Speaker adaptation of an acoustic-to-articulatory inversion model using cascaded Gaussian mixture regressions

Thomas Hueber, Gérard Bailly, Pierre Badin, Frédéric Elisei

Speaker Adaptation of an Acoustic-Articulatory Inversion Model using Cascaded Gaussian Mixture Regressions

Thomas Hueber, Gérard Bailly, Pierre Badin, Frédéric Elisei

GIPSA-lab, UMR 5216/CNRS/INP/UJF/U.Stendhal, Grenoble, France

(firstname.lastname)@gipsa-lab.grenoble-inp.fr

Abstract

The article presents a method for adapting a GMM-based acoustic-articulatory inversion model trained on a reference speaker to another speaker. The goal is to estimate the articulatory trajectories in the geometrical space of a reference speaker from the speech audio signal of another speaker. This method is developed in the context of a system of visual biofeedback, aimed at pronunciation training. This system provides a speaker with visual information about his/her own articulation, via a 3D orofacial clone. In previous work, we proposed to use GMM-based voice conversion for speaker adaptation. Acoustic-articulatory mapping was achieved in 2 consecutive steps: 1) converting the spectral trajectories of the target speaker (i.e. the system user) into spectral trajectories of the reference speaker (voice conversion), and 2) estimating the most likely articulatory trajectories of the reference speaker from the converted spectral features (acoustic-articulatory inversion). In this work, we propose to combine these two steps into the same statistical mapping framework, by fusing multiple regressions based on trajectory GMM and maximum likelihood criterion (MLE). The proposed technique is compared to two standard speaker adaptation techniques based respectively on MAP and MLLR.

Index Terms: articulatory inversion, voice conversion, speaker adaptation, GMM, computer assisted pronunciation training, biofeedback

1. Introduction

In the context of speech therapy and computer-assisted pronunciation training (CAPT), systems of visual biofeedback can be used to increase the articulatory awareness of a learner by displaying the position of his/her tongue and lips. These systems can be divided into two categories:

- Systems using motion capture instrumentation to capture directly the motion of the speech articulators (mainly the tongue) such as electro-palatography as in [1], or ultrasound as in [2], [3].
- Systems aiming at estimating articulatory trajectories directly from the speech audio signal. In [4], Engvall proposed a semi-automatic system in which the learner's pronunciation was first evaluated by an expert in phonetics. Then, the corresponding articulatory trajectories were automatically presented via a virtual orofacial clone, able to display simultaneously the motion of tongue and lips. In our previous work [6], we described a fully automatic system of visual biofeedback, based also on an orofacial clone [7]. In this approach, the orofacial clone, composed of tongue, lips, and jaw 3D models, is animated automatically from the speech audio signal using acoustic-articulatory inversion (figure 1). The present work focuses specifically on the speaker adaptation problem.

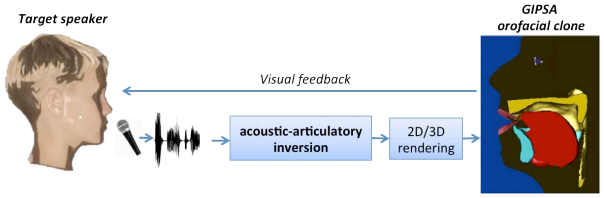


Figure 1: *System of visual articulatory feedback.*

We address the problem of adapting a statistical acoustic-articulatory model trained on a reference speaker to any other speaker, referred to as the *target speaker*. This is a critical issue for the design of a multi-speaker system of visual biofeedback, based on acoustic-articulatory inversion.

The problem of acoustic-articulatory inversion, which consists in recovering the dynamics of the main speech articulators (tongue, lips, velum, jaw) from the speech audio signal, has been addressed in many studies using either codebook-based approaches, as in [8], or statistical mapping techniques, as in [9] [10] [11] [12] (based respectively on ANN, SVM, GMM and HMM). However, only a few studies have addressed the problem of speaker adaptation of the acoustic-articulatory inversion model. In [19], Dusan and Deng proposed to compensate the difference of vocal tract length between the target speaker and the reference speaker on which the inversion model was trained. In [20], Hiroya proposed a speaker adaptation technique for an HMM-based acoustic-articulatory inversion model (initially introduced in [12]). The adaptation method is an iterative procedure composed of the 2 following steps: 1) estimating the articulatory trajectories from the target speaker acoustics and the reference inversion model and 2) finding the parameters that maximize the likelihood of the inversion model for both the target speaker acoustics and the estimated articulatory trajectories. In [21], we described a statistical inversion technique also based on trajectory HMM. Unlike [12], local dependencies between acoustic and articulatory parameters were modeled for each HMM state by Gaussian distributions with full covariance matrix rather than linear regression functions. In that study, the problem of speaker adaptation was preliminary addressed by introducing a GMM-based spectral conversion step before the acoustic-articulatory inversion step. The goal was to adapt the acoustic observations rather than the model parameters (*feature-based* instead of *model-based* speaker adaptation). The spectral features of the target speaker were mapped onto the reference speaker's acoustic space.

In this paper, we propose a new approach which merges both the voice conversion step and the acoustic-articulatory inversion step into a single GMM-based mapping framework. In this work, we adopted the framework introduced by Toda in [17] which is based on an explicit modeling of the parameter dynamics (trajectory GMM) and maximum likelihood

criterion (MLE). The proposed technique is compared to two standard speaker adaptation techniques based respectively on maximum-a-posteriori (MAP) and maximum likelihood linear regression (MLLR).

The article is organized as follows. Section 2 details the acoustic-articulatory inversion technique based on trajectory GMM. The theoretical aspects of the proposed speaker adaptation techniques are presented in section 3 (state-of-the-art speaker adaptation techniques based on MAP and MLLR are also briefly recalled). Section 4 describes the data acquisition protocol and details the practical implementation of the mapping techniques. Experimental results are presented and discussed in section 5. Conclusions and perspectives are presented in the last section.

2. Acoustic-articulatory inversion based on trajectory GMM

The following section briefly recalls the theoretical aspects of the acoustic-articulatory inversion technique based on trajectory GMM [11]. Sequences of spectral and articulatory feature vectors for the reference speaker are noted respectively $\mathbf{x}$ and $\mathbf{y}$, and are defined as: $\mathbf{x} = [\mathbf{x}_1,..,\mathbf{x}_t,...,\mathbf{x}_T]$ and $\mathbf{y} = [\mathbf{y}_1,..,\mathbf{y}_t,...,\mathbf{y}_T]$, where $\mathbf{x}_t$ / $\mathbf{y}_t$, are D_x/D_y dimensional vectors of acoustic/articulatory features observed at the time t (T is the sequence length). $\mathbf{Y}$ is built by augmenting static features with their first derivatives, such as $\mathbf{Y} = [\mathbf{Y}_1,..,\mathbf{Y}_t,...,\mathbf{Y}_T]$ with $\mathbf{Y}_t = [\mathbf{y}_t\ \Delta\mathbf{y}_t]$. The joint probability density function (pdf) of acoustic and articulatory features is modeled by a Gaussian Mixture Model (GMM), such as:

$$p(\mathbf{z} \mid \Theta) = p(\mathbf{x},\mathbf{Y}) = \sum_{m=1}^{M} \alpha_m N(\mathbf{z}, \mu_m, \Sigma_m) \text{ with}$$

$$\mathbf{z} = [\mathbf{x},\mathbf{Y}],\ \mu_m = \begin{bmatrix} \mu_m^{\mathbf{x}} \\ \mu_m^{\mathbf{Y}} \end{bmatrix},\ \Sigma_m = \begin{bmatrix} \Sigma_m^{\mathbf{xx}} & \Sigma_m^{\mathbf{xY}} \\ \Sigma_m^{\mathbf{Yx}} & \Sigma_m^{\mathbf{YY}} \end{bmatrix} \quad (1)$$

where Θ is the parameter set of the model, $N(.,\mu,\Sigma)$ is a normal distribution with mean μ and covariance matrix Σ, M is the number of mixture components, and α_m is the weight associated with the m^{th} mixture component. Given a training dataset of feature vectors for the reference speaker, the parameters of a GMM (weights, mean vectors and covariance matrices for each component) are estimated using the expectation-maximization (EM) algorithm (the initial clustering of acoustic-articulatory space is obtained using the k-means algorithm). For the mapping stage, a conditional pdf $p(\mathbf{Y}_t \mid \mathbf{x}_t, \hat{m}_t, \Theta)$ is derived, for each frame t, from the joint pdf $p(\mathbf{x}_t, \mathbf{Y}_t)$ estimated during training, such as:

$$p(\mathbf{Y}_t \mid \mathbf{x}_t, \hat{m}_t, \Theta) = N(\mathbf{Y}_t, E^{\mathbf{y}}_{\hat{m}_t,t}, D^{\mathbf{y}}_{\hat{m}_t})$$

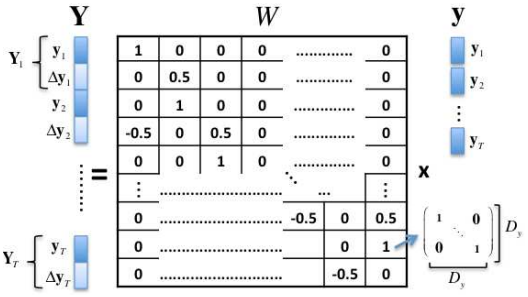
$$\text{with } \begin{cases} E^{\mathbf{y}}_{\hat{m}_t,t} = \mu^{\mathbf{y}}_{\hat{m}_t} + \Sigma^{\mathbf{yx}}_{\hat{m}_t}\Sigma^{\mathbf{xx}^{-1}}_{\hat{m}_t}(\mathbf{x}_t - \mu^{\mathbf{x}}_{\hat{m}_t}) \\ D^{\mathbf{y}}_{\hat{m}_t} = \Sigma^{\mathbf{yy}}_{\hat{m}_t} - \Sigma^{\mathbf{yx}}_{\hat{m}_t}\Sigma^{\mathbf{xx}^{-1}}_{\hat{m}_t}\Sigma^{\mathbf{xy}}_{\hat{m}_t} \end{cases} \quad (2)$$

(the mathematical basis of this derivation can be found in [18], p.337) where $\hat{m} = [\hat{m}_1,..,\hat{m}_t,...,\hat{m}_T]$ is the suboptimum sequence of mixture component defined as $\hat{m} = \arg\max_m \{P(m \mid \mathbf{x},\Theta\}$ and determined using the Viterbi algorithm (in our experiment, and similarly to what was reported in [11], similar results were obtained using a forward-backward approach which takes into account in a probabilistic manner the contributions of all mixture components). Articulatory trajectories $\hat{y}$ are finally estimated by solving the following equation:

$$\hat{\mathbf{y}} = \left(W^T D^{-1}_{\hat{m}} W\right)^{-1} W^T D^{-1}_{\hat{m}} E_{\hat{m}} \quad (3)$$

$$\text{with } E_{\hat{q}} = [E_{\hat{m}_1},...,E_{\hat{m}_T}],\ D^{-1}_{\hat{m}} = diag[D^{-1}_{\hat{m}_1},...,D^{-1}_{\hat{m}_T}]$$

and W, a [$2D_xT$-by-D_yT] matrix representing the relationship between static and dynamic feature vectors, defined as:



(4)

Like the MLPG algorithm introduced by Tokuda in [13] for HMM-based speech synthesis, this method determines the sequence of feature vectors that maximizes the likelihood of the model with respect to a continuity constraint on the predicted trajectories.

3. Speaker adaptation of an acoustic-articulatory inversion model

Sequences of spectral feature vectors for the target speaker are noted $\tilde{\mathbf{x}}$ and are defined as: $\tilde{\mathbf{x}} = [\tilde{\mathbf{x}}_1,...,\tilde{\mathbf{x}}_t,...,\tilde{\mathbf{x}}_{\tilde{T}}]$ where $\tilde{\mathbf{x}}_t$ is a D_x dimensional vectors of spectral features observed at the time t ($\tilde{T}$ is the sequence length). In this paper, we focus on a *supervised* mode of speaker adaption, i.e. we assume that a set of audio recordings of both the target and the reference speaker pronouncing the same text is available. The adaptation dataset contains audio data only; no articulatory data of the target speaker is available.

3.1. Speaker adaptation of an acoustic-articulatory model based on MAP and MLLR

We investigate first the use of state-of-the-art speaker adaptation techniques to adapt the acoustic parameters of the acoustic-articulatory inversion model of the reference speaker, i.e. the mean sub-vector $\mu_m^{\mathbf{x}}$ and the covariance sub-matrix $\Sigma_m^{\mathbf{xx}}$ for each component m of the GMM (this parameter set is called $\Theta^{\mathbf{x}}$). We focus on MAP-based and MLLR adaptation techniques.

The basic principle of the MAP-based adaptation [22] is to find the model parameter set $\Theta^{\tilde{\mathbf{x}}}$ that maximizes the posterior

probability $P(\Theta^{\mathbf{x}} \mid \tilde{\mathbf{x}})$ ($\tilde{\mathbf{x}}$ are the adaptation data) considering the acoustic GMM trained on reference speaker data $\Theta^{\mathbf{x}}$, as a prior probability distribution over models parameters, such as: $\Theta^{\tilde{\mathbf{x}}} = \arg\max_{\Theta^{\mathbf{x}}} \{P(\Theta^{\mathbf{x}} \mid \tilde{\mathbf{x}})\} = \arg\max_{\Theta^{\mathbf{x}}} \{P(\tilde{\mathbf{x}} \mid \Theta^{\mathbf{x}}) P(\Theta^{\mathbf{x}})\}$.

Parameter set $\Theta^{\tilde{\mathbf{x}}}$ is determined using the EM algorithm, using the following re-estimation equations (to be concise, we recall only the equation used to update the mean vectors; see [22] for the priors and covariance-related equations):

$$\tilde{\mu}_m^{\mathbf{x}} = \frac{\tau \cdot \mu_m^{\mathbf{x}} + \sum_{t=1}^{T} c_{mt} \tilde{\mathbf{x}}_t}{\tau + \sum_{t=1}^{T} c_{mt}} \text{ with } c_{mt} = \frac{\tilde{\alpha}_m N(\tilde{\mathbf{x}}_t, \tilde{\mu}_m^{\mathbf{x}}, \tilde{\Sigma}_m^{\mathbf{x}})}{\sum_{l=1}^{l=M} \tilde{\alpha}_l N(\tilde{\mathbf{x}}_t, \tilde{\mu}_l^{\mathbf{x}}, \tilde{\Sigma}_l^{\mathbf{x}})} \quad (5)$$

The hyperparameter τ is a heuristic factor which controls the balance between the ML (maximum likelihood) estimate of the mean using the adaptation data, and its initial value. In our implementation, the value of this parameter is shared across all the GMM components.

Maximum-Likelihood Linear Regression (MLLR) is another standard adaptation technique typically used in speech recognition systems [23]. In this technique, model parameters are adapted using an affine transform, such as:

$$\tilde{\mu}_m^{\mathbf{x}} = \mathbf{A}\mu_m^{\mathbf{x}} + \mathbf{b} \text{ and } \tilde{\Sigma}_m^{\mathbf{xx}} = \mathbf{H}\Sigma_m^{\mathbf{xx}}\mathbf{H}^T \quad (6)$$

The likelihood function of adaptation data is maximized with respect to transform parameters **(A**,**b**,**H)** using the EM algorithm. In our implementation, the parameters of the affine transform are shared across all GMM components. This results in a significant reduction of the amount of parameters to be estimated compared to the MAP approach.

3.2. Proposed speaker adaptation technique

The proposed speaker adaptation technique consists in combining spectral conversion and acoustic-articulatory inversion into a single mapping framework. The basic idea of the proposed method is to use the acoustic-articulatory model of the reference speaker, as prior knowledge's for clustering the adaptation data and estimating the parameters of the spectral mapping model.

In the adaptation stage, time-alignment is performed for each target/reference speaker pair ($\tilde{\mathbf{x}}$ / $\mathbf{x}$), using dynamic time warping (DTW). We note $q(t)$ the warped time axis given by DTW. For each target acoustic observation $\tilde{\mathbf{x}}_t$, the conditional probabilities of the acoustic-articulatory GMM given the corresponding acoustic-articulatory observation $z_{q(t)} = [\mathbf{x}_{q(t)} \mathbf{y}_{q(t)}]$ of the reference speaker, is then calculated:

$$P(m \mid z_{q(t)}, \Theta) = \frac{\alpha_m N(z_{q(t)}, \mu_m, \Sigma_m)}{\sum_{l=1}^{l=M} \alpha_l N(z_{q(t)}, \mu_l, \Sigma_l)} \quad (7)$$

where Θ is the parameter set of the acoustic-articulatory GMM (we use the same notation *m* for representing both the m^{th} acoustic-articulatory class and the m^{th} GMM component). The adaptation dataset is then clustered by assigning to class *m* all the acoustic observations for which the conditional probabilities $P(m \mid z_{q(t)}, \Theta)$ is maximum across all mixture components. This clustering is used to initialize a so-called *spectral mapping GMM* $\tilde{\Theta}$, which models the joint pdf of target/reference speaker's acoustic observations, such as:

$$p(\mathbf{x}, \mathbf{x} \mid \tilde{\Theta}) = \sum_{m=1}^{M} \alpha_m N([\tilde{\mathbf{x}}, \mathbf{x}], \tilde{\mu}_m, \tilde{\Sigma}_m) \text{ with}$$

$$\tilde{\mu}_m = \begin{bmatrix} \tilde{\mu}_m^{\tilde{\mathbf{x}}} \\ \mu_m^{\mathbf{x}} \end{bmatrix}, \tilde{\Sigma}_m = \begin{bmatrix} \tilde{\Sigma}_m^{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} & \Sigma_m^{\tilde{\mathbf{x}}\mathbf{x}} \\ \Sigma_m^{\mathbf{x}\tilde{\mathbf{x}}} & \Sigma_m^{\mathbf{xx}} \end{bmatrix} \quad (8)$$

In order to have an acceptable (i.e. well-conditioned) estimation of the covariance matrix for classes with few adaptation data, we use the shrinkage method described by Ledoit and Wolf in [24]. Besides, an iterative procedure is used to refine the DTW-alignment and thus the estimation of *spectral mapping GMM*. At each iteration, a spectral adaptation of the target signal $\tilde{\mathbf{x}}$ is achieved using the current estimation of the *spectral mapping GMM*. This spectral conversion step reduces the acoustic distance between the target and reference speaker and facilitates the DTW-alignment.

The presented training procedure imposes that the *spectral mapping* GMM $\tilde{\Theta}$ and the acoustic-articulatory inversion GMM Θ share the same structure, i.e. both models have the same number of components (*M*) and there is a one-to-one correspondence between each component of the two models.

In the inversion stage, the suboptimum sequence of mixture component $\hat{m} = [\hat{m}_1, .., \hat{m}_t, ..., \hat{m}_T]$ defined as $\hat{m} = \arg\max_m \{P(m \mid \tilde{\mathbf{x}}, \tilde{\Theta}\}$ is first determined using the Viterbi algorithm, from the acoustic observations of the target speaker and the spectral mapping GMM. A conditional pdf $p(\mathbf{x}_t \mid \tilde{\mathbf{x}}_t, \hat{m}_t, \tilde{\Theta}) = N(\mathbf{x}_t, E_{\hat{m}_t}^{\mathbf{x}}, D_{\hat{m}_t}^{\mathbf{x}})$ is derived, for each frame *t*, such as:

$$E_{\hat{m}_t}^{\mathbf{x}} = \mu_{\hat{m}_t}^{\mathbf{x}} + \Sigma_{\hat{m}_t}^{\mathbf{x}\tilde{\mathbf{x}}} \Sigma_{\hat{m}_t}^{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{-1}} (\tilde{\mathbf{x}}_t - \tilde{\mu}_{\hat{m}_t}^{\tilde{\mathbf{x}}}). \quad (9)$$

By combining equation 9 (spectral mapping) and equation 2 (acoustic-articulatory inversion), we derive a conditional pdf of the articulatory observation of the reference speaker $\mathbf{Y}_t$ given an acoustic observation of the target speaker $\tilde{\mathbf{x}}_t$ $p(\mathbf{Y}_t \mid \tilde{\mathbf{x}}_t, \hat{m}_t, \Theta, \tilde{\Theta}) = N(\mathbf{Y}_t, \tilde{E}_{\hat{m}_t}^{\mathbf{y}}, D_{\hat{m}_t}^{\mathbf{y}})$ such as:

$$\tilde{E}_{\hat{m}_t}^{\mathbf{Y}} = \mu_{\hat{m}_t}^{\mathbf{Y}} + \Sigma_{\hat{m}_t}^{\mathbf{Yx}} \Sigma_{\hat{m}_t}^{\mathbf{xx}^{-1}} (E_{\hat{m}_t}^{\mathbf{x}} - \mu_{\hat{m}_t}^{\mathbf{x}})$$

$$= \mu_{\hat{m}_t}^{\mathbf{Y}} + \Sigma_{\hat{m}_t}^{\mathbf{Yx}} \Sigma_{\hat{m}_t}^{\mathbf{xx}^{-1}} ((\mu_{\hat{m}_t}^{\mathbf{x}} + \Sigma_{\hat{m}_t}^{\mathbf{x}\tilde{\mathbf{x}}} \Sigma_{\hat{m}_t}^{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{-1}} (\tilde{\mathbf{x}}_t - \tilde{\mu}_{\hat{m}_t}^{\tilde{\mathbf{x}}})) - \mu_{\hat{m}_t}^{\mathbf{x}})$$

finally, we obtain:

$$\tilde{E}_{\hat{m}_t}^{\mathbf{Y}} = \mu_{\hat{m}_t}^{\mathbf{Y}} + \Sigma_{\hat{m}_t}^{\mathbf{Yx}} \Sigma_{\hat{m}_t}^{\mathbf{xx}^{-1}} \Sigma_{\hat{m}_t}^{\mathbf{x}\tilde{\mathbf{x}}} \Sigma_{\hat{m}_t}^{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{-1}} (\tilde{\mathbf{x}}_t - \tilde{\mu}_{\hat{m}_t}^{\tilde{\mathbf{x}}}) \quad (10)$$

The proposed mapping method is called *cascaded Gaussian mixture regressions* in reference to the product of cross-covariance matrices $\Sigma_{\hat{m}_t}^{\mathbf{Yx}}$ and $\Sigma_{\hat{m}_t}^{\mathbf{x}\tilde{\mathbf{x}}}$. It projects an acoustic observation directly from the acoustic space of the target speaker, to the articulatory space of the reference speaker.

4. Experimental protocol

Articulatory data of the reference speaker were recorded synchronously with the audio signal using the Carstens 2D

EMA system (AG200). Six coils were glued on the tongue tip, blade, and dorsum, as well as on the upper lip, the lower lip and the jaw. Sequence of articulatory features were downsampled from 200 Hz to 100 Hz and low-pass filtered at 20 Hz. The recorded database consisted of two repetitions of 224 VCVs, two repetitions of 109 pairs of CVC real French words, and 88 sentences (approximately 17 minutes of speech, long pauses being excluded). In order to evaluate the speaker adaptation technique, a second database of audio data only, was recorded by two target speakers (one male M1 and one female F1). These speakers were asked to pronounce the same text material as described above. The audio speech signal was downsampled to 16 kHz and parameterized by 13 MFCC coefficients (Blackman window, 25 frame length, 10 ms frame shift). In order to take into account the dynamic constraints on acoustic parameters, we adopted the approach described in [11] which consists in concatenating consecutive acoustic frames in one single feature vector. The optimal number of frames to concatenate was found to be 5. The dimensionality of the constructed vector was reduced to 25 using PCA, by keeping the eigenvectors carrying at least 90% of the variance of the training set.

For the inversion experiment on the reference speaker, the acoustic-articulatory database was divided into 5 partitions of equal size. A 5-fold cross-validation technique was employed for evaluation. One list was used for testing and the remaining 4 lists were used for training the acoustic-articulatory GMM and estimating its hyperparameters (the number of consecutive acoustic frames to take into account in the acoustic feature extraction process, the number of GMM components which was found to be 128 (16,32,64,128,256 were tested), the hyperparameter τ for MAP-based adaptation). The accuracy of the acoustic-articulatory inversion for the reference speaker was measured by calculating, for each partition, the root mean square error (referred to as μRMS) between the measured and the estimated EMA parameters. However, this quantity could not be calculated for the speaker adaptation experiments since no articulatory data was available for the target speakers. Therefore, the *articulatory recognition* paradigm, described in [6], was used: an HMM-based phonetic decoder trained on the articulatory data of the reference speaker (using a standard training procedure of context-dependent (triphone) tied-state HMM), was used to decode the synthetic articulatory trajectory at the phonetic level. The obtained recognition accuracy $Acc_{art}\% = 100 \cdot (N - D - S - I) / N$ (where N is the total number of phones in the test set, S, D and I are respectively the number of substitution, deletion, and insertion errors) was considered as a measure of the accuracy of the synthetic trajectory. In order to alleviate the problem of insertion/deletion errors due to the absence of a language model, this evaluation procedure was used only on VCV and CVC sequences (in that case, the decoder was forced to recognize VCV and CVC only).

5. Results & Discussion

For the acoustic-articulatory inversion experiment on the reference speaker, we obtained μRMS=1.3mm and Acc_{art}=94%. This result is compatible with the literature on acoustic-articulatory inversion using statistical approaches (such as [16] or [17]). Two series of speaker adaption experiments were conducted. In the first one, the audio-only database was divided into 5 partitions of equal size. One partition was used for adaption (i.e. 1/5 of the recorded database; ~2mn of speech); another partition was used for test. Results are presented in Table 1.

Table 1. *Recognition accuracy for speaker M1 and F1 (confidence interval was ±2%).*

Speaker	No adaptation	MLLR	MAP	**Cascaded -GMR**
M1	50%	85%	87%	**90%**
F1	56,78%	73%	78%	**89%**

Best performance for both speakers was obtained using the proposed speaker adaptation technique based on cascaded Gaussian mixture regressions (GMR). The most important improvement was observed for the female speaker F1 for whom the acoustic distance with the reference speaker (a male) was likely to be the greatest. The second series of experiments focused on how the performance was affected by amount of available adaptation data. Figure 2 shows the performance for different sizes of the adaptation dataset for speaker F1 (similar tendencies were observed with speaker M1).

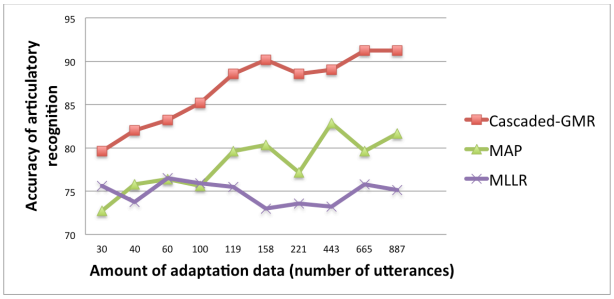


Figure 2: *Recognition accuracy as a function of the amount of adaptation data (speaker F1).*

The reduction of the amount of adaptation data affected more the proposed cascaded-GMR technique and the MAP-based technique than the MLLR approach. However, even with a very small amount of adaptation data (less than 30 VCV sequences), the proposed technique still outperformed MAP-based and MLLR approaches.

6. Conclusions and Perspectives

The article introduces a new method for adapting a GMM-based acoustic-articulatory inversion model trained on a reference speaker, to another speaker. The goal is to estimate the articulatory trajectories in the geometrical space of a reference speaker from the speech audio signal of another speaker. This method is developed in the context of a system of pronunciation training based on a 3D orofacial clone. The proposed technique which combines spectral conversion and acoustic-articulatory inversion into a single GMM-based mapping framework outperforms standard speaker adaptation techniques such as MAP and MLLR.

In future work, the proposed approach will be evaluated on a larger number of target speakers. Objective evaluation of the estimated articulatory trajectories by expert phoneticians will also be conducted. Finally, the adaptation of acoustic-articulatory inversion model to foreign speakers and speakers with speech disorders will be investigated.

7. References

[1] Wrench, A., Gibbon, F., McNeill, A.M., Wood, S., “An EPG therapy protocol for remediation and assessment of articulation disorders”, in Proc. of ICSLP, Denver, USA, pp. 965-968, 2002

[2] Bernhardt, B.M., Gick, B., Bacsfalvi, P., Adler-Bock, M. “Ultrasound in speech therapy with adolescents and adults”, Clinical Linguistics & Phonetics, vol. 19, pp. 605-617, 2005.

[3] Cleland, J., Scobbie, J.M. & Wrench, A., “Visual Feedback for Children with Speech Sound Disorders”, Poster presented at the 3rd Colloquium of British Association of Clinical Linguistics, 2011.

[4] Engwall, O., “Can audio-visual instructions help learners improve their articulation? - An ultrasound study of short term changes”, in Proc. of Interspeech, pp. 2631-2634, 2008.

[5] Massaro, D. W., Liu, Y., Chen, T. H., Perfetti, C. A. “A Multilingual Embodied Conversational Agent for Tutoring Speech and Language Learning”, in Proc. of Interspeech, Pittsburg, USA, pp. 825-828, 2006.

[6] Ben Youssef A., Hueber T., Badin P., Bailly G., "Toward a multi-speaker visual articulatory feedback system", in Proc. of Interspeech, Firenze, Italia, pp. 489-492, 2011.

[7] Badin, P., Elisei, F., Bailly, G., Tarabalka, Y., “An audiovisual talking head for augmented speech generation: models and animations based on a real speaker's articulatory data”, in 5th Conf. on Articulated Motion and Deformable Objects, Eds.: F.J. Perales & R.B. Fisher, Berlin, Heidelberg, pp. 132-143, 2008.

[8] Ouni, S., Laprie, Y., “Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion”, J. Acoustical Society of America, vol. 118, pp. 444-460, 2005.

[9] Richmond, K., “Estimating Articulatory Parameters from the Acoustic Speech Signal”, PhD thesis, CSTR Edinburgh, 2002.

[10] Toutios, A., Margaritis, K. “A support vector approach to the acoustic-to-articulatory mapping”, in Proc. of Interspeech, pp. 3221-3224, 2005.

[11] Toda, T., Black, A.W., Tokuda, K., “Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model”, Speech Comm. vol. 50, no. 3, pp. 215-227, 2007.

[12] Hiroya, S., Honda, M., “Estimation of Articulatory Movements from Speech Acoustics Using an HMM-Based Speech Production Model”, IEEE Transactions on Speech and Audio Processing, vol. 12, no. 2, pp. 175-185, 2004.

[13] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., “Speech parameter generation algorithms for HMM-based speech synthesis”, in Proc. of ICASSP, pp. 1315-1318, 2000.

[14] Hiroya, S. and M. Honda, “Speaker Adaptation Method for Acoustic-to-Articulatory Inversion using an HMM-Based Speech Production Model”, IEICE Transactions On Information And Systems, E87-D(5), pp. 1071-1078, 2004.

[15] Ghosh, P., Narayanan, S., “A subject-independent acoustic-to-articulatory inversion”, in Proc. of ICASSP, pp. 4624-4627, 2011.

[16] Zen, H., Nankaku, Y., Tokuda, K., “Continuous Stochastic Feature Mapping Based on Trajectory HMMs”, IEEE Trans. on Audio, Speech, and Lang. Proc., vol. 19, no. 2, pp. 417- 430, 2011.

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

[18] M. Kay., S, “Fundamentals of Statistical Signal Processing: Estimation Theory”, Prentice Hall, 1993.

[19] S. Dusan and L. Deng. Vocal-tract length normalization for acoustic-to-articulatory mapping using neural networks. J. Acoust. Soc. Am., 106(4):2181, 1999.

[20] Hiroya, S., Honda, M., “Speaker adaptation method for acoustic-to-articulatory inversion using an HMM-based speech production model”, IEICE Trans. Inf. & Syst., E87-D(5), pp. 1071-1078, 2004.

[21] Hueber T., Ben Youssef A., Bailly G., Badin P., Elisei, F., "Cross-speaker Acoustic-to-Articulatory Inversion using Phone-based Trajectory HMM for Pronunciation Training", in Proc. of Interspeech, Portland, USA, 2012.

[22] Gauvain, J. L., Lee, C. H., “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains”, IEEE Transactions on Speech and Audio Processing, vol. 2, issue 2, pp. 291-298, 1994.

[23] Leggetter, C. and Woodland, P., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer, Speech and Language, vol. 9, pp. 171-185, 1995.

[24] Ledoit, O., Wolf, M. “A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices”, Journal of Multivariate Analysis, vol. 88, issue 2, pp. 365-411, 2004.