Statistical Conversion of Silent Articulation into Audible Speech using Full-Covariance HMM

Thomas Hueber, Gérard Bailly

GIPSA-lab, CNRS/Univ. Grenoble-Alpes, Grenoble, France

Corresponding author: Thomas Hueber, Ph.D.

Address: GIPSA-lab, 11 rue des Mathématiques, 38402 Saint Martin d'Hères, France

E-mail: thomas.hueber@gipsa-lab.fr / Tel: +33 4 76 57 49 40

Suggested running title: Statistical Conversion of Silent Articulation into Audible Speech

Submitted: June 30, 2014; revision: November 14, 2014, February 21, 2015

ABSTRACT

This article investigates the use of statistical mapping techniques for the conversion of articulatory movements into audible speech with no restriction on the vocabulary, in the context of a silent speech interface driven by ultrasound and video imaging. As a baseline, we first evaluated the *GMM-based mapping considering dynamic features*, proposed by (Toda et al., 2007) for voice conversion. Then, we proposed a 'phonetically-informed' version of this technique, based on *full-covariance HMM*. This approach aims 1) at modeling explicitly the articulatory timing for each phonetic class, and 2) at exploiting linguistic knowledge to regularize the problem of silent speech conversion. Both techniques were compared on continuous speech, for two French speakers (one male, one female). For modal speech, the HMM-based technique showed a lower spectral distortion (objective evaluation). However, perceptual tests (transcription and XAB discrimination tests) showed a better intelligibility of the GMM-based technique, probably related to its less fluctuant quality. For silent speech, a perceptual identification test revealed a better segmental intelligibility for the HMM-based technique on consonants.

Keywords: silent speech interface, GMM, HMM, ultrasound, articulatory-acoustic mapping.

1. INTRODUCTION

Silent Speech Interfaces (SSIs) have emerged as a new research field in the last few years (Denby et al., 2010). SSI can be defined as devices that enable oral speech communication without vocalization. With a SSI, the 'silent' speaker articulates normally but does not produce any sound. SSI could be used to preserve the privacy of conversations, for discreet hand-free communication (as in a military operation), or on the contrary, in very noisy environments (where the audio speech signal is too degraded). Since silent speech does not involve vocal folds vibration, SSI could potentially be used after a total laryngectomy, as a temporary alternative to the esophageal voice, which takes time to master, or to the tracheoesophageal voice, which may require an additional surgery. So far, different technologies have been proposed to capture the articulatory activity during silent speech, such as surface electromyography (sEMG) with sensors placed on the face and neck (Schultz and Wand, 2010), or permanent-magnetic articulography (PEMA) with magnets glued on the tongue and lips (Fagan et al., 2008). Another approach is to capture and post-process a so-called Non-Audible-Murmur (NAM) using a stethoscopic microphone (Nakajima et al., 2003). In our approach (Denby et al., 2006; Hueber et al., 2010b), articulatory movements are captured using a medical ultrasound transducer placed beneath the chin, and a video camera in front of the lips, as shown in Figure 1. This sensor set provides relatively complete information on tongue (via ultrasound), lips and jaw movements[1], while remaining totally non-invasive.

Several studies addressed the problem of *silent speech recognition*, i.e. the identification of a sequence of words from silent articulation: (Wand and Schultz, 2011) for

[1] but no systematic information about the velum position, as discussed in Section 4.1.

sEMG, (Nakajima et al., 2006) for NAM, (Gilbert et al., 2010) for PEMA and (Hueber et al., 2009) for ultrasound. In this study, we addressed the problem of *silent speech conversion*, i.e. the direct reconstruction of the speaker's voice from his/her silent articulation, without any restriction on the vocabulary size.

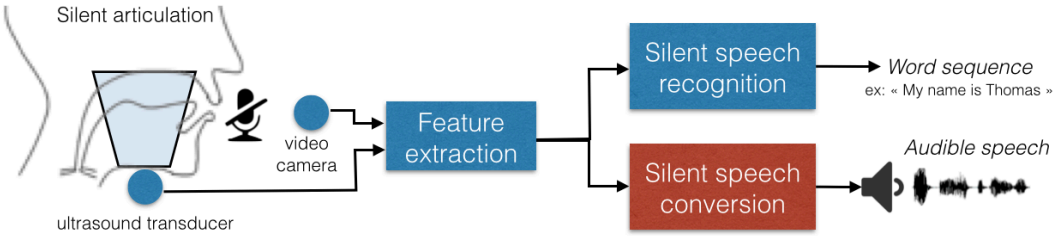

Figure 1: Silent speech interface driven by ultrasound and video imaging. The present study focuses on the direct conversion of silent articulation into audible speech without any restriction on the vocabulary size (contrary to silent speech recognition).

In our previous work (Hueber et al., 2010b), this problem was addressed using a 'recognition-followed-by-synthesis' approach. The system was composed of two chained modules: 1) a HMM-based decoder that predicts the most likely phonetic sequence from the observed articulatory movements, and 2) a unit selection algorithm that generates the spectral trajectories from the decoded phonetic sequence. The intermediate phonetic decoding step was motivated by the introduction of linguistic knowledge to regularize the problem of silent speech conversion. Such information might help recover some of the missing information in the silent articulatory data, such as the voicing feature. This approach gave encouraging results but presented some drawbacks. First, the quality of the synthesis depended strongly on the performance of the phonetic decoding: an error during that step systematically corrupted the synthesis. Second, since articulatory and acoustic modalities were processed separately during training, the dependencies between articulatory and spectral features were not explicitly modeled. As a consequence, the spectral targets depended on the decoded phonetic

labels only, and did not take into account the articulatory variability within each phonetic class. Therefore the first goal of this new study was to investigate mapping techniques that should be able to explicitly model these local acoustic-articulatory relationships.

Furthermore, in all our previous studies, articulatory-to-acoustic mapping was not performed on actual silent speech: the converted articulatory data were acquired while the speaker was still vocalizing. However, recent studies such as (Hueber et al., 2010a) and (Janke et al., 2010) suggested that silent speech articulation differs from that of modal speech, probably due to the lack of acoustic feedback. Therefore, the second goal of this new study was to evaluate our system on actual silent speech.

The problem of speech synthesis from articulatory movements, commonly called "articulatory synthesis" has been originally addressed by the use of a two or three-dimensional articulatory model of the vocal tract (Birkholz et al., 2006; Maeda, 1990), coupled with an acoustic simulation method (Sondhi and Schroeter, 1987). In the past few years, supervised machine learning techniques have brought significant improvements in articulatory-to-acoustic mapping. These techniques seem to be well adapted to tackle the non-uniqueness and the non-linear aspects of the acoustic-articulatory relationships. Most studies exploit articulatory data recorded using electromagnetic articulography (EMA) (Hiroya and Honda, 2004; Richmond, 2006; Toda et al., 2008; Zhang and Renals, 2008; Zen et al., 2011; Youssef et al., 2011; Hueber et al., 2012). This motion-capture device enables the very accurate tracking of a set of sensors glued on the main speech articulators (tongue, lips, jaw, velum). Several approaches have been proposed in the literature to model the relationship between articulatory positions captured by EMA and the corresponding speech spectrum. They aim at addressing either the direct mapping problem (articulatory synthesis) or the inverse mapping problem (acoustic-to-articulatory inversion). Some of them are based on discriminative models, such as artificial neural networks (ANN) as in (Kello and Plaut, 2004)

(direct mapping) or (Richmond, 2006) (inversion). Others are based on generative models, such as Gaussian Mixture Model (GMM) (Toda et al., 2008) (both direct mapping and inversion), Hidden Markov Models (HMM) (Hiroya and Honda, 2004; Youssef et al., 2011; Hueber et al., 2012) (inversion), and trajectory HMM (Zhang and Renals, 2008) (Zen et al., 2011) (inversion).

In this study, we investigated the use of statistical mapping techniques to convert the silent articulation captured by ultrasound and video imaging into audible speech, without any restriction on vocabulary size. First, we investigate the *GMM-based mapping considering dynamic features* proposed by (Toda et al., 2007) for voice conversion. This technique, to which we refer in this paper as GMM+dyn, is able to generate smoother and more accurate parameter trajectories than the conventional GMM-based regression. However, this approach does not use any linguistic knowledge on the articulatory-acoustic data, such as the underlying phonetic structure. In order to take such information into account, and in line with our previous work, we investigated another regression technique based on the joint modeling of acoustic and articulatory trajectories by *full-covariance HMMs*. This technique, which also guarantees smooth trajectories, is referred to as the HMM+dyn technique in this paper. Contrary to GMM+dyn, HMM+dyn aims 1) at explicitly modeling the phoneme-specific dynamics of articulation, and 2) at exploiting linguistic knowledge in order to regularize the ill-posed problem of silent speech conversion.

Both techniques were evaluated in the context of open-vocabulary and continuous speech, for two French speakers (one male and one female). They were first evaluated on modal speech (i.e. from non silent articulation) using both objective measurements and perceptual tests (transcription test and XAB preference test). The segmental intelligibility of the two techniques was then compared on actual silent speech using transcription and identification tests.

The article is organized as follows. Section 2 details the theoretical basis of GMM+dyn and HMM+dyn techniques. Section 3 describes the data acquisition and the extraction of articulatory and acoustic features. The experimental protocol and results are detailed in section 4. Conclusions and perspectives are presented in the last section.

2. Theoretical basis of considered mapping techniques

In this article, input (i.e. articulatory) and output (i.e. acoustic) feature vectors observed at time t are noted respectively $\mathbf{x}_t$ and $\mathbf{y}_t$ (the dimension of these column vectors is noted respectively D_x and D_y) and are considered as realizations of the random variables $\mathbf{X}$ and $\mathbf{Y}$. In the two mapping techniques considered in this study (GMM+dyn and HMM+dyn), output feature vectors are augmented by their *N*-first derivatives, which are referred to as 'dynamic features'. The resulting feature vector is noted $\tilde{\mathbf{y}}_t$ with $\tilde{\mathbf{y}}_t = [\mathbf{y}_t^\dagger, \Delta\mathbf{y}_t^\dagger]^\dagger$ (we considered here only the first derivative) where † denotes the transpose operator (the associated random variable is noted $\tilde{\mathbf{Y}}$). For a given sequence of T observations, the linear relation between static and dynamic feature vectors can be expressed by introducing the $[2D_yT\text{-by-}D_yT]$ matrix W defined as:

$$\begin{bmatrix} \tilde{\mathbf{y}}_1 \\ \vdots \\ \tilde{\mathbf{y}}_T \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1 \\ \Delta\mathbf{y}_1 \\ \mathbf{y}_2 \\ \Delta\mathbf{y}_2 \\ \vdots \\ \mathbf{y}_T \\ \Delta\mathbf{y}_T \end{bmatrix} = \underbrace{\begin{bmatrix} I_{D_y} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \cdots \\ -0.5I_{D_y} & \mathbf{0} & 0.5I_{D_y} & \mathbf{0} & \cdots & \cdots \\ \mathbf{0} & I_{D_y} & \mathbf{0} & \mathbf{0} & \cdots & \cdots \\ -0.5I_{D_y} & \mathbf{0} & 0.5I_{D_y} & \mathbf{0} & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & I_{D_y} & \mathbf{0} & \cdots & \cdots \\ \vdots & \cdots & & \vdots & \ddots & \cdots \\ \mathbf{0} & \cdots & & & & \mathbf{0} \\ \mathbf{0} & \cdots & & -0.5I_{D_y} & & \mathbf{0} \\ \mathbf{0} & \cdots & & & & I_{D_y} \end{bmatrix}}_{W} \times \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_T \end{bmatrix} \quad (1)$$

with I_{D_y} the D_y-dimensional identity matrix, so that $\Delta\mathbf{y}_t = 0.5\mathbf{y}_{t+1} - 0.5\mathbf{y}_{t-1}$ (in this study, the first derivatives of output features for the first and last frames were defined as: $\Delta\mathbf{y}_1 = \Delta\mathbf{y}_2$ and $\Delta\mathbf{y}_T = \Delta\mathbf{y}_{T-1}$).

2.1. GMM-based mapping considering dynamic features (GMM+dyn)

The following section recalls the theoretical aspects of the mapping technique based on Gaussian Mixture Model (GMM) with a continuity constraint on dynamic features. This technique was initially proposed in the context of voice conversion (Toda et al., 2007), and was then applied to articulatory-acoustic mapping (Toda et al., 2008). In the training stage, the joint probability density function (pdf) of input and output features is modeled by a GMM such as:

$$p(\mathbf{z}\,|\,\Theta) = p(\mathbf{x},\tilde{\mathbf{y}}\,|\,\Theta) = \sum_{m=1}^{M} \alpha_m N(\mathbf{z},\mu_m^{\mathbf{Z}},\Sigma_m^{\mathbf{Z}})$$
$$\text{with } \mathbf{z} = \begin{bmatrix}\mathbf{x}\\ \tilde{\mathbf{y}}\end{bmatrix},\ \mu_m^{\mathbf{Z}} = \begin{bmatrix}\mu_m^{\mathbf{X}}\\ \mu_m^{\tilde{\mathbf{Y}}}\end{bmatrix},\ \Sigma_m^{\mathbf{Z}} = \begin{bmatrix}\Sigma_m^{\mathbf{XX}} & \Sigma_m^{\mathbf{X}\tilde{\mathbf{Y}}}\\ \Sigma_m^{\tilde{\mathbf{Y}}\mathbf{X}} & \Sigma_m^{\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}}\end{bmatrix} \tag{2}$$

where Θ is the parameter set of the GMM[2], $N(.,\mu,\Sigma)$ 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 (prior probability). Given a training dataset of input and output observations, the Maximum Likelihood estimation (ML-estimation) of the GMM parameters Θ_{ML} is determined using the expectation-maximization algorithm (EM), such as: $\Theta_{ML} = \arg\max_{\Theta} p(\mathbf{x},\tilde{\mathbf{y}}\,|\,\Theta)$. A graphical representation of the GMM considered during training is represented in Figure 2 (left).

---

[2] $p(\mathbf{z}\,|\,\Theta) = p(\mathbf{x},\tilde{\mathbf{y}}\,|\,\Theta)$ is an abuse of notation meaning $p(\mathbf{Z}=\mathbf{z}\,|\,\Theta) = p(\mathbf{X}=\mathbf{x},\tilde{\mathbf{Y}}=\tilde{\mathbf{y}}\,|\,\Theta)$.

In the mapping stage, a conditional pdf $p(\tilde{\mathbf{y}}_t \mid \mathbf{x}_t, \Theta)$ is calculated for each $\mathbf{x}_t$ vector, from the joint pdf $p(\mathbf{x}, \tilde{\mathbf{y}} \mid \Theta)$ estimated during training, such as:

$$p(\tilde{\mathbf{y}}_t \mid \mathbf{x}_t, \Theta) = \sum_{m=1}^{M} P(m \mid \mathbf{x}_t) p(\tilde{\mathbf{y}}_t \mid \mathbf{x}_t, m, \Theta) \tag{3}$$

where $P(m \mid \mathbf{x}_t)$ is a posterior probability which can be seen as the *responsibility* that the m^{th} mixture component takes for 'explaining' the input feature vector $\mathbf{x}_t$, defined such as:

$$P(m \mid \mathbf{x}_t) = \frac{\alpha_m N(\mathbf{x}_t, \mu_m^{\mathbf{X}}, \Sigma_m^{\mathbf{XX}})}{\sum_{l=1}^{M} \alpha_l N(\mathbf{x}_t, \mu_l^{\mathbf{X}}, \Sigma_l^{\mathbf{XX}})} \tag{4}$$

and $p(\tilde{\mathbf{y}}_t \mid \mathbf{x}_t, m, \Theta)$ is the conditional probability of $\tilde{\mathbf{y}}_t$ given both $\mathbf{x}_t$ and the m^{th} mixture component, which is also a Gaussian and is defined as:

$$\begin{aligned} &p(\tilde{\mathbf{y}}_t \mid \mathbf{x}_t, m, \Theta) = N(\tilde{\mathbf{y}}_t, E_m^{\tilde{\mathbf{Y}}}(\mathbf{x}_t), D_m^{\tilde{\mathbf{Y}}}) \\ &\text{with } \begin{cases} E_m^{\tilde{\mathbf{Y}}}(\mathbf{x}_t) = \mu_m^{\tilde{\mathbf{Y}}} + \Sigma_m^{\tilde{\mathbf{Y}}\mathbf{X}} \Sigma_m^{\mathbf{XX}^{-1}} (\mathbf{x}_t - \mu_m^{\mathbf{X}}) \\ D_m^{\tilde{\mathbf{Y}}} = \Sigma_m^{\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}} - \Sigma_m^{\tilde{\mathbf{Y}}\mathbf{X}} \Sigma_m^{\mathbf{XX}^{-1}} \Sigma_m^{\mathbf{X}\tilde{\mathbf{Y}}} \end{cases} \end{aligned} \tag{5}$$

In conventional GMM-based mapping, the sequence of output vector is estimated frame-by-frame, such as:

$$\hat{\mathbf{y}}_t = E[\mathbf{y}_t \mid \mathbf{x}_t] = \sum_{m=1}^{M} P(m \mid \mathbf{x}_t) E_m^{\mathbf{Y}}(\mathbf{x}_t) \tag{6}$$

As mentioned in (Toda et al., 2007), the main drawback of this approach is that the estimated trajectory can present some abnormal discontinuities. These discontinuities can be due to instable posterior probabilities $P(m \mid \mathbf{x}_t)$ (responsibilities) when $\mathbf{x}_t$ becomes equidistant to the centroid of several mixture components. The GMM+dyn approach aims at addressing this issue.

In the GMM+dyn framework, the sequence of output feature vectors is not estimated frame-by-frame, but rather in one single operation from the entire sequence of input feature vectors $[\mathbf{x}_1,...,\mathbf{x}_t,...,\mathbf{x}_T]$. This sequence is written as a D_yT-dimensional column vector $\mathbf{x}_{seq}$ such as $\mathbf{x}_{seq} = [\mathbf{x}_1^\dagger,...,\mathbf{x}_T^\dagger]^\dagger$. Similarly, the sequences of output feature vectors $[\tilde{\mathbf{y}}_1,...,\tilde{\mathbf{y}}_T]$ (static and dynamic features) and estimated feature vectors $[\hat{\mathbf{y}}_1,...,\hat{\mathbf{y}}_T]$ (static features only) are also written as column vectors such as $\tilde{\mathbf{y}}_{seq} = [\tilde{\mathbf{y}}_1^\dagger,...,\tilde{\mathbf{y}}_T^\dagger]^\dagger$ ($2D_yT$-dimensional vector) and $\hat{\mathbf{y}}_{seq} = [\hat{\mathbf{y}}_1^\dagger,...,\hat{\mathbf{y}}_T^\dagger]^\dagger$ (D_yT-dimensional vector). In this framework, $\hat{\mathbf{y}}_{seq}$ is estimated by maximizing the conditional pdf $p(\tilde{\mathbf{y}}_{seq} \mid \mathbf{x}_{seq}, \Theta)$ with respect to static features $\mathbf{y}_{seq} = [\mathbf{y}_1^\dagger,...,\mathbf{y}_T^\dagger]^\dagger$ such as $\hat{\mathbf{y}}_{seq} = \arg\max_{\mathbf{y}_{seq}} p(\tilde{\mathbf{y}}_{seq} \mid \mathbf{x}_{seq}, \Theta)$. By searching for a consistent relationship between static and dynamic features, this approach guarantees the smoothness of the estimated trajectory. This approach can be seen as an adaptation of the 'maximum likelihood parameter generation' algorithm (MLPG) proposed in (Tokuda et al., 2000) for HMM-based synthesis, to the GMM-based mapping. Similarly to the MLPG algorithm, $\hat{\mathbf{y}}_{seq}$ can be estimated by solving a closed-form equation (Toda et al., 2007), given by:

$$\hat{\mathbf{y}}_{seq} = (W^T D^{-1} W)^{-1} W^T D^{-1} E \tag{7}$$

where $E = [E_{\mathbf{x}_1}^{\tilde{\mathbf{Y}}\dagger},...,E_{\mathbf{x}_t}^{\tilde{\mathbf{Y}}\dagger},...,E_{\mathbf{x}_T}^{\tilde{\mathbf{Y}}\dagger}]^\dagger$ is a $2D_yT$-dimensional column vector and $D = diag[D_{\mathbf{x}_1}^{\tilde{\mathbf{Y}}},...,D_{\mathbf{x}_t}^{\tilde{\mathbf{Y}}},...,D_{\mathbf{x}_T}^{\tilde{\mathbf{Y}}}]$ a $2D_yT \times 2D_yT$ matrix with:

$$E_{\mathbf{x}_t}^{\tilde{\mathbf{Y}}} = \sum_{m=1}^{M} P(m \mid \mathbf{x}_t) E_m^{\tilde{\mathbf{Y}}}(\mathbf{x}_t) \text{ and } D_{\mathbf{x}_t}^{\tilde{\mathbf{Y}}} = \sum_{m=1}^{M} P(m \mid \mathbf{x}_t) D_m^{\tilde{\mathbf{Y}}} \tag{8}$$

In our implementation, and similarly to (Toda et al., 2008), the calculation of E and D was simplified by keeping only, for each input feature vector, the mixture component with the

highest responsibility. The column vector E and the matrix D of Equation 7 become respectively $\hat{E} = [E^{\tilde{\mathbf{Y}}}_{\hat{m}_1}(\mathbf{x}_1)^\dagger, ..., E^{\tilde{\mathbf{Y}}}_{\hat{m}_t}(\mathbf{x}_t)^\dagger, ..., E^{\tilde{\mathbf{Y}}}_{\hat{m}_T}(\mathbf{x}_T)^\dagger]^\dagger$ and $\hat{D} = diag[D^{\tilde{\mathbf{Y}}}_{\hat{m}_1}, ..., D^{\tilde{\mathbf{Y}}}_{\hat{m}_t}, ..., D^{\tilde{\mathbf{Y}}}_{\hat{m}_T}]$ with $\hat{\mathbf{m}}$ a suboptimal sequence of mixture components defined as $\hat{\mathbf{m}} = [\hat{m}_1, ..., \hat{m}_t, ..., \hat{m}_T]$ with $\hat{m}_t = \arg\max_m \{P(m \mid \mathbf{x}_t)\}$. Contrary to the conventional GMM-based mapping (Equation 6), the GMM+dyn technique uses both the expectation $E^{\tilde{\mathbf{Y}}}_{\mathbf{x}_t}$ and the variance $D^{\tilde{\mathbf{Y}}}_{\mathbf{x}_t}$ of the conditional probability $p(\tilde{\mathbf{y}}_t \mid \mathbf{x}_t)$, for all input frames. Since D is a block-diagonal matrix, D^{-1} is also block-diagonal and generally full-rank. Thus, via the product $D^{-1}E$ of Equation 7, all input vectors of sequence $[\mathbf{x}_1, ..., \mathbf{x}_t, ..., \mathbf{x}_T]$ contribute to the estimation of each output vector $\hat{\mathbf{y}}_t$. As a consequence, contextual information is naturally taken into account in the mapping. In our application, such information can potentially be helpful to disambiguate partially observed articulatory gestures. A graphical representation of this conversion method is represented in Figure 2 (right).

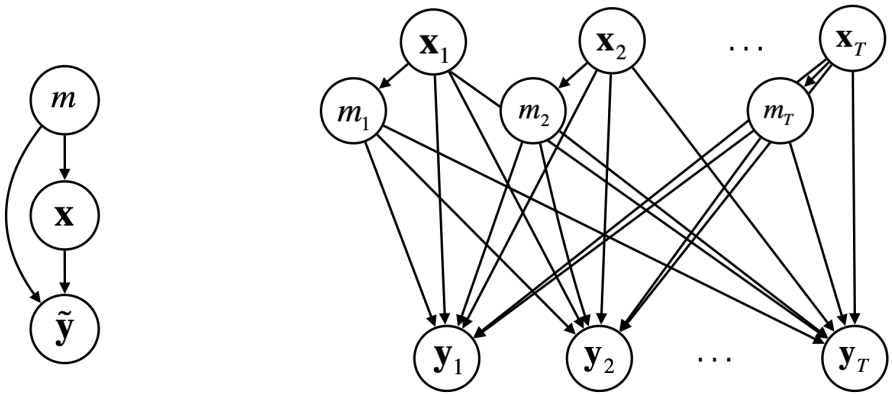

Figure 2 left: Directed graphical model associated with the GMM considered during training ($\mathbf{x}$ and $\tilde{\mathbf{y}}$ are observed variables, m is a latent variable) / right: Graphical representation of the GMM-based mapping process considering dynamic features (GMM+dyn).

As shown in Figure 2, the graphical model used for training (left) is different from the one used for the mapping (right). (Zen et al., 2011) addressed this inconsistency issue and

proposed a training algorithm that explicitly takes into account the relationships between static and dynamic features. This approach, called trajectory GMM, was not used in this study but remains an interesting perspective.

2.2. HMM-based mapping considering dynamic features (HMM+dyn)

The following section describes the theoretical aspects of the proposed HMM-based mapping technique. This technique, which is referred to as the HMM+dyn technique, can be seen as a straightforward adaptation of the GMM+dyn technique to the framework of Hidden Markov Models. In the training stage, parallel sequences of articulatory and acoustic observations are modeled by a *full-covariance HMM*, i.e. a HMM for which state emission probability is modeled by a multivariate Gaussian distribution with a full-covariance matrix. The optimal parameter set Θ_{ML} is estimated similarly to a standard HMM, by maximizing the likelihood of the joint pdf such as $\Theta_{ML} = \arg\max_{\Theta} p(\mathbf{x},\tilde{\mathbf{y}} \mid \Theta)$ with:

$$p(\mathbf{x},\tilde{\mathbf{y}} \mid \Theta) = p(\mathbf{z} \mid \Theta) = \sum_{\mathbf{q}} P(\mathbf{q} \mid \Theta) p(\mathbf{z} \mid \mathbf{q}, \Theta) \tag{9}$$

and

$$P(\mathbf{q} \mid \Theta) = \pi_{q_1} \prod_{t=2}^{T} a_{q_{t-1}q_t} \text{ and } p(\mathbf{z} \mid q_t, \Theta) = N(.,\mu_{q_t}^{\mathbf{Z}}, \Sigma_{q_t}^{\mathbf{Z}}) \tag{10}$$

where $\mathbf{q} = [q_1,...,q_t,...,q_T]$ is a sequence of T states, π_{q_1} are the initial state probabilities, $a_{q_{t-1}q_t}$ are the state transition probabilities, and $\mu_{q_t}^{\mathbf{Z}}$ / $\Sigma_{q_t}^{\mathbf{Z}}$ are the mean/covariance matrix of emission probability associated with the HMM state q_t (with $\mathbf{z}$ defined as in Equation 2). These parameters are determined from the training dataset using the Baum-Welch algorithm. The directed graphical model associated with the full-covariance HMM considered during training is represented in Figure 3 (left).

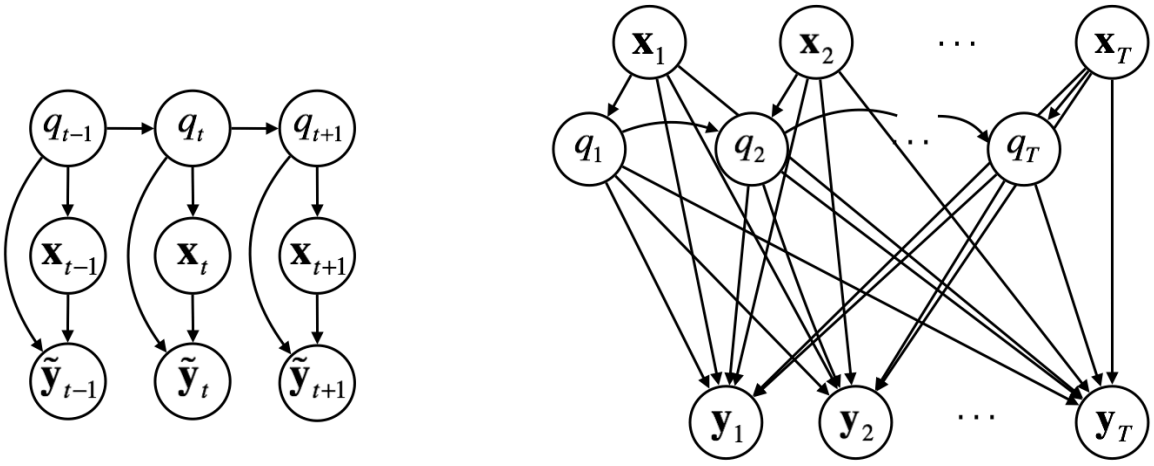

Figure 3: left: Directed graphical model associated with the full-covariance HMM considered during training ($\mathbf{x}$ and $\tilde{\mathbf{y}}$ are observed variables, q_t is a latent variable) / right: Graphical representation of the HMM-based mapping process considering dynamic features (HMM+dyn).

As shown in Figure 3 (left), this model can be seen as a GMM for which only some transitions between mixture components are possible and can occur with a certain (transition) probability. Therefore, this approach aims at modeling more explicitly the time organization of articulatory gestures, compared to a GMM. In this approach, the HMM state duration model is 'implicit'; it is given by a geometric distribution, as shown by Equation 10. However, as discussed in (Ostendorf et al., 1996) in the context of automatic speech recognition, this duration modeling might not be optimal. The use of an explicit duration model, in the framework of Hidden Semi-Markov Models (Russell and Moore, 1985), remains an interesting perspective that should be addressed in future work.

Contrary to our previous work, the local relationships between articulatory and acoustic observations are here explicitly modeled, through the full-covariance matrices of HMM emission probabilities. For a given state q_t, this local relationship is given by the conditional expectation $E(\tilde{\mathbf{y}}_t \mid \mathbf{x}_t, q_t, \Theta)$ given by Equation 5, which can be seen as linear regression function such as:

$$E(\tilde{\mathbf{y}}_t \mid \mathbf{x}_t, q_t, \Theta) = A_{q_t}\mathbf{x}_t + b_{q_t} \text{ with } A_{q_t} = \Sigma_{q_t}^{\tilde{\mathbf{Y}}\mathbf{X}}\Sigma_{q_t}^{\mathbf{X}\mathbf{X}^{-1}} \text{ and } b_q = \mu_q^{\tilde{\mathbf{Y}}} - A_q\mu_q^{\mathbf{X}} \quad (11)$$

In this study, we investigated a conversion technique based on full-covariance HMM that explicitly considers the dynamic features, similarly to the MLPG algorithm (Tokuda et al., 2000) and the GMM+dyn technique. The sequence of target feature vector $\hat{\mathbf{y}}$ is estimated from the input sequence $[\mathbf{x}_1,...,\mathbf{x}_t,...,\mathbf{x}_T]$ in a single mapping process (and not frame-by-frame). As for the GMM+dyn case, this sequence is defined as the one that maximizes the conditional pdf $p(\tilde{\mathbf{y}} \mid \mathbf{x}, \Theta)$ with respect to the static features. In the context of HMM, this conditional pdf can be expressed as a function of the state sequence $\mathbf{q}$, such as:

$$p(\tilde{\mathbf{y}} \mid \mathbf{x}, \Theta) = \sum_{\forall \mathbf{q}} p(\tilde{\mathbf{y}} \mid \mathbf{x}, \mathbf{q}, \Theta) P(\mathbf{q} \mid \mathbf{x}, \Theta) \quad (12)$$

Similarly to the GMM+dyn technique, this conditional pdf can be approximated such as $p(\tilde{\mathbf{y}} \mid \mathbf{x}, \Theta) \sim p(\tilde{\mathbf{y}} \mid \mathbf{x}, \hat{\mathbf{q}}, \Theta)$, where $\hat{\mathbf{q}}$ is a suboptimal state sequence defined as $\hat{\mathbf{q}} = \arg\max_{\mathbf{q}}\{P(\mathbf{q} \mid \mathbf{x}, \Theta)\}$. In the HMM framework, this state sequence can be determined from the input sequence $[\mathbf{x}_1,...,\mathbf{x}_t,...,\mathbf{x}_T]$ using the Viterbi algorithm. The generation of the output sequence $\hat{\mathbf{y}}_{seq}$ is then similar to the GMM+dyn approach. First, a conditional pdf $p(\tilde{\mathbf{y}}_t \mid \hat{q}_t, \mathbf{x}_t, \Theta)$ is estimated for each vector $\mathbf{x}_t$ and decoded state $\hat{q}_t$ using Equation 5. Output sequence $\hat{\mathbf{y}}_{seq}$ is then computed by solving Equation 7 with $\hat{E} = [E_{\hat{q}_1}^{\tilde{\mathbf{Y}}}(\mathbf{x}_1),...,E_{\hat{q}_t}^{\tilde{\mathbf{Y}}}(\mathbf{x}_t),...,E_{\hat{q}_T}^{\tilde{\mathbf{Y}}}(\mathbf{x}_T)]^{\dagger}$ and $\hat{D} = diag[D_{\hat{q}_1}^{\tilde{\mathbf{Y}}},...,D_{\hat{q}_t}^{\tilde{\mathbf{Y}}},...,D_{\hat{q}_T}^{\tilde{\mathbf{Y}}}]$ (i.e. the HMM state sequence $\hat{\mathbf{q}}$ stands for the suboptimal sequence of GMM component $\hat{\mathbf{m}}$). A graphical representation of this conversion process is presented in Figure 3 (right).

In line with our previous work, we trained a set of *phone HMMs*, i.e. one HMM for each phonetic class. Similarly to HMM-based recognition or synthesis systems, phone HMMs can be concatenated together to build models of higher linguistic level (syllable, word,

sentence, etc.). The goal of using phone HMMs was twofold. First, the phonetic segmentation of the data is used as prior knowledge in the training stage. Second, external linguistic knowledge can be introduced in the mapping for regularization purpose, similarly to our previous work (Hueber et al., 2010b). This linguistic knowledge $P(\mathbf{q},\Theta)$ can be introduced during the intermediate state-decoding step, which can be re-written as: $P(\mathbf{q}\,|\,\mathbf{x},\Theta) = P(\mathbf{q},\Theta)p(\mathbf{x}\,|\,\mathbf{q},\Theta) / p(\mathbf{x},\Theta)$. It can consist of a set of phonotactic rules, or a statistical language model at phonetic or syllabic level. Later, this step is referred to as the *phonetic decoding step*.

2.3. Relation to previous work on HMM-based mapping

Similar feature mapping algorithms based on HMM have been proposed in the literature. The following section describes the most relevant examples and discusses the differences with our approach. In (Chen, 2001), HMM was used to drive a lip model from speech acoustics for lip synchronization purpose. In that study, the training stage was identical to the one used in the present study. However, the conversion stage was different since it was done frame-by-frame (i.e. the estimated feature vector was defined as $\hat{\mathbf{y}}_t = E^{\mathbf{Y}}_{q_t}(\mathbf{x}_t)$) without considering dynamic features. In (Hiroya and Honda, 2004), an HMM-based mapping algorithm considering dynamic features was proposed in the context of acoustic-to-articulatory inversion. A similar approach was used in (Ling et al., 2009) in the context of speech synthesis driven both by text and articulatory inputs, and in (Ling et al., 2010), for the estimation of articulatory movement from text (eventually completed with acoustic inputs). To our best understanding, our approach differs from the one described in (Hiroya and Honda, 2004) in two aspects. First, in the training stage of (Hiroya and Honda, 2004), only the input stream was considered when estimating HMM state responsibilities $P(q_t\,|\,\mathbf{x}_t,\Theta)$, whereas both input and output stream were considered in our approach (as shown in Equation 9).

Second, the parameters of the state-dependent linear regression functions A_{q_t} and b_{q_t} (Equation 11) were estimated iteratively during the training stage. In our approach, they are deduced at conversion time from the conditional expectation associated with the decoded state $\hat{q}_t$ as explained previously, and similarly to the GMM-based mapping technique.

3. Data acquisition and parameterization

Articulatory-to-acoustic mapping experiments were conducted to evaluate GMM+dyn and HMM+dyn techniques, in the context of a silent speech interface driven by ultrasound and video imaging. Both techniques were evaluated on a continuous speech database recorded specifically for this study, with two French speakers (one male, one female).

3.1 Experimental setup

Ultrasound scans were acquired using the portable Terason T3000 system, with a 128 elements microconvex transducer. For both speakers, ultrasound frequency range was set to 3-5 MHz, scanning angle to 140°, and penetration depth to 7 cm. Video images of the speaker's face were recorded using the industrial 1/3" CMOS color camera (Imaging Source DFM 22BUC03-ML). Exposure time was set constant to 1/128 second (automatic gain and white-balance correction were disabled). Ultrasound and video sensors were attached to the speaker's head using a slightly modified version of the probe stabilization helmet designed by (Wrench et al., 2007). The experimental setup used for data acquisition is shown in Figure 4.

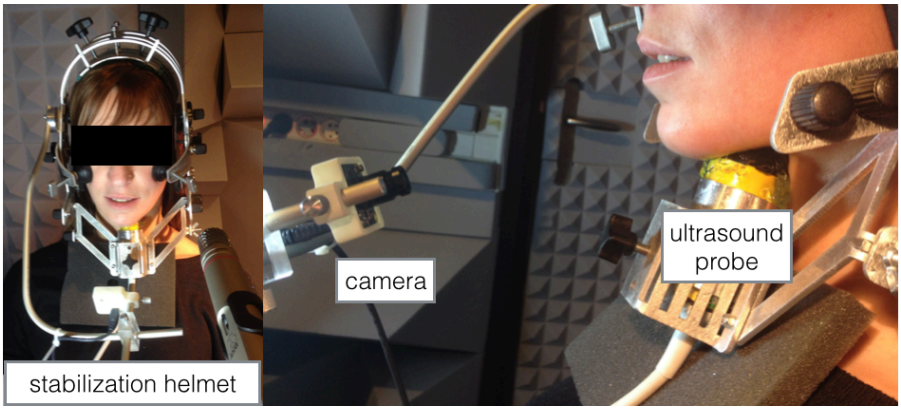

Figure 4: Experimental setup used for data acquisition (speaker B)

In our previous studies (Hueber et al., 2011, 2010b), both ultrasound and video data were acquired at 29.97 fps (NTSC format), using the video analog output of ultrasound system, and a consumer video camera. However, (Wrench and Scobbie, 2006) showed that this digital-to-analog conversion of ultrasound scans (also called rasterization) could cause inaccurate measurements of tongue location (with an error up to 10mm for tongue tip and 7mm for tongue body). In order to avoid this conversion, we developed a software named *Ultraspeech*[3] able to record high quality ultrasound scans at their maximum spatial and temporal resolution (with no distortion) by accessing directly to the internal buffer of the ultrasound system (cineloop). More details about this software can be found in (Hueber et al., 2008). In this study, ultrasound and video images were both recorded at 60 fps, which was twice as high as in our previous work. Audio was recorded synchronously at 32 bits, 44.1 kHz. Typical examples of the recorded ultrasound and video images are given in Figure 5.

[3] The software is free to download at http://www.ultraspeech.com

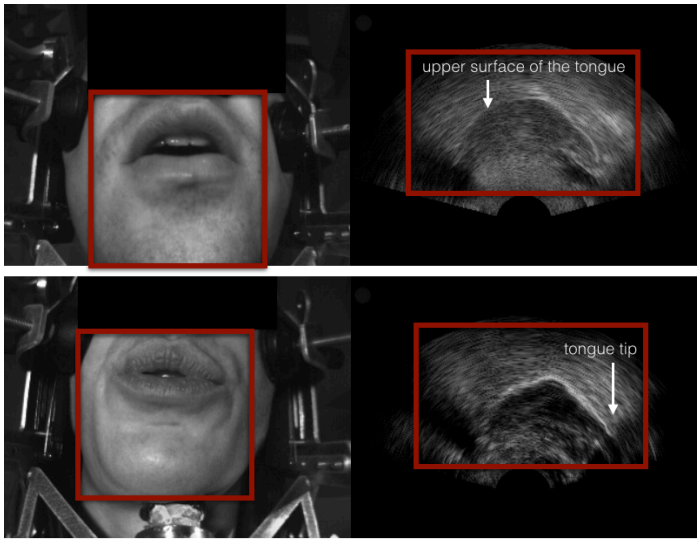

Figure 5: Typical examples of recorded ultrasound and video images (320x240 and 640x480 pixels respectively) for male speaker A (up) and female speaker B (bottom). Red bounding boxes delimit the regions of interest considered for feature extraction.

3.2 Recorded database

The recorded corpus was divided into five parts:

- P1: a set of 288 'phonetically-dense' sentences selected from a large text corpus and recorded in modal speech. These sentences were selected using the following procedure. First, a set of 50,000 sentences was extracted from the text corpus of "Le Monde" (2003 edition) distributed by ELRA[4]. These sentences have a simple syntactic structure (e.g. only one verb), and contains from 5 to 12 syllables. Second, a final set of 288 sentences was extracted from this initial set using a greedy algorithm (Buchsbaum and van Santen, 1996). The sentences were selected in order to have at least one occurrence of each French diphone in the final set.

[4] http://catalog.elra.info/product_info.php?products_id=438

- P2: The first 100 ‘phonetically-balanced’ sentences of (Combescure, 1981), recorded in modal speech[5].
- P3: The first 30 sentences of P2, but recorded in silent speech (i.e. articulated normally but without vocalization).
- P4: 100 carrier sentences “Tu t’appelles VCV, c’est ça ?” (“Your name is VCV, that’s right ?”), with vowel V selected from the set {a,i,u,y,e,ø,o,ɑ̃,ɛ̃,ɔ̃} and consonant C from {p,t,k,f,s,ʃ,m,n,ʁ,l}. Carrier sentences were preferred to isolated VCV in order to evaluate the segmental intelligibility of the system on continuous speech while controlling the phonetic context. VCV were placed in the middle of the carrier sentence in order to avoid vowel reduction typically observed when it is placed at the end of a sentence[6].
- P5: same as P4 but recorded in silent speech[7].

In order to measure how “silent” was the speech recorded in P3 and P5, the absolute sound pressure level measurement (SPL) was measured using a calibrated Brüel & Kjær microphone placed one meter away from the speaker’s face and a conditioning amplifier (Nexus B&K). The mean SPL were of 44.0 and 62.0 dBSPL for respectively silent, and normal speech while the mean SPL for the ambient noise of the anechoic room was of 43.9 dBSPL.

---

[5] A video example of one recorded sentence is provided as electronic supplementary material (SpeakerA_list1_sent302.mov).

[6] A video example of one recorded sentence for VCV [asa] is provided as electronic supplementary material (SpeakerA_asa_carrier_modal.mov).

[7] A video example of one of the P5 sentences for VCV [asa] is provided as electronic supplementary material (SpeakerA_asa_carrier_silent.mov).

All sentences recorded in vocalized speech were phonetically transcribed, using the following phone set for French {a,i,u,y,e,ɛ,ə,ø,œ,o,ɔ,ɑ̃,ɛ̃,ɔ̃,p,t,k,b,d,g,f,s,ʃ,v,z,ʒ,m,n,ʁ,l,ɲ,j,w} (33 phonetic classes). Phonetic transcriptions were then corrected manually in order to match the pronunciation differences of the two speakers. Finally, they were time-aligned with the audio signal, using an HMM-based speech recognizer and a forced-alignment procedure.

3.3 Extraction of articulatory and acoustic features

To decrease the effect of speckle, ultrasound images were filtered using the anisotropic diffusion filter proposed by (Yu and Acton, 2002). This iterative filter introduces intra-region smoothing while inhibiting inter-region smoothing, using the local coefficient of variation as edge detector. Speckle was thus reduced while important image features (such as tongue contour) were preserved. Similarly to our previous studies, the EigenTongues decomposition (Hueber et al., 2007) was then used to parameterize each ultrasound image. The technique is a straight-forward adaptation of the Eigenfaces method proposed by (Turk and Pentland, 1991) in the context of automatic face recognition and can be summarized as follows. In the training stage, a subset of ultrasound frames was selected from the recorded dataset. In order to maximize the phonetic coverage of the training set, we kept approximately 60 frames for each of the 33 phonetic classes used to describe French (~2000 frames). A region of interest (ROI) shown in Figure 5 was first determined manually. It was delimited by the highest point of the ultrasound probe (bottom), the maximum penetration depth (up), the acoustic shadow of the hyoid bone (left), and the acoustic shadow of the mandible (right). Since the helmet maintains the ultrasound probe and the camera fixed relatively to the skull, the same ROI was used for all the frames of the recorded database.

The region of interest (ROI) shown in Figure 5 was then resized to 32x32 pixels. A decomposition basis that best explains the variation of pixel intensity in the training frames

was then extracted using a Principal Component Analysis (PCA). Basis vectors for ultrasound are called EigenTongues. In the feature extraction stage, the resized ROI of each new ultrasound frame was projected onto the set of EigenTongues. Articulatory features used for the mapping experiments were defined as the first N coordinates in that space. The number N was determined by keeping the eigenvectors that carry 80% of the variance. For both speakers, 30 coefficients were used as static features for ultrasound. A similar approach (EigenLips decomposition) was adopted to parameterize video images (ROI contains the lips and the bottom part of the face); the optimal number of projections was found to be 25.

The spectral content of the audio speech signal was parameterized by 12 mel-cepstrum coefficients (Blackman window, 25ms frame length, 5ms frame shift), using the SPTK toolkit[8]. Articulatory feature trajectories were oversampled from 60Hz to 200Hz, in order to be compatible with the speech analysis rate.

4. Articulatory-to-acoustic mapping experiments

4.1 Articulatory-to-acoustic mapping experiments on modal speech

4.1.1 Practical implementation and experimental protocol

The performance of GMM+dyn and HMM+dyn mapping techniques was first evaluated on modal speech, using a 8-fold cross-validation procedure. The partitioning of the recorded database was the following. The 488 sentences recorded in modal speech (which correspond to the parts P1, P2 and P4, described in Section 3.2) were divided into 8 subsets of 61 sentences. For each of the 8 repetitions of the cross-validation, one subset was used for test, while the 7 other subsets were used for training (427 sentences). Among these training sentences, 400 of them were actually used for the estimation of the GMM and HMMs

[8] SPTK toolkit: http://sp-tk.sourceforge.net/

parameters, and 27 sentences were used to adjust: a) the model insertion penalty of the phonetic decoding step for HMM+dyn and b) the optimal number of components of the model for GMM+dyn. That number was found equal to 100 for both speakers (16, 32, 64, 100, 128 components were tested; K-means algorithm was used for initialization). No improvement was observed with 128 components. With more components (>128), the covariance matrices of the GMM were badly conditioned, probably because of the size of the training set. Interestingly, the optimal number of GMM components corresponded roughly to the total number of states for HMM+dyn experiments (33 models x 3 states = 99).

The dynamic component of articulatory features (i.e. $\Delta\mathbf{x}_t$) was not considered when training GMM and full-covariance HMM, in order to avoid conditioning problems of covariance matrices. We recall also that the dimensionality of the articulatory observations D_x considered in this study is equal to 55, which is much higher than the dimensionality of EMA feature vectors considered in (Toda et al., 2008) (which was 14). For the GMM, no significant improvements were observed when using $\Delta\mathbf{x}_t$, for a number of components inferior to 64. With more components (and *a fortiori* for 100 components which was found to be the optimal number of components), conditioning problems of covariance matrices were observed when using $\Delta\mathbf{x}_t$. This can be explained by a too large number of parameter to estimate (equals to $M(1+(D_\mathbf{x}+2D_\mathbf{y})+(D_\mathbf{x}+2D_\mathbf{y})^2/2)$ with M the number of mixture components) which was equals to 540,850 when considering only $\mathbf{x}_t$ and increased up to 1,264,100 when considering both $\mathbf{x}_t$ and $\Delta\mathbf{x}_t$ (while the size of the training material remained fixed to ~ 230,000 observations). The same conditioning problems were observed also with full-covariance HMM when considering both $\mathbf{x}_t$ and $\Delta\mathbf{x}_t$. However, $\Delta\mathbf{x}_t$ was used for the intermediate phonetic decoding stage of HMM+dyn experiments. Its impact on the final performance is discussed in Section 4.1.2.

For the GMM+dyn experiments, silent frames at beginning and end of each sentence were removed from the training set. For the HMM+dyn experiments, they were kept and used to train a so-called *silence model*.

HMMs were trained as follows. First, sequences of articulatory and acoustic feature vectors were modeled for each phonetic class by a single stream HMM. At this stage, the covariance matrix of each state emission pdf was forced to be diagonal. HMM topology was left-to-right with 3 emitting states and a possible transition between state 1 and state 3. HMMs were first trained separately, using the Baum-Welch algorithm and then processed simultaneously, using an embedded training strategy. Two other model sets were then initialized. The first one consisted of context-dependent models (triphone) 2-stream HMMs, trained only on articulatory features (including their dynamic component $\Delta\mathbf{x}_t$). For this model, the emission pdf associated with state q is expressed as:

$b_q(\mathbf{x}_t) = \prod_{S\in\{US,VIDEO\}} N(\mathbf{x}_{tS}, \mu^{\mathbf{x}}_{qS}, \Sigma^{\mathbf{x}}_{qS})^{\gamma_s}$ where $\mathbf{x}_{tS}$, $\mu^{\mathbf{x}}_{qS}$ and $\Sigma^{\mathbf{x}}_{qS}$ are extracted from $\mathbf{x}_t$, $\mu^{\mathbf{x}}_q$ and $\Sigma^{\mathbf{x}}_q$ by keeping only the feature related to the S stream (either ultrasound or video), and γ_s are stream-specific weighting parameters. In this study, these parameters were determined on a validation set. For each weighting parameter, we evaluated the following values {0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8} with the constraint $\gamma_{US} + \gamma_{VIDEO} = 1$. The optimal values were $\gamma_{US} = 0.7$ and $\gamma_{VIDEO} = 0.3$, which confirm that the tongue carries the most important part of the accessible articulatory information. Interestingly, the same optimal weights were found in (Hueber et al., 2009) for English language. A tree-based state-tying strategy was used to cope with data sparsity. The second set of HMMs consisted in context-independent (monophone) HMMs, trained on joint articulatory features (without their dynamic component, i.e. $\mathbf{x}_t$ but not $\Delta\mathbf{x}_t$) and acoustic features (with both their static and dynamic components, i.e. $\tilde{\mathbf{y}}_t$).

Covariance matrix of state emission pdf was here forced to be full in order to explicitly model local acoustic-articulatory relationships.

The first model set (context-dependent HMMs with diagonal covariance matrices) was used to determine the target phonetic sequence from the input articulatory observations $[\mathbf{x}_1,...,\mathbf{x}_t,...,\mathbf{x}_T]$ (phonetic decoding step). The target state sequence $\hat{\mathbf{q}}$ was then obtained by force-aligning the decoded phonetic sequence on the input articulatory observations using the second model set (i.e. context-independent full-covariance HMMs). These models were further used to effectively generate the output spectral features $\hat{\mathbf{y}}$ given both $\hat{\mathbf{q}}$ and $[\mathbf{x}_1,...,\mathbf{x}_t,...,\mathbf{x}_T]$ (mapping stage). (Ling et al., 2010) proposed an iterative procedure to refine the state sequence $\hat{\mathbf{q}}$ by re-decoding the sequence of joint observations $[\mathbf{x}_t^\dagger, \hat{\mathbf{y}}_t^\dagger]^\dagger$ (for $t = 0,...,T$) (i.e. original input / estimated output observation) using the Viterbi algorithm. This refinement was experimented in our context (i.e. articulatory-to-acoustic mapping) but it did not bring any improvement. Therefore, we kept the initial state sequence $\hat{\mathbf{q}}$ for the generation of $\hat{\mathbf{y}}$.

The use of two different model sets for phonetic decoding and parameter generation can be explained by the impossibility to train context-dependent full-covariance HMMs on this database, due to the lack of training data (even a tree-based state-tying strategy did not make the training feasible). The training of HMMs and the phonetic decoding stage were done with the HTK toolkit (Young, 2005). GMM training and GMM+dyn/HMM+dyn mapping were implemented in MATLAB.

For the intermediate phonetic decoding step of HMM+dyn, we investigated the use of a statistical language model at phonetic level. In this study, we trained a simple phonetic bigram on the set of 50,000 sentences extracted from *Le Monde* French newspaper (detailed in Section 3.2), using CMU SLM toolkit (Clarkson and Rosenfeld, 1997).

4.1.2 Objective evaluation

For the experiment based on HMM+dyn, the performance of the intermediate phonetic decoding stage was first measured by evaluating the *phone accuracy* ($PAcc$) defined as $PAcc = 100 \cdot (N_p - D_p - S_p - I_p) / N_p$, where N_p was the number of phones in the test set, S_p the number of substitution errors, D_p deletion errors, and I_p insertion errors. The 95% confidence half-interval of phone accuracy measurement Δ was defined as the Wilson Score such as $\Delta = t_\alpha \sqrt{\left(PAcc(1-PAcc)/N_p\right) + \left(t_\alpha^2/(4N_p^2)\right)} \Big/ (1 + t_\alpha^2/N_p)$ (with $t_\alpha = 1.95$ and a Normal assumption). Since we focused in this study only on the estimation of the spectral content of the reconstructed speech, confusions between consonants that differ only in the voicing feature {[p]-[b]}, {[t]-[d]}, {[s]-[z]}, {[f]-[v]} and {[ʃ]-[ʒ]} were not counted as errors. Table 1 details the performance of the intermediate phonetic decoding step of HMM+dyn mapping technique, for both speakers, with no linguistic priors (a) and when using a phonetic bigram for the intermediate decoding step (b).

Table 1: Performance of the intermediate phonetic decoding step of HMM+dyn. N_p , S_p , D_p and I_p are respectively the number of phones, substitution errors, deletion errors, and insertion errors in the test set (Δ is the 95 % confidence half-interval).

	Speaker A		Speaker B	
	No priors (a)	Phonetic bigram (b)	No priors (a)	Phonetic bigram (b)
PAcc (%)	77.6	77.9	75.2	75.4
2Δ	1.8	1.8	1.9	1.8
D_p	549	499	1172	971
S_p	976	989	931	990
I_p	768	770	412	534
N_p	10263	10263	10154	10154

With a phone accuracy superior to 75% for both speakers, the performance obtained here for French is 8% higher than in our previous work on English (Hueber et al., 2010b, 2009). This difference may come from the more accurate acquisition system used in this study. Furthermore, with much less deletion errors (but a comparable number of substitution errors), the performance is significantly higher on speaker A compared to speaker B. Speaker A seemed to speak a bit more clearly than speaker B, which may explain this difference, amongst other possible causes. Performances reported in Table 1 were obtained when considering both static and dynamic components of articulatory features (i.e. $\mathbf{x}_t$ and $\Delta\mathbf{x}_t$). When considering only the static component, the performance was 5.1% lower for speaker A and 5.9% for speaker B.

However, the use of a bigram phone model does not influence significantly the system performance. Its only benefit is to result in a more balanced repartition between insertion and deletion errors, notably for speaker B. This lack of improvement can be explained by an inconsistency between the language model and the phonetic content of the test sentences. As explained in section 3.2, test sentences extracted from part P1 have been selected from a larger text corpus (‘Le Monde’), in order to have at least one example of each diphone (with a speech synthesis application in mind). Therefore, they contains very ‘rare’ phonetic patterns; such as the consonantal cluster [z+s] of the sentence ‘Pieranunzi jaz[z s]a vie’. However, the phonetic bigram was trained on a much larger set (50,000 for the text corpus ‘Le Monde’). Therefore, it models the ‘typical’ frequency of phone sequences in French and may consider rare phonetic patterns as very ‘unlikely’. This consistency issue between a language model and an applicative domain is well known in automatic speech recognition (ASR). The same issue exists also in silent speech. It is likely that a domain-specific language model (such as a n-gram at phonetic or lexical level) will lead to a better performance. The evaluation of such

models was however out of the scope of this study which focused on the conversion of silent articulation with no prior information on the applicative domain.

In order to analyze in more details the output of the phonetic decoder (without phonetic bigram), we computed the confusion matrix for speaker A displayed in Figure 6 (most of the errors of speaker B were of the same type as those discussed for speaker A).

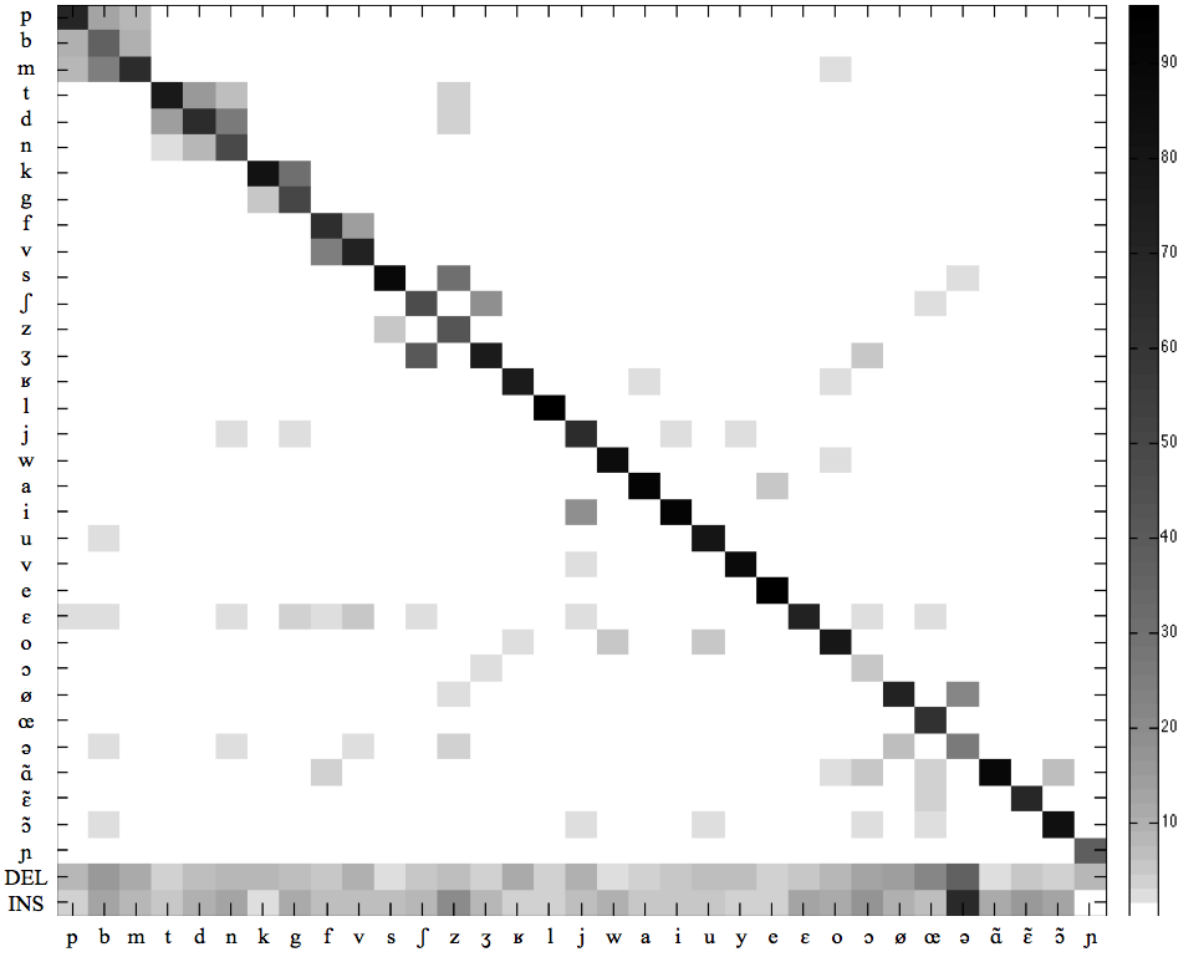

Figure 6: Confusion matrix for the intermediate phonetic decoding step of the HMM+dyn mapping experiment for speaker A. The color space map was chosen to emphasize errors (DEL and INS stand for deletion and insertion errors).

As expected, many confusions are made between phonemes that distinguish by the voicing feature (such as {[k],[g]} or {[f],[v]}), or by the nasality feature (such as {[p],[b]} vs. [m], and {[t],[d]} vs. [n]). Similarly, some errors concern nasal vowels (that do not exist in

English), such as [ɔ] being confused with [ɔ̃] and [œ] confused with [ɛ̃], [ɑ̃], or [ɔ̃]. This could be explained by the lack of information on the position of the velum, which most of the time cannot be observed from ultrasound scans. In addition, many confusions are made on dental and alveolar sounds {[t],[d],[n],[s],[z],[ʃ],[ʒ]}. This can be explained by the lack of information on the position of the tongue tip (apex), which is sometimes hidden by the ultrasonic shadow of the mandible. Finally, some confusions made on vowels - such as [ø] being confused with the schwa [ə] - may not necessarily have a strong impact on the intelligibility of the synthesized speech.

For the mapping experiments on modal speech, the quality of the spectrum estimated from the articulatory movements using GMM+dyn and HMM+dyn techniques, was evaluated by calculating for each estimated vector $\hat{\mathbf{y}}_t$ the Mel-Cepstral distortion defined as:

$MCD_{dB}(\hat{\mathbf{y}}_t, \mathbf{y}_t) = (10/\ln 10)\sqrt{2.\sum_{d=1}^{D}(\hat{y}_{t,d} - y_{t,d})^2}$ (with $\hat{y}_{t,d}$ and $y_{t,d}$ respectively the d^{th} estimated and original mel-cepstral coefficient). The statistical significance of all the different comparisons between experimental conditions (speaker, technique) was assessed by paired t-tests. No phonetic bigram was used for the HMM+dyn experiments, in order to compare the two techniques only on their ability to model the articulatory-acoustic relationships, without the help of external linguistic information. However, in order to still evaluate the benefit of such information, we also measured the spectral distortion obtained with the HMM+dyn when the target phonetic sequence is known (and forced-aligned on the articulatory movements). This experimental condition is referred to as HMM+dyn*. Results are presented in table 2.

Table 2: Objective evaluation of the quality of the spectrum derived from articulatory movements acquired in modal speech using GMM+dyn and HMM+dyn. For HMM+dyn*, the

phonetic target was known (*PAcc* = 100%). MCD_{dB} is the mel-cepstral distortion in dB averaged over the test set, and $\sigma_{MCD_{dB}}$ its standard deviation.

	Speaker A			Speaker B		
	GMM+dyn	HMM+dyn	HMM+dyn*	GMM+dyn	HMM+dyn	HMM+dyn*
MCD_{dB}	5.56	5.21	4.57	6.57	5.74	5.69
$\sigma_{MCD_{dB}}$	2.68	3.14	3.74	2.89	3.36	3.01

The distortion range from 4.5 to 6.5dB is compatible with other studies based on EMA articulatory data (e.g. see Table 1 of (Toda et al., 2008)). This result is encouraging because ultrasound and video articulatory data can be considered as much noisier than EMA data. For both speakers, and especially for speaker B, HMM+dyn outperforms GMM+dyn (5.21dB vs. 5.56dB for speaker A, and 5.74dB vs. 6.57dB for speaker B, $p<0.005$). The distortion obtained for the HMM+dyn* experiments was 0.7dB lower than HMM+dyn for speaker A ($p<0.005$), and 0.05dB for speaker B ($p<0.05$). As expected, the accuracy of the estimated spectrum depends on the amount of linguistic knowledge (which can be adapted to the target application). Performances reported in Table 2 for HMM+dyn and HMM+dyn* were obtained when considering both static and dynamic components of articulatory features for the intermediate phonetic decoding stage (i.e. $\mathbf{x}_t$ and $\Delta\mathbf{x}_t$, while only $\mathbf{x}_t$ was considered for the conversion stage). When considering only $\mathbf{x}_t$ for the phonetic decoding stage, the distortion increases to 5.37 dB for speaker A and 6dB for speaker B, but remains significantly lower than for GMM+dyn ($p<0.05$). Similarly to our previous experiments on phonetic decoding, we observed a much lower performance for the female speaker B compared to male speaker A (~1dB, $p<0.005$). Again, this difference is difficult to interpret since it can have many causes, such as a varying speaking rate and style, a varying quality of ultrasound images (Stone, 2005), or a slight displacement of the sensors during data acquisition. A similar difference

between female and male on a articulatory-to-acoustic mapping experiment was also reported in (Toda et al., 2008).

4.1.3 Perceptual evaluation

In our previous study (Hueber et al., 2010b), perceptual evaluation of the proposed silent speech interface was conducted by considering the target phonetic sequence as known (similarly to the HMM+dyn* experimental condition detailed in Section 4.1.2). In this study, we propose the very first perceptual evaluation of the entire conversion process: from the articulatory gestures captured by ultrasound and video imaging, to audible speech, and without any limitation on the vocabulary. However, with a mel-Cepstral distortion superior to 5dB for both speakers, the synthetic speech was unlikely to be systematically intelligible. For perceptual evaluations, we considered therefore the 'best case scenario' by keeping only the speaker who gave the best results during the objective evaluation (i.e. speaker A). Two perceptual tests were conducted: a transcription test and a discrimination test.

4.1.3.1 Participants

Ten native speakers of French, with no particular expertise in speech synthesis or in phonetics, were asked to evaluate synthetic speech stimuli in an anechoic room, using the same open headphones. They were allowed to listen several times to each sample.

4.1.3.2 Transcription test

For this first transcription test, the stimuli consisted in two distinct sets of 15 sentences each, selected from part P2 of the database. In the first set, articulatory movements were converted into a target spectrum using GMM+dyn, whereas the HMM+dyn was used for the second set. Each sentence was synthesized using a MLSA filter (Imai et al., 1983) derived from the estimated spectrum and excited with white noise (the stimuli sounded like whispered speech). In this study, we did not use any post-processing technique aiming at alleviating the

commonly accepted 'over-smoothing effect' of MLPG approach, such as the global variance (GV) (Toda and Tokuda, 2007), the LSPA GV (Shannon and Byrne, 2013), the variance scaling or the histogram normalization (Silén et al., 2012). However, the use of such techniques in the context of articulatory-to-acoustic mapping remains an interesting perspective for increasing both the intelligibility and the naturalness of the synthetic speech.

Participants were asked to transcribe the resulting 30 sentences, with absolutely no prior information on their linguistic content. After the test, all the transcriptions were manually checked in order to remove misspellings. The accuracy of the transcription was then evaluated similarly as in automatic speech recognition, by calculating the word-accuracy *WAcc*, such as $WAcc = 100 \cdot (N_w - D_w - S_w - I_w) / N_w$, where N_w was the number of words in the test set, D_w the number of deletion errors, S_w the number of substitution errors, and I_w the number of insertion errors. This evaluation methodology is relatively severe since it penalizes word insertion and deletion, and treats content words and grammatical words equally.

Results showed that neither GMM+dyn nor HMM+dyn was able to generate intelligible speech in a systematic manner. However, some sentences were perfectly transcribed by most of the participants, especially the ones with a simple syntactic structure and a common vocabulary, such as "la voiture s'est arrêtée au feu rouge" (*the car stopped at red light*) or "mon père m'a donné l'autorisation" (*my father gave me permission*)[9]. Interestingly, the best results were obtained with the GMM+dyn technique, with a word

[9] The audio stimuli of these two sentences are provided as electronic supplementary material. The files are labeled *SpeakerA_list1_sentK_method.wav* with *K=302/308* and *method* is *gmm_dyn* for GMM+dyn, *hmm_dyn* for HMM+dyn or *anasyn* for the analysis-(re-)synthesis of the original signal using MLSA digital filter (excited by white noise).

accuracy of 60.4%, compared to 54% for the HMM+dyn approach. However, with a 95% confidence interval of 10% (due to the size of the test set which contains 210 words), the difference was not statistically significant. On this transcription task, the performance of both techniques should be considered as equivalent.

4.1.3.3 Discrimination test

In order to compare the two mapping methods on the same sentences, we conducted a XAB discrimination test, where X was the *reference* built by analyzing and (re)-synthesizing the original speech signal, while A and B were synthetic versions of the same sentence obtained using respectively the GMM+dyn and HMM+dyn techniques. The test stimuli consisted of the first 30 sentences of part P2 of the database recorded by speaker A (described in Section 3.2). The ten participants were asked to determine which sound of A and B was the most similar to X (A and B were presented in a random order). Surprisingly, and despite a higher spectral distortion of 0.7dB, stimuli obtained with the GMM+dyn approach were preferred 69% of the time to those obtained with the HMM+dyn approach. Most of the participants reported that for several stimuli X, it was difficult to decide between A and B since one was of lower quality but constant over the sentence, while the other one was judged of higher but more fluctuant overall quality. Most participants reported to have privileged stability over quality. A fluctuant quality is typically observed with the HMM+dyn approach. This perceptual feeling is consistent with objective measurements of the spectral distortion, which is lower with HMM+dyn compared to GMM+dyn, but has a higher standard deviation (See Table 2). This larger variability is likely to come from the spurious errors that can be made during the intermediate phonetic decoding step.

4.2 Articulatory-to-acoustic mapping experiments on silent speech

4.2.1 Accuracy of the phonetic decoding step between modal and silent speech

Recent studies suggested differences in terms of articulatory strategies between modal and silent speech, probably due to the lack of acoustic feedback. These studies were based on EMA data (Hueber et al., 2010a) and EMG data (Janke et al., 2010). In order to measure the impact of these differences on articulatory-to-acoustic mapping, we compared the accuracy of the phonetic decoding step of the HMM+dyn technique, between modal and silent speech. For this experiment, the training set was composed of 388 sentences recorded in modal speech only (i.e. P1 and P2), and the test set was either 100 VCV in a carrier sentence pronounced in modal speech (part P4) or the same sentences but pronounced in silent speech (part P5). The decoding graph of the HMM articulatory recognizer was constrained to sequences such as [t u t a p ɛ l **V_1 C V_2** s e s a] with V_1 potentially different from V_2 (ex: “tu t’appelles [olɛ], c’est ca ?”) and the performance was evaluated on V_1CV_2 sequences only. Results are presented in Table 3.

Table 3: Comparison of the performance of the phonetic decoding step of HMM+dyn mapping, between modal and silent speech, for VCV sequences embedded in a carrier sentence (95% confidence interval was 6% for vowels and 8% for consonants).

	Vocalized speech	Silent speech
PAcc on VCV (N_p = 300)	59.3%	44.3%
PAcc on Vowels (N_p = 200)	60%	45%
PAcc on Consonants (N_p = 100)	58%	43%

With a phone accuracy of only 59.3% on modal speech, the performance is much lower than that presented in Section 4.1.2 (i.e. ~78%, Table 1). This difference can partly be explained by

a much smaller training set (388 sentences against 688 sentences in the previous experiment). Indeed, we observed a degradation of the performance by 15% in silent speech compared to modal speech, for both vowels and consonants. These results are in line with the literature and suggest that silent speech may be articulated differently than modal speech.

4.2.2 Perceptual evaluation

Two perceptual tests were conducted to measure the intelligibility of the synthetic speech converted from silent articulation using the GMM+dyn and HMM+dyn techniques.

4.2.2.1 Transcription test

First we replicated on silent speech the transcription test conducted for modal speech, described in Section 4.1.3.2. The stimuli consisted of the same sentences that the one used for the previous test, but pronounced in silent speech (extracted from dataset P3) and converted into audio speech using GMM+dyn for the first set, and HMM+dyn for the second set (and MLSA filter excited by white noise). The accuracy of the transcription was evaluated using the same protocol as the one described in Section 4.1.3.2. As expected, with a *WAcc* approximately equaled to 30% for both methods (29.5% for GMM+dyn and 31% for HMM+dyn), the performance was much lower (almost twice as low) compared to modal speech (~60% *WAcc*). These results confirmed that the conversion of an unspecific 'full sentence' pronounced silently, into a perfectly intelligible speech audio signal, is not feasible yet without prior information or constrains on the linguistic context[10].

[10] The audio stimuli corresponding to the same two sentences mentioned in section 4.1.3.1 but converted from silent articulation are provided as electronic supplementary material. These files are labeled SpeakerA_list1_sentK_method_silent.wav with K=302/308 and method=gmm_dyn/hmm_dyn.

4.2.2.2 Identification test

In order to understand in more details which phonemes were badly reconstructed from silent articulation, we conducted another series of perceptual tests. In these tests, the listener was asked to identify a specific phoneme, inserted in a carrier sentence. The test stimuli consisted of 80 sentences, selected from part P5 of the corpus recorded by speaker A. It was divided in 4 series of 10 sentences, converted into audible speech using GMM+dyn and HMM+dyn techniques[11]. In the first two series, the ten participants were asked to identify the central consonant of a VCV sequence among the following set {p,t,k,f,s,ʃ,m,n,ʁ,l}, in the vocalic context [a] (i.e. [apa], [ata], [aka]) or [u] (i.e. [upu], [utu], [uku]). In the two other series, they were asked to identify the vowel among the following set {ɑ,i,u,y,e,ø,o,ɑ̃,ɛ̃,ɔ̃}, in consonantal context [p] (i.e. [apa], [ipi], etc.) or [ʃ] (i.e. [aʃa], [iʃi], etc.). Consonantal contexts were chosen for their different places and manners of articulation (bilabial vs. post-alveolar and plosive vs. fricative) and thus, for their different coarticulation patterns.

The segmental intelligibility of the synthetic speech was measured by calculating the mean percentage of correct identification, for each phonetic class. In order to assess the statistical significance of these results, we conducted a binary logistic regression test (also called *logit regression*). The test was conducted from a *generalized linear mixed-effects model* (using the package *glmer* in R software), considering:

[11] Two stimuli examples are provided as electronic supplementary material. The files are labeled *SpeakerA_VCV_carrier_method.wav*: *VCV is asa or ishi (iʃi*) and *method* is *gmm_dyn* for GMM+dyn or *hmm_dyn* for HMM+dyn.

- one binary variable to explain which was the success of phoneme identification (2-level: 0 the phoneme was not correctly identified and 1 otherwise)
- two explicative variables: one categorical factor called *SegmentType* (with 2-levels: Vowel/Consonant), a second categorical factor called *MappingMethod* (with 2-levels: GMM+dyn/HMM+dyn)
- a random *Listener* effect on the intercept.

Results are presented in Figure 7 and the corresponding statistical analyses are summarized in Table 4.

Since the interaction between *SegmentType* and *MappingMethod* was statistically significant (assessed using a likelihood ratio test), we conducted post-hoc analyses to test more specifically the contrast between the two mapping methods (GMM+dyn and HMM+dyn) for vowels, then for consonants. GMM+dyn mapping tends to slightly outperformed HMM+dyn on vowels (66.5% vs. 59.5%), but without statistical significance. However, HMM+dyn outperformed significantly GMM+dyn on consonants (61.5% vs. 44.5%). This latter result supports the benefit of an explicit modeling of the timing organization of articulatory gestures. The confusion matrices obtained for the stimuli synthesized with the HMM+dyn are detailed in Table 5 for both vowels and consonants.

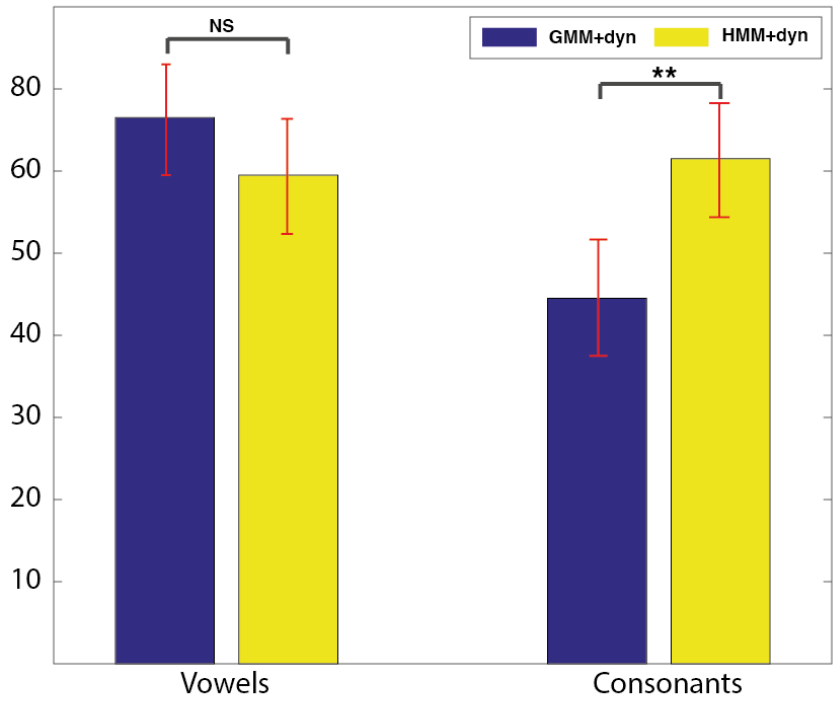

Figure 7: Perceptual identification task of vowel/consonant in carrier sentences, recorded in silent speech and converted into audible speech using either GMM+dyn or HMM+dyn technique. Error bars are calculated using a binomial approximation.

Table 4: Logistic regression of the effect of *SegmentType* (Vowel or Consonant) and the *MappingMethod* (GMM+dyn or HMM+dyn) on segment identification (perceptual listening test based on VCV sequences embedded in carrier sentences), using a generalized linear mixed-effects model (package *glmer* in R software).

Fixed effects	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	0.6857	0.1498	4.577	4.72e-06 ***
SegmentType	-0.9066	0.2066	-4.388	1.15e-05 ***
MappingMethod	-0.3133	0.2075	-1.510	0.131075
SegmentType:MappingMethod	1.0025	0.2905	3.451	0.000559 ***
Likelihood Ratio Test (LRT)				
		df	X^2	P(>X^2)
SegmentType*MappingMethod		1	12.005	0.0005305 ***
Post-hoc analyses				
Contrast	Estimate	Std. Error	z-value	Pr(>|z|)
HMM+dyn - GMM+dyn / Vowel	-0.3133	0.2075	-1.510	0.2450
HMM+dyn - GMM+dyn / Consonant	6893	0.2034	3.389	0.0014 **

Table 5: Confusion matrices of the perceptual identification test for stimuli synthesized using HMM+dyn. These values indicate the average percentage of correct answers over the two considered vocalic or consonantal contexts. Chance level was 10%.

	[a]	[i]	[u]	[y]	[e]	[o]	[ø]	[ɑ̃]	[ɛ̃]	[ɔ̃]
[a]	85	0	0	0	0	5	0	10	0	0
[i]	0	100	0	0	0	0	0	0	0	0
[u]	0	0	85	0	0	5	10	0	0	0
[y]	0	5	0	95	0	0	0	0	0	0
[e]	25	10	0	5	50	0	5	5	0	0
[o]	0	0	40	0	0	45	10	0	0	10
[ø]	0	30	20	5	0	0	35	0	5	5
[ɑ̃]	5	0	0	0	10	15	0	65	5	0
[ɛ̃]	50	0	10	0	0	0	0	35	5	0
[ɔ̃]	0	0	45	0	0	10	10	5	0	30

	[p]	[t]	[k]	[f]	[s]	[ʃ]	[m]	[n]	[r]	[l]
[p]	45	0	10	20	0	0	10	0	10	5
[t]	5	40	0	0	0	0	5	50	0	0
[k]	0	0	55	5	0	5	5	5	25	0
[f]	0	0	30	50	5	0	0	5	10	0
[s]	0	0	0	0	100	0	0	0	0	0
[ʃ]	0	0	0	0	0	100	0	0	0	0
[m]	5	5	0	5	0	5	50	5	5	20
[n]	0	0	0	5	5	30	30	10	0	20
[r]	0	0	0	0	0	0	0	0	100	0
[l]	0	0	0	15	10	0	5	5	0	65

Consonants [s], [ʃ] and [r] are systematically identified correctly by all participants. As in the objective evaluation of HMM+dyn on modal speech, confusions are made between [t] and [n] and between [p] and [m] (likely due to the lack of information of the velum position). The consonant [n] is sometimes perceived as [ʃ], [l], or [s], whereas this confusion never occurs in modal speech (see Figure 6). This result could suggest a more anterior place of articulation for [n] in silent speech. A similar tendency was observed in (Hueber et al., 2010a). Substitutions of [r] by [k] (25%) might be explained by similar tongue shapes in the mid-sagittal plane, especially in a back context [u]. However, substitutions of [k] by [f] (30%), and [f] by [p] (20%) remain difficult to interpret. As concerns the vowels, good performances were observed for the extreme vowels [a], [i], [u]. Most of the remaining errors came from the misperception of nasal vowels, consistent with the results of the objective evaluation. Thus, [ɔ̃] was sometimes perceived as [o] (10%), and [ɑ̃] was identified as [ɛ̃] (35%).

5. Conclusions and Perspectives

In this article, we compared two statistical mapping techniques for the conversion of silent articulation into audible speech, with no restriction on the vocabulary size, for a silent speech interface application. First, we investigated the GMM-based mapping technique considering dynamic features proposed by (Toda et al., 2007) (which was referred to as GMM+dyn). Then, we adapted this technique to the framework of HMM. Similarly to the GMM+dyn technique, the proposed method models explicitly the local correlations between articulatory and acoustic observations using full-covariance phone HMMs. It also considers explicitly the relationship between static and dynamic features to guarantee smooth output feature trajectories (therefore the proposed method was referred to as HMM+dyn). Contrary to the GMM+dyn technique, it aims at modeling more explicitly the timing organization of articulatory gestures, and exploiting linguistic knowledge to regularize the problem of silent speech conversion. Both techniques were evaluated in the context of a silent speech interface (SSI) driven by ultrasound and video imaging, on continuous speech, for two French speakers producing both modal and silent speech. The main results of objective and perceptual evaluations are summarized in the next paragraphs.

For modal speech, the performance of the intermediate phonetic decoding step of HMM+dyn was more than 75% for both speakers. Furthermore, HMM+dyn outperformed GMM+dyn in terms of average spectral distortion (objective evaluation). However, perceptual tests indicated that naive listeners preferred the stimuli generated by GMM+dyn, probably because they were of lower, but more stable quality. Nevertheless, with a word accuracy of 60%, transcription tests showed that none of these techniques was yet able to synthesize perfectly intelligible speech, in a systematic manner.

For silent speech, the performance of the intermediate phonetic decoding step of HMM+dyn decreased by 15%. This suggested that silent speech might be articulated differently than modal speech, likely due to the lack of audio feedback. Finally, perceptual identification tests showed an improved segmental intelligibility for HMM+dyn compared to GMM+dyn for consonants. This result supports the interest of the HMM+dyn method of modeling explicitly the timing of speech articulation.

The introduction of linguistic knowledge is a way to regularize the ill-posed problem of silent speech conversion. In future work, the use of more informative prior information in the phonetic decoding step of the HMM+dyn technique should be envisioned. Depending the targeted application, such information could consist in a limited vocabulary and/or a domain-specific language model. Besides, the use of model adaptation techniques could be envisioned to tackle the problem of articulatory differences between modal and silent speech. Another way to limit these articulatory differences could consist in synthesizing speech in real-time and providing the user with this acoustic feedback. To that purpose, a reactive implementation of the HMM+dyn technique should be developed. It could be based on the short-term MLPG algorithm (Muramatsu et al., 2008). Such a system would be necessary to study how people use a silent speech interface in a realistic communicative situation and how they adapt their articulation to maximize the efficiency of their 'silent' communication.

ACKNOWLEDGEMENTS

This study was conducted in the context of the Ultraspeech2 project funded by the 6th Christian Benoit Award (ACB/AFCP/ISCA) attributed to Thomas Hueber. The authors would like to thank the speakers who contributed to the data acquisition, the participants of the perceptual listening tests, Laurent Girin for his very useful comments, and the reviewers for their work on this article and the fruitful exchange.

REFERENCES

Birkholz, P., Jackèl, D., Kroger, B., 2006. Construction and control of a three-dimensional vocal tract model, in: Proceedings of ICASSP. Toulouse, France, pp. 873–876.

Buchsbaum, A.L., van Santen, J.P., 1996. Selecting training inputs via greedy rank covering, in: Proceedings of the 7th Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia, PA, USA, pp. 288–295.

Chen, T., 2001. Audiovisual speech processing. Signal Process. Mag. IEEE 18(1), 9–21.

Clarkson, P., Rosenfeld, R., 1997. Statistical language modeling using the CMU-cambridge toolkit., in: Proceedings of Eurospeech. Rhodes, Greece, pp. 2707–2710.

Combescure, P., 1981. Vingt listes de dix phrases phonétiquement équilibrées. Rev. Acoust. 14(56), 34–38.

Denby, B., Oussar, Y., Dreyfus, G., Stone, M., 2006. Prospects for a silent speech interface using ultrasound imaging, in: Proceedings of ICASSP. Toulouse, France, pp. 365–368.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J., Brumberg, J., 2010. Silent speech interfaces. Speech Commun. 52(4), 270–287.

Fagan, M.J., Ell, S.R., Gilbert, J.M., Sarrazin, E., Chapman, P.M., 2008. Development of a (silent) speech recognition system for patients following laryngectomy. Med. Eng. Phys. 30(4), 419–425.

Gilbert, J.M., Rybchenko, S.I., Hofe, R., Ell, S.R., Fagan, M.J., Moore, R.K., Green, P., 2010. Isolated word recognition of silent speech using magnetic implants and sensors. Med. Eng. Phys. 32(10), 1189–1197.

Hiroya, S., Honda, M., 2004. Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. IEEE Trans. Speech Audio Process. 12(2), 175–185.

Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., Stone, M., 2007. Eigentongue feature extraction for an ultrasound-based silent speech interface, in: Proceedings of ICASSP. Honolulu, USA, pp. 1245–1248.

Hueber, T., Badin, P., Savariaux, C., Vilain, C., Bailly, G., 2010a. Differences in articulatory strategies between silent, whispered and normal speech? a pilot study using electromagnetic articulography, in: Proceedings of International Seminar on Speech Production (ISSP). Montreal, Canada.

Hueber, T., Benaroya, E.-L., Chollet, G., Denby, B., Stone, M., 2010b. Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips. Speech Commun. 52(4), 288–300.

Hueber, T., Benaroya, E.-L., Denby, B., Chollet, G., 2011. Statistical Mapping Between Articulatory and Acoustic Data for an Ultrasound-Based Silent Speech Interface., in: Proceedings of Interspeech. Firenze, Italia, pp. 593–596.

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

Hueber, T., Chollet, G., Denby, B., Dreyfus, G., Stone, M., 2009. Visuo-Phonetic Decoding using Multi-Stream and Context-Dependent Models for an Ultrasound-based Silent

Speech Interface, in: Proceedings of Interspeech. Brighton, England, pp. 640–643.

Hueber, T., Chollet, G., Denby, B., Stone, M., 2008. Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application, in: Proceedings of International Seminar on Speech Production. Strasbourg, France, pp. 365–369.

Imai, S., Sumita, K., Furuichi, C., 1983. Mel Log Spectrum Approximation (MLSA) filter for speech synthesis. Electron. Commun. Jpn. Part Commun. 66, 10–18.

Janke, M., Wand, M., Schultz, T., 2010. Impact of lack of acoustic feedback in EMG-based silent speech recognition., in: Proceedings of Interspeech. Makuhari, Japan, pp. 2686–2689.

Kello, C.T., Plaut, D.C., 2004. A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters. J. Acoust. Soc. Am. 116(4), 2354–2364.

Ling, Z.-H., Richmond, K., Yamagishi, J., 2010. An Analysis of HMM-based prediction of articulatory movements. Speech Commun. 52(10), 834–846.

Ling, Z.-H., Richmond, K., Yamagishi, J., Wang, R.-H., 2009. Integrating Articulatory Features Into HMM-Based Parametric Speech Synthesis. Audio Speech Lang. Process. IEEE Trans. On 17(6), 1171–1185.

Maeda, S., 1990. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model, in: Speech Production and Speech Modelling. Springer, pp. 131–149.

Muramatsu, T., Ohtani, Y., Toda, T., Saruwatari, H., Shikano, K., 2008. Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory., in: Proceedings of Interspeech. Brisbane, Australia, pp. 1076–1079.

Nakajima, Y., Kashioka, H., Campbell, N., Shikano, K., 2006. Non-audible murmur (NAM) recognition. IEICE Trans. Inf. Syst. 89(1), 1–8.

Nakajima, Y., Kashioka, H., Shikano, K., Campbell, N., 2003. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin, in: Proceedings of ICASSP. Hong Kong, Hong Kong, pp. 708–711.

Ostendorf, M., Digalakis, V.V., Kimball, O.A., 1996. From HMM's to segment models: a unified view of stochastic modeling for speech recognition. Speech Audio Process. IEEE Trans. On 4(5), 360–378.

Richmond, K., 2006. A trajectory mixture density network for the acoustic-articulatory inversion mapping., in: Proceedings of Interspeech. Pittsburgh, Pennsylvania, USA, pp. 577–580.

Russell, M., Moore, R., 1985. Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition, in: Proceedings of ICASSP. Detroit, Michigan, USA, pp. 5–8.

Schultz, T., Wand, M., 2010. Modeling coarticulation in EMG-based continuous speech recognition. Speech Commun. 52(4), 341–353.

Shannon, M., Byrne, W., 2013. Fast, low-artifact speech synthesis considering global variance, in: Proceedings of ICASSP. Vancouver, British Columbia, Canada, pp. 7869–7873.

Silén, H., Helander, E., Nurminen, J., Gabbouj, M., 2012. Ways to Implement Global

Variance in Statistical Speech Synthesis., in: Proceedings of Interspeech. Portland, USA.

Sondhi, M.M., Schroeter, J., 1987. A hybrid time-frequency domain articulatory speech synthesizer. IEEE Trans. Acoust. Speech Signal Process. 35(7), 955–967.

Stone, M., 2005. A guide to analysing tongue motion from ultrasound images. Clin. Linguist. Phon. 19(6-7), 455 – 501.

Toda, T., Black, A.W., Tokuda, K., 2007. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Trans. Audio Speech Lang. Process. 15(8), 2222–2235.

Toda, T., Black, A.W., Tokuda, K., 2008. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. Speech Commun. 50(3), 215–227.

Toda, T., Tokuda, K., 2007. A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis. IEICE Trans. Inf. Syst. E90-D, 816–824.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., 2000. Speech parameter generation algorithms for HMM-based speech synthesis, in: Proceedings of ICASSP. Istanbul, Turkey, pp. 1315–1318.

Turk, M., Pentland, A., 1991. Eigenfaces for Recognition. J. Cogn. Neurosci. 3, 71–86.

Wand, M., Schultz, T., 2011. Session-independent EMG-based Speech Recognition., in: Proceedings of Biosignals. Rome, Italy, pp. 295–300.

Wrench, A., Scobbie, J., Linden, M., 2007. Evaluation of a helmet to hold an ultrasound probe. Presented at the Ultrafest IV, New York, USA.

Wrench, A., Scobbie, J.M., 2006. Spatio-temporal inaccuracies of video-based ultrasound images of the tongue., in: Proceedings of the International Seminar on Speech Production. Ubatuba, Bresil, pp. 451–458.

Young, S., 2005. The HTK Book. (http://htk.eng.cam.ac.uk/).

Youssef, A.B., Hueber, T., Badin, P., Bailly, G., 2011. Toward a multi-speaker visual articulatory feedback system, in: Proceedings of Interspeech. Firenze, Italia, pp. 589–592.

Yu, Y.J., Acton, S.T., 2002. Speckle reducing anisotropic diffusion. IEEE Trans. Image Process. 11(11), 1260–1270.

Zen, H., Nankaku, Y., Tokuda, K., 2011. Continuous stochastic feature mapping based on trajectory hmms. IEEE Trans. Audio Speech Lang. Process. 19(2), 417–430.

Zhang, L., Renals, S., 2008. Acoustic-Articulatory Modelling with the Trajectory HMM. IEEE Signal Process. Lett. 15, 245–248.

FIGURE CAPTIONS

Figure 1: Silent speech interface driven by ultrasound and video imaging. The present study focuses on the direct conversion of silent articulation into audible speech without any restriction on the vocabulary size (contrary to silent speech recognition).

Figure 2 left: Directed graphical model associated with the GMM considered during training ($\mathbf{x}$ and $\tilde{\mathbf{y}}$ are observed variables, m is a latent variable) / right: Graphical representation of the GMM-based mapping process considering dynamic features (GMM+dyn).

Figure 3: left: Directed graphical model associated with the full-covariance HMM considered during training ($\mathbf{x}$ and $\tilde{\mathbf{y}}$ are observed variables, q_t is a latent variable) / right: Graphical representation of the HMM-based mapping process considering dynamic features (HMM+dyn).

Figure 4: Experimental setup used for data acquisition (speaker B)

Figure 5: Typical examples of recorded ultrasound and video images (320x240 and 640x480 pixels respectively) for male speaker A (up) and female speaker B (bottom). Red bounding boxes delimit the regions of interest considered for feature extraction.

Figure 6: Confusion matrix for the intermediate phonetic decoding step of the HMM+dyn mapping experiment for speaker A. The color space map was chosen to emphasize errors (DEL and INS stand for deletion and insertion errors).

Figure 7: Perceptual identification task of vowel/consonant in carrier sentences, recorded in silent speech and converted into audible speech using either GMM+dyn or HMM+dyn technique. Error bars are calculated using a binomial approximation.