3 mm for M1. The gain is even higher at smaller N_0 values but the absolute performances of all models decrease significantly. In terms of gain on data, the RMSE obtained by the IC-GMR on speaker F1 at about 1 min of enrollment data is close to the RMSE obtained by the D-GMR (and the SC-GMR) at about 2.5 min of enrollment data. These results demonstrate the effectiveness of the IC-GMR approach and the benefit of exploiting prior information on the reference speaker to cope with the lack of knowledge on the source speaker. Interestingly, the IC-GMR outperforms systematically (and significantly) the SC-GMR, for both speakers. Moreover, the gain of SC-GMR over D-GMR at small N_0 values is noticeable for speaker M1, but more moderate for speaker F1 (for F1, the gain of SC-GMR over D-GMR is significant only for the two smallest sizes of enrollment dataset; the gain of IC-GMR over D-GMR is here much larger than the gain of SC-GMR over D-GMR). To explain the difference between IC-GMR and SC-GMR, we recall that for the SC-GMR, the enrollment dataset is used to train the **Z**-to-**X** GMR "from scratch" while keeping the reference **X**-to-**Y** GMR unchanged. In the IC-GMR, the statistical relationships between all available source and reference speakers' data (i.e. $\mathbf{z}$, $\mathbf{x}$, and $\mathbf{y}$) are jointly exploited. Therefore, addressing the cross-speaker and cross-modality mapping issues in the same probabilistic framework, as it it the case for the IC-GMR, appears to be the most effective strategy.

Now we discuss the performance of the "SC-GMR (no-audio)" and "IC-GMR (no-audio)" models for which the enrollment stage is done without exploiting the audio signal to align source and reference speaker's data. The performance is lower than the one obtained with the SC-GMR and IC-GMR using audio for data alignment. The difference is significant and amounts to 10% on average (for both speakers) for the SC-GMR, and about 5% for the IC-GMR. This decrease of performance can be explained by the difficulty to encode both inter-speaker and intra-speaker variability in a simple linear model such as the joint EigenTongue model proposed in Section 3.3. However, the IC-GMR (no-audio) remains significantly better than the D-GMR for less than 1 min of enrollment data (which is not the case for the SC-GMR (no audio)). Therefore, in a scenario where the patient or learner is not able to phonate, the IC-GMR remains the best mapping approach to design a robust system of visual biofeedback.

4.4. Generalization capability of C-GMR techniques

As mentioned in the introduction, this study also aims at evaluating the generalization capability of the D-GMR, SC-GMR and IC-GMR conversion models. This would be of significant interest in the context of pathological speech or of speech produced by a L2 learner. Ideally, the model should be able to generalize to articulations not seen during the enrollment stage, but that should be acquired during the therapy / L2 training.

In a first attempt to test this property, we have simulated such a scenario using the data of speakers F1 and M1 (none of them being a pathological speaker). For each speaker, we conducted a series of simulations where one phoneme (or one class of phonemes) was removed from the enrollment set and used as testing set. More specifically, for the enrollment data, we used the VCV sequences from the corpus described in Section 3.1, V being in {a i u ɛ o}, and C being in {t k ʁ l s ʃ} (two repetitions of each sequence). These two sets were selected to maximize the coverage of lingual articulations for French. For each simulation, we generated an enrollment dataset composed of the above VCV sequences *where one phoneme was excluded* (this results in about 1 min of enrollment data). Hence, either V belongs to a subset of 4 vowels taken from {a i u ɛ o} or C belongs to a subset of 5 consonants taken from {t k ʁ l s ʃ}. This was done independently for each of the eleven phonemes in {a i u ɛ o t k ʁ l s ʃ}. For each enrollment dataset (i.e. for each missing phoneme), a corresponding test dataset was generated: We first selected in the VCV dataset the sequences where V or C is the phoneme in {a i u ɛ o t k ʁ l s ʃ} *not selected in the training set*. From these sequences, only the vectors corresponding to the tested phoneme (i.e. neither taken from the V part nor the C part of the sequence) were selected to compose the test dataset. Moreover, in the case of consonants, we added to the test dataset the feature vectors corresponding to the C part of VCV sequences where C is in {t d n k g ʁ l s ʃ z ʒ} and with the same place of articulation as the tested phoneme. This allowed us to increase the size of each test set, while maintaining phonetic consistency. Let us give an example: when testing the generalization of the model to phoneme /t/, the training set was composed of all vectors from VCV sequences with V in {a i u ɛ o} and C in {k ʁ l s ʃ}, and the test set was composed of the vectors extracted from the C part of VCV sequences with V not in {a i u ɛ o} and C in {t d n}. For each of the eleven phonemes {a i u ɛ o t k ʁ l s ʃ}, the D-GMR, SC-GMR and IC-GMR models were trained with the corresponding missing phoneme enrollment dataset, following the procedure described in Section 2.3, and were

evaluated on the corresponding test set. The same reference speaker model as the one considered in Section 4.3 was used for the **X**-to-**Y** GMR of the SC-GMR, and for the initialization of the IC-GMR. For comparison, we also carried out a set of baseline experiments, one for each model (D-GMR, SC-GMR and IC-GMR), where all VCV sequences with the eleven phonemes (and only those phonemes) were used for enrollment training (i.e. no phoneme among the eleven ones was missing; testing was made with the same test dataset as for the generalization experiments). These baseline experiments differ from the ones described in Section 4.3 by the use of VCV sequences only in the enrollment dataset, instead of VCV and full sentences as done in Section 4.3. Note that, due to the relatively small size of the VCV corpus, no cross-validation procedure was performed in all the following experiments.

Fig. 7 displays the boxplots of the RMSE obtained for the generalization experiments and for the baselines. These boxplots represent the RMSE values obtained on the complete dataset of test sequences for the 11 considered phonemes (per-phoneme results will be detailed later), for both the generalization experiments and the baselines, and both F1 and M1 speakers. First, we observe that the RMSE of the baselines is slightly higher than the one presented in Fig. 5 and 6 for a similar amount of adaptation data ($\approx 1mn$) (for D-GMR, SC-GMR and IC-GMR). This may be explained by the fact that the baseline models are here trained and evaluated on a sparser dataset (VCV sequences with a limited set of phonemes) as compared to the experiments described in Section 4.3 (where the training and test sets were composed of a randomly selected set of VCV, words and sentences). Second, the errors obtained in the generalization experiments are higher than those obtained with the baselines, for all models and for both speakers. As expected, removing a phoneme from the enrollment dataset decreases the performance for this phoneme, since the model has to extrapolate from the other phonemes.

Let us now analyze the performance of all models by distinguishing their *absolute performance*, that is the RMSE obtained either in the generalization experiments or the baseline experiments, and the *relative performance*, that is the difference between the generalization and the baseline performances. Similarly to the experiments described in Section 4.3, the best absolute performance is obtained with the IC-GMR: for speaker F1, the median RMSE obtained by the IC-GMR is about 0.7 mm lower than that obtained with the SC-GMR (the second best model); for speaker M1, it is about 0.9 mm below the median error of the D-GMR (the second best model for this

speaker). Similarly, the IC-GMR has also the best relative performance for both speakers. Indeed, the degradation observed between the baseline and the generalization conditions is 1.1 mm for F1 and 0.9 mm for M1 for the IC-GMR, whereas if it is 1.4 mm for F1 and 1.3 mm for M1 for the SC-GMR, and 3.2 mm for F1 and 1.7 mm for M1 for the D-GMR. Hence, the IC-GMR appears to be the preferential choice for designing an accurate and robust conversion model.

Fig. 7 also displays the output outliers. It can be seen that the outliers produced by the IC-GMR are less spread than for the other two methods. This is another illustration of the benefit of relying on a well-estimated reference model. In contrast, building a direct **Z-Y** (cross-speaker and cross-modality) model on a very small number of feature vectors may be expected to lead to a large variability of the estimations. The situation is quite similar for the **Z-X** (cross-speaker) GMR model of the SC-GMR.

4.5. Examples of talking head animations

Fig. 8 displays snapshots of animations produced with the proposed system for two VCV sequences. These animations were built by converting the ultrasound image sequences of speaker M1 uttering /ata/ and /uku/ into EMA control parameters. The conversion was performed with the IC-GMR approach and approximately 1 min of enrollment data. In these two examples, the tongue position for the initial and final vowels, and well as the place of articulation for the middle consonant (palatal for /k/ and alveolar for /t/) appear to be correctly mapped into the vocal tract of the ATH. Video clips of these two animations are provided as supplementary material. We also provide the same sequences (i.e. /ata/ and /uku/ pronounced by M1) obtained with the D-GMR and SC-GMR.

5. Discussion

In this section, we discuss some important issues that could be raised when applying the proposed model to speech therapy or L2 language learning. In particular, in such a context it is expected that the patient or L2 learner will not be able to articulate all phonemes correctly. Two important examples of mispronunciation patterns must be mentioned:

- a phonological substitution (e.g. substitution of /t/ in /k/ in /ʁ/ context as in “trotinette”

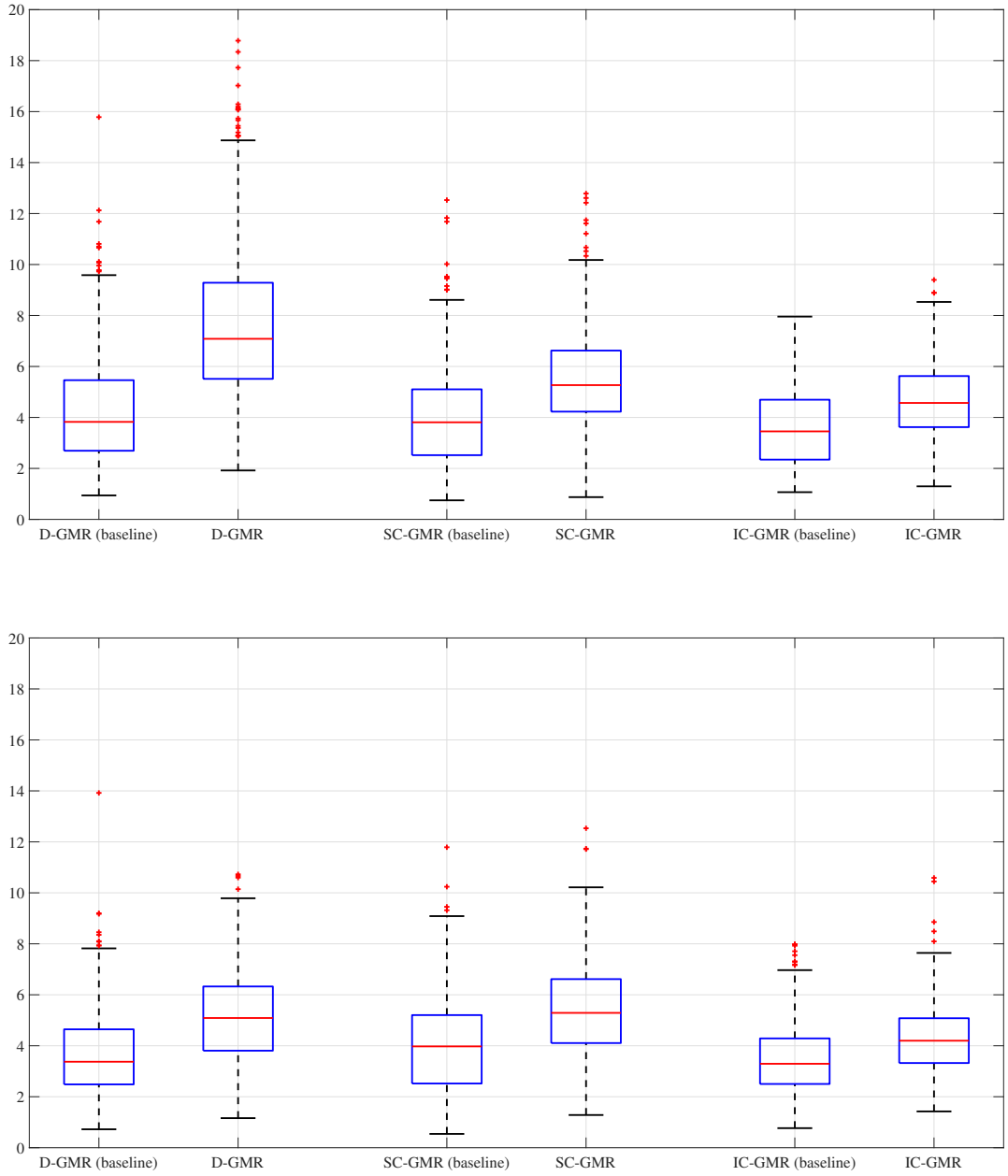

Figure 7: Boxplots of the RMSE (in mm) obtained for the different experiments for Speaker F1 (top) and Speaker M1 (bottom)

which becomes "krotinette"). In that case, each phoneme of the sequence, considered individually, exists in the articulatory repertoire of the reference speaker.

- an altered articulation of one or several phonemes (e.g., a voiceless alveolar *lateral* fricative

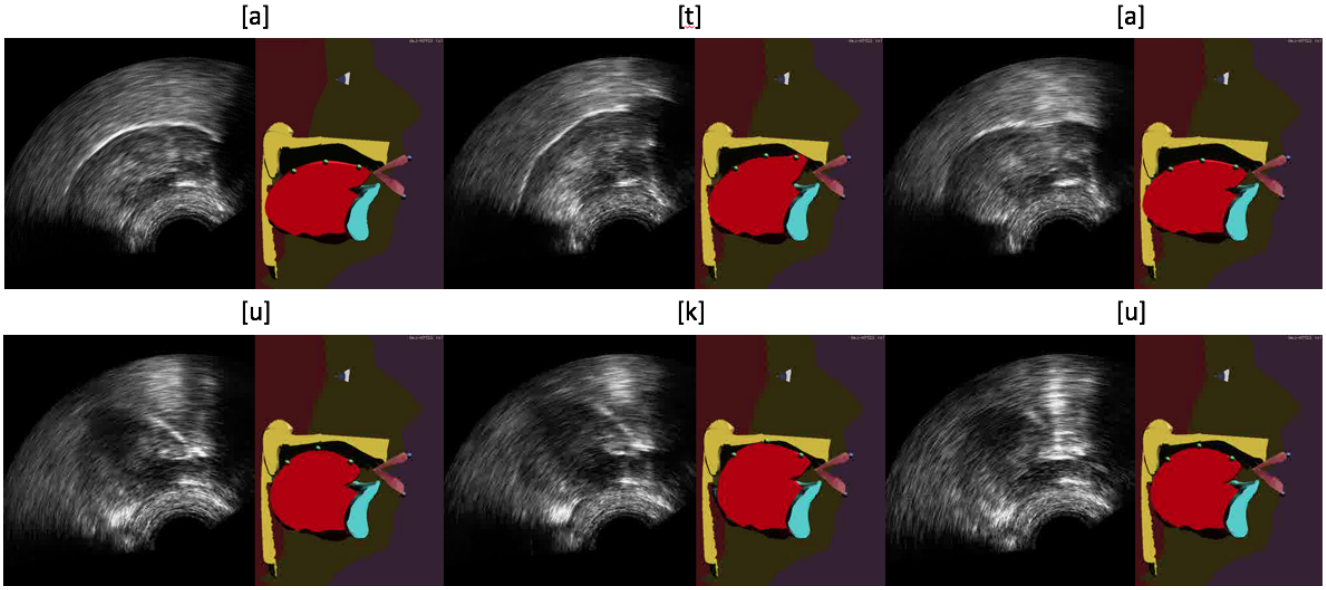

Figure 8: Snapshots of talking head animations generated from raw ultrasound images, for VCV /ata/ (top) and /uku/ (bottom) for speaker M1, with the IC-GMR approach, and considering approximately 1 min of enrollment data.

instead of /s/). Such articulation is likely to be not covered by the reference speaker's model.

In the first case (i.e. phonological substitution), the therapist/teacher could manually re-label the erroneous item. However, this may be inconvenient (time-consuming) and we thus exclude this approach. Therefore, in both cases, the only practical approach would be to exclude all erroneous items from the enrollment dataset for the adaptation of the reference model, since the latter is a supervised process and any labeling error could degrade the system's performance. In other words, the adaptation of the reference model (i.e. the training of the IC-GMR given both the source speaker adaptation data and the reference speaker data) is to be performed by exploiting only tongue articulations that can be considered phonetically correct. The increased sparsity of the resulting enrollment dataset is technically tackled in our approach by exploiting a missing data methodology that allows models to be trained from incomplete data. Note that this is what we do in the generalization experiments of Section 4.4: some phonemes are missing but all phonetic realizations do actually correspond to intended targets, without labeling errors. In practice, the speech therapist or the language teacher is expected to carefully monitor the ultrasound recordings and to exclude erroneous items from the enrollment dataset. The therapist/teacher should thus consider the quality of the phonetic realization produced by the patient/learner, i.e. the articulation

produced, but not the phonological target.

Then, in order to be used with pathological or non-native speakers, the system should be able to deal with articulations which were not retained during the enrollment session (i.e. because considered as phonetically incorrect and thus excluded from the enrollment dataset) for training the IC-GMR model. In the case of a phonological substitution, we can conjecture an adequate behavior of the conversion model as long as these phonemes considered individually are covered by the reference model (even partially). Indeed, the system might be able to provide the correct feedback even for a target phoneme absent from the enrollment dataset, as demonstrated by the generalization experiment proposed in Section 4.4 (*when the user is finally able to articulate this phoneme* at the end of a set of pronunciation training sessions). In the case of an altered phoneme, not covered by the reference model, the behavior of the system is difficult to predict, since it likely depends on many factors. Some of these factors are intrinsic, e.g. the inner interpolation capability of the machine learning mapping technique. Some of these factors are extrinsic, e.g., the phonetic proximity between the reference speaker language and the source speaker language, the type of pathology of the patient, the ability of the patient/learner to integrate the visual information to adjust his/her motor strategies, etc. Properly evaluating the final interpolation capabilities of the system (after training with consistent but incomplete data, and taking into account erroneous realizations of the source speaker) is a problem on its own. Because of the many factors mentioned above, this deserves a full dedicated study based on multi-speaker, and multi-language or multi-pathology, ultrasound databases. In a general manner, the capability of GMM-like models (and of other families of models) to correctly interpolate features between data classes in more or less void regions of the data space is a topic of research on its own in the machine learning community. In addition, one way to enhance the overall performance of the system would be to perform the model adaptation following a close-loop/incremental paradigm: the conversion model could be first trained using a few "anchor phonemes" that the source speaker is clearly able to pronounce correctly. Then, during the practical system use, the therapist/teacher could enrich the enrollment dataset with any new phonetic realization that he/she would consider as valid (resulting from the patient/learner progress) and retrain (automatically) the conversion model to improve the visual feedback. Note that, in contrast to the D-GMR model, a very interesting property of the IC-GMR model is that, even if the patient/learner is unable to pronounce some phonemes at the beginning of the training

sessions, the components corresponding to these phonemes do exist in the model, since the IC-GMR is based on the reference speaker's GMR. During this incremental adaptation process, we can thus expect these components to be "activated" as soon as the patient/learner's articulation will be able to approach them, and thus to behave like "attractors".

6. Conclusion

In this article, we presented a new method for automatically animating the tongue model of an articulatory talking head from ultrasound images. The proposed method was developed in the context of visual biofeedback which aims at showing a speaker his/her own tongue movements when speaking. The proposed method is based on the statistical mapping between a set of visual features extracted from raw ultrasound images of the speaker and a set of EMA control parameters of the talking head. This cross-modality and cross-speaker mapping problem was addressed using supervised machine learning, in the framework of Gaussian Mixture Models. We compared different approaches and found that the best results were obtained with the Integrated-Cascaded Gaussian Mixture Regression (IC-GMR) recently proposed in the context of cross-speaker acoustic-articulatory inversion. The core idea of this approach is to exploit a conversion model pre-trained on a large dataset recorded on a reference speaker, and to adapt this model to the user.

The main results obtained after an experimental evaluation based on a multi-speaker database in a simplified usage scenario are that, compared to the other tested models (D-GMR and SC-GMR):

- The IC-GMR approach leads to the best trade-off between conversion performance and amount of enrollment data,
- The IC-GMR generalizes better to articulatory movements not seen during training.

These two features make this model appropriate for implementing a robust system of visual articulatory biofeedback.

Future works will mainly consist in testing the system in more realistic scenarios. First, we plan to record non-native learners, test (offline) the behavior of the system on source speaker realizations that are not present in the enrollment dataset, and test the influence on the system performance of erroneous realizations that are not discarded or not properly re-labeled. Then, we will evaluate the proposed system in realistic practical scenarios, involving speakers with phonetic or phonological

disorders, as well as L2 learners, in the spirit of what is discussed in Section 5. Besides, a long-term perspective could be the automatic adaptation of the geometry of the reference ATH to the geometry of the source speaker's vocal tract, as recently proposed by [48] (for MRI images).

7. Supplementary material

The source code of the C-GMR technique is publicly available and can be downloaded on the following Git repository https://git.gipsa-lab.grenoble-inp.fr/cgmr.git . All datasets (audio, ultrasound and EMA for both the reference speaker and source speakers M1 and F1) used in this study will be made publicly available if the paper is accepted. Examples of ATH animations from ultrasound images are also provided in the following video file: supplementary_material_fabre_specom_2016.mp4

8. Acknowledgments

The authors acknowledge the support of the French Région Rhône-Alpes through doctoral funding to Diandra Fabre in the framework of the ARC6 program "Information and communication technologies and innovative computer uses".

9. References

[1] P. Badin, A. B. Youssef, G. Bailly, F. Elisei, T. Hueber, Visual articulatory feedback for phonetic correction in second language learning, in: Proc. of L2WS, Tokyo, Japan, 2010, pp. P1–10.

[2] T. M. Byun, E. R. Hitchcock, M. T. Swartz, Retroflex versus bunched in treatment for rhotic misarticulation: Evidence from ultrasound biofeedback intervention, Journal of Speech, Language, and Hearing Research 57 (2014) 2116–2130.

[3] Z. Roxburgh, J. M. Scobbie, J. Cleland, Articulation therapy for children with cleft palate using visual articulatory models and ultrasound biofeedback, in: Proc. of ICPhS, Glasgow, Scotland, 2015.

[4] J. Cleland, J. M. Scobbie, A. A. Wrench, Using ultrasound visual biofeedback to treat persistent primary speech sound disorders, Clinical Linguistics & Phonetics 29 (2015) 1–23 575–597.

[5] H. M. Lipetz, B. M. Bernhardt, A multi-modal approach to intervention for one adolescent's frontal lisp, Clinical linguistics & phonetics 27 (2013) 1–17.

[6] S. Ouni, Tongue control and its implication in pronunciation training, Computer Assisted Language Learning 27 (2014) 439–453.

[7] F. Gibbon, A. Lee, Electropalatography for older children and adults with residual speech errors, in: Seminars in Speech and Language, volume 36, Thieme Medical Publishers, 2015, 2015, pp. 271–282.

[8] S. Wood, Electropalatography in the assessment and treatment of speech difficulties in children with Down syndrome, Down Syndrome Research and Practice 12 (2010) 98–102.

[9] F. Gibbon, W. J. Hardcastle, H. Suzuki, An electropalatographic study of the /r/,/l/ distinction for Japanese learners of English, Computer Assisted Language Learning 4 (1991) 153–171.

[10] A. Lee, N. Zharkova, F. Gibbon, Vowel imaging, in: M. Ball, F. Gibbon (Eds.), Handbook of vowels and vowel disorders. 2nd edition, Hove: Psychology Press, 2013, pp. 138–159.

[11] A. Lee, F. E. Gibbon, E. Kearney, D. Murphy, Tongue–palate contact during selected vowels in children with speech sound disorders, International Journal of Speech-Language Pathology 16 (2014) 562–570.

[12] M. A. Epstein, Ultrasound and the IRB, Clinical Linguistics & Phonetics 19 (2005) 567–572.

[13] M. Adler-Bock, B. M. Bernhardt, B. Gick, P. Bacsfalvi, The use of ultrasound in remediation of north American English /r/ in 2 adolescents, American Journal of Speech-Language Pathology 16 (2007) 128–139.

[14] M. Cavin, The use of ultrasound biofeedback for improving English /r/, Working Papers of the Linguistics Circle 25 (2015) 32–41.

[15] J. L. Preston, N. Brick, N. Landi, Ultrasound biofeedback treatment for persisting childhood apraxia of speech, American Journal of Speech-Language Pathology 22 (2013) 627–643.

[16] M. Stone, A guide to analysing tongue motion from ultrasound images, Clinical linguistics & phonetics 19 (2005) 455–501.

[17] J. Cleland, C. McCron, J. M. Scobbie, Tongue reading: Comparing the interpretation of visual information from inside the mouth, from electropalatographic and ultrasound displays of speech sounds, Clinical Linguistics & Phonetics 27 (2013) 299–311.

[18] B. S. Atal, J. Chang, M. V. Mathews, J. W. Tukey, Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique, The Journal of the Acoustical Society of America 63 (1978) 1535–1555.

[19] J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, E. Saltzman, Accurate recovery of articulator positions from acoustics: New conclusions based on human data, The Journal of the Acoustical Society of America 100 (1996) 1819–1834.

[20] S. Ouni, Y. Laprie, Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion, The Journal of the Acoustical Society of America 118 (2005) 444–460.

[21] M. Rahim, W. Keijn, J. Schroeter, C. Goodyear, Acoustic to articulatory parameter mapping using an assembly of neural networks, in: Proc. of ICASSP, Toronto, Ontario, Canada, 1991, pp. 485–488.

[22] K. Richmond, Estimating articulatory parameters from the acoustic speech signal, Ph.D. thesis, University of Edinburgh, 2002.

[23] G. Ananthakrishnan, O. Engwall, Mapping between acoustic and articulatory gestures, Speech Communication 53 (2011) 567–589.

[24] T. Toda, A. W. Black, K. Tokuda, Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model, Speech Communication 50 (2008) 215–227.

[25] S. Hiroya, M. Honda, Estimation of articulatory movements from speech acoustics using an HMM-based speech production model, IEEE Transactions on Speech and Audio Processing 12 (2004) 175–185.

[26] H. Zen, Y. Nankaku, K. Tokuda, Continuous stochastic feature mapping based on trajectory HMMs, IEEE Transactions on Audio, Speech, and Language Processing 19 (2011) 417–430.

[27] S. Ouni, M. M. Cohen, D. W. Massaro, Training Baldi to be multilingual: A case study for an Arabic Badr, Speech Communication 45 (2005) 115 – 137.

[28] P. Badin, F. Elisei, G. Bailly, Y. Tarabalka, An audiovisual talking head for augmented speech generation: Models and animations based on a real speakers articulatory data, in: Articulated Motion and Deformable Objects, Springer, 2008, pp. 132–143.

[29] O. Engwall, Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher, Computer Assisted Language Learning 25 (2012) 37–64.

[30] S. Fagel, A. Hilbert, C. C. Mayer, M. Morandell, M. Gira, M. Petzold, Avatar user interfaces in an OSGi-based system for health care services, in: Proc. of AVSP, Annecy, France, 2013, pp. 173–174.

[31] A. Ben Youssef, P. Badin, G. Bailly, P. Heracleous, Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden Markov models, in: Proc. of Interspeech, Brighton, England, 2009, pp. 2255–2258.

[32] T. Hueber, L. Girin, X. Alameda-Pineda, G. Bailly, Speaker-adaptive acoustic-articulatory inversion using cascaded Gaussian mixture regression, IEEE/ACM Transactions on Audio, Speech, and Language Processing 23 (2015) 2246–2259.

[33] K. Xu, Y. Yang, A. Jaumard-Hakoun, C. Leboullenger, G. Dreyfus, P. Roussel, M. Stone, B. Denby, Development of a 3D tongue motion visualization platform based on ultrasound image sequences, in: Proc. of ICPhS, Glasgow, Scotland, 2015.

[34] M. Li, C. Kambhamettu, M. Stone, Automatic contour tracking in ultrasound images, Clinical linguistics & phonetics 19 (2005) 545–554.

[35] P. Badin, Y. Tarabalka, F. Elisei, G. Bailly, Can you read tongue movements? Evaluation of the contribution of tongue display to speech understanding, Speech Communication 52 (2010) 493–503.

[36] Y. Stylianou, O. Cappé, E. Moulines, Continuous probabilistic transform for voice conversion, IEEE Transactions on Speech and Audio Processing 6 (1998) 131–142.

[37] T. Toda, A. W. Black, K. Tokuda, Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory, IEEE Transactions on Audio, Speech, and Language Processing 15 (2007) 2222–2235.

[38] D. Fabre, T. Hueber, P. Badin, Automatic animation of an articulatory tongue model from ultrasound images using gaussian mixture regression, in: Proc. of Interspeech, Singapour, Malaysia, 2014, pp. 2293–2297.

[39] C. M. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag New York, USA, 2006.

[40] Z. Ghahramani, M. I. Jordan, Supervised learning from incomplete data via an EM approach, Advances in Neural Information Processing Systems 6 (1994) 120–127.

[41] Z. Ghahramani, M. I. Jordan, Learning from incomplete data, Technical Report, Cambridge, MA, USA, 1995.

[42] M. Aron, M.-O. Berger, E. Kerrien, B. Wrobel-Dautcourt, B. Potard, Y. Laprie, Multimodal acquisition of articulatory data: Geometrical and temporal registration, The Journal of the Acoustical Society of America 139 (2016) 636–648.

[43] T. Hueber, G. Chollet, B. Denby, M. Stone, Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application, in: Proc. of ISSP, Strasbourg, France, 2008, pp. 365–369.

[44] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1991) 71–86.

[45] T. Hueber, G. Aversano, G. Chollet, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel, M. Stone, Eigentongue feature extraction for an ultrasound-based silent speech interface, in: Proc. of ICASSP, Honolulu, Hawaii, USA) pages=1245-1248,, 2007.

[46] S. Young, P. Woodland, G. Evermann, M. Gales, The HTK toolkit 3.4.1, 2013.

[47] T. Toda, A. W. Black, K. Tokuda, Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model, Speech Communication 50 (2008) 215–227.

[48] J. A. Valdés Vargas, P. Badin, L. Lamalle, Articulatory speaker normalisation based on MRI-data using three-way linear decomposition methods, in: Proc. of Interspeech, Portland, Oregon, USA, 2012, pp. 2186–2189.