ARTICULATORY PARAMETER GENERATION USING UNSUPERVISED HIDDEN MARKOV MODELS

Hélène Lachambre, Lionel Koenig, and Régine André-Obrecht

IRIT - Université de Toulouse
118 route de Narbonne, 31062 Toulouse Cedex 9, France
{lachambre, koenig, obrecht}@irit.fr

ABSTRACT

We present an acoustic-to-articulatory inversion method based on unsupervised Hidden Markov Models. A global HMM is first trained from the acoustic and articulatory data. This model is then split in two sub-models which represent the acoustic part and the articulatory part of the data. These two sub-models are linked through the fact that they are deduced from the same global model.

The articulatory generation is made by decoding a sequence of acoustic vector with the acoustic model, and by transposing the results in the articulatory model: two alternative estimation processes are assessed.

Over our 18 minutes corpus, the RMS error is 2.25 mm. The results of this new approach are very encouraging, since no optimisation is done in the generation process.

1. INTRODUCTION

In acoustic-to-articulatory inversion, the aim is to recover the vocal tract shape from the acoustic parameters of a speech sound. This may be very useful for applications such as augmented speech (to help hearing-impaired persons), or foreign languages learning (to shown the student how the pronunciation should be done or corrected).

This problem has now been studied for more than 30 years, and recent methods have proved to be very efficient. The main two approaches are the GMM (Gaussian Mixture Models) approach [1, 2], and the HMM (Hidden Markov Models) approach [3, 4, 5].

In the GMM approach, the joint probability of the acoustic and articulatory parameters is modeled by a Gaussian Mixture Model. A mapping is then done to enable the inversion, this mapping is done either using the MMSE (Minimum Mean Square Error) criterion [1] or the MLE (Maximum Likelihood Estimation) criterion [1, 2].

In the HMM approach, the main idea is to take into account the temporal dimension of speech, in the acoustic space as well as in the articulatory space. The speech acoustics is modeled with HMM, classically trained on a phoneme-labeled corpus. In [3], the articulatory part is modeled by a state-dependent linear regression between acoustic and articulatory parameters. In [4, 5], the articulatory part and the acoustic part are jointly modeled with multi-stream HMM, which finally gives an acoustic HMM and an articulatory HMM. In the recognition stage, an acoustic signal is first decoded by the acoustic HMM, giving a sequence of states, which is then converted into articulatory parameters, either with linear regression [3] or with the articulatory HMM [4, 5]. In this last case, the articulatory parameter generation is provided by using the trajectory model proposed by the HTS system [6].

In this article, we explore the use of an unsupervised HMM, which can allow to take advantage of both the statistical approach (GMM) and the HMM approach. As the GMM approach, the training is unsupervised, it does not need any expert data (e.g. phoneme labelling of the corpus). At the same time, as the HMM approach, we can take into account the temporal dimension of speech to preserve the continuity.

The article is organised as follows: in part 2 we describe the corpus and features, then an overview of our method is proposed in part 3. The training (resp. generation) process is detailed in part 4 (resp. 5). We finally expose the experimental protocol and give some results in part 6.

2. CORPUS

The major difficulty is to collect a corpus large enough, in which acoustic and articulatory data are collected with a perfect synchronisation. As we are part of the french ANR project ARTIS [7], we have the access to the database developed by the Gipsa-Lab in Grenoble, France. Experiments and results have already been published by this laboratory [2, 8, 9]. We remind in the following lines the composition of this corpus.

The corpus is in French, and it is pronounced by a male speaker, used to this exercise. It is composed of several utterances of vowels alone, VCV nonsense sequences, CVC real french words, and full sentences. The initial and final long pauses being removed, the corpus is approximately 18 min long.

The articulatory data is recorded using an Electro-Magnetic Articulograph (EMA), that makes a tracking of flesh points. In this corpus, six coils are used, located on jaw, upper lip, lower lip, tongue tip, tongue middle and tongue back. Each coil is known by two coordinates in a sagittal plan, resulting into a 12-dimension vector.

The acoustic data are recorded at a 44,100 Hz sampling rate. The acoustic signal is parametered the following way: 12 MFCC and Energy, with their derivatives, leading to a 26-dimension acoustic vector.

The acoustic features are classically extracted every 10 ms. The EMA data are low-pass filtered at 20 Hz and down sampled at 100 Hz, this final operation making both articulatory and acoustic data synchronous.

The 12-dimension articulatory vector is completed with the 12 derivatives, which finally gives every 10 ms a 50-dimension vector $\boldsymbol{O} = [\boldsymbol{O}^{ac^T}\boldsymbol{O}^{art^T}]^T = [O_1^{ac}, ..., O_{26}^{ac}, O_1^{art}, ..., O_{24}^{art}]^T$.

3. INVERSION SYSTEM OVERVIEW

Our inversion system is based on three tied Hidden Markov Models (figure 1).

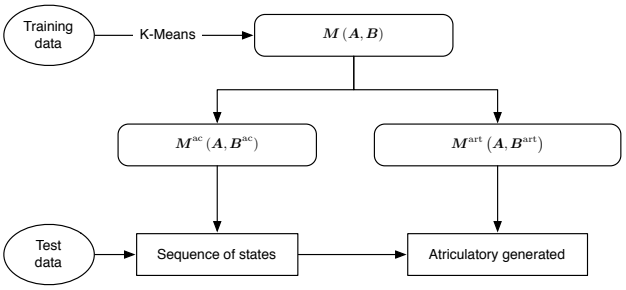

Figure 1: Global scheme of the inversion system.

First, a global HMM $\boldsymbol{M}(A, B)$ (noted $\boldsymbol{M}$) is trained using a clustering algorithm as explained in part 4.1. This model is composed of :

- Q states and a $Q * Q$ transition matrix A.
- A set B of Gaussian probability density functions $b_i \sim \mathcal{N}(\mu_i, \Sigma_i))$, one for each state i.

From this model $\boldsymbol{M}$, we deduce two "sub-models": $\boldsymbol{M}_{ac}(A, B_{ac})$ (noted $\boldsymbol{M}_{ac}$) and $\boldsymbol{M}_{art}(A, B_{art})$ (noted $\boldsymbol{M}_{art}$), representing the acoustic part and the articulatory part of $\boldsymbol{M}$. $\boldsymbol{M}_{ac}$ and $\boldsymbol{M}_{art}$ are precised in part 4.2.

Given a sequence of acoustic vectors, a classical recognition process is made with the acoustic model $\boldsymbol{M}_{ac}$, which gives a sequence of states. This sequence of states is then transposed in the articulatory model $\boldsymbol{M}_{art}$ to generate the corresponding articulatory vector sequence. These transposition and estimation processes are described in details in part 5.

4. UNSUPERVISED MODELS TRAINING

As said above, a global model $\boldsymbol{M}$ is trained before building the two sub-models $\boldsymbol{M}_{ac}$ and $\boldsymbol{M}_{art}$. We propose an unsupervised approach for this phase. The advantage of our proposal is to be independent of any expert knowledge or any manual transcription as with the GMM approach and simultaneously to exploit the temporal dimension as with the HMM approach.

4.1 Global model training

The global model $\boldsymbol{M}$ is trained in three steps; the number Q of states is decided *a priori*, but no *a priori* structure is supposed:

- The training vectors are clustered with an unsupervised algorithm, resulting into Q classes. Each class is assimilated to a state i of the global HMM. Each training vector is therefore assigned *a posteriori* to a state and labelled.
- The probability density of each state i, is modeled by a Gaussian distribution $\mathcal{N}(\mu_i, \Sigma_i)$. This distribution is estimated with the training vectors assigned *a posteriori* to the state i.
- The transition matrix A is classically empirically estimated by counting the number of occurrences of the transitions between states, a transition between two states being a transition between two vectors *a posteriori* assigned to these states.

In this study, the unsupervised clustering is done with the K-means algorithm which can be refined with a GMM modelling. In this last case, the GMM density is estimated with the EM algorithm, initialised with the density found by the K-means algorithm.

4.2 Sub-models definition

From the global model $\boldsymbol{M}$, we build two sub-models $\boldsymbol{M}_{ac}$ and $\boldsymbol{M}_{art}$, to model respectively the acoustic and the articulatory parts of the data. Their estimation is described below.

4.2.1 States and transition matrices

Both $\boldsymbol{M}_{ac}$ and $\boldsymbol{M}_{art}$ have the same number Q of states as $\boldsymbol{M}$. The states of the sub-models being deduced from the global model $\boldsymbol{M}$, each training sub-vector is assigned to the same state in $\boldsymbol{M}_{ac}$ or in $\boldsymbol{M}_{art}$ than in $\boldsymbol{M}$. The transition matrices of the sub-model are consequently exactly the same as A.

4.2.2 Probability emission densities

The probability emission density b_i^{ac} (resp. b_i^{art}) of each state of $\boldsymbol{M}_{ac}$ (resp. $\boldsymbol{M}_{art}$) is modeled by a Gaussian distribution $b_i^{ac} \sim \mathcal{N}(\mu_i^{ac}, \Sigma_i^{ac})$ (resp. $b_i^{art} \sim \mathcal{N}(\mu_i^{art}, \Sigma_i^{art})$). They are estimated with the training sub-vectors assigned *a posteriori* to the state i.

Note that b_i^{ac} and b_i^{art} can be easily deduced from b_i: in the observation probability $\mathcal{N}(\mu_i, \Sigma_i)$ the mean vector μ_i can be written as

$$[\mu_i^{ac^T}, \mu_i^{art^T}]^T$$

with μ_i^{ac} the mean vector of the acoustic part, and μ_i^{art} the mean vector of the articulatory part. Σ_i can be as well written as

$$\Sigma_i = \begin{bmatrix} \Sigma_i^{ac} & \Sigma_i^{ac,art} \\ \Sigma_i^{art,ac} & \Sigma_i^{art} \end{bmatrix}$$

5. ARTICULATORY VECTORS GENERATION

Two extremely simple generation processes are used to generate an articulatory vector sequence from an acoustic one.

First, the acoustic signal is parametered the same way as for the training phase (see part 2): Energy, MFCC, ΔEnergy and ΔMFCC are extracted every 10 ms. We therefore get a sequence of K vectors: $\boldsymbol{O}_1^{ac}, ..., \boldsymbol{O}_K^{ac}$.

5.1 GMM generation

As an extension of the method proposed by [10], the generated observation $\hat{\boldsymbol{O}}_t^{art^{GMM}}$ at instant t is supposed to follow a Gaussian mixture model:

$$\hat{\boldsymbol{O}}_t^{art^{GMM}} \sim \sum_{i=1}^{Q} P(s_t^{ac} = i|\boldsymbol{O}_1^{ac}, ..., \boldsymbol{O}_K^{ac}) b_i^{art} \quad (1)$$

with $b_i^{art} \sim \mathcal{N}(\mu_i^{art}, \Sigma_i^{art})$ the probability density function of the state s_i^{art} of $\boldsymbol{M}^{art}$.

Using the usual notations [11], it gives:

$$\hat{\boldsymbol{O}}_t^{art} = \sum_{i=1}^{Q} \gamma_t^{ac}(i) \mu_i^{art} \quad (2)$$

where $\gamma_t^{ac}(i)$ is:

$$\gamma_t^{ac}(i) = \frac{\alpha_t^{ac}(i)\beta_t^{ac}(i)}{\sum_{i=1}^{Q} \alpha_t^{ac}(i)\beta_t^{ac}(i)} \quad (3)$$

with:

$$\alpha_t^{ac}(i) = P(\boldsymbol{O}_1^{ac}, ..., \boldsymbol{O}_t^{ac}, s_t^{ac} = i)$$
$$\beta_t^{ac}(i) = P(\boldsymbol{O}_{t+1}^{ac}, ..., \boldsymbol{O}_K^{ac}|s_t^{ac} = i)$$

5.2 Best State (BS) generation

We alternatively propose to replace the sum by its predominant term, corresponding to the most probable state at instant t:

$$\begin{aligned} \hat{\boldsymbol{O}}_t^{art^{BS}} &= \mu_{\hat{s}_t}^{art} \\ \hat{s}_t &= \underset{i=1,...,Q}{\operatorname{argmax}} \gamma_t^{ac}(i) \end{aligned} \quad (4)$$

Note that in all cases, to fit as well as possible the learning data (see section 2), the generated vectors are finally low-pass filtered at 20 Hz.

6. EXPERIMENTS AND RESULTS

For the experiments, the corpus is split in two: the training data base is composed of ⅔ of the corpus, the remaining ⅓ being used for the tests. The corpus is split in such a way that each kind of sounds (vowels alone, VCV non-sense sequences, CVC real words, and full sentences) is present in the training set as well as in the test set.

6.1 Experiments

In this section, we present the different configurations we tested (for the learning of the models as well as for the generation of the articulatory data) before presenting and commenting the results in the following section 6.2.

6.1.1 Experimental protocol

For the learning of the global and of the two sub-models, we have first tested the influence of the number Q of clusters produced by the K-means algorithm. Two values have been tested: $Q = 32$ and $Q = 128$. Note that the french language contains 36 phonemes. In classical HMM models, each phoneme is modeled with a 3-state model, resulting in an HMM with $3 * 36 = 108$ states, a value close to one of our choice with 128 states.

Using the GMM-EM clustering, only the 128-GMM configuration was tested

Those three models will be noted:

- 32 states with no re-estimation : $\boldsymbol{M}_{32}^{K_{means}}$
- 128 states with no re-estimation : $\boldsymbol{M}_{128}^{K_{means}}$
- 128 states with EM-re-estimation : $\boldsymbol{M}_{128}^{EM}$

6.1.2 Generation configurations

As presented in part 5 two generation configurations are tested. They are noted as follows:

- Using the GMM approach : GMM
- Taking only the best state : BS

In the next section, each experiment will be referenced by its training configuration and its generation configuration. For example, the configuration $\boldsymbol{M}_{32}^{K_{means}}$-BS corresponds to models with 32 states and no re-estimation, and a generation using only the best state.

6.2 Results

Before giving the quantitative results with the different configurations, we give an idea of the structure of the HMMs.

6.2.1 Structure of the unsupervised trained model

In supervised HMM training, each phoneme is classically modeled by a 3-states HMM, the phonemes being linked between each other. The transition matrix of such a HMM is block diagonal, with only some other transitions (inter-phoneme transitions) allowed.

In our experiments, it is interesting to observe that the models trained with this method are qualitatively similar, in terms of structure and transition matrix, to classical models. As shown on figure 2, the transition matrix of this 128-states HMM is diagonal with only some other transition getting a significant transition probability.

Figure 3 shows the topology of a 32-states HMM model. Only the most probable transitions between states (probability $> 10^{-2}$) are represented. This second figure confirms that only some transitions are significant.

6.2.2 Quantitative results

Only the more significant results are presented in table 1. The performances are classically given in terms of Root Mean Square Error (RMSE) between the measured and the generated articulatory vectors, and in terms of Pearson Product-Moment Correlation Coefficient (PMCC). The RMSE measures the mean error of the coil positions in millimetres, while the PMCC measures the similarity of the trajectories.

As we could have inferred, increasing the number of states of the HMM model significantly improves the

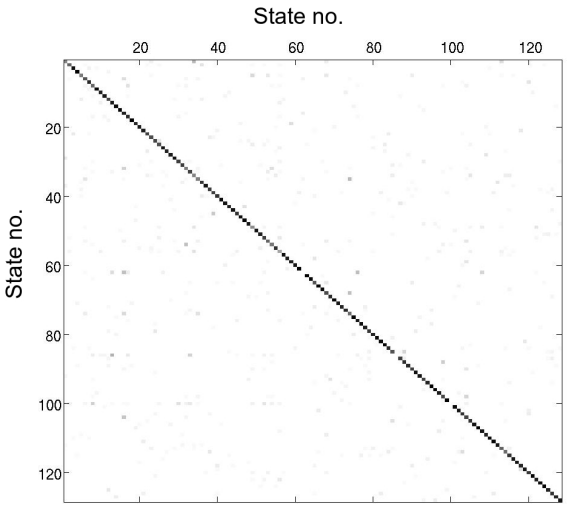

Figure 2: Example of a transition matrix trained with the unsupervised method: white is a 0 probability, black is a probability equal to 1. The state number Q is 128.

Table 1: RMSE (in mm) and PMCC on the test corpus, with the different configurations.

	RMSE	PMCC
$\boldsymbol{M}_{32}^{K_{means}}$-BS	2.78	0.50
$\boldsymbol{M}_{128}^{K_{means}}$-BS	2.48	0.55
$\boldsymbol{M}_{128}^{K_{means}}$-GMM	2.47	0.56
$\boldsymbol{M}_{128}^{EM}$-GMM	2.25	0.59

results. We noticed that for 256 states, the K-means algorithm gives empty clusters, the training database is certainly not sufficient and the number of 128 states is a good compromise. We therefore conducted the other experiments with a 128-states HMM.

The second conclusion is that in the generation phase, taking only the most probable state is almost as good as taking a weighted combination of all states. This is probably because most of the time, the best state really has a greater probability than any other state.

Finally, re-estimating the density probability with the EM algorithm to fit a 128-GMM on the 128 states during the training phase allows to fit more precisely the training data and to significantly improve the results.

As this corpus is used by other laboratories, a comparison may be done with some precaution. As described in literature [4], the generation process is often done using a trajectory model for the articulatory data, for example with the HTS system [6]. The best results obtained on this corpus by such methods give a RMSE around 1.7 mm for the HMM approach and 2.25 mm for the GMM one [2]. Comparatively, our method gives very encouraging results, in terms of RMSE (we achieve 2.25 mm) as well as in terms of PMCC, especially when taking into account that no trajectory model is added in the generation phase.

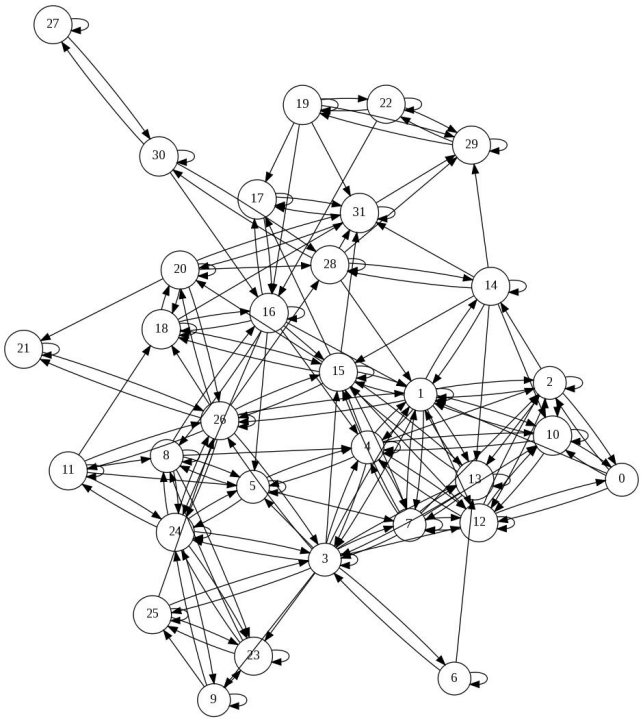

Figure 3: Visualisation of the topology of a model trained with the unsupervised method. The state number Q is 32. Only the transitions with a probability greater than 10^{-2} are represented

7. CONCLUSION AND PERSPECTIVES

We have presented a very simple method to generate articulatory parameters from acoustical data, based on HMM modeling. Its originality is that the HMM is trained in an unsupervised way, with no *a priori* structure and no expert data. The results are very satisfying in the sense that they are near the literature results with a great potential.

A phase of Baum-Welch re-estimation must complete efficiently this training by taking the contextual information into account; in the literature, the HMM are context dependent.

We will also have to explore the fact that the articulatory space is not completely covered during the generation, as can be seen on figure 4: the repartition of the generated articulatory data (in black) is less spread than the repartition of the measured articulatory data (in grey); this fact is tied to the probabilistic approach.

Nevertheless, it will be necessary to add some trajectory models to improve significantly the generated articulatory vectors.

8. ACKNOWLEDGEMENTS

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

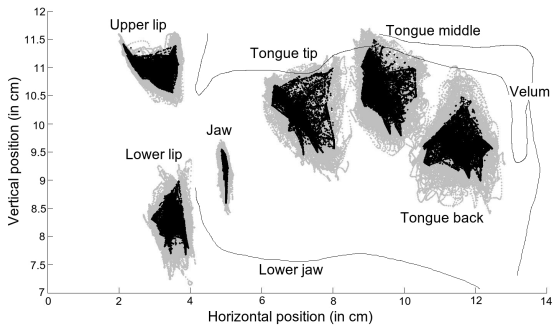

Figure 4: Articulatory space for EMA. Each area corresponds to the different positions of a coil. Grey: measured values, black: generated values

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*, 50:215–227, 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)*, pages 160–165, 2010.

[3] S. Hiroya and M. Honda. Estimation of articulatory movements from speech acoustics using an hmm-based speech production model. *IEEE Transactions on Audio, Speech, and Language Processing*, 12(2):175–185, 2004.

[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*, pages 2255–2258, 2009.

[5] L. Zhang and S. Renals. Acoustic-articulatory modeling with the trajectory HMM. *IEEE Signal Processing Letters*, 15:245–258, 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] French ANR project. *ARTIS: Articulatory inversion from audio-visual speech for augmented speech presentation*, 2008-2012.

[8] 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*, pages 2002–2005, 2010.

[9] P. Badin, Y. Tarabalka, F. Elisei, and G. Bailly. Can you "read tongue movements"? In *Interspeech - European Conference on Speech Communication and Technology*, pages 2635–2638, 2008.

[10] C. A. Rødbro, M. N. Murthi, S. V. Andersen, and S. H. Jensen. Hidden Markov Model-Based Packet Loss Concealment for Voice Ovec IP. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1609–1623, 2006.

[11] L. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Upper Saddle River, NJ, USA, 1993.