Acoustic-to-articulatory inversion in speech based on statistical models

Atef Ben Youssef, Pierre Badin, Gérard Bailly

GIPSA-lab (Département Parole & Cognition / ICP), UMR 5216 CNRS – Grenoble University
961 rue de la Houille Blanche, BP 46, F-38402 Saint Martin d'Hères cedex, France
{Atef.Ben-Youssef, Pierre.Badin, Gerard.Bailly}@gipsa-lab.grenoble-inp.

Abstract

Two speech inversion methods are implemented and compared. In the first, multistream Hidden Markov Models (HMMs) of phonemes are jointly trained from synchronous streams of articulatory data acquired by EMA and speech spectral parameters; an acoustic recognition system uses the acoustic part of the HMMs to deliver a phoneme chain and the states durations; this information is then used by a trajectory formation procedure based on the articulatory part of the HMMs to resynthesise the articulatory movements. In the second, Gaussian Mixture Models (GMMs) are trained on these streams to directly associate articulatory frames with acoustic frames in context, using Maximum Likelihood Estimation. Over a corpus of 17 minutes uttered by a French speaker, the RMS error was 1.66 mm with the HMMs and 2.25 mm with the GMMs.

Index Terms: Speech inversion, ElectroMagnetic Articulography (EMA), Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Maximum Likelihood Estimation (MLE).

1. Introduction

Speech inversion is a long-standing problem, as testified by the famous work by Atal *et al.* [1] in the seventies. Speech inversion was traditionally based on analysis-by-synthesis, as implemented by [2], or by [3] who optimised codebooks to recover vocal tract shapes from formants. But since a decade, more sophisticated data-driven techniques have appeared, thanks to the availability of large corpora of articulatory and acoustic data provided by devices such as the ElectroMagnetic Articulograph (EMA) or motion tracking devices based on classical or infrared video.

Our laboratory is thus involved in the development of an *inversion* system that allows producing *augmented speech* from the sound signal alone, possibly associated with video images of the speaker's face. *Augmented speech* consists of audio speech supplemented with signals such as the display of usually hidden articulators such (e.g. tongue or velum) by means of a virtual talking head, or with hand gestures as used in *cued speech* by hearing-impaired people.

2. State-of-the-art

At least, two classes of statistical models of the speech production mechanisms can be found in the recent literature: Hidden Markov Models (HMMs) (cf. [4], [5] or [6]), and Gaussian Mixture Models (GMMs) (cf. [7]). In addition to the structural differences between HMMs and GMMs, an important difference is that HMMs explicitly use phonetic information and temporal ordering while the GMMs simply cluster the multimodal behaviour of similar speech chunks.

Hiroya & Honda [4] developed a method that determines articulatory movements from speech acoustics using a HMM-based speech production model. After proper labelling of the training corpus, each allophone is modelled by a context-dependent HMM, and the proper inversion is performed by a state-dependent linear regression between the observed acoustic and the corresponding articulatory parameters. The articulatory parameters of the statistical model are then determined for a given speech spectrum by maximizing a posteriori estimation. In order to assess the importance of phonetics, they tested their method under two experimental conditions, namely *with* and *without* phonemic information. In the former, the phone HMMs were assigned according to the correct phoneme sequence for each test utterance. In the latter, the optimal state sequence was determined among all possible state sequences of the phone HMMs and silence model. They found that the average RMS errors of the estimated articulatory parameters were 1.50 mm from the speech acoustics and the phonemic information in the utterance and 1.73 mm from the speech acoustics only.

Zhang & Renals [5] developed a similar approach. Their system jointly optimises multi-stream phone-sized HMMs on synchronous acoustic and articulatory frames. The inversion is carried out in two stages : first a representative HMM state alignment is derived from the acoustic channel ; a smoothed mean trajectory is generated from the HMM state sequence by an articulatory trajectory formation model using the same HMMs. Depending on the availability of the phone labels for the test utterance, the state sequence can be either returned by an HMM decoder, or by forced alignment derived from phone labels, leading to RMS errors of respectively 1.70 mm and 1.58 mm.

Toda and coll. [7] described a statistical approach for both articulatory-to-acoustic mapping and acoustic-to-articulatory inversion mapping without phonetic information. Such an approach interestingly enables language-independent speech modification and coding. They modelled the joint probability density of articulatory and acoustic frames in context using a Gaussian mixture model (GMM) based on a parallel acoustic-articulatory speech database. They employed two different techniques to establish the GMM mappings. Using a minimum mean-square error (MMSE) criterion with an 11 frames acoustic window and 32 mixture components, they obtained RMS inversion errors of 1.61 mm for one female speaker, and of 1.53 mm for a male speaker. Using a maximum likelihood estimation (MLE) method and 64 mixture components, they improved their results to 1.45 mm for the female speaker, and 1.36 mm for the male speaker.

The studies described above do not allow concluding about the optimal inversion method since data, speakers and languages are not comparable. Hiroya & Honda [4] and Zhang & Renals [5] have shown that using explicit phonetic information to built HMMs gives better results. Toda and coll. [7], using GMMs and no phonetic information, get lower

RMS errors. However, the corpora as well as training and testing conditions are not completely comparable. Therefore, the aim of the present work is to compare, *ceteris paribus*, the HMM-based method used in [6] with a GMM-based method similar to that of [7] using the minimum mean-square error (MMSE) criterion and subsequent MLE optimisation for the GMM-based mapping method.

3. Articulatory and acoustic data

3.1. The corpus

For this study, a corpus already recorded was used [8]. It consists of a set of two repetitions of 224 nonsense vowel-consonant-vowel (VCV) sequences (uttered in a slow and controlled way), 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, differing only by a single cue (the French version of the Diagnostic Rhyme Test); 68 short French sentences, 9 longer phonetically balanced French sentences, and 11 long arbitrary sentences. The corpus was recorded on a single male French subject, which means that no speaker adaptation / normalisation problems will be dealt with in this study.
The phones have initially been labelled for each utterance using a forced alignment procedure based on the audio signal and the corresponding phonetic transcription based on HMMs. Subsequent manual correction of both phoneme labels and phoneme boundaries were performed using the *Praat* software [9]. The centres of allophones were automatically chosen as the average between beginning and end of the phonemes. Altogether the corpus, from which long pauses were excluded, contains approximately 100,000 frames, *i.e.* about 17 minutes of speech, corresponding to 5132 allophones. The 36 phonemes are: [a ɛ e i y u o ø ɔ œ ɑ̃ ɛ̃ œ̃ ɔ̃ p t k f s ʃ b d g v z ʒ m n ʁ l w ɥ j ə _ _], where _ and _ are internal short and utterance initial and final long pauses respectively.

3.2. The acoustic and articulatory data

The articulatory data have been recorded by means of an ElectroMagnetic Articulograph (EMA) that tracks motion of flesh points of the articulators thanks to small electromagnetic receiver coils glued on the organs. Studies have shown that the number of degrees of freedom of speech articulators (jaw, lips, tongue …) for speech is limited, and that a small but sufficient number of carefully selected measurement locations can allow retrieving them with a good accuracy [8, 10]. In the present study, six coils are used: a jaw coil is attached to the lower incisors (jaw), whereas three coils are attached to the tongue tip (tip), the tongue middle (mid), and the tongue back (bck) at approximately 1.2 cm, 4.2 cm, and 7.3 cm, respectively, from the extremity of the tongue; an upper lip coil (upl) and a lower lip coil (lwl) are attached to the boundaries between the vermilion and the skin in the midsagittal plane. Extra coils attached to the upper incisor and to the nose served as references to compensate for head movements in the midsagittal plane. The audio-speech signal was recorded at a sampling frequency of 22,050 Hz, in synchronization with the EMA coordinates, which were recorded at a 500 Hz sampling frequency, low-pass filtered at 20 Hz in order to reduce noise, and down sampled to 100 Hz.

3.3. Overview of the data

We verified that the general articulatory characteristics of each phoneme were in accordance with our expectation by displaying, in the midsagittal plane, the dispersion ellipses of the six coils estimated over the sets of all the instances. The minimum and maximum number of instances per phoneme was 17 (for short pauses) and 348 (for /a/). This illustrates the coherence and the validity of the data. Figure 1, which displays these ellipses for phoneme /t/, illustrates the very low variability of the tongue tip and jaw coils for /t/, as could be expected since the tongue is in contact with the hard palate for this articulation. It should however be reminded that the articulations were sampled at the instant midway between the phone boundaries, which does not completely ensure that it corresponds to the actual centre of the phone if the trajectories are not symmetrical.

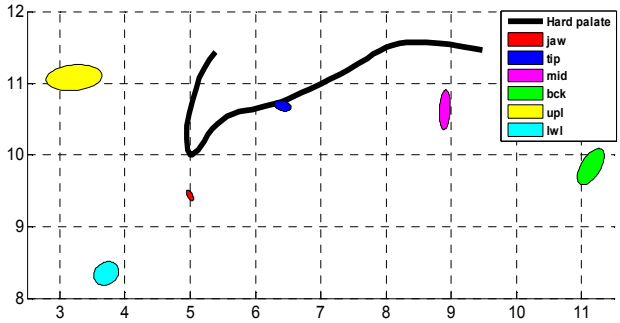

Figure 1. *Dispersion ellipses of the measured coordinates of the six EMA coils for phoneme /t/. These ellipses are computed from the samples taken at the middle of the 231 instances of /t/ in the corpus.*

3.4. Context classes for phonemes

Due to coarticulatory effects, it is unlikely that a single context-independent HMM could optimally represent a given allophone. Therefore, context-dependent HMMs were trained. Rather than using a priori phonetic knowledge to define such classes, confusion trees have been built for both vowels and consonants, based on the matrix of Manhattan distances of the coils coordinates between the centre frame of each pair of phone. Each allophone was represented by its mean over all the associated instances. Using hierarchical clustering to generate dendrograms we define six coherent classes for vocalic contexts ([a ɛ ɛ̃ | ø œ œ̃ | e i | y | u | o ɔ ɑ̃ ɔ̃]), and ten coherent classes for consonantal contexts ([p b m | f v | ʁ | ʃ ʒ | l | t d s z n | j | ɥ | k g | w]). The schwa, the short and the long pauses ([ə _ _]) are ignored in the context classes. Using acoustic spectral distances did lead to classes less satisfactory from the point of view of phonetic knowledge.

4. HMM models

We recall the experiments published previously by Ben Youssef et al. [6]. For the training of the HMMs, acoustic feature vectors consisted of the 12 Mel-Frequency Cepstral Coefficients (MFCC) and of the logarithm of the energy, along with the first time derivatives, computed from the signal over 25 ms windows at a frame rate of 100 Hz to match the EMA sampling frequency. Articulatory feature vectors consisted of the x and y coordinates of the six active coils.

Their first time derivatives are also added. The EMA traces were down sampled to match the 100 Hz shift rate of the acoustic feature vectors.

Various contextual schemes were tested: phonemes without context *(no-ctx)*, with left *(L-ctx)* or right context *(ctx-R)*, and with both left and right contexts *(L-ctx-R)*.

Left-to-right, 3-state phoneme HMMs with one Gaussian per state and a diagonal covariance matrix are used. For training and test the HTK3.4 toolkit is used [11]. The training is performed using the Expectation Maximization (EM) algorithm based on the Maximum Likelihood (ML) criterion.

The acoustic and articulatory features vectors are considered as two streams in the HTK multi-stream training procedure. Subsequently, the HMMs obtained are split into *articulatory HMMs* and *acoustic HMMs*.

A bigram language model considering sequences of phones in context is trained over the complete corpus. No prosodic constraints such as a duration model are added. The acoustic-to-articulatory inversion is achieved in two stages. The first stage performs phoneme recognition, based on the acoustic HMMs. The result is the sequence of recognised allophones together with the duration of each state in each HMM. An inheritance procedure allows to replace a missing HMM by the closest one aims to compensate for the too small size of the training set [6].

The second stage of the inversion aims at reconstructing the articulatory trajectories from the chain of phoneme labels and state durations delivered by the recognition procedure. As described in [12], the synthesis is performed using the trajectory formation procedure proposed by [13] with the software developed by the HTS group [14-15]. A linear sequence of HMM states is built by concatenating the corresponding phone HMMs, and a sequence of observation parameters is generated using a specific ML-based parameter generation algorithm [15].

4.1. Evaluation of the HMM-based inversion

Three criteria have been used to assess the inversion results: (1) the square root of the mean quadratic error (RMSE) between the measured and recovered coordinates, (2) the Pearson Product-Moment Correlation Coefficient (PMCC), a less conservative criterion that measures only the level of amplitude similarity and of synchrony of the trajectories, and (3) the recognition rates (percent correct and precision) are used to assess specifically the recognition stage.

A jack-knife training procedure is used: the data are split into five partitions approximately homogeneous from the point of view of phone distribution; each partition is used in turn to assess the performances of the HMM models trained with the four remaining partitions. The RMSE and PMCC are calculated over the five test partitions – therefore the whole corpus –, excluding the long pauses at the beginning and the end of each utterance. The recognition rates are also aggregated over the five partitions.

Table 1, which displays the recognition rates, the RMSE and correlation coefficients for the HMM-based inversion, shows that the use of phones in context increases the performances of the inversion. The best results are however not obtained for the phones with both right and left contexts, but for the phones with the right context. This is likely due to the limited size of the corpus (the ratio of the number of missing test phone HMMs over the total number of train phones is on the average over the five training partitions is 4, 4, and 12 % for the *L-ctx*, *ctx-R*, and *L-ctx-R* contexts, respectively).

We found that the use of state durations produced by the recognition stage results in an improvement of about 10 % for RMSE and about 4% for PMCC, compared to the previously used z-scoring method. We found also that the missing HMMs inheritance mechanism increases the recognition performances by 1 to 5 %. The language model increase rates of recognition / accuracy from 72.29 / 34.22 % to 93.66 / 80.90 %. This spectacular improvement has however a low influence on the performances since, in right context, the RMSE goes from 1.83 to 1.66 mm and the correlation from 0.90 to 0.92.
Besides, in order to assess the contribution of the trajectory formation to errors of the complete inversion procedure, we also synthesized these trajectories using a forced alignment of the states based on the original labels, emulating a perfect acoustic recognition stage. From Table 1, we can estimate that the contribution of the trajectory formation stage to the overall RMSE amounts to nearly 90 %. This relatively high level of errors can likely be explained by the fact that the trajectory formation model tends to oversmooth the predicted movements and does not capture properly coarticulation patterns.

Table 1. *Recognition rates (Percent **Cor**rect, **Acc**uracy) aggregated over the whole corpus (1). RMSE (mm) and PMCC for the HMM-based inversion: full inversion (2), with perfect recognition step (3).*

	no-ctx		**L-ctx**		**ctx-R**		**L-ctx-R**	
	Cor	Acc	Cor	Acc	Cor	Acc	Cor	Acc
(1)	88.90	68.99	92.61	78.14	**93.66**	**80.90**	87.12	80.83
	RMSE	PMCC	RMSE	PMCC	RMSE	PMCC	RMSE	PMCC
(2)	2,07	0,87	1,72	0,91	**1,66**	**0,92**	1,91	0,89
(3)	1,91	0,90	1,55	0,93	1,55	0,93	**1,40**	**0,94**

5. Multimodal GMM models

The GMM was trained using the expectation–maximization (EM) algorithm with joint acoustic-articulatory vectors as feature vectors. The GMM-based mapping is then applied using the minimum mean-square error (MMSE) criterion, which has been often used for voice conversion [16] or in acoustic-to-articulatory inversion [7]. Moreover, to improve the mapping performance, the maximum likelihood estimation (MLE) was applied to the GMM-based mapping method as in [7]. The determination of a target parameter trajectory with appropriate static and dynamic properties is obtained by combining local estimates of the mean and variance for each frame $p(t)$ and its derivative $\Delta p(t)$ with the explicit relationship between static and dynamic features (*e.g.* $\Delta p(t) = p(t) - p(t-1)$) in the MLE-based mapping. In order to take into account coarticulation [7] [17], the acoustic information is taken from some time span around the instant of interest. Besides, the dynamics of the articulators is taken into account by considering the time derivate of the articulatory trajectories. Thus, if we denote by $\boldsymbol{Y_{Ac}}(:, 1{:}n_{Ac})$ the matrix of the 12 measured MFCC + log-energy coefficients ($n_{Ac} = 13$) and by $\boldsymbol{Y_{EMA}}(:, 1{:}n_{EMA})$ the matrix of EMA coil coordinates, the feature vector at each time instant indexed by j is the concatenation of '$2n$+1' of vectors of acoustic parameters and of EMA coordinates [PCA($\boldsymbol{Y_{Ac}}(J, 1{:}n_{Ac})$); $\boldsymbol{Y_{EMA}}(j, 1{:}n_{EMA})$; $\boldsymbol{\Delta Y_{EMA}}(j, 1{:}n_{EMA})$], where $\boldsymbol{\Delta}$ denotes first

time derivation, and $J = j+[-n:+n]$ denotes the time instant indices of the set of input frames used for contextual information. The number of input frames was varied from phoneme size (n=4, ~90 ms) to diphone size (n=8, ~170 ms), but the dimension '$(2n+1)\times n_{Ac}$' of the resulting vector was reduced to a fixed value of 24 by Principal Component Analysis (PCA). The number of mixture components was varied from 8 à 64. Each Gaussian is represented by full covariance matrix (48×48), a vector of means (48) and an associated weighting coefficient.

Table 2 displays the performances of the GMM-based inversion for different parameters, using the *jack-knife* method on the same partitions as for the HMMs. The RMSE decreases when the number of mixtures increases and reaches a minimum for a context window of 110 ms. The more likely explanation is that a diphone size window optimally contains the local phonetic features necessary for inversion. The best inversion precision is finally obtained for a combination of a 110 ms window with 64 Gaussians that seems to constitute the best representation of the 36 phonemes. Moreover, we have found that the extra MLE optimisation stage increases the performances by about 5 %.

Table 2. *RMSE (mm) and PMCC for the GMM-based inversion as a fonction of nombre of Gaussians (# mix) and size of context ctw (ms).*

#mix	**8**		**16**		**32**		**64**	
ctw	RMSE	PMCC	RMSE	PMCC	RMSE	PMCC	RMSE	PMCC
90	2,68	0,78	2,61	0,80	2,38	0,83	2,32	0,84
110	2,68	0,78	2,54	0,80	2,37	0,83	**2,25**	**0,85**
130	2,66	0,78	2,51	0,81	2,36	0,83	2,27	0,85
150	2,66	0,78	2,50	0,81	2,44	0,82	2,32	0,84
170	2,65	0,78	2,44	0,82	2,41	0,82	2,29	0,84

6. Comparisons and discussion

Figure 2 displays the statistics of the RMSE of each phoneme for the HMM-based and GMM-based methods. It confirms that the global RMSE obtained with the HMM-based inversion is lower than that obtained with the GMM-based one (the difference is highly significant, $p<10^{-6}$). This result is surprising if we refer to two of the most elaborate experiments available in the literature: Hiroya & Honda [4] found 1.73 mm with HMMs (which is close to our results) whereas Toda *et al.* [7] found 1.36 – 1.45 mm with GMMs. Even taking into account the fact that these experiments were based on different speakers and languages, we did not expect such a difference. A possible explanation for this contrastive behaviour lays perhaps in the fact that GMM-based techniques are more appropriate to deal with unimodal mappings where events in source and targets are largely synchronous, whereas HMM-based techniques are able to deal with context-dependent mappings and delays between frames structured by state transitions.

A more detailed analysis can be found in Figure 4 that displays the phoneme-specific RMSE computed over the centres of all occurrences of each phoneme, sorted in ascending order for the HMMs. It can be observed that the error is higher for back articulations than for coronal ones. No specific trend was observed for the individual RMSE for each coil coordinates, except a lower error for the jaw than for other articulators (see Figure 5).

Another interesting way to analyse the characteristics of the HMM and GMM inversion methods is to compare the measured and resynthesised articulatory spaces of the EMA coils, as done in Figure 3. We see that the space resynthesised by the HMM-based inversion covers almost completely the original space, while the space generated by the GMM-based inversion is quite smaller, especially for the back, mid and lower lip coils. These centralisation effects could be related to the smoothing effects possibly due to the MLE criterion used in both the HMMs and the GMMs.

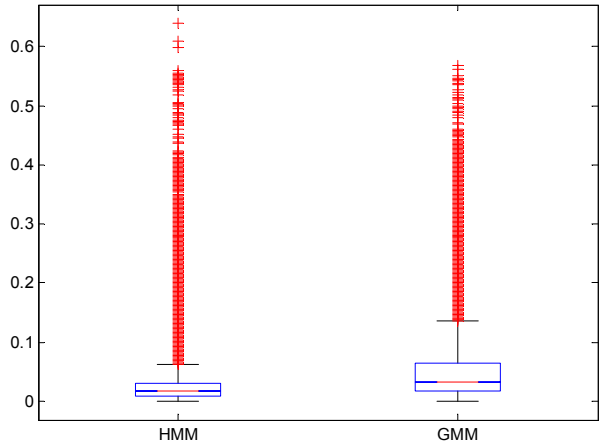

Figure 2. *Comparing RMSE of HMM and GMM reconstruction using anova.*

7. Conclusions and perspectives

We have implemented and compared two acoustic-to-articulatory speech inversion techniques, which contrast in the way they capture and exploit a priori multimodal coherence.

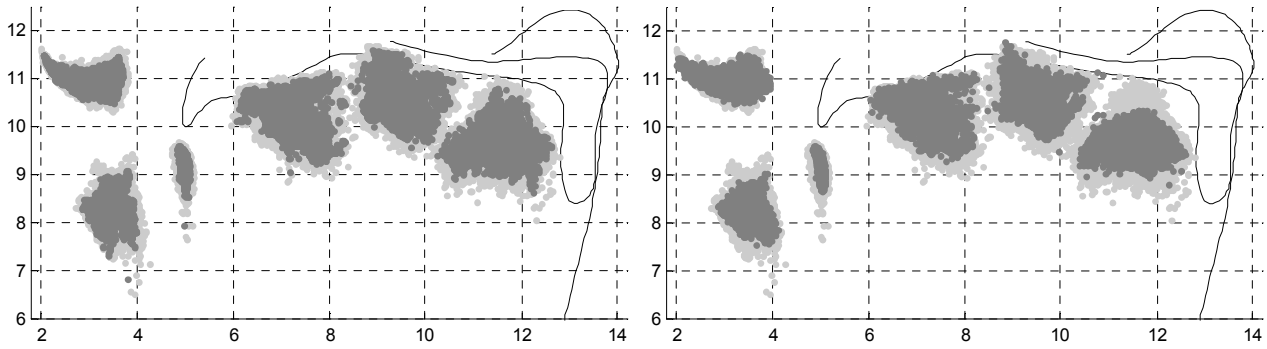

Figure 3. *Articulatory spaces of the EMA coils for the phones sampled at centre. Light grey: measured coordinates. Dark grey synthesized coordinates (top: HMM, bottom: GMM).*

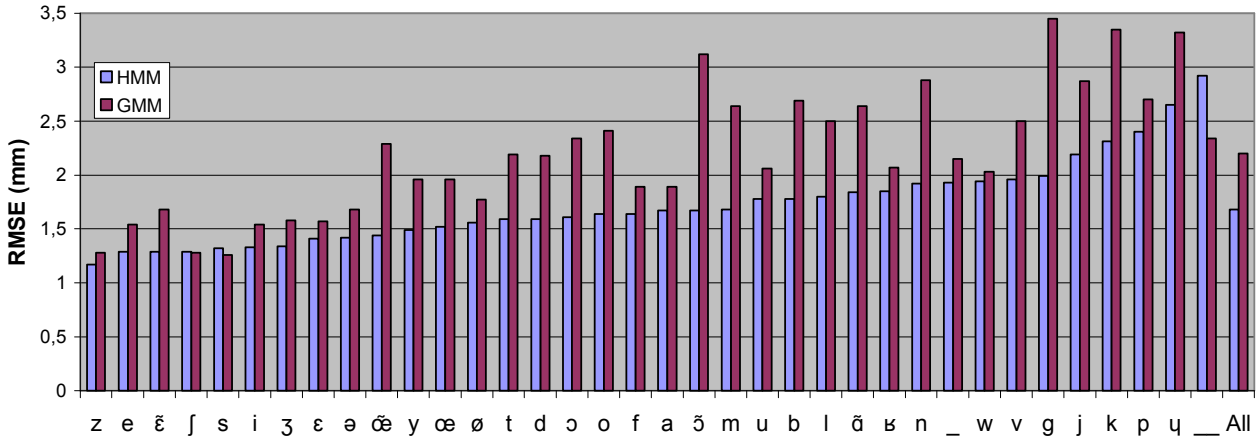

Figure 4. *Individual RMSE for each phoneme.*

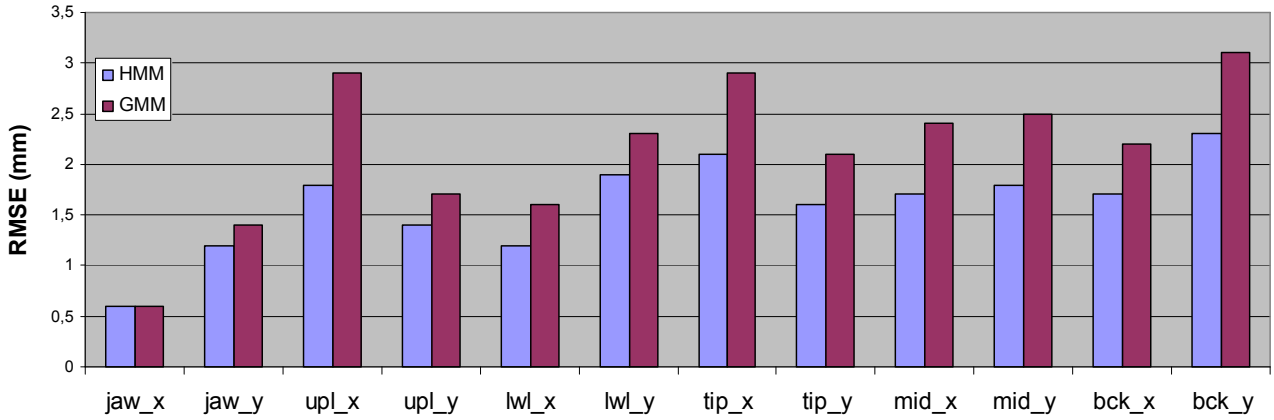

Figure 5. *Individual RMSE for each EMA coil.*

Both systems could be improved. HMM-based inversion can include more sophisticated treatment of articulatory-to-acoustic asynchrony by introducing delay models that have been quite effective in HMM-based multimodal synthesis [18] as well as other optimization criteria such as minimization of reconstruction error [19]. The GMM-based system could be improved by considering other dimensionality reduction techniques such as Linear Discriminant Analysis (LDA) that are quite effective in HMM-based inversion [17]. Both systems could also be improved by incorporating visual information as input and including this additional information more intimately in the optimization process that will consider multimodal coherence between input and output parameters: lips are clearly visible and jaw is indirectly available in facial movements.

This work tends to show that the inversion process should be "phonetic-aware". Several reserves can however be made on these first experiments.

The HMM system benefits from the phonotactics of the target language. Note however that French has a rich syllabic inventory: we can imagine that results obtained with languages such as Japanese, Polish or Spanish with various syllabic complexities may lead to different results.

Global objective measurements may not entirely mirror phone-specific behaviour that may drastically impact subjective rating of generated articulation. The precision of the recovery is of course a highly important element for the evaluation but other elements such as the precision of the recovery of crucial elements such as vocal tract constrictions are naturally also very important.

We have shown elsewhere [20] that viewers have various performance for *tongue reading* and that performance increases with training. Note also that the realism of motion may compensate for inaccurate detailed shaping: the kinematics of the computed trajectories could be more important for perception that the accuracy of the trajectories themselves.

Finally, the results of this study will allow us to develop a tutoring system for on-line phonetic correction [21], in which recovered articulatory movements will be used to drive a virtual 3D talking head with all possible articulatory degrees-of-freedom [22-23].

8. Acknowledgements

We sincerely thank Christophe Savariaux and Coriandre Vilain for the EMA recordings, Tomoki Toda (NAIST, Japan) for making his GMM toolbox available, and Thomas Hueber for fruitful discussions. This work was partially supported by the French ANR-08-EMER-001-02 *ARTIS* and the French-Japanese PHC SAKURA *CASSIS* projects.

9. References

[1] Atal, B.S., Chang, J.J., Mathews, M.V., and Tukey, J.W., "Inversion of articulatory-to-acoustic transformation in the vocal

tract by a computer-sorting technique," *Journal of the Acoustical Society of America*, vol. 63, pp. 1535-1555, 1978.

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

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

[4] 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, 2004.

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

[6] Ben Youssef, A., Badin, P., Bailly, G., and Heracleous, P., "Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden Markov models," presented at Interspeech 2009, Brighton, UK, 2009.

[7] 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.

[8] Badin, P., Tarabalka, Y., Elisei, F., and Bailly, G., "Can you "read tongue movements"?," presented at Interspeech 2008 (Special Session: Talking Heads and Pronunciation Training), Brisbane, Australia, 2008.

[9] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer (Version 4.3.14) [Computer program]. Retrieved May 26, 2005, from http://www.praat.org/," 2005.

[10] Badin, P. and Serrurier, A., "Three-dimensional linear modeling of tongue: Articulatory data and models," presented at 7th International Seminar on Speech Production, ISSP7, Ubatuba, SP, Brazil, 2006.

[11] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X.A., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P., "The HTK Book. Revised for HTK Version 3.4 December 2006," 2006.

[12] Govokhina, O., Bailly, G., Breton, G., and Bagshaw, P., "TDA: A new trainable trajectory formation system for facial animation," presented at InterSpeech, Pittsburgh, PE, 2006.

[13] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T., "Speech parameter generation algorithms for HMM-based speech synthesis," presented at IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 2000.

[14] Tamura, M., Kondo, S., Masuko, T., and Kobayashi, T., "Text-to-audio-visual speech synthesis based on parameter generation from HMM," presented at EUROSPEECH'99, Budapest, Hungary, 1999.

[15] Zen, H., Tokuda, K., and Kitamura, T., "An introduction of trajectory model into HMM-based speech synthesis," presented at Fifth ISCA ITRW on Speech Synthesis (SSW5), Pittsburgh, PA, USA, 2004.

[16] Stylianou, Y., Cappé, O., and Moulines, E., "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 131-142, 1998.

[17] Tran, V.-A., Bailly, G., Loevenbruck, H., and Jutten, C., "Improvement to a NAM captured whisper-to-speech system," presented at Interspeech, Brisbane, Australia, 2008.

[18] Govokhina, O., Bailly, G., and Breton, G., "Learning optimal audiovisual phasing for a HMM-based control model for facial animation," presented at 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, 2007.

[19] Wu, Y.J., Zen, H., Nankaku, Y., and Tokuda, K., "Minimum generation error criterion considering global/local variance for HMM-based speech synthesis," presented at ICASSP, Las Vegas, NE, USA, 2008.

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

[21] Badin, P., Bailly, G., and Boë, L.-J., "Towards the use of a virtual talking head and of speech mapping tools for pronunciation training," presented at Proceedings of the ESCA Tutorial and Research Workshop on Speech Technology in Language Learning, Stockholm - Sweden, 1998.

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

[23] Badin, P., Elisei, F., Bailly, G., and Tarabalka, Y., "An audiovisual talking head for augmented speech generation: Models and animations based on a real speaker's articulatory data," presented at Conference on Articulated Motion and Deformable Objects, Mallorca, Spain, 2008.