Toward a multi-speaker visual articulatory feedback system

Atef Ben Youssef, Thomas Hueber, Pierre Badin, Gérard Bailly

GIPSA-lab (DPC / ICP), UMR 5216 CNRS / INP / UJF / U. Stendhal, Grenoble, France
`{Atef.Ben-Youssef, Thomas.Hueber, Pierre.Badin, Gerard.Bailly}@gipsa-lab.grenoble-inp.fr`

Abstract

In this paper, we present recent developments on the HMM-based acoustic-to-articulatory inversion approach that we develop for a "visual articulatory feedback" system. In this approach, multi-stream phoneme HMMs are trained jointly on synchronous streams of acoustic and articulatory data, acquired by electromagnetic articulography (EMA). Acoustic-to-articulatory inversion is achieved in two steps. Phonetic and state decoding is first performed. Then articulatory trajectories are inferred from the decoded phone and state sequence using the maximum-likelihood parameter generation algorithm (MLPG). We introduce here a new procedure for the re-estimation of the HMM parameters, based on the Minimum Generation Error criterion (MGE). We also investigate the use of model adaptation techniques based on maximum likelihood linear regression (MLLR), as a first step toward a multi-speaker visual articulatory feedback system.

Index Terms: Acoustic-articulatory inversion, ElectroMagnetic Articulography (EMA), Hidden Markov Model (HMM), Minimum Generation Error (MGE), Speaker adaptation, Maximum Likelihood Linear Regression (MLLR).

1. Introduction

Systems of visual articulatory feedback aim at providing the speaker with 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) [1]. The visual articulatory feedback system developed at GIPSA-lab is based on a 3D talking head used in an augmented speech scenario, *i.e.* it displays all speech articulators including usually non visible articulators such as the tongue. In this system, the talking head is animated automatically from the audio speech signal, using acoustic-to-articulatory inversion. For this purpose, we developed different inversion methods based on the joint modeling of acoustic and articulatory data (acquired by electromagnetic articulography - EMA), using statistical models such as Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) [2].

The use of supervised machine learning techniques for acoustic-to-articulatory inversion has already been proposed in the literature. HMMs trained on parallel acoustic and articulatory speech data segmented at the phonetic level, have been used in [3], [4], [5] and [2]. GMMs have been used in [2] and [6]. Artificial neural network (ANN) and Support Vector Machines (SVM) have been used respectively in [7] and [8] However, these studies do not allow to conclude on the optimal inversion method since data, speakers and languages are not comparable.

Since our goal is to provide "any" speaker with visual articulatory feedback, the inversion system need to be robust and easy to adapt. With that in mind, we present here recent developments on our HMM-based inversion system that have lead to significant improvements. Among them, we introduce a new training procedure based on the optimization of the Minimum Generation criterion (MGE). We also present a first approach to address the speaker adaptation problem. Speaker adaptation can be done in both acoustic and articulatory domains. In the present work, we propose to adapt our multimodal HMM models in the acoustic domain by using maximum likelihood linear regression.

This article is organized as follows. The HMM-based acoustic-to-articulatory inversion system is described in Section 2. The proposed speaker adaptation technique is presented in Section 3. The methodology used for evaluation and experimental results are presented in section 4. Conclusions and perspectives are presented in the last section.

2. HMM-based speech inversion system

2.1. Baseline system

In the proposed HMM-based mapping approach (previously published by Ben Youssef et al. in [2]), the sequence of articulatory vectors, predicted from the given sequence of acoustic vectors x, is defined as[1] $\hat{y} = \arg\max_{y}\left\{p(y|x)\right\}$ with

$$p(y \mid x) = p(y \mid \lambda, q)P(\lambda, q \mid x) \qquad (1)$$

where λ represents the parameters set of the HMM and q the HMM state sequence. By applying the Bayes rule, we obtain

$$p(y \mid x) = p(y \mid \lambda, q)p(x \mid \lambda, q)P(\lambda) \qquad (2)$$

As shown in Equation 2, the HMM-based mapping can be achieved by a recognition stage followed by a synthesis stage, which means: (1) finding the most likely phonetic and state sequence for a given source vector (and a set of a priori information provided by a statistical language model), and (2) inferring the target vector from the decoded state sequence.

In the training stage, a left-to-right, 3-state multi-stream HMM is trained on articulatory-acoustic data for each phonetic class. The first stream is dedicated to the modeling of acoustic feature; the second stream is used to model the articulatory features. For each stream, the emission probability density of each state is modeled by a GMM with diagonal covariance matrix. HMM were initialised and trained by the Baum Welch algorithm based on the Maximum Likelihood (ML) criterion.

Due to coarticulatory effects, it is unlikely that a single context-independent HMM could optimally represent any given allophone. Contexts were therefore grouped in *context classes* for both vowels and consonants separately. Based on the matrix of Mahalanobis distances of the coils coordinates between the centre frame of each pair of phoneme means, hierarchical clustering generated six coherent classes for vowels (/a ɛ ɛ̃/, /ø œ œ̃/, /e i/, /o ɔ ɑ̃ ɔ̃/, /y/, and /u/), and ten coherent classes for consonants (/p b m/, /t d n/, /s z/, /ʃ ʒ/, /k g/, /f v/, /l/, /ʁ/, /j ɥ/, and /w/). Context-dependent

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

HMMs were then trained, using various contextual schemes: phonemes without context (no-ctx), with left (L-ctx) or right context (ctx-R), and with both left and right contexts (L-ctx-R). A tree-based state-tying strategy based on the Minimum Description Length (MDL) criterion, was adopted to address the problem of data sparsity (biphones or triphones having only a few occurrences in the training dataset).

Each resulting multi-stream HMM were then split into two distinct HMMs: an "acoustic HMM" and an "articulatory HMM". Acoustic HMMs were finally refined by increasing incrementally the number of Gaussian mixture components.

The prediction of the sequence of articulatory feature vectors**,** for a given test sequence of acoustic feature vectors**,** was achieved in two stages. First, phonetic and state decoding was performed by the Viterbi algorithm using the acoustic HMMs. A bigram phonetic language model trained on one year of the newspaper "Le Monde" (year 2003) was used (thus, the recognised phoneme sequences respect French phonotactics). Second, given the predicted sequence of phones and the decoded HMM state sequence, the target vector sequence was inferred by the maximum-likelihood parameter generation algorithm (MLPG) [9], using the articulatory HMMs.

2.2. Minimum Generation Error (MGE) criterion

In this paper, we introduce a new training procedure, based on the MGE approach (Minimum Generation Error) initially proposed by Wu *et al.* in [10] for HMM-based text-to-speech synthesis. We propose to adapt this technique to the acoustic-to-articulatory mapping problem. The training procedure is performed as follows. The parameters of single Gaussian articulatory HMMs are first estimated by maximising the likelihood of the model given the training data (using the standard Baum-Welch algorithm). Then, the articulatory trajectories which maximize the likelihood of the current set of articulatory HMMs are generated using the MLPG algorithm. The state sequence used to drive this intermediate synthesis stage is obtained by forced-alignment of the acoustic data at the phonetic level. The *generation error* is defined as the Euclidean distance between the generated and the measured articulatory trajectories. Given this error, the parameter of the articulatory HMMs (mean and variance) are finally updated using the equations detailed in [10]. In our implementation, this procedure is iterated 5 times.

3. Speaker adaptation

Compared to other approaches (based on ANNs or GMMs for instance), the mapping between acoustic and articulatory modalities is not performed at the feature level, but at the phonetic level. Based on this consideration, we investigated the possibility to perform the inversion by directly decoding the new speaker's speech at this level. Because the accuracy of the inversion process depends strongly on the performance of this decoding stage, it is crucial to adapt the reference speaker models (*i.e.* the speaker used to build the original speech inversion system). This additional stage makes the models of the reference speaker compatible with the new speaker's voice, but also with a different acoustic environment.

To build the adaptation database, the new speaker is asked to utter a corpus of adaptation sentences. The adaptation procedure is performed as follows. First, the speech signal is automatically segmented at the phonetic level using forced-alignment and the acoustic models trained on the reference subject. Second, Maximum Likelihood Linear Regression (MLLR) technique is used to adapt each acoustic HMMs. MLLR estimates linear transformations for models parameters to maximise the likelihood of the adaptation data [11].

4. Evaluation

4.1. Databases

The database used in this study (and also in our previous study [5]) 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 (referred to as the *reference speaker PB*) (approximately 5100 phones). Articulatory movements were recorded synchronously with the audio signal using the Carstens 2D EMA system (AG200). Six coils were used to measure articulators kinetics: a jaw coil was attached to the lower incisors, whereas three coils were attached to the tongue tip, the tongue middle, and the tongue back; upper and lower lip coils were attached to the boundaries between the vermilion and the skin in the midsagittal plane. Two coils were used for head alignment. The audio-speech signal was recorded at a sampling frequency of 22 kHz and was parameterized by 13 MFCC (Blackman window, 25 ms frame length, 10 ms frame shift). EMA coordinates were recorded at 500 Hz, low-pass filtered at 20 Hz in order to reduce noise, and down sampled to 100 Hz to fit the analysis rate of the acoustic signal. The database, which consists of approximately 17 minutes of speech, long pauses being excluded, was labelled at the phonetic level (using a force-alignement procedure and a manual check).

In order to evaluate the proposed speaker adaptation technique, audio database were recorded from three native French speakers: male speaker TH recorded the same speech material as the reference speaker PB; another male speaker GB and female speaker AC recorded a different corpus, consisting of 240 sentences initially designed for speech synthesis purpose.

4.2. Evaluation

The accuracy of the inversion was measured in different ways. First, we calculated the root mean square (RMS) error between the measured and the estimated EMA parameters, such as:

$$RMS = \sqrt{\frac{1}{D}\frac{1}{T}\sum_{d=1}^{D}\sum_{t=1}^{T}\left(\hat{y}_{d,t} - y_{d,t}\right)^2} \quad (3)$$

where T is the number of frames in the test set, D is the number of EMA parameters (12 in this study), $\hat{y}_t$ and y_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 (as in [3], [4], [5] or [6]). This RMS is called here µRMS and is defined as:

$$\mu RMS = \frac{1}{D}\sum_{d=1}^{D}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\hat{y}_{d,t} - y_{d,t}\right)^2} \quad (4)$$

We also calculated the "Pearson Product-Moment Correlation Coefficient" (PMCC) which measures the level of amplitude similarity and synchrony of the trajectories. Finally, we calculated the "recognition accuracy" to assess specifically the phonetic decoding stage.

A 5-fold cross-validation procedure was used for the evaluation: the database was split into 5 partitions approximately homogeneous from the point of view of phone

Table 1. *µRMSE, RMSE (mm) and PMCC for the HMM-based inversion: (a) inversion from audio and labels input (perfect recognition); (b) inversion from audio only.*

		no-ctx			**L-ctx**			**ctx-R**			**L-ctx-R**		
		µRMSE	RMSE	PMCC	µRMSE	RMSE	PMCC	µRMSE	RMSE	PMCC	µRMSE	RMSE	PMCC
(a)	MLE	1,80	1,87	0,89	1,49	1,54	0,93	1,50	1,55	0,93	1,39	1,44	0,94
	MGE	1,55	1,61	0,92	1,34	1,38	0,94	1,35	1,39	0,94	1,31	1,34	0,94
(b)	MLE	1,88	1,96	0,78	1,64	1,70	0,91	1,59	1,65	0,92	1,63	1,69	0,91
	MGE	1,71	1,78	0,90	1,54	1,60	0,92	**1,48**	**1,53**	**0,93**	1,57	1,63	0,92

distribution. Each partition was used once as the test set while the other 4 partitions composed the training set. RMS, µRMS PMCC, and recognition rates were averaged over the five test partitions.

In order to estimate the contribution of the new MGE-based training procedure, it is needed to evaluate the synthesis stage independently from the recognition stage. In that purpose, we simulated a "perfect recognition" by aligning the original phonetic labels on the acoustic waveform. Table 1a shows that using MGE decreases the RMSE by an amount between 0.08 mm and 0.26 mm, depending on the context type.

Table 1.b displays the performances of the inversion from audio signal alone. Best results were obtained with context-dependant models including information about the right phonetic context (ctx-R). In this case, µRMSE was found to be 1.48 mm (RMSE = 1.53 mm, PMCC = 0.93, recognition accuracy = 86.20%).

4.3. Articulatory recognition

Since no articulatory data were acquired for 3 of the 4 speakers used in this study, it is impossible to determine the RMSE between the measured and the predicted articulatory trajectories. Therefore, we have based the evaluation on the automatic "articulatory recognition" of the predicted trajectories. In that purpose, we have trained an HMM-based phonetic decoder on the articulatory data of the reference speaker PB.

Contrarily to the acoustic recognition stage which determines *phonemes*, this articulatory recognition procedure was designed to recognise *phoneme classes* (groups of phonemes). As the context classes, described in section 2.1, were established based on articulatory distances, they take naturally into account the fact that all features are not exhaustively present in the EMA data (voicing cannot be measured; no velum coil was available in our recording setup). Therefore, these 16 context classes have been used as *phoneme classes* for the articulatory recognition. Note that phonemes differing only by voicing or velum position are grouped in the same classes (*e.g.* /p b m/, /t d n/, /k g/, etc.). In addition, two extra *phoneme classes* were used: one for the schwa and the short pause, and the other for the long pause at the boundaries of sentences. Finally, these 18 articulatory *phoneme classes* were used to train and to recognize the articulatory trajectories.

The HMM-based articulatory recognition system was built using a procedure similar to the one described in section 2.1. The performance of this system was evaluated on the articulatory data of the reference speaker PB, using the same 5-fold cross-validation procedure that the one described previously. Best performance was obtained using context dependent model (with right context) and 8 Gaussians per state. In this case, the recognition accuracy was found to be 84.84 %. These articulatory HMMs are used to evaluate the articulatory trajectories generated from the acoustic signal of any new speaker.

4.3.1. Evaluation of the predicted articulatory trajectories of the reference speaker

In order to establish a baseline for the assessment of inversion of new speakers by automatic articulatory recognition, we have computed the articulatory recognition rates for the original speaker. It was also deemed important to decide which data should be used for training this reference articulatory recognition system: (1) original articulatory trajectories, (2) articulatory trajectories recovered by inversion from the audio signal alone, or (3) articulatory trajectories recovered by inversion from both audio signal and labels (perfect acoustic recognition). All combinations were finally evaluated, for cxt-R contexts, using 5-fold cross validation for each combination, as can be seen in Table 2. Interestingly, we observed that recognition rate were always higher for synthesised trajectories than for measured ones, whatever the training corpus: this might be ascribed to the fact that synthesised articulatory trajectories are more lawful or regular than measured ones, since they are produced by models that constitute simplified representations of data, good though they can be. We observe also that rates are higher for models trained on measured data (except in the case of testing and training with data obtained by perfect recognition). It was therefore decided to use models trained on measured data.

Table 2. *"Phoneme class" articulatory recognition accuracy (with 5-fold cross-validation) for speaker PB using ctx-R; measured EMA (1), EMA synthesized from audio only (2), and from both audio and labels (perfect recognition) (3).*

Context « ctx-R »		**Test**		
		(1)	(2)	(3)
Train	(1)	**84.84**	**84.56**	**88.39**
	(2)	**58.12**	**79.25**	**83.78**
	(3)	**57.06**	**84.44**	**90.04**

4.3.2. Evaluation of the predicted articulatory trajectories of new speakers

The acoustic adaptation technique described at section 3 was applied to the acoustic HMMs trained on 4/5 of the original speaker's corpus using 4/5 of the new speaker's corpus; the remaining 1/5 of the new speaker's corpus was used to test both acoustic recognition and articulatory recognition. Note that in order to avoid the complexity and possible overtraining that may occur when using 5-fold cross-validation for both the reference articulatory training and the new speaker adaptation, all the test have been applied using the first 4/5 of the corpus for training or adaptation and the last 1/5 for testing. For subject TH, the sentences used for the adaptation were the same as those used for training the initial acoustic HMMs on PB.

Table 3 shows the various acoustic recognition rates and articulatory recognition rates of the inversed trajectories. We observe that subject TH has performances very close to those of reference PB; this could be explained by the fact that his corpus was recorded in an imitation mode: he imitated each sentence after being prompted by the audio recording from PB, which would favour similar dynamics. Oppositely, the worst performances are obtained for female speaker AC, both at acoustic and articulatory levels, which may be ascribed to the sex difference, and the difference in size and content of the corpus – allowing only 192 adaptation sentences. Intermediary results are obtained for speaker GB, with the intriguing degradation of the articulatory score compared to the fairly good acoustic one. However, a more thorough analysis of the acoustic recognition has shown that the accuracy rates for the set of 631 allophones in right context (ctx-R) were much lower than for the 36 French phonemes for this speaker (see Table 3). This was confirmed by the observation of the detailed recognition rates for vowels which showed some confusion between different contexts.

Table 3. *Acoustic and articulatory recognition accuracy for all the speakers, using 1/5 of the corpus for testing.*

Accuracy	**PB**	**TH**	**GB**	**AC**
Acoust. Phonemes	85.92	83.77	79.12	62.81
Articulation	83.70	82.23	69.46	56.77
Acoust. Allophones	79.88	76.53	66.77	48.01

5. Conclusions and perspectives

This paper presents latest developments on our HMM-based acoustic-to-articulatory inversion system that we develop for a "visual articulatory feedback" system. The introduction of a new training procedure based on the optimization of the Minimum Generation Error (MGE) criterion has lead to significant improvements (about 10%). As a first step toward a multi-speaker system, we also investigated the use of a MLLR model adaptation technique. The quality of the articulatory trajectories was evaluated by measuring the performance of an "articulatory HMM-based phonetic decoder". Recognition accuracies range between 56.8 % and 82.2 % for three speakers, compared to 83.7 % for the original speaker, demonstrating the interest of the method.

The next step of our development will be to test more speakers, and to study more explicitly the influence of the nature and size of the adaptation corpus. It will also be of great importance to investigate adaptation methods in the case of non-native speaker adaptation (*e.g.* [12]).

Finally, in the framework of Computer Aided Pronunciation Training (CAPT), we aim to use this speaker adaptation approach in our visual articulatory feedback system, based on acoustic-to-articulatory speech inversion.

6. Acknowledgements

The authors would like to acknowledge Christophe Savariaux for its involvement in the recording of EMA data. This work was partially supported by the French ANR-08-EMER-001-02 project *ARTIS*.

7. References

[1] Badin, P., Ben Youssef, A., Bailly, G., Elisei, F., and Hueber, T., "Visual articulatory feedback for phonetic correction in second language learning," in *L2SW, Workshop on "Second Language Studies: Acquisition, Learning, Education and Technology"*, Tokyo, Japan, 2010, pp. P1-10.

[2] Ben Youssef, A., Badin, P., and Bailly, G., "Acoustic-to-articulatory inversion in speech based on statistical models," in *AVSP 2010*, Hakone, Kanagawa, Japon, 2010, pp. 160-165.

[3] Hiroya, S. and Honda, M., "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," *IEEE Trans. Speech and Audio Processing,* vol. 12, pp. 175-185, March 2004.

[4] Zhang, L. and Renals, S., "Acoustic-articulatory modeling with the trajectory HMM," *IEEE Signal Processing Letters,* vol. 15, pp. 245-248, 2008.

[5] Ling, Z.-H., Richmond, K., and Yamagishi, J., "An Analysis of HMM-based prediction of articulatory movements," *Speech Communication,* vol. 52, pp. 834-846, 2010.

[6] Toda, T., Black, A. W., and Tokuda, K., "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication,* vol. 50, pp. 215-227, 2008/3 2008.

[7] Richmond, K., "Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion," in *Advances in Nonlinear Speech Processing (Lecture Notes in Computer Science 4885)*. vol. 4885/2007 Berlin, Heidelberg, Germany: Springer Verlag, 2007, pp. 263-272.

[8] Toutios, A. and Margaritis, K., "A support vector approach to the acoustic-to-articulatory mapping," in *Interspeech 2005*, Lisbon, Portugal, 2005, pp. 3221-3224.

[9] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T., "Speech parameter generation algorithms for HMM-based speech synthesis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, 2000, pp. 1315-1318.

[10] Wu, Y.-J. and Wang, R.-H., "Minimum generation error criterion for tree-based clustering of context dependent HMMs," in *interspeech 2006* Pittsburgh, USA, 2006, pp. 2046-2049.

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

[12] Ohkawa, Y., Suzuki, M., Ogasawara, H., Ito, A., and Makino, S., "A speaker adaptation method for non-native speech using learners' native utterances for computer-assisted language learning systems," *Speech Communication,* vol. 51, pp. 875-882, 2009.