Statistical Mapping between Articulatory and Acoustic Data, Application to Silent Speech Interface and Visual Articulatory Feedback

Thomas Hueber, Pierre Badin, Gérard Bailly,
Atef Ben-Youssef, Frédéric Elisei
GIPSA-lab, UMR 5216/CNRS/INP/UJF/U.Stendhal
961 rue de la Houille Blanche - BP 46
38402 Saint Martin d'Hères, France
thomas.hueber,pierre.badin,gerard.bailly,
atef.ben-youssef,frederic.elisei
@gipsa-lab.grenoble-inp.fr

Bruce Denby
Université Pierre et Marie Curie
SIGMA Laboratory / ESPCI ParisTech
10 rue Vauquelin, 75005, Paris France
denby@ieee.org

Gérard Chollet
LTCI / CNRS, Telecom ParisTech
46 rue Barrault, 75634, Paris, France
gerard.chollet@tsi.enst.fr

Abstract

This paper reviews some theoretical and practical aspects of different statistical mapping techniques used to model the relationships between the articulatory gestures and the resulting speech sound. These techniques are based on the joint modeling of articulatory and acoustic data using Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM). These methods are implemented in two systems: (1) the *silent speech interface* developed at SIGMA and LTCI laboratories which converts tongue and lip motions, captured during silent articulation by ultrasound and video imaging, into audible speech, and (2) the *visual articulatory feedback* system, developed at GIPSA-lab, which automatically animates, from the speech sound, a 3D orofacial clone displaying all articulators (including the tongue). These mapping techniques are also discussed in terms of real-time implementation.

Keywords: statistical mapping, silent speech, ultrasound, visual articulatory feedback, talking head, HMM, GMM

1. Introduction

Speech can be defined as a set of gestures made audible and visible. Speech sounds result from the coordinated movements of all the vocal organs, *i.e* the larynx and the supra-glottal articulators (tongue, lips, jaw, velum). Various approaches have been proposed in the literature to model the relationship between articulatory gestures and acoustic characteristics of speech. Some of them address the problem of *articulatory-to-acoustic* mapping, *i.e* the generation of speech sounds from a set of varying vocal tract configurations. Others address the inverse problem, commonly called *acoustic-to-articulatory inversion*, which consists in recovering the articulatory trajectories from the speech audio signal.

A first approach to the articulatory-acoustic mapping problem consists in modeling the vocal tract geometry and its corresponding acoustic transfer function. Different types of articulatory models have been proposed to control the vocal tract geometry:

- *in geometrical models* such as [1] and [2], the positions and shapes of the articulators (and therefore the vocal tract geometry) are controlled by a set of vocal tract parameters that are defined *a priori*.
- *in statistical models* such as [3] and [4], the positions and shapes of the articulators are defined as a combination of standard vocal tract configurations. These standard configurations are derived from statistical analyses of large articulatory datasets, acquired using MRI, electromagnetic articulography (EMA), or X-Ray.
- *in biomechanical models* such as [5] and [6], the motion of the articulators is predicted from their physiological structure and their neuromuscular control.

Different methods have been proposed to characterize the acoustic properties of the vocal tract from its estimated geometry. Some of them address the problem in the frequency domain [7], in the time domain [8], or in both [9] (using an hybrid approach). Such a geometric and acoustic modeling of the vocal tract can also be used for acoustic-to-articulatory inversion, by using an analysis-by-synthesis paradigm, as in [10] and [11].

The articulatory-acoustic mapping problem can also be addressed using a *corpus-based* approach by exploiting *in vivo* databases containing “parallel” articulatory-acoustic data (speech sounds recorded simultaneously with articulatory movements). Two types of corpus-based approaches can be found in the literature: *codebook-based* and *statistical model-based approaches*.

In the first one, a subset of articulatory-acoustic pairs is extracted from the database using vector quantization techniques. This subset is called the codebook. For a given source vector, for example a vector of acoustic parameters, the mapping consists (in its simplest implementation) in finding the closest acoustic vector in the codebook and in retrieving the corresponding articulatory vector. Codebook-based approaches have been used for both acoustic-to-articulatory inversion [12] [13] and articulatory-to-acoustic mapping [14].

Statistical model-based approaches aim at estimating the articulatory-acoustic mapping function using supervised machine learning techniques. Two types of statistical models can be envisioned:

- a *discriminative model* that describes directly the posterior probability $p(y \mid x)$ of the target vector y for a given source vector *x*. Such a model can be obtained using artificial neural networks (ANN) as in [15], [16] (for acoustic-to-articulatory inversion) and [17] (for articulatory-to-acoustic mapping), or a support vector machine (SVM), as in [18] (for acoustic-to-articulatory inversion).
- a *generative model* that describes the joint probability $p(x,y)$ and uses Bayes theorem to recover the posterior probability $p(y \mid x)$. Toda [19] and Hiroya [20] propose to use respectively a Gaussian Mixture Model (GMM), and a Hidden Markov Model (HMM), two classical types of generative models, for both acoustic-to-articulatory and acoustic-to-articulatory mapping.

In this paper, we illustrate the use of statistical model-based approaches, especially GMM-based and HMM-based ones, for two research domains: *silent speech interface* which requires the conversion of articulatory data into acoustic data, and *visual articulatory feedback* which is based, in our approach, on acoustic-to-articulatory inversion.

This article is organized as follows. Theoretical and practical aspects of GMM-based and HMM-based mapping techniques are presented in section 2. Their implementation in the context of silent speech interface and visual articulatory feedback are presented in section 3 and 4, respectively. The feasibility of a real-time implementation of these mapping techniques is discussed in section 5. Conclusions and perspectives are presented in the last section.

2. GMM and HMM-based Mapping

2.1 GMM-based mapping

In the GMM framework, the probability density function of a continuous random variable O is defined as a sum of normal distributions as:

$$p(\mathbf{o} \mid \Theta) = \sum_{m=1}^{M} \alpha_m N\left(\mathbf{o}, \mu_m, \Sigma_m\right) \quad (1)$$

$$\text{with } \Sigma_{m=1}^{M} \alpha_m = 1 \text{ and } \forall m \in \left[1..M\right], \alpha_m \geq 0$$

where $\mathbf{o}$ is a realization of O (a feature vector), d is the dimension of $\mathbf{o}$, Θ is the parameter set of the model, $N\left(., \mu, \Sigma\right)$ is a normal (Gaussian) distribution with mean μ and covariance matrix Σ, M is the number of mixture components, α_m is the weight associated with the m^{th} mixture component and T is the transpose operator. Given a training dataset of feature vectors, the parameters of a GMM (weights, mean vectors and covariance matrices for each component) can be efficiently estimated using the expectation-maximisation (EM) algorithm. This algorithm finds the parameters set which maximizes the likelihood of the model given the training data.

In this work, we adopt the approach proposed by Kain for voice conversion [21]. This approach is based on the modeling of the joint probability density of source and target vectors $p(Z) = p(X,Y)$ with:

$$\mathbf{Z} = \left[\mathbf{X}\ \mathbf{Y}\right] = \begin{pmatrix} x_{11} & \cdots & x_{1d_x} & y_{11} & \cdots & y_{1d_y} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nd_x} & y_{N1} & \cdots & y_{Nd_y} \end{pmatrix} \quad (2)$$

where X and Y are respectively the sequence of N source and target vectors (d_x and d_y are respectively the dimensions of the source and target vectors).

The mapping function that predicts the target vector $\hat{\mathbf{y}}_t$ from the given source vector $\mathbf{x}_t$, observed at time t, is formulated as a weighted sum of linear models such as:

$$\hat{\mathbf{y}}_t = F(\mathbf{x}_t) = \sum_{m=1}^{M} (W_m \mathbf{x}_t + b_m) \cdot P(c_m \mid \mathbf{x}_t) \quad (3)$$

with W_m and μ_m the transformation matrix and bias vector associated with the m^{th} component of the model, defined as:

$$W_m = \Sigma_m^{YX} (\Sigma_m^{XX})^{-1}, \ \ b_i = \mu_m^{Y} - \Sigma_m^{YX} (\Sigma_m^{XX})^{-1} \mu_m^{X}$$

$$\text{with } \Sigma_m = \begin{bmatrix} \Sigma_m^{XX} & \Sigma_m^{XY} \\ \Sigma_m^{YX} & \Sigma_m^{YY} \end{bmatrix} \text{ and } \mu_m = \begin{bmatrix} \mu_m^{X} \\ \mu_m^{Y} \end{bmatrix} \quad (4)$$

and $P(\mathrm{c}_m|\mathbf{x}_t)$, the probability that the source vector "belongs" to the m^{th} component, defined as[1]:

$$P(c_m \mid \mathbf{x}_t) = \frac{\alpha_m N(\mathbf{x}_t, \mu_m^{X}, \Sigma_m^{XX})}{\sum_{p=1}^{M} \alpha_p N(\mathbf{x}_t, \mu_p^{X}, \Sigma_p^{XX})} \quad (5)$$

In our implementation, the GMM model is initialized using the *k-means* algorithm.

[1] The notations "P" and "p" are used for discrete and continuous probability distributions, respectively.

2.2 HMM-based mapping

In the proposed HMM-based mapping approach[2], the sequence of target vectors $\hat{\mathbf{y}}$, predicted from the given sequence of source vectors **x,** is defined as:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \left\{ p(\mathbf{y} \mid \mathbf{x}) \right\} \tag{6}$$

with

$$p(\mathbf{y} \mid \mathbf{x}) = p(\mathbf{y} \mid q, \lambda) \cdot P(q \mid \mathbf{x}, \lambda) \tag{7}$$

where λ is the parameter set of the HMM and q the HMM state sequence. As shown in Equation 7, an HMM-based mapping can be achieved with a *recognition followed by synthesis* approach which means: 1) finding the optimal state sequence for a given source vector, and 2) inferring the target vector from the decoded state sequence.

The HMM can be defined and trained in different ways. In this paper, we describe a method based on the use of phonetic information. In the training stage, a multistream HMM (MSHMM) is trained on articulatory-acoustic data for each phonetic class. In a MSHMM, each stream has, for each state, its own Gaussian mixture and thus its own emission probability density function.

The initialization of the HMMs requires temporal segmentation of the training data at phonetic level. As articulatory and acoustic data are recorded synchronously, this segmentation can be obtained by annotating the acoustic data, either manually, or by using an initial set of "audio-only HMMs" and a forced-alignment procedure. After initialization, HMMs are then trained using a standard procedure (similar to that described in [22]): HMMs are trained first separately, using the standard Baum-Welch re-estimation algorithm and then processed simultaneously, using an *embedded training* strategy. Since articulatory and acoustic features are naturally sensitive to context effects such as co-articulation and anticipation, context-dependency is then introduced in the modeling. Triphone models are created by adding information about left and right contexts to the phone models. A tree-based state-tying strategy is adopted to address the problem of data sparsity (triphones having only a few occurrences in the training dataset). Finally, tied-state models are refined by increasing incrementally the number of Gaussian mixture components.

The prediction of the target (feature-)vector sequence **y,** for a given test sequence of source feature vectors **x,** is achieved in two stages. First, phonetic decoding is performed using the Viterbi algorithm (only the parameters of the MSHMM related to the source stream are used for the decoding stage). Second, given the predicted sequence of phones and the decoded HMM state sequence, target vector sequence is inferred using the speech parameter generation algorithm proposed by Tokuda for HMM-based speech synthesis [23]. This algorithm determines the vector sequence that maximizes the likelihood of the model with respect to a continuity constraint on the predicted feature trajectories. The key steps of this algorithm are described below.

In this approach, an *observation vector* $\mathbf{o}_t$ is hypothesized to be composed of static features $\mathbf{y}_t$ (the target features), and dynamic features, the first and second derivatives of static features, so that $o_t = [y_t, \Delta^{(1)} y_t, \Delta^{(2)} y_t]$. By using derivative discretization techniques, the relationship between a sequence of observation vectors $\mathbf{o} = [\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_N]$ and the sequence of static feature vectors $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N]$ can be written in a matrix form such as:

$$\mathbf{o} = W\mathbf{y} \quad \text{with} \quad W = [W_1^T, ..., W_N^T]^T$$

$$W_t = \begin{cases} \mathbf{0}_{3d \times dN} & (t = 1,2) \\ \left[\mathbf{0}_{3d \times (t-3)d}, W_{base}, \mathbf{0}_{3d \times (N-t)d} \right] & (N \geq t \geq 3) \end{cases} \tag{8}$$

$$W_{base} = \begin{pmatrix} \mathbf{0}_{d \times d} & \mathbf{0}_{d \times d} & \mathrm{I}_{d \times d} \\ \mathbf{0}_{d \times d} & -0{,}5 \cdot \mathrm{I}_{d \times d} & -0{,}5 \cdot \mathrm{I}_{d \times d} \\ \mathrm{I}_{d \times d} & -2 \cdot \mathrm{I}_{d \times d} & \mathrm{I}_{d \times d} \end{pmatrix}$$

with *N*, the number of vector in the sequence, *d* the dimension of the target features, **I** and **0** respectively the identity and zero matrices. Given the decoded HMM state sequence **q** = $\{q_1, \ldots, q_N\}$, Tokuda shows that the target vector sequence $\hat{\mathbf{y}}$ can be obtained by solving:

$$\hat{\mathbf{y}} = \left(W^T \Sigma_q^{-1} W \right)^{-1} W^T \Sigma_q^{-1} M_q$$
$$\text{with } M_q = \left[\mu_{q_1}, \ldots, \mu_{q_N} \right], \; \Sigma_q^{-1} = diag\left[\Sigma_{q_1}^{-1}, \ldots, \Sigma_{q_N}^{-1} \right] \tag{9}$$

where μ_k and Σ_k are respectively the mean and diagonal covariance matrix of the Gaussian emission probability density associated with state *k* (in this approach, only single-Gaussian densities can be used to model the distribution of target data).

3. Ultrasound-based Silent Speech Interface

A "silent speech interface" (SSI) is a device that allows speech communication without the necessity of vocalizing. SSI could be used in situations where silence is required (as a silent cell phone), or for communication in very noisy environments. Further applications are possible in the medical field. For example, SSI could be used by laryngectomized patients as an alternative to electrolarynx which provides a very robotic voice; to oesophageal speech, which is difficult to master; or to tracheo-oesoephageal speech, which requires additional surgery.

[2] All the procedures involving HMM manipulations, described in this paper, are done using the HTK and HTS toolkits.

The design of a SSI has recently received considerable attention from the speech research community [24]. Different approaches have been proposed in the literature. A speaker may for example produce small airflow in his vocal tract and capture the resulting “murmur” with a stethoscopic (or NAM) microphone as in [25] and [26]. Other approaches, based on completely non-acoustic features have also been proposed, as for example in [27] where electromyographic electrodes placed on the speaker’s face (or on his neck in [28]) record muscular activity. In our approach, articulatory movements are captured by a non-invasive multimodal imaging system composed of an ultrasound transducer placed beneath the chin and a video camera in front of the lips [24]. The *articulatory-to-acoustic* mapping problem, *i.e* the synthesis of an intelligible speech signal from visual articulatory data (only), is addressed using the statistical-model based approaches described in section 2.

3.1 Data acquisition

The experimental setup used to record parallel articulatory-acoustic data is presented in figure 1. The hardware component of the system is based on the portable Terason T3000 ultrasound system, a 140° microconvex transducer, an industrial USB Bayer color camera and a standard sound system. In order to automate the two imaging devices (ultrasound system and video camera) and to record the different streams, we developed a dedicated software, named Ultraspeech[3] [31], which processes the ultrasound, video and audio streams in parallel using multithreading programming techniques and prevent data loss using a FIFO-based buffer management approach. This system was used to record simultaneously, and synchronously: the ultrasound stream at 60 fps (320x240 pixels); the video stream at 60 fps (640x480 pixels); and the acoustic signal (16 bits, 16 kHz).

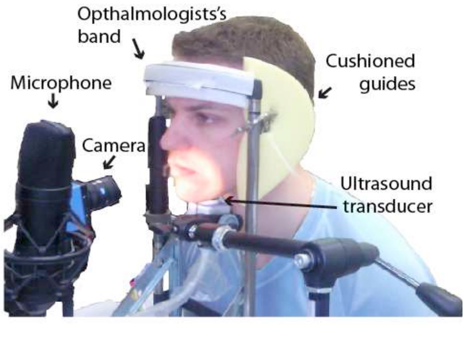

Figure 1. Experimental setup (hardware component).

The recorded dataset used in this work consists of the 1132 sentences of CMU ARCTIC corpus [10], uttered by a female native English speaker. To prevent speaker fatigue, the acquisition was split into 10 sessions, spaced in time. An inter-session re-calibration mechanism (detailed in [31]), was used to maintain the positioning accuracy of the sensors across all sessions (and thus the data consistency). A typical pair of ultrasound and video images is shown in figure 2.

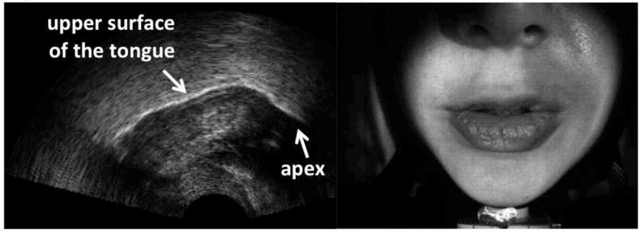

Figure 2. Example of an ultrasound vocal tract image (in the midsagittal plane) with lip frontal view

3.2 Visual feature extraction

Regions of interest (ROI) selected in ultrasound and video images were first resized to 64x64 pixels. The *EigenTongues* [32] decomposition technique was used to encode each ultrasound frame. In this method, the vocal tract configuration is interpreted as a linear combination of standard configurations, the *EigenTongues*, obtained by performing a Principal Component Analysis (PCA) on a phonetically balanced subset of frames. A similar technique was used to encode lip images (*EigenLips*). The numbers of projections onto the set of EigenTongues/EigenLips used for coding were determined by keeping the eigenvectors carrying at least 80% of the variance of the training set; typical values used on this database were 30 coefficients for each of the two streams. In order to be compatible with the speech analysis rate, the EigenTongues/EigenLips coefficient sequences were oversampled from 60 Hz to 100 Hz. Finally, they were concatenated with their first and second derivative in one and same *visual feature vector*.

3.3 GMM-based visuo-acoustic mapping

In the GMM-based mapping experiment, the spectral content of the audio speech signal was parameterized by 25 mel-cepstrum coefficients (Blackman window, 25 frame length, 10 ms frame shift).

The first 1110 sentences of the recorded database were divided into 37 lists of 30 sentences. In order to increase the statistical relevance of the mapping performance, a jackknife (leave-one-out) technique was employed: each list was used once as the test set while the other 34 lists composed the training set. Two test lists were excluded from this jackknife procedure to be used as a validation set for the determination of the optimal number of Gaussians (parameter M in equation 1), which was found to be 32. Silence frames were removed from the training set using an automatic (threshold-based) silence detection method.

The quality of the mapping between visual and spectral features was evaluated by calculating the *Mel-cepstral*

[3] http://www.ultraspeech.com

distortion between the target and the predicted mel-ceptrum coefficients, defined as:

$$Mel - CD[dB] = \frac{10}{\ln 10}\sqrt{2\sum_{d=0}^{24}(\hat{m}_d - m_d)^2} \tag{10}$$

The Mel-cepstral distortion was found to be *7.8 dB* if the the 0^{th} cepstral dimension, *i.e* the component known to correspond to overall signal power, was taken into account. It was *6.2 dB* if this term was ignored.

In this experiment, the audio speech signal was synthesized using a MLSA digital filter [33] derived from the predicted mel-cepstrum coefficients. The generation of the excitation signal requires the prediction of the voiced/unvoiced parameters as well as the pitch for voiced frames. A feed-forward neural network was used to perform the mapping between the visual features and the voiced/unvoiced parameter. The network was trained using a standard gradient descent algorithm; the log-sigmoid function was used as the activation function for the hidden neurons and the output layer. The optimal number of hidden neurons, determined by cross-validation, was found to be 10. The accuracy of the classifier, its sensibility and its specificity were respectively 0.82, 0.80 and 0.84. This means that about 80% of the frames were correctly classified. However, this relative good performance should be interpreted carefully. Since there is no direct relationship between voicing and articulatory configuration, the performance may be partially explained by *indirect* relationships; for instance, stable vocal tract configurations are likely to correspond to vowels and thus to voiced frames.

GMM-based mapping between visual features and pitch (for voiced frames) has also been investigated. The performance was measured by evaluating the root mean squared error (RMSE) between the estimated f0 and the target f0. Because the error was greater to 50 Hz, it was not possible to generate an acceptable excitation signal with this approach; a constant pitch value was finally used for synthesis.

A preliminary subjective evaluation revealed that it was not possible to synthesize intelligible speech consistently with this GMM-based mapping approach. However, the lack of intelligibility seems to be related to the quality of the vocoder, rather than to the accuracy of the articulatory-to-acoustic mapping. We intend to use more robust synthesis techniques (such as the Harmonic plus Noise Model used in the HMM-based mapping experiment described in the next section).

3.4 HMM-based visuo-acoustic mapping

In the HMM-based mapping experiment, a *Harmonic plus Noise Model* (HNM) decomposition techniques was used to parametrize the speech signal [34]. In our implementation, harmonic and noise components were represented by an auto-regressive model, described respectively by 12 and 16 LSF coefficients (Line Spectral Frequencies). For each of the 40 phonetic classes, the sequences of visual and acoustic features were modeled by a left-to-right, 5-state (3 emitting states), continuous multistream HMMs (with diagonal covariance matrices). For the visual part, the optimal number of Gaussian per state was found to be 4. For the audio part, harmonic component and pitch (defined only for voiced frame) were modeled using the *Multi-Space probability Distribution* approach (MSD) described in [35].

In the proposed HMM-based mapping approach, linguistic constraints can be introduced to help the phonetic decoding. With that in mind, we implemented two decoding scenarios. In the first, considered “unconstrained”, the structure of the decoding network was a simple loop in which all phones loop back to each other. In the second, or “constrained” scenario, the phonetic decoder was forced to recognize words contained in the CMU Arctic sentences. In that case, the decoding network allows all possible word combinations which can be built from a 3k word dictionary. No statistical language model was used in the present study.

The performance of the phonetic decoding stage was measured by evaluating the *recognition accuracy* defined as:

$$P = 100 \cdot \frac{N - D - S - I}{N} \tag{11}$$

where N is the total number of phones in the test set, S the number of substitution errors, D deletion errors, and I insertion errors. The recognition accuracy was found to be *70.8%* for the unconstrained scenario and *83.3%* for the constrained scenario. Quite naturally, most of the substitution errors were made on phones with similar tongue and lip movements, such as {p,b,m}, {t,d,n}, {f,v}, {k,g,ŋ}, {ʃ,ʒ]}. However, some of these mismatches in the phonetic decoding would not necessarily lead to unintelligible synthesis; context effects could be used to advantage in a real communicative situation.

In order to evaluate the intelligibility of the synthesized speech, 15 test sentences were chosen randomly from the sentences of the database for which the performance of the phonetic decoder was similar to that on the entire dataset (P=80%). Seven native speakers of American English were asked to transcribe the 15 synthesized sentences. The quality of the transcription was evaluated with a word-based accuracy criterion, which is traditionally used in speech recognition, and is similar to the criterion defined in equation 11 (with N now the number of words). Even if the global quality of the synthesis was found to be much more acceptable with this approach compared to the GMM-based approach, only 50% of the words were correctly transcribed. This relative poor intelligibility may be partially explained by the non-realistic prosody often obtained with this approach. When evaluated at the phonetic level, the quality of transcription was relatively good though. Sentences that are relatively short (one prosodic group) and contain "common words", were often perfectly transcribed. This shows that, even consistently intelligible synthesis is as yet not possible, the system is, in some cases, able to generate an intelligible speech waveform from visual articulatory data only.

4. Visual Articulatory Feedback

Systems of visual articulatory feedback (VAF) aim at providing the speaker visual information about his/her own articulation. Several studies show that this kind of system can be useful for both speech therapy and *Computer Aided Pronunciation Training* (CAPT). The use of different types of sensors has been proposed in the literature. In [36], Wrench et al. used electro-palatography (EPG) to capture tongue-palate contact points. Patients can then observe derived visual patterns to place their tongue correctly in velar and alveolar regions. In [37], Bernhardt et al. proposed the use of ultrasound imaging. This way, speech therapists can freeze the image stream to show the patient the articulatory target to reach. In [38], Engvall proposed to use a talking-head for pronunciation training. The talking head was used in an *augmented speech scenario*, *i.e.* it displayed all speech articulators including the tongue and the velum. In this study, a wizard-of-Oz setup was used: a human listener chose the adequate pre-generated feedback, based on the user's pronunciation.

The visual articulatory feedback system developed at GIPSA-lab is also based on a 3D talking head used in an augmented speech scenario [39]. However, in our system, the talking head is animated *automatically* from the audio speech signal, using acoustic-to-articulatory inversion.

4.1 Talking head

The talking head used in the study consists of 3D models of the speech organs of a same speaker, built from MRI, X-ray and video data. The *jaw/lips/face* model is controlled by five parameters (*jaw height*, *jaw advance*, *lip protrusion*, *lower/upper lip heights*). The *jaw/tongue* model is also controlled by five parameters (*jaw height*, *tongue body* and *tongue dorsum* which control respectively the front-back and flattening-arching movements of the tongue, *tongue tip vertical/horizontal* which control the shape of the tongue tip). Figure 3 gives an example of different types of display of this talking head. As shown in [39], this 3D clone can be efficiently animated from a 2D EMA data stream (from the same speaker): the information provided by the location of the EMA coils is sufficient to inverse the articulatory models of the talking head, *i.e.* to estimate the control parameters that provide the best fit between the modeled 3D surfaces and the measured coordinates of the coils.

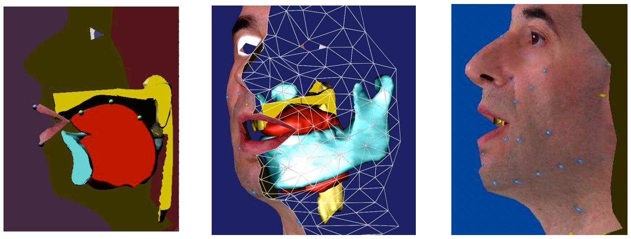

Figure 3. Talking head for different types of displays. Left: "augmented 2D view", center: "augmented 3D view", right: "complete face in 3D with skin texture"

4.2 Acoustic-to-articulatory inversion

In order to animate the talking head, trajectories of the EMA coils were recovered from the audio speech signal, using acoustic-to-articulatory inversion. Our previous work on acoustic-to-articulatory inversion is described in [40].

The recorded database consists of two repetitions of 224 VCVs (where C is one of the 16 French consonants and V is one of 14 French oral and nasal vowels), two repetitions of 109 pairs of CVC real French words, and 88 sentences, uttered by a male native French speaker. Articulatory movements 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. The database consisted of approximately 17 minutes of speech, long pauses being excluded.

The audio speech signal was parameterized by 25 mel-cepstrum coefficients (Blackman window, 25 frame length, 10 ms frame shift). For the GMM-based mapping, dynamic features were extracted by concatenating $2N+1$ adjacent acoustic frames $\mathbf{x}_t$, such as:

$$\overline{\mathbf{x}}_t = \left[\mathbf{x}_{t-N} \dots \mathbf{x}_t \dots \mathbf{x}_{t+N}\right] \quad (12)$$

The resulting vector $\overline{\mathbf{x}}_t$ is called here a *contextual feature vector*. This approach was found to be more efficient than the one described at section 3.2, based on the computation of the first and second derivatives[4]. As proposed in [19], principal component analysis was used to reduce the dimensionality of the contextual vectors. The numbers of principal components used for coding were determined by keeping the eigenvectors that account for 80% of the variance of the training set; typical value used on this database was 25 coefficients. The optimal value of N (see equation 12) was found to be 5, which corresponds to a 110 ms length window. For the HMM-based mapping, acoustic feature vector were simply completed with their first and second derivatives. EMA data were downsampled

[4] This approach has also been tested for the articulatory-to-acoustic mapping experiments described in section 3. However, it did not bring any improvement.

from 200 Hz to 100 Hz (in order to be compatible with the speech analysis rate) and low-pass filtered at 20 Hz.

The dataset was divided into 5 partitions. A jackknife technique was employed in order to increase the statistical relevance of our results. The accuracy of the inversion was measured by calculating, for each partition, the root mean square (RMS) error between the measured and the estimated EMA parameters, such as:

$$RMS_p = \sqrt{\frac{1}{D}\frac{1}{T_p}\sum_{d=1}^{D}\sum_{t=1}^{T_p}\left(\hat{y}_{d,t} - y_{d,t}\right)^2} \tag{13}$$

where T_p is the number of frames in partition p, D is the number of EMA parameters (12 in this study), $\hat{y}_{d,t}$ and $y_{d,t}$ are respectively the estimated and the measured position of the d^{th} EMA parameters at time t. A different formulation of the RMS error, in which the RMS is averaged over all the features, can be found in the literature. This RMS is called here *µRMS* and is defined as:

$$\mu RMS_p = \frac{1}{D}\sum_{d=1}^{D}\sqrt{\frac{1}{T_p}\sum_{t=1}^{T_p}\left(\hat{y}_{d,t} - y_{d,t}\right)^2} \tag{14}$$

For the GMM-based experiment, the optimal number of components in the mixture was found to be 64. The estimated articulatory data were smoothed by low-pass filtering (20 Hz cutoff frequency). The averaged *RMS and µRMS* (over the 5 partitions) were respectively *1.9 mm* and *1.7 mm*.

For the HMM-based mapping, articulatory-acoustic features were modeled, for each of the 34 phonetic classes, by a left-to-right, 5-state (3 emitting states), continuous multistream HMMs (with diagonal covariance matrices). For the HMM part related to acoustic stream, the optimal number of Gaussians per state was found to be 32. The performance of the mapping was evaluated using the jackknife procedure described above. With a mean recognition accuracy of 90%, the averaged *RMS* and *µRMS* were *1.7 mm* and *1.5 mm*, respectively.

Figure 4 gives an example of articulatory trajectories estimated from the audio signal with GMM and HMM-based mapping techniques. It also illustrated the animation of the talking head, derived from the estimated parameters (with the HMM-based mapping technique).

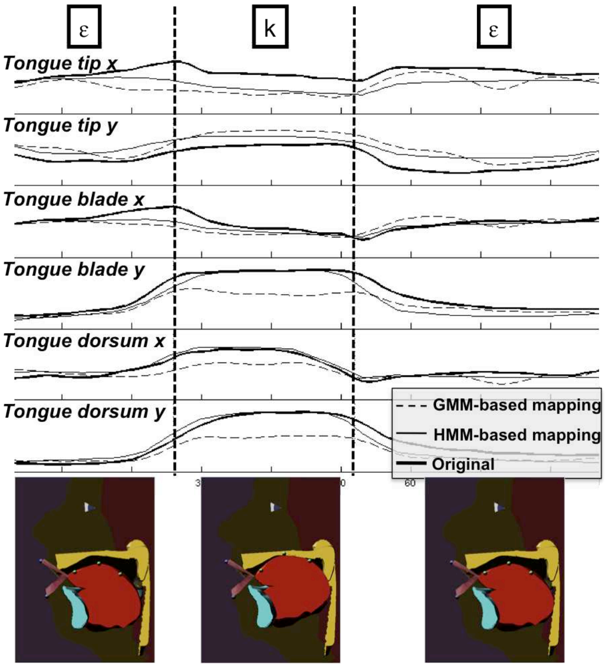

**Figure 4. Top: Example of articulatory trajectories estimated from the audio speech signal for the VCV [ɛkɛ] (only tongue EMA parameters are displayed).
Bottom: Corresponding animation of the talking head (only one frame per phoneme is displayed).**

5. Toward a Real-time Implementation

In order to be used in a realistic communicative situation, the different systems described in this paper have to run in real-time. Real-time implementation of the GMM and HMM-based mapping techniques is discussed in the following paragraphs.

The GMM-based mapping approach is a *frame-by-frame* process: as shown in equation 3, the estimated target features at time t ($\mathbf{y}_t$) depends only on the source features observed at the same time ($\mathbf{x}_t$). In a GMM-based mapping approach, the overall latency of the processing chain would mainly come from the feature extraction (and not from the conversion itself). We note $1/l$ the *feature extraction rate*; typical values for l are 16.6ms for the SSI (since the frame rate of the ultrasound and video streams is 60 fps) and 10ms for the VAF system (since the frame-shift parameter for the analysis of the audio signal is 10ms). The extraction of *dynamic features* introduces an additional delay in the processing chain. The computation of the second derivatives of the features extracted at time t (as described in section 3.2 for the SSI), requires an additional delay of l milliseconds, since it is based on the features extracted at time t-l but also at time t+l. The building of *contextual feature vectors* (as described in section 4.2) introduces an additional delay of $N.l$, milliseconds, where $2N+1$ is the number of adjacent frames taken into account. Using the GMM-based mapping approach, the minimum latency would be *l+l=33.2ms* for the SSI and *l+5l=60ms*, for the VAF system.

A real-time implementation of the HMM-based mapping approach is not as straightforward as for the GMM-based approach. As shown in equation 10, the HMM-based mapping is not a *frame by frame* process. The estimation of the target features requires first the decoding of the most likely HMM state sequence (for the given sequence of source vectors). This task is achieved by the Viterbi algorithm. However, this algorithm is based on a backtracking procedure, which requires both the first and the last observation to be available. Thus, this algorithm is not well adapted to a real-time implementation. Different approaches have been proposed in the literature to decode HMM online. In [42], the Viterbi algorithm is applied on a sliding window of consecutive observations. The advantage of this method is that the additional delay it adds to the processing chain is constant (and equal to the length of the sliding window). However, this method does not guarantee that the sequence of successive "local" paths is identical to the optimal path, *i.e* the path that would have been obtained if all the observations were taken into account. In [43], Bloit and Rodet proposed the *short-time Viterbi algorithm*, in which the Viterbi algorithm is applied on a sliding window of *variable length*. Under certain constraints on the HMM topology, the proposed algorithm guarantees that the successive decoded paths are identical to the optimal path. In this method, a constant maximum latency can also be obtained by forcing a suboptimal decoding when the window length exceeds a predefined threshold. We intend to implement the *short-time Viterbi* algorithm in the SSI and in the VAF system.

6. Conclusions and Perspectives

The paper presents two statistical mapping techniques, used to model the relationships between articulatory movements and the resulting speech sound. These techniques are based on the joint modeling of articulatory-acoustic data using respectively Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM). These methods were implemented (1) in an ultrasound-based silent speech interface for the conversion of tongue and lip images into speech and (2) in a visual articulatory feedback system that automatically animates a 3D talking-head from the speech sound.

For both systems, the best performance was obtained with the HMM-based method. In this method, external linguistic information (such as phonological or morphological information) can be introduced to constrain the mapping. We intend to implement a real-time version of the HMM-based mapping method, based on the *short time Viterbi* algorithm [43]. Future work will also investigate different mapping techniques recently described in the literature, such as (1) the low-delay implementation of the GMM-based mapping approach proposed by Toda et al. [44], which is based on the maximum likelihood estimation of the feature trajectories, and (2) the approach based on *trajectory HMM* proposed by Zen et al. in [45].

7. Acknowledgments

The authors would like to acknowledge useful discussions with Elie-Laurent Beneroya, Pierre Roussel, Gérard Dreyfus, Maureen Stone and Christophe Savariaux.

References

[1] O. Engwall, "Combining MRI, EMA and EPG measurements in a three-dimensional tongue model," *Speech Communication*, vol. 41, pp. 303-329, 2003.

[2] P. Birkholz, D. Jackèl, B.J Kröger. "Construction and control of a three-dimensional vocal tract model," in *Proc. of ICASSP*, 2006, pp. 873–876.

[3] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Modelling*, W.J. Hardcastle, A. Marchal, Kluwer: Academic Publishers,1990, pp. 131-149.

[4] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, C. Savariaux, "Three-dimensional articulatory modeling of tongue, lips and face, based on MRI and video images," *Journal of Phonetics*, vol. 30, pp. 533-553, 2002.

[5] S. Buchaillard, P. Perrier, Y. Payan, "A biomechanical model of cardinal vowel production: Muscle activations and the impact of gravity on tongue positioning," *J. Acoustical Society of America*, vol. 126, no. 4, pp. 2033-2051, 2009.

[6] J. Dang, K. Honda, "Construction and control of a physiological articulatory model," *J. Acoustical Society of America*, vol. 115, pp. 853-870, 2004.

[7] P. Badin, G. Fant, "Notes on vocal tract computation," *Speech Transmission Laboratory - Quaterly Progress Status Report - Stockholm*, vol. 2-3, pp.54-108, 1984.

[8] S. Maeda, "A digital simulation method of the vocal tract system," *Speech Communication*, vol. 1, pp. 199-229, 1982.

[9] M.M. Sondhi, J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 7, pp. 955-967, 1987.

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

[11] K. Mawass, P. Badin, G. Bailly, "Synthesis of French fricatives by audio-video to articulatory inversion," *Acta Acustica*, vol. 86, pp. 136-146, 2000.

[12] J. Hogden, A. Löfqvist, V. L. Gracco, I. Zlokarnik, P. Rubin, E. Saltzman, "Accurate recovery of articulator positions from acoustics: New conclusions based on human data," *J. Acoustical Society of America*, vol. 100, no. 3, pp. 1819-1834, 1996.

[13] S. Suzuki, T. Okadome, M. Honda, "Determination of articulatory positions from speech acoustics by applying

dynamic articulatory constraints," in *Proc. of ICSLP*, 1998, pp. 2251-2254.

[14] T. Kaburagi, M. Honda, "Determination of the vocal tract spectrum from the articulatory movements based on the search of an articulatory-acoustic database," in *Proc. of ICSLP*, 1998, pp. 433-436.

[15] J. Papcun, T.R. Hochberg, T.R. Thomas, F. Larouche, J. Zacks, S. Levy, "Inferring articulation and recognizing gestures from acoustics with a neural network trained on X-ray microbeam data," *J. Acoustical Society of America*, pp. 688-700, 1992.

[16] K. Richmond, "Estimating Articulatory Parameters from the Acoustic Speech Signal," *PhD thesis*, CSTR Edinburgh, 2002.

[17] C. T. Kello, D. C. Plaut, "A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters," *J. Acoustical Society of America*, vol. 116, pp. 2354-2364, 2004.

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

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

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

[21] A. Kain, "High-resolution voice transformation," *PhD thesis*, OGI School of Science & Engineering, Oregon Health & Science University, 2001.

[22] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, "The HTK Book," 2005, http://htk.eng.cam.ac.uk.

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

[24] B. Denby, T. Schultz, K. Honda, T. Hueber, et al., "Silent Speech Interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270-287, 2010.

[25] Y. Nakajima, H. Kashioka, K. Shikano, N. Campbell, "Non-audible murmur recognition," in *Proc. of Eurospeech*, pp. 2601-2604, 2003.

[26] V.-A. Tran, G. Bailly, H. Loevenbruck, T. Toda "Improvement to a NAM-captured whisper-to-speech system," *Speech Communication*, vol. 52, no. 4, pp. 314-326, 2010.

[27] T. Schultz, M. Wand, "Modeling coarticulation in EMG-based continuous speech recognition," *Speech Communication*, vol. 52, no. 4, pp. 341-353, 2010.

[28] C. Jorgensen, S. Dusan, "Speech interfaces based upon surface electromyography," *Speech Communication*, vol. 52, no. 4, pp. 354-366, 2010.

[29] M.J. Fagan, S.R Ell, J.M.Gilbert, E. Sarrazin, P.M. Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy," *Medical Engineering & Physics*, vol. 30, no. 4, pp. 419-425, 2008.

[30] T. Hueber, E. L. Benaroya, G. Chollet, et al., "Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips," *Speech Communication*, vol. 52, no. 4, pp. 288-300, 2010.

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

[32] T. Hueber, G. Aversano, G. Chollet, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel, M. Stone, "Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface," in *Proc. of ICASSP*, 2007, pp. I1245-I1248.

[33] S. Imai, K. Sumita, C. Furuichi, "Mel Log Spectrum Approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, pp. 10-18, 1983.

[34] I. Stylianou, "Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification," *PhD thesis*, Signal et Image, ENST Paris, Paris, 1990.

[35] K. Tokuda, T. Mausko, N. Miyazaki, T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Transactions on Information and Systems*, vol. E85-D, pp. 455-464, 2002.

[36] A., Wrench, F. Gibbon, A.M. McNeill, S. Wood, "An EPG therapy protocol for remediation and assessment of articulation disorders," in *Proc. of ICSLP*, 2002, pp. 965-968.

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

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

[39] P. Badin, F. Elisei, G. Bailly, Y. Tarabalka, "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.

[40] A. Ben Youssef, P. Badin, G. Bailly, "Acoustic-to-articulatory inversion in speech based on statistical models," in *Proc. of AVSP*, 2010, pp. 160-165.

[41] P. Badin, Y. Tarabalka, F. Elisei, G. Bailly, "Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding," *Speech Communication*, vol. 52, pp. 493-503, 2010.

[42] A., Seward, "Low-latency incremental speech transcription in the Synface project," in *Proc. of EuroSpeech*, 2003, pp. 1141-1144.

[43] J. Bloit, X. Rodet, "Short-time Viterbi for online HMM decoding: Evaluation on a real-time phone recognition task," in *Proc. of ICASSP*, 2008, pp. 2121-2124.

[44] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory," in *Proc. of Interspeech*, 2008, pp. 1076-1079.

[45] H. Zen, Y. Nankaku, K. Tokuda, "Continuous Stochastic Feature Mapping Based on Trajectory HMMs," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 417-430.