PERFORMANCES OF UNSUPERVISED HMM IN ACOUSTIC-TO-ARTICULATORY INVERSION

Hélène Lachambre, Lionel Koenig, Régine André-Obrecht

IRIT - University of Toulouse - 118 route de Narbonne, F-31062 Toulouse, France

lachambre@irit.fr, lionel.koenig@gmail.com, obrecht@irit.fr

ABSTRACT

In the context of the acoustic-to-articulatory inversion, various unsupervised HMM based feature-mapping methods are assessed and compared. In a previous study we introduced an unsupervised HMM as an alternative model to the phone-HMM. We propose here to evaluate this approach using different inversion methods, in order to assess the behavior of our model and its compatibility with the most efficient inversion algorithms available. The best configuration leads to similar root mean square error (up to 1.44 mm) than phoneme-based HMM.

***Index Terms*—** Acoustic-to-articulatory inversion, Unsupervised Hidden Markov Models, Trajectory models

1. INTRODUCTION AND RELATION TO PREVIOUS WORK

Acoustic-to-articulatory inversion aims to estimate the shape of the mouth - position of jaw, tongue, lips - from an audio speech signal.

Most strategies consist of jointly modeling acoustic and articulatory data, and using the correlation between these data to generate the articulatory parameters from the acoustic ones.

Two main approaches have been explored for the joint modeling of acoustic and articulatory data: Gaussian Mixture Models (GMM), and Hidden Markov Models (HMM).

In the GMM approach [1, 2, 3], the joint distribution of acoustic and articulatory parameters is a Gaussian Mixture Model. The inversion is performed either by a classical mapping using minimum mean square error (MMSE) [1] or maximum likelihood estimation (MLE) criterion [1, 2], or by considering an explicit relationship between static and dynamic features with trajectory GMMs [3].

The HMM approach [3, 4, 5, 6] lies on two HMMs, which model respectively the acoustic and the articulatory parameters. Both HMMs are usually jointly trained using multi-stream HMMs. The inversion step always begins using an audio decoding with the acoustic HMM and determining a sequence of phone states. The articulatory parameter sequence is synthesized using a trajectory HMM [4, 5, 6, 7]. An improvement [3] is obtained by training a trajectory HMM to decode the acoustic observations as well as synthesizing the articulatory parameters. All these approaches need to have a phoneme labeled acoustic-articulatory database.

In a previous study [8], we proposed an unsupervised joint HMM as an alternative model. This method has shown encouraging results using very basic inversion methods. To complete this study, we explore different acoustic decoding, and different inversion methods, in order to assess the behavior of the unsupervised model and its compatibility with the most efficient state-of-the-art inversion algorithms.

The description of the experimental corpora in section 2 is followed by a summary of the basics of the unsupervised HMM training. Then we present in section 4 the different strategies used for the inversion procedure with a discussion of their assessment in section 5.

2. ACOUSTIC-ARTICULATORY DATASETS

2.1. ARTIS corpus

This acoustic-articulatory corpus has been developed by the GIPSA-Lab in Grenoble, France. It has already been used to assess several studies on acoustic-to-articulatory inversion [4, 8, 9, 10].

This corpus contains data recorded by one French male speaker. It is composed of 224 Vowel-Consonant-Vowel (VCV), two repetitions of 109 short Consonant-Vowel-Consonant (CVC) French words, 68 short sentences and 20 long sentences.

Articulatory data are recorded thanks to an Electromagnetic Articulograph (EMA), and consists of two coordinates (x, y) in six points, in a mid-sagittal plane. These points are positioned on the upper and lower lip, jaw, tongue tip, middle and back. Their 12 derivatives are added to these coordinates, to give the articulatory observation vector $\boldsymbol{O}_t^{(art)}$.

Monophonic audio data are recorded simultaneously with the articulatory coordinates at a 16 kHz sample rate. They are parametrized with 12 MFCCs, log energy and their first derivatives. All these parameters are computed every 10ms,

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

which gives synchronous acoustic and articulatory data. Throughout this article, the 26-coefficients audio vector at frame-time t will be noted $\boldsymbol{O}_t^{(ac)}$.

2.2. MOCHA-TIMIT corpus

The Multichanel Articulatory Database is an acoustic-to-articulatory corpus developed by the "Centre for Speech Technology Research," University of Edinburgh, Scotland [11]. This corpus has also been used in several international publications [1, 3, 12], and is available online[1].

This corpus contains data from two English speakers, one male (with a Northern English accent) and one female (with a Southern English accent). It is composed, for each speaker, of 460 long sentences.

As for the ARTIS database, it is composed of synchronous acoustic and articulatory (EMA) data, computed every 10 ms. Articulatory data are represented by 7 coordinates (x,y) positioned on the upper and lower lips, jaw, velum, tongue tip, middle, and back. Audio data are also recorded in WAVE format and parametrized with 12 MFCCs, log energy and their derivatives.

2.3. Notations

$\boldsymbol{O}_t$ denotes the concatenation of these two vectors, named "global vector" at time t: $\boldsymbol{O}_t = \left[\boldsymbol{O}_t^{(ac)^\dagger}, \boldsymbol{O}_t^{(art)^\dagger}\right]^\dagger$ with $(\cdot)^\dagger$ the transposition operator. $\boldsymbol{O} = \left[\boldsymbol{O}_1^\dagger, \cdots, \boldsymbol{O}_T^\dagger\right]^\dagger$ is the global observation vector.

3. MODEL DESCRIPTION

To be independent of any phonetic labeling, an unsupervised training procedure [8] is used to estimate the global HMM which represent the global observation vector $\boldsymbol{O}$.

The training procedure is divided into three steps:

1. An unsupervised clustering is performed to distribute the training data into Q classes. This clustering is done with a Q mixtures GMM. Each vector $\boldsymbol{O}_t$ gets *a posteriori* a label i ($1 \leq i \leq Q$), and the temporal sequence of the so obtained labels is stored.

2. A Q-states HMM is built. The observation probability density function (PDF) b_i of each state i, is assumed to be Gaussian with $\boldsymbol{\mu}_i = \left[\boldsymbol{\mu}_i^{(ac)^\dagger}, \boldsymbol{\mu}_i^{(art)^\dagger}\right]^\dagger$ the mean vector and $\boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(ac)} & \boldsymbol{\Sigma}_i^{(ac,art)} \\ \boldsymbol{\Sigma}_i^{(art,ac)} & \boldsymbol{\Sigma}_i^{(art)} \end{bmatrix}$ the full covariance matrix. This PDF is estimated using the observation vectors belonging to the class i. The transition matrix $\boldsymbol{A}$ is estimated considering the transitions between successive labels present in the training set.

3. The acoustic and articulatory sub-models $\boldsymbol{M}^{(ac)}$ and $\boldsymbol{M}^{(art)}$ are easily deduced from the global model: they both have the same transition matrix $\boldsymbol{A}$. Their observation probability density function (PDF) $b_i^{(ac)}$ and $b_i^{(art)}$ for each state i are marginal Gaussian laws of the global Gaussian pdf with $\boldsymbol{\mu}_i^{(ac)}$ and $\boldsymbol{\mu}_i^{(art)}$ their mean vectors, and $\boldsymbol{\Sigma}_i^{(ac)}$ and $\boldsymbol{\Sigma}_i^{(art)}$ their covariance matrices.

As shown previously [8], using $Q = 128$ states works quite well. We note that, as the French language is roughly made of 35 phones (44 phones for English language), and as a classic phone-based approach is to take three states per phones, our HMM modeling has approximately the same number of states as phonetic HMMs.

4. ARTICULATORY VECTORS GENERATION

In order to fully assess our unsupervised modeling, we test the performances of several known inversion methods.

As in most HMM approaches, the first step is a decoding with the acoustic HMM, to provide a sequence of states corresponding to an alignment of the sequence of the acoustic observation vectors. This state sequence is then transposed in the articulatory model. We will first describe the two acoustic decoding alternatives, then the different inversion methods.

4.1. Acoustic decoding - Single path approximation

The state sequence is done by considering either at each frame t the most likely state ("Gamma" path) or the most likely global state sequence ("Viterbi" path).

With the usual notations [13], the probability to be in each state i at each time step t, $\gamma_t^{ac}(i) = \mathrm{P}\left(q_t = i | \boldsymbol{O}_1, \cdots, \boldsymbol{O}_T\right)$ is expressed using the α (forward) and β (backward) variables:

$$\begin{aligned} \gamma_t^{(ac)}(i) &= \tfrac{\alpha_t^{(ac)}(i)\beta_t^{(ac)}(i)}{\sum_{i=1}^{Q} \alpha_t^{(ac)}(i)\beta_t^{(ac)}(i)} \\ \alpha_t^{(ac)}(i) &= \mathrm{P}\left(\boldsymbol{O}_1^{(ac)}, \cdots, \boldsymbol{O}_t^{(ac)}, q_t = i\right) \\ \beta_t^{(ac)}(i) &= \mathrm{P}\left(\boldsymbol{O}_{t+1}^{(ac)}, \cdots, \boldsymbol{O}_T^{(ac)} | q_t = i\right) \end{aligned} \tag{1}$$

The **Gamma** path is then determined by:

$$\boldsymbol{q} = \left[\tilde{q}_1 = \underset{1 \leq i \leq Q}{\operatorname{argmax}}\, \gamma_1(i), \cdots, \tilde{q}_T = \underset{1 \leq i \leq Q}{\operatorname{argmax}}\, \gamma_T(i)\right]$$

The **Viterbi** path $\boldsymbol{q} = [\hat{q}_1, \cdots, \hat{q}_T]$ is determined with the Viterbi algorithm. It maximizes the probability of the whole state sequence knowing the observations and the model.

In both cases, $\boldsymbol{q} = [q_1, \cdots, q_T]$ denotes the single path. Articulatory parameter generation algorithm performances will be assessed using their single path approximation.

[1] http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html

4.2. Minimum mean square error estimator

This estimator assumes that articulatory parameters follow a Gaussian mixture model composed of all the PDFs of the articulatory HMM:

$$\boldsymbol{O}_t^{(art)} \sim \sum_{i=1}^{Q} w_t(i)\mathcal{N}\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)$$

where $w_t(i)$ are the weights.

The mean square error is expressed as:

$$E\left(\left\|\widehat{\boldsymbol{O}}^{(art)} - \boldsymbol{O}^{(art)}\right\|^2\right) \tag{2}$$

where E is the stochastic expectation.

Minimizing the mean square error in Equ. (2) gives the **Minimum Mean Square Error (MMSE)** estimator:

$$\widehat{\boldsymbol{O}}_t^{(art)} = \sum_{i=1}^{Q} \gamma_t^{(ac)}(i)\boldsymbol{\mu}_i^{(art)} \tag{3}$$

An approximation of the MMSE estimator is obtained by pruning the sum in Equ. (3) to its most prominent term: $\tilde{q}_t = \underset{i=1,\dots,Q}{\operatorname{argmax}}\,\gamma_t^{(ac)}(i)$. This simplified estimator consists in using the "Gamma" path, so it is thereafter called the **Single Path (SP)** estimator with a "Gamma" decoding.

By substituting the "Gamma" path by the "Viterbi" path, we obtain another approximation of the MMSE estimator, thereafter named **Single Path (SP)** with a "Viterbi" decoding; it is given by:

$$\widehat{\boldsymbol{O}}_t^{(art)^{SP}} = \boldsymbol{\mu}_{\hat{q}_t} \tag{4}$$

Experiments in [8] have shown that the SP approximations have performances similar to the MMSE method. Therefore, in the following inversion methods, we will consider the single path approximation and its two path decoders: "Gamma" and "Viterbi".

4.3. Maximum Likelihood

An extension of the previous method is to take into account the value of the acoustic vector. This is done by using the classical GMM mapping function [1]: the **Maximum Likelihood (ML)**. In the framework of a single path approximation, this estimator is given by equations 5:

$$\widehat{\boldsymbol{O}}_t^{(art)} = \boldsymbol{\mu}_{q_t} + \boldsymbol{\Sigma}_i^{(ac,art)}\boldsymbol{\Sigma}_i^{(ac)^{-1}}(\boldsymbol{O}_t^{(ac)} - \boldsymbol{\mu}_{q_t}) \tag{5}$$

where $\boldsymbol{\Sigma}_i^{(ac,art)}$ and $\boldsymbol{\Sigma}_i^{(ac)}$ are the cross acoustic/articulatory covariance and the acoustic covariance of the observation pdf b_i of the i^{th} state.

4.4. Trajectory feature mapping

We propose here to test our model with the **Trajectory** models [7], in which the link between static features and their derivatives are explicitly modeled.

If we note $\boldsymbol{C}_t^{(ac)}$ and $\boldsymbol{C}_t^{(art)}$ the static parts of $\boldsymbol{O}_t^{(ac)}$ and $\boldsymbol{O}_t^{(art)}$, and $\boldsymbol{C}_t = \left[\boldsymbol{C}_t^{(ac)^\dagger}, \boldsymbol{C}_t^{(art)^\dagger}\right]^\dagger$ thus we can write the following equation:

$$\boldsymbol{O} = \boldsymbol{W}\left[\boldsymbol{C}_1^\dagger, \cdots, \boldsymbol{C}_T^\dagger\right]^\dagger \tag{6}$$

$$\boldsymbol{W} = \begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 0 & 1 & 0 & 0 & \cdots \\ \cdots & -1 & 1 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 1 & \cdots \\ \cdots & 0 & 0 & -1 & 1 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$\boldsymbol{W}$ is the derivation matrix and $\boldsymbol{O}$ is the global observation vector.

The generated articulatory parameters $\boldsymbol{C}^{(art)}$ are a solution of the maximization of the log likelihood $\log \mathrm{P}\left(\boldsymbol{O}^{(\boldsymbol{art})}|\boldsymbol{q}\right)$ subject to the dynamic constraint expressed in Equ. (6) where $\boldsymbol{q}$ is the state sequence determined using the acoustic HMM decoding.

An extension which takes into account the value of the acoustic vector is proposed by Zen *et al.* [3]. The **Corrected Trajectory** estimator is the solution of the equation:

$$\widehat{\boldsymbol{C}}^{(art)} = E\left(\boldsymbol{C}^{(art)}|\boldsymbol{C}^{(ac)}\right) \tag{7}$$

5. EXPERIMENTS AND RESULTS

5.1. Experimental protocol and Metrics

In every experiment, a 5-fold cross-validation method is used. The corpus is split into five phonetically equivalent parts.

The models used in every experiment are exactly the same ones, which allows proper comparison of the different decoding and inversion methods.

Results are classically given in terms of root mean square error (RMSE) in millimeters, and Pearson product-moment correlation coefficient (PMCC).

5.2. Results

The results presented in this section show that our model can achieve state-of-the-art performances.

Table 1 and 2 present results obtained with each system configuration, on both corpora. A configuration is defined by the nature of the state path ("Viterbi" or "Gamma") and the inversion method ("SP", "MMSE", "ML", "Trajectory" or "Corrected Trajectory").

For comparison purposes, we have given the best results obtained by the GipsaLab on the ARTIS corpus [10]. This method uses phonetic HMMs, trained with the Minimum Generation Error (MGE) criterion, and an inversion step using Trajectory HMMs achieves a RMSE of **1.49 mm**.

On the MOCHA-TIMIT corpus, Zen *et al.* [3] reach, on the male speaker a RMSE of **1.52 mm** with a trajectory phonetic HMM, and a RMSE of **1.13 mm** with a trajectory GMM.

A maybe surprising fact is that there is no difference between the MMSE inversion method and its Gamma single path approximation in term of precision. This confirms that the unsupervised model used truly has a HMM behavior: in most instants only one state has a great probability. The Viterbi single path approximation gives even better results, which allows us to consider either the Gamma or the Viterbi single path approximations in other inversion methods.

Table 1. *RMSE (in mm) and PMCC on the ARTIS corpus, with the different decoding and inversion methods.*

Decoding	Inversion	RMSE	PMCC
Gamma	SP	2.28	0.54
Viterbi	SP	2.18	0.56
None	MMSE	2.27	0.54
Gamma	ML	1.87	0.61
Viterbi	ML	1.77	0.64
Gamma	Trajectory	1.86	0.66
Viterbi	Trajectory	1.78	0.68
Gamma	**Corrected Trajectory**	**1.49**	**0.72**
Viterbi	**Corrected Trajectory**	**1.49**	**0.72**

Taking into account the value of the acoustic vector is clearly a great improvement of any method, since it improves the RMSE from 16 to 35 percents (0.3 - 0.5 mm). In fact, this allows us to generate various articulatory vectors from the same state sequence, by taking into account the acoustic variability.

With all inversion methods a Viterbi decoding leads to better results, probably because considering the sequence altogether in the decoding phase is more coherent with the notion of trajectory, which also considers the sequence altogether.

Finally, as shown before by other studies, using trajectory HMM to explicitly represent the link between static and dynamic component of the same vector is a great improvement, in both RMSE and in terms of correlation.

Table 2. *RMSE (in mm) and PMCC on the MOCHA-TIMIT corpus for both speakers, with the different decoding and inversion methods.*

		male / female	
Decoding	Inversion	RMSE	PMCC
Gamma	SP	1.85 / 1.92	0.63 / 0.64
Viterbi	SP	1.73 / 1.83	0.69 / 0.68
None	MMSE	1.83 / 1.91	0.64 / 0.65
Gamma	ML	1.64 / 1.71	0.71 / 0.71
Viterbi	ML	1.55 / 1.64	0.75 / 0.73
Gamma	Trajectory	1.85 / 1.92	0.63 / 0.64
Viterbi	Trajectory	1.73 / 1.82	0.69 / 0.68
Gamma	**Cor. Traj.**	**1.44 / 1.52**	**0.82 / 0.80**
Viterbi	**Cor. Traj.**	**1.44 / 1.52**	**0.82 / 0.80**

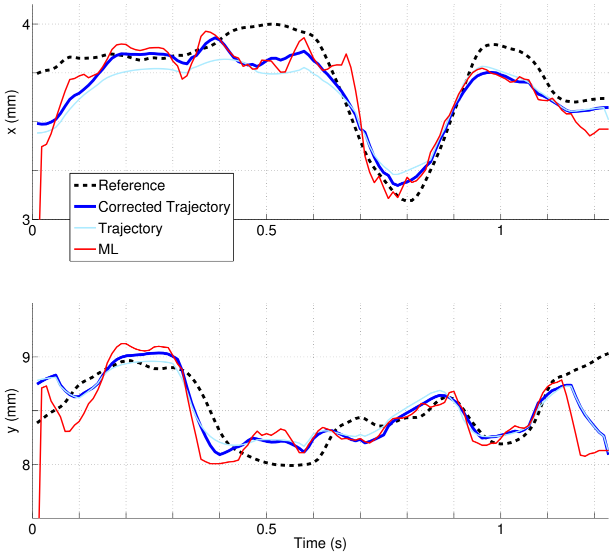

Fig. 1. *Reconstructed trajectory for the lower lip for the sentence "Faire la nouba" [fɛʁlanuba] with various methods using **"Viterbi"** decoding.*

6. CONCLUSION

In this article we fully compared different acoustic decoding and articulatory parameter generation methods with an unsupervised HMM modeling. We therefore now have a precise idea of the improvement brought by each algorithm. It also validates the potential of such an unsupervised HMM approach in articulatory parameter generation.

Our main interest will now be to improve the training, using other criterion such as Minimum Generation Error, and to take the static/dynamic relationship into consideration during the training phase. We will also adapt our method for inter-speaker and inter-language acoustic-to-articulatory inversion.

7. ACKNOWLEDGMENTS

The authors thank the GIPSA-Lab (Grenoble, France) for making their acoustic-to-articulatory corpus available. This work is supported by the French National Research Agency under contract number ANR-08-EMER- 001-02 (ARTIS).

8. REFERENCES

[1] T. Toda, A. W. Black, and K. Tokuda, "Statistical Mapping between Articulatory Movements and Acoustic Spectrum Using a Gaussian Mixture Model," *Speech Communication*, vol. 50, pp. 215–227, 2008.

[2] A. Ben Youssef, P. Badin, and G. Bailly, "Acoustic-to-articulatory inversion in speech based on statistical models," in *9th International Conference on Auditory-Visual Speech Processing (AVSP)*, 2010, pp. 160–165.

[3] H. Zen, Y. Nankaku, and K. Tokuda, "Continuous Stochastic Feature Mapping Based on Trajectory HMMs," *IEEE Transaction on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 417–430, 2010.

[4] A. Ben Youssef, P. Badin, G. Bailly, and P. Heracleous, "Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme Hidden Markov Models," in *Interspeech - European Conference on Speech Communication and Technology*, 2009, pp. 2255–2258.

[5] L. Zhang and S. Renals, "Acoustic-articulatory modeling with the trajectory HMM," *IEEE Signal Processing Letters*, vol. 15, pp. 245–258, 2008.

[6] H. Zen, K. Tokuda, and T. Kitamura, "An introduction of trajectory model into HMM-based speech synthesis," in *Fifth ISCA ITRW on Speech Synthesis*, 2004.

[7] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech Parameter Generation Algorithms for HMM-based Speech Synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2000, pp. 1315–1318.

[8] H. Lachambre, L. Koenig, and R. André-Obrecht, "Articulatory Parameter Generation using Unsupervised Hidden Markov Models," in *European Signal Processing Conference (EUSIPCO)*, 2011, pp. 456–459.

[9] A. Ben Youssef, P. Badin, and G. Bailly, "Can tongue be recovered from face? The answer of data-driven statistical models," in *Interspeech - European Conference on Speech Communication and Technology*, 2010, pp. 2002–2005.

[10] A. Ben Youssef, T. Hueber, P. Badin, and G. Bailly, "Toward a multi-speaker visual articulatory feedback system," in *Interspeech - European Conference on Speech Communication and Technology*, 2011, pp. 589–592.

[11] A. Wrench, "A multichannel articulatory database and its application for automatic speech recognition," in *In Proceedings 5 th Seminar of Speech Production*, 2000, pp. 305–308.

[12] A. Gutkin and S. King, "Detection of Symbolic Gestural Events in Articulatory Data for Use in Structural Representations of Continuous Speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2005, pp. 885–888.

[13] Lawrence Rabiner and Biing-Hwang Juang, *Fundamentals of speech recognition*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, PTR Prentice Hall edition, 1993.