Statistical Mapping between Articulatory and Acoustic Data for an Ultrasound-based Silent Speech Interface

Thomas Hueber [1], *Elie-Laurent Benaroya*[2], *Bruce Denby* [3,2], *Gérard Chollet* [4]

[1]GIPSA-lab, UMR 5216/CNRS/INP/UJF/U.Stendhal, Grenoble, France
[2]Sigma Laboratory, ESPCI Paristech, Paris, France
[2]Université Pierre et Marie Curie, Paris, France
[3]LTCI/CNRS, Telecom ParisTech, Paris, France
thomas.hueber@gipsa-lab.grenoble-inp.fr, laurent.benaroya@espci.fr, denby@ieee.org, gerard.chollet@tsi.enst.fr

Abstract

This paper presents recent developments on our "silent speech interface" which converts tongue and lip motions, captured by ultrasound and video imaging, into audible speech. We present here two approaches to model the relationships between the observed articulatory movements and the resulting speech sound, which are based on the joint modeling of visual and spectral features using respectively Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM). The prediction of the voiced/unvoiced parameter from visual articulatory data only is also investigated using an artificial neural network (ANN). The proposed mapping techniques are evaluated on a continuous speech database containing one-hour of high-speed ultrasound and video sequences.

Index Terms: silent speech interface, GMM, HMM, ultrasound, video, multimodal, statistical mapping

1. Introduction

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. The design of SSIs has recently received considerable attention from the speech research community [1]. 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 [2] and [3]. Other approaches, based on completely non-acoustic features have also been proposed, as for example in [4] where electromyographic electrodes placed on the speaker's face (or on his neck in [5]) 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.

In our previous work ([6] [7]), the "visuo-acoustic" mapping problem, *i.e* the synthesis of an audible speech signal from visual articulatory data only, has been addressed using a concatenative synthesis approach. The system was composed of two distinct modules: a HMM-based "visual" phonetic decoder and a segmental vocoder exploiting an audiovisual unit dictionary in which each visual unit has an equivalent in the acoustic domain. Given a test sequence of visual features, a phonetic target sequence was first predicted. Then, a unit selection algorithm found in the dictionary the optimal sequence of units that best matched the input test data. Finally, the speech waveform was generated by concatenating the acoustic segments for all selected units. This approach gives encouraging results but presents some drawbacks. First, the quality of the synthesis depends strongly on the performance of the phonetic decoding and an error during the recognition stage corrupts necessarily the synthesis. Second, since the visual and the audio modality are treated separately, this approach does not model explicitly the dependency between the articulatory and the acoustic variables. In this paper, we investigate the use of statistical mapping techniques to address the visuo-acoustic conversion. We describe two techniques based on the joint modeling of articulatory and acoustic data using respectively Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM). We also address the problem of the prediction of the voiced/unvoiced parameter using an artificial neural network (ANN).

This article is organized as follows. The data acquisition and the feature extraction are described respectively in Section 2 and 3. Theoretical and practical aspects of the proposed GMM, HMM and ANN-based mapping techniques are detailed in Section 4. Experimental results are presented in section 5. Conclusions and perspectives are presented in the last section.

2. Data acquisition

The experimental setup used for data acquisition 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 (the ultrasound system and the video camera), we developed a dedicated software, named Ultraspeech[1]. Ultraspeech processes the ultrasound, video and audio streams in parallel using multithreading programming techniques and prevents data loss using a FIFO-based buffer management approach. This software was used to record simultaneously, and synchronously: the acoustic signal (16 bits, 16 kHz); the ultrasound stream (320x240 pixels) and the video stream (640x480 pixels). The ultrasound and video stream were both recorded at a frame rate of 60 fps (frames per second), which was 2 times higher than in our previous studies [6] [7] (for which a different acquisition setup were used).

The recorded dataset used in this work consists of the 1132 sentences of CMU ARCTIC corpus [8], uttered by a female

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

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 [9]), 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.

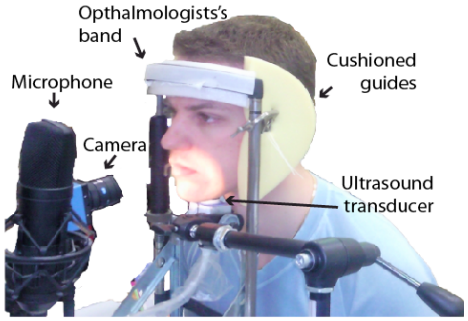

Figure 1: *Experimental setup used for data acquisition.*

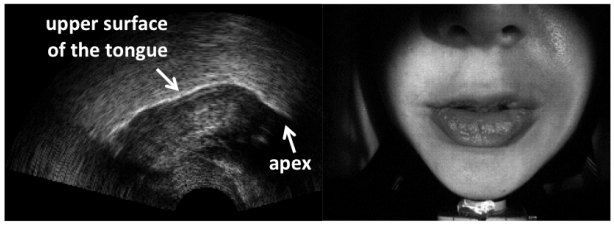

Figure 2: *Example of an ultrasound vocal tract image (in the midsagittal plane) with lip frontal view.*

3. Feature extraction

Regions of interest (ROI) selected in ultrasound and video images were first resized to 64x64 pixels. The *EigenTongues* decomposition technique was used to encode each ultrasound frame [10]. 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.

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 voiced/unvoiced characteristic and the fundamental frequency were also extracted. All the audio manipulations were done using the SPTK tools. Silence frames were removed from the training set using an automatic (threshold-based) silence detection method.

4. Visuo-acoustic mapping

4.1. GMM-based mapping

We investigate the use of the GMM mapping framework originally proposed by Stylianou [11] for voice conversion. In this study, we used the implementation proposed by Kain [12] which is based on the modeling of the joint probability density of source and target vectors $p(Z) = p(X,Y)$ with:

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

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:

$$W_m = \Sigma_m^{YX} (\Sigma_m^{XX})^{-1}, \quad b_i = \mu_m^Y - \Sigma_m^{YX} (\Sigma_m^{XX})^{-1} \mu_m^X$$
$$\text{with } \Sigma_m = \begin{bmatrix} \Sigma_m^{XX} & \Sigma_m^{XY} \\ \Sigma_m^{YX} & \Sigma_m^{YY} \end{bmatrix} \text{ and } \mu_m = \begin{bmatrix} \mu_m^X \\ \mu_m^Y \end{bmatrix} \quad (4)$$

and $P(c_m \mid \mathbf{x}_t)$, the probability that the source vector "belongs" to the m^{th} component, defined as:

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

where $N(., \mu, \Sigma)$ is a normal (Gaussian) distribution with mean μ and covariance matrix Σ. In our implementation, the GMM is initialized using the *k-means* algorithm.

4.2. HMM-based mapping

In the proposed HMM-based mapping approach, the sequence of target vectors $\hat{\mathbf{y}}$, predicted from the given sequence of source vectors **x,** is defined as $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \{ p(\mathbf{y} \mid \mathbf{x}) \}$ with:

$$p(\mathbf{y} \mid \mathbf{x}) = p(\mathbf{y} \mid q, \lambda) \cdot P(q \mid \mathbf{x}, \lambda) \quad (6)$$

where λ is the parameters set of the HMM and q the HMM state sequence. As shown in Equation 7, the 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 of the 40 phonetic classes. Two streams are dedicated to the modeling

of the visual features (ultrasound/video), one stream is used to model the spectral features (mel-cepstrum coefficients). For each stream, the emission probability density of each state is modeled by a GMM with diagonal covariance matrix. The initialization of the HMMs requires temporal segmentation of the training data at phonetic level. As articulatory and acoustic data were recorded synchronously, this segmentation was obtained by annotating the acoustic data. This was done automatically using a forced-alignment procedure and an initial set of acoustic HMMs trained on the multi-speaker TIMIT database. After initialization, HMM parameters were estimated using a standard procedure (similar to that described in [13]): models 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 was then introduced in the modeling. Triphone models were created by adding information about left and right contexts to the phone models. A tree-based state-tying strategy based on the Minimum Description Length (MDL) criterion, was adopted to address the problem of data sparsity (triphones having only a few occurrences in the training dataset). Each resulting multistream HMM were then split into two distinct HMMs: a 2-streams "visual HMM" and a 1-stream "acoustic HMM". Visual HMMs were finally refined by increasing incrementally the number of Gaussian mixture components.

The prediction of the sequence of acoustic feature vector **y,** for a given test sequence of visual feature vectors **x,** was achieved in two stages. First, phonetic and state decoding was performed by the visual HMMs, using the Viterbi algorithm. Second, given the predicted sequence of phones and the decoded HMM state sequence, target vector sequence was inferred by the acoustic HMMs, using the speech parameter generation algorithm proposed by Tokuda for HMM-based speech synthesis [14]. This algorithm determines the vector sequence that maximizes the likelihood of the model with respect to a continuity constraint on the predicted feature trajectories. 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. All the procedures involving HMM manipulations described in this paper, are done using the HTK and HTS toolkits.

4.3. Prediction of the voiced/unvoiced parameter

In this study, the synthesis of the audio speech signal is achieved using a MLSA digital filter derived from the predicted mel-cepstrum coefficients [15]. The generation of the excitation signal requires the prediction of the voiced/unvoiced parameter as well as the pitch for voiced frames. In this paper, we investigate the prediction of the voiced/unvoiced parameter from visual articulatory data, using an artificial neural network (ANN). A feed-forward neural 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 mean squared error (MSE) was used as the cost function.

5. Results & Discussion

The partitioning of the 1132 recorded sentences was done as follow. 82 sentences were used as a validation set for the determination of the model hyper-parameters which are: (a) the optimal number of Gaussians for the GMM/HMM models (which was found to be 32 for the GMM and 4 for the HMM), (b) the model insertion penalty for the phonetic decoding stage in the HMM-based mapping experiment (which was found to be -20 for the unconstrained scenario and -150 for constrained scenario), (c) the optimal number of hidden neurons for the prediction of the voiced/unvoiced parameter (which was found to be 10). 900 sentences where used for training, the remaining 150 sentences composed the test set.

The quality of the mapping between visual and spectral features was evaluated by calculating the *Mel-cepstral distortion* between the target and the predicted mel-cepstrum coefficients, defined as:

$$Mel-CD[dB] = \frac{10}{\ln 10}\sqrt{2\sum_{d=0}^{24}(\hat{m}_d - m_d)^2} \qquad (7)$$

For the GMM-based mapping experiment, the Mel-cepstral distortion (with the 95 % confidence interval calculated with a normal approximation) was found to be 7.6 ± 0.03 dB if the 0th cepstral dimension, *i.e* the component known to correspond to overall signal power, was taken into account, and 6 dB ± 0.02 dB if this term was ignored. As expected, it was difficult to estimate correctly the speech signal power only from the corresponding articulatory motions.

For the HMM-based mapping experiment, the performance of the intermediate phonetic decoding stage was measured by evaluating the *recognition accuracy* defined as:

$$P = 100 \cdot \frac{N - D - S - I}{N} \qquad (8)$$

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 62% for the unconstrained scenario and 70% 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. The mel-cepstral distortion obtained in the unconstrained and constrained decoding were respectively 7.2 ± 0.03 dB and 7.1 ± 0.03 dB; 5.8 ± 0.02 dB and 5.6 ± 0.02 dB if the 0th cepstral dimension was ignored. The mel-cepstral distortion obtained when the phonetic target is given (*i.e* P = 100%), was found to be 5.4 ± 0.01 dB and 4.6 ± 0.01 dB when excluding the first mel-cepstrum coefficient. The HMM-based approach outperforms the GMM-based approach, even if the decoded phonetic sequence contains some errors.

The accuracy of the voiced/unvoiced binary classifier (section 4.3), 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; and "corpus-effects", since the CMU Arctic

corpus does not contain the same number of examples for each phonetic class.

A constant pitch was used here for the synthesis of the audio signal (for the frames predicted as "voiced"). In order to evaluate the intelligibility of the synthesized speech, 3 native speakers of American English were asked to transcribe the synthetic speech signals corresponding to 15 sentences randomly extracted from the test set. The global quality of the synthesis was found to be more acceptable with the HMM-based approach compared to the GMM-based approach. However, even if some sentences were well transcribed (especially the short ones and those containing "common words"), this preliminary subjective evaluation revealed that it was not possible to synthesize intelligible speech "consistently", neither with the GMM-based mapping approach, nor with the HMM-based approach. For now, statistical approaches based on the joint modeling of the visual and acoustic data do not outperform our previous concatenative approach, in which the two modalities were modeled separately. To measure the impact of the "joint modeling" on the phonetic decoding stage (for the HMM-based mapping approach), we evaluated the performance of a HMM-based decoder trained only on the two visual modalities (following the procedure described in section 4.2). The recognition accuracy was found to be 70.8% for the unconstrained scenario and 83.3% for the constrained scenario, *i.e* approximately 10% higher than the performance obtained with the joint modeling approach. The use of alternative strategies to combine the visual and acoustic modalities at the classifier level is envisioned.

6. Conclusions and Perspectives

The paper presents recent developments on our "silent speech interface", driven by ultrasound and video images of the vocal tract. Two techniques, based respectively on the joint modeling of articulatory-acoustic data using Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) have been proposed to model the relationships between articulatory movements and the resulting speech sound. These techniques have been evaluated on a database containing one-hour of high-speed ultrasound and video data. The best mapping was obtained with the HMM-based method in which external linguistic information (such as phonological or morphological information) can be introduced to constrain the mapping.

Future work will focus on the improvement of the visuo-acoustic mapping. We will investigate the adaptation of different mapping techniques recently described in the literature, such as the GMM-based approach proposed by Toda et al. [16] based on the maximum likelihood estimation of the feature trajectories, and the approach proposed by Zen et al. in [17], which is based on trajectory HMM.

The specificities of silent articulation will also be studied. Preliminary tests showed that the performance of statistical models trained on "vocalized" visual speech decreases when they are used to decode "silent" visual speech (if no model adaptation scheme is applied). This may reveal some differences in terms of articulatory strategies between these two production modes, which we have started to describe in a pilot study [18], using electromagnetic articulography (EMA).

7. Acknowledgments

The authors would like to acknowledge useful discussions with Pierre Roussel, Gérard Dreyfus and Maureen Stone.

8. References

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

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

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

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

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

[6] Hueber, T., Benaroya, E.L., Chollet, G., Denby, B., Dreyfus, G., Stone, M., "Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips", Speech Communication, 52(4), pp. 288-300, 2010.

[7] Hueber, T., Chollet, G., Denby, B., Dreyfus, G., Stone, M., "Towards a Segmental Vocoder Driven by Ultrasound and Optical Images of the Tongue and Lips", in Proc. of Interspeech, pp. 2028-2031, 2008.

[8] Black, A. W., Lenzo, K., "Building voices in the Festival speech synthesis system", http://festvox.org/bsv, 2000.

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

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

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

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

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

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

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

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

[17] Zen, H., Nankaku, Y., Tokuda, K., "Continuous Stochastic Feature Mapping Based on Trajectory HMMs", IEEE Trans. on Audio, Speech, and Language Processing, vol. 19, no. 2, pp. 417-430, 2011.

[18] Hueber, T., Badin, P., Savariaux, C., Vilain, C., Bailly, G., "Differences in articulatory strategies between silent, whispered and normal speech ? a pilot study using electromagnetic articulography", in Proc. of ISSP, to appear, 2010.