HMMs and GMMs based methods in acoustic-to-articulatory speech inversion

Atef Ben Youssef, Viet-Anh Tran, Pierre Badin, Gérard Bailly

HMMs and GMMs based methods for acoustic-to-articulatory speech inversion

Atef Ben Youssef *Viet-Anh Tran* *Pierre Badin* *Gérard Bailly*

GIPSA-lab (Département Parole & Cognition / ICP), UMR 5216 CNRS – Université de Grenoble
961 rue de la Houille Blanche, D.U. - BP 46, F-38402 Saint Martin d'Hères cedex, France
Courriel : {Atef.BenYoussef, Viet-Anh.Tran, Pierre.Badin, Gerard.Bailly}@gipsa-lab.grenoble-inp.fr

RÉSUMÉ

Afin de récupérer les mouvements des articulateurs tels que les lèvres, la mâchoire ou la langue, à partir du son de parole, nous avons développé et comparé deux méthodes d'inversion basées l'une sur les modèles de Markov cachés (HMMs) et l'autre sur les modèles de mélanges de gaussiennes (GMMs). Les mouvements des articulateurs sont caractérisés par les coordonnées médiosagittales de bobines d'un articulographe électromagnétique (EMA) fixées sur les articulateurs. Dans la première méthode, des HMMs à deux flux, acoustique et articulatoire, sont entrainés à partir de signaux acoustique et articulatoire synchrones. Le HMM acoustique sert à reconnaitre les phones, ainsi que leurs durées. Ces informations sont ensuite utilisées par le HMM articulatoire pour synthétiser les trajectoires articulatoires. Pour la deuxième méthode, un GMM d'association directe entre traits acoustiques et articulatoires est entrainé sur le même corpus suivant le critère de minimum d'erreur quadratique moyenne (MMSE) à partir des trames acoustiques d'empan temporel plus ou moins grand. Pour un corpus de données EMA mono-locuteur enregistré par un locuteur français, l'erreur RMS de reconstruction sur le corpus de test pour la méthode fondée sur les HMMs se situe entre 1.96 et 2.32 mm, tandis qu'elle se situe entre 2.46 et 2.95 mm pour la méthode basée sur les GMMs.

1. INTRODUCTION

Speech inversion is a long-standing problem, as testified by the famous work by Atal *et al.* [Ata78] in the seventies. Speech inversion was traditionally based on analysis-by-synthesis, as implemented by [Maw00], or by [Oun05] who optimised codebooks to recover vocal tract shapes from formants. But since a decade, more sophisticated learning techniques have appeared, thanks to the advent of the availability of large corpora of articulatory and acoustic data provided by devices such as the ElectroMagnetic Articulograph (EMA) or marker 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. [Hir04], [Zha08] or [Ben09]), and Gaussian Mixture Models (GMMs) (cf. [Tod08]). 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 [Hir04] 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 phoneme is modelled by a context-dependent HMM, and a separate linear regression mapping is trained at each HMM state 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 [Zha08] 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 steps : 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. [Tod08] 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 [Hir04] and Zhang & Renals [Zha08] have shown that using explicit phonetic information to built HMMs gives better results. Toda and coll. [Tod08], 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 the HMM-based method used in [Ben09] with a GMM-based method similar to that of [Tod08] using the minimum mean-square error (MMSE) criterion for the GMM-based mapping method, everything else being comparable.

3. ARTICULATORY AND ACOUSTIC DATA

3.1. The corpus

For this preliminary study, a corpus already recorded was used [Bad08]. 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 string based on HMMs. Subsequent manual correction of both phoneme labels and phoneme boundaries were performed using the *Praat* software [Boe05]. The centres of allophones were automatically chosen as the average between beginning and end of the phonemes. Altogether the corpus contained 7350 allophones, i.e. about 22 minutes of speech. 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 allows tracking flesh points of the articulators thanks to small electromagnetic receiver coils. 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 [Bad06; Bad08]. 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 22050 Hz, in synchronization with the EMA coordinates, which were recorded at a 500 Hz sampling frequency.

3.3. Overview of the data

In order to reduce the noise of the EMA data, the articulatory trajectories were first low pass filtered with a cut-off frequency of 50 Hz. Then 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 18 (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 coil 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.

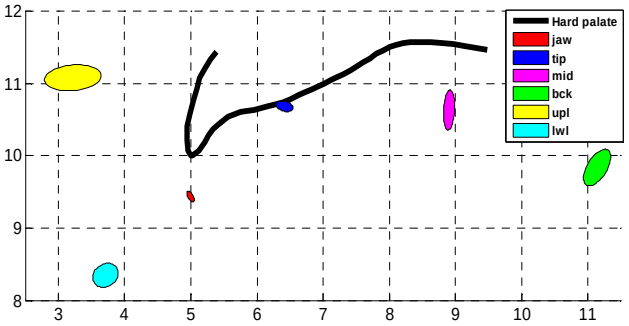

Figure 1: Dispersion ellipses of the measured coordinates of the six EMA coils for phoneme /t/. These ellipses are computed from the samples taken at the middle of the 231 instances of /t/ in the corpus.

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 City-Block distances of the coils coordinates between each pair of phone. Each allophone was represented by its mean over all the associated instances. Using hierarchical clustering to generate dendrograms allowed to define six coherent classes for vocalic contexts ([a ɛ ɛ̃ | ø œ œ̃ | e i | y | u | o ɔ ɑ̃ ɔ̃]) shown as a confusion tree in Figure 2 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]) shown as a confusion tree in Figure 3. 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.

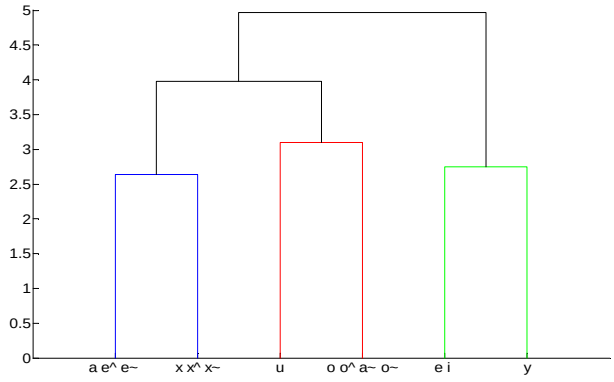

Figure **2:** Dendrogram of 6 classes for vocalic contexts

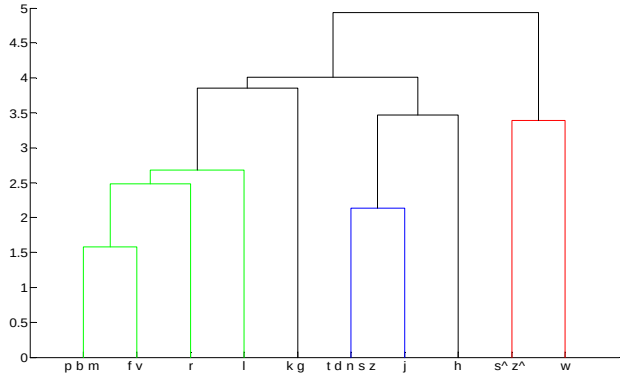

Figure 3: Dendrogram of 10 classes for consonantal contexts

4. ARTICULATORY AND ACOUSTIC HMMS MODELS

We recall the experiments made in [Ben09]. 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 down sampled to 16 kHz over 25 ms windows at a frame rate of 100 Hz. 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 [You06]. The training is performed using the Expectation Maximization (EM) algorithm based on the Maximum Likelihood (ML) criterion.

The acoustic and articulatory features vectors are here 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. Thus, the recognised phoneme sequences respect French phonotactics. No prosodic constraints such as a duration model are added. The acoustic-to-articulatory inversion is achieved in two steps. The first step performs phoneme recognition, based on the acoustic HMMs. The result is a sequence of recognised allophones together with their durations. The recognition results are given in the Table 1. The recognition performances are increased by the use of phonemes in context. The procedure of replacement of the HMMs, which aims to compensate for the too small size of the training sets (cf. [Ben09]), increases the recognition rate from 76.9 to 84.0 %.

Table 1 : Recognition rates (Percent Correct, Accuracy) for the experiments with different types of contexts. The star * indicates the series of experiments for which missing HMMs were replaced by the closest model.

Train - Test	no-ctx		L-ctx		ctx-R		L-ctx-R	
Nb phones	Nb		Nb		Nb		Nb	
Cor, Acc (%)	Cor	Acc	Cor	Acc	Cor	Acc	Cor	Acc
1 - 1	36		392		387		1376	
	89,67	70,41	97,63	91,29	98,12	92,95	99,28	97,88
2/3 - 1/3	36		366		358		1159	
	88.09	67.91	85.56	64.66	87.57	66.19	76.90	68.59
2/3 - 1/3 *			385		379		1312	
			89,09	72.30	**91,46**	77,18	**84,04**	75.69

The second step of the inversion aims at reconstructing the articulatory trajectories from the chain of phoneme labels and boundaries delivered by the recognition procedure. As described in [Gov06], the synthesis is performed as follows, using the software developed by the HTS group [Tam99; Zen04]. A linear sequence of HMM states is built by concatenating the corresponding phone HMMs. The proper state durations are estimated by z-scoring. A sequence of observation parameters is generated using a specific ML-based parameter generation algorithm [Zen04]. In order to assess the contribution of the trajectory formation to RMS errors of

the complete inversion method, we also synthesised these trajectories directly from the original labels, simulating a perfect acoustic recognition step.

The root-mean-square (RMS) error is calculated for the difference between the measured and the estimated articulatory coordinates, excluding the long pauses at the beginning and the end of each utterance. The mean correlation coefficient (Corr) measures the degree of amplitude similarity and the synchrony of the trajectories. Table 2 shows the RMS errors. The results are consistent with those in Table 1. In addition, Table 2 displays the errors corresponding to cases where the recognition step is bypassed by synthesising the articulatory trajectories directly from the original phoneme sequences; the relatively high level of these errors shows that a significant part of the overall error is due to the trajectory formation model that often oversmooths the predicted movements and does not capture properly coarticulation patterns. Note that all the differences in Table 2 are significant ($p < 0.03$).

Table 2: RMS errors (mm) and correlation coefficients for the experiments with different types of contexts. The star * indicates the series of experiments for which missing HMMs are replaced by the closest model. The ^ indicate that the synthesis is generated from the original labels.

Train - Test		no-ctx	L-ctx	ctx-R	L-ctx-R
1 - 1	RMS	2.26	1.62	1.62	**1.05**
	Corr	0.72	0.82	0.83	**0.90**
2/3 - 1/3	RMS	2.32	2.15	**2.06**	2.31
	Corr	0.70	0.71	**0.73**	0.69
2/3 - 1/3 *	RMS		2.07	**1.96**	2.08
	Corr		0.72	**0.75**	0.73
2/3 - 1/3 ^	RMS	2.21	1.86	1.87	**1.74**
	Corr	0.75	0.77	0.75	**0.82**

5. MULTIMODAL GMM MODELS

We apply the GMM-based mapping using the minimum mean-square error (MMSE) criterion, which has been often used for voice conversion. The determination of a target parameter trajectory with appropriate static and dynamic properties is obtained here by combining local estimates of the mean and variance for each frame p(t) and its derivative Δp(t) with the explicit relationship between static and dynamic features (e.g. $\Delta p(t) = p(t) - p(t-1)$) in the MMSE-based mapping.

The 1st to 13th Mel-cepstral coefficients are used as a spectral representation of the speech signal. The shift duration is also 10 ms. The 12-dimensional EMA data is accordingly down sampled to match this 100 Hz sampling rate.

The number of mixture components is varied from 8 to 32. The number of input acoustic frames is fixed to 9 but the size of context window is varied from a phoneme size (~100 ms) to a syllable size (330 ms) (by picking one frame every 1-4 frames). A reduction of the acoustic dimension (9*13 = 117) is performed by Principal Component Analysis (PCA): the number of principal components is set to 24.

It has been reported in the literature that the low pass filtering of training as well as estimated articulatory trajectories improves the mapping performance (e.g., Richmond, 2001). The optimal cut-off frequency of the low pass filter is deemed to be 25Hz.

The RMS error and the correlation coefficient are calculated over the same training and test data used for evaluating the HMM system.

Table 3 shows the RMS error (in mm) and the correlation coefficient for the different experiments using different numbers of mixtures and different sizes of context window. The RMS error decreases when the number of mixtures increases and when the size of context window decreases. The most plausible interpretation is that a phoneme-sized window optimally contains necessary local phonetic cues for inversion. The 32 mixtures appear to constitute the best representation of the 36 phonemes.

The better mapping accuracy is finally achieved when the size of the context window is set to 90 ms and the number of mixture components is set to 32 in this experiment. In that case, the RMS error is 2.46 mm.

Table 3: RMS errors and correlation coefficient with different numbers of mixtures (# mix) and sizes of context window (ctw)

ctw	90		170		250		330	
#mix	RMS	Corr	RMS	Corr	RMS	Corr	RMS	Corr
8	2.87	0.56	2.77	0.59	2.84	0.57	2.95	0.55
16	2.76	0.58	2.61	0.62	2.67	0.60	2.94	0.55
32	**2.46**	**0.63**	2.55	0.63	2.61	0.60	2.87	0.57

6. COMPARISON AND COMMENTS

Figure 4 displays the measured and reconstructed articulatory trajectories of the Y-coordinates for the two competing systems. It seems that the GMM-based method has difficulty in dealing with the asynchronous behaviour of inter-articulatory coordination.

Our HMM-based system generates an RMS error of 1.96 mm for the same data. Surprisingly, the performance of the HMM-based inversion mapping is significantly more accurate than that of the GMM-based system although results published on voice–conversion experiments seem to suggest the opposite. 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.

Both systems can however be improved. HMM-based inversion can include more sophisticated treatment of articulatory-to-acoustic asynchrony by introducing delay

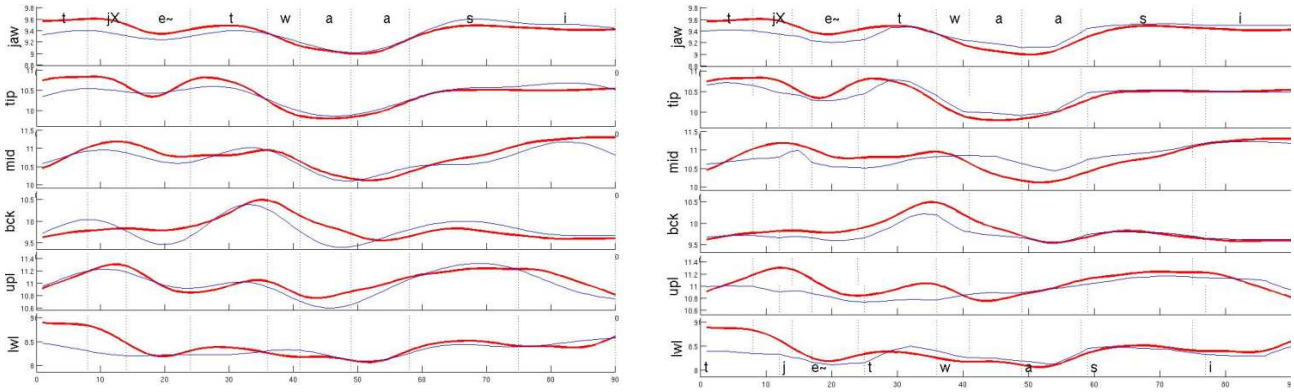

Figure 4: Comparing original (thick lines) and synthesized (thin lines) trajectories of ordinates of 6 EMA fleshpoints computed from the acoustic signal. Left: HMM-based inversion and trajectory formation using phone-sized Markov models with right context; right: synthesis by GMM-mapping using a context window of 90 ms and a mixture of 32 Gaussians.

models that have been quite effective in HMM-based multimodal synthesis [Gov07]. 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 [Tra08]. Both systems can 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.

Figure 5 displays the statistics of the RMS errors for the HMM-based and GMM-based methods. The difference is highly significant ($p<10^{-7}$). Figure 5 shows that the HMM-based system produces a global RMS error lower than that produced by the GMM-based one, but produces in some cases errors that are higher than the highest errors obtained with the GMM system.

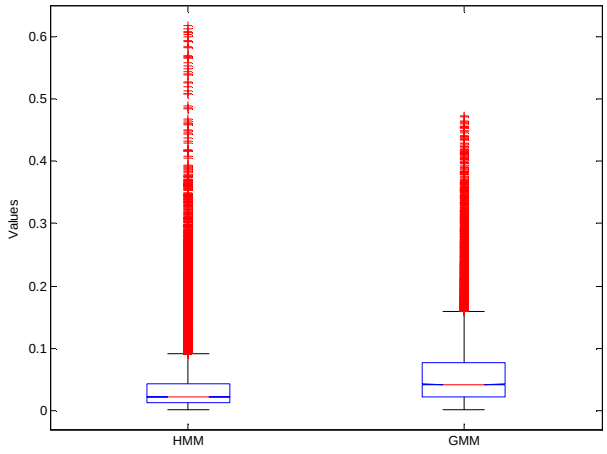

Figure 5: Comparing RMS error of HMM and GMM reconstruction using anova1.

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

First, the HMM system benefits from the phonotactics of the target language. Despite the fact that the language model has not been trained on test data, the phonological structures of test and training utterances are very similar. Note however that French has a rich syllabic inventory and that we can imagine that results obtained with languages such as Japanese, Polish or Spanish with various syllabic complexities may lead to different results.

Secondly, 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.

Thirdly, we have shown elsewhere [Tar07] 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 [Bad98], in which recovered articulatory movements will be used to drive a virtual 3D talking head with all possible articulatory degrees-of-freedom [Bad02; Bad08].

8. Acknowledgments

We sincerely thank Christophe Savariaux for helping us with the EMA recordings. This work has been partially supported by the French ANR ARTIS grant and the French-Japanese PHC SAKURA CASSIS project.

REFERENCES

[Ata78] Atal, B. S., J. J. Chang, M. V. Mathews and J. W. Tukey (1978) Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *Journal of the Acoustical Society of America*, **63**, pp. 1535-1555.

[Bad98] Badin, P., G. Bailly and L.-J. Boë (1998) Towards the use of a virtual talking head and of speech mapping tools for pronunciation training *Proceedings of the ESCA Tutorial and Research Workshop on Speech Technology in Language Learning* (Stockholm - Sweden.

[Bad02] Badin, P., G. Bailly, L. Revéret, M. Baciu, C. Segebarth and C. Savariaux (2002) Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images. *Journal of Phonetics*, **30: 3**, pp. 533-553.

[Bad08] Badin, P., F. Elisei, G. Bailly and Y. Tarabalka (2008) An audiovisual talking head for augmented speech generation: Models and animations based on a real speaker's articulatory data *Conference on Articulated Motion and Deformable Objects* (Mallorca, Spain, Springer LNCS) pp. 132-143.

[Bad06] Badin, P. and A. Serrurier (2006) Three-dimensional linear modeling of tongue: Articulatory data and models, in: H. C. Yehia, D. Demolin and R. Laboissière (eds.) *Proceedings of the 7th International Seminar on Speech Production, ISSP7* (Ubatuba, SP, Brazil, UFMG, Belo Horizonte, Brazil) pp. 395-402.

[Bad08] Badin, P., Y. Tarabalka, F. Elisei and G. Bailly (2008) Can you "read tongue movements"? *Interspeech 2008* (Brisbane, Australia pp. 2635-2638.

[Ben09] Ben Youssef, A., P. Badin, G. Bailly and P. Heracleous (2009) Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden Markov models *Interspeech 2009* (Brighton, UK.

[Boe05] Boersma, P. and D. Weenink (2005) Praat: doing phonetics by computer (Version 4.3.14) [Computer program]. Retrieved May 26, 2005, from http://www.praat.org/.

[Gov07] Govokhina, O., G. Bailly and G. Breton (2007) Learning optimal audiovisual phasing for a HMM-based control model for facial animation *6th ISCA Workshop on Speech Synthesis* (Bonn, Germany.

[Gov06] Govokhina, O., G. Bailly, G. Breton and P. Bagshaw (2006) TDA: A new trainable trajectory formation system for facial animation *InterSpeech* (Pittsburgh, PE pp. 2474-2477.

[Hir04] Hiroya, S. and M. Honda (2004) Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. *IEEE Trans. Speech and Audio Processing*, **12: 2**, pp. 175-185.

[Maw00] Mawass, K., P. Badin and G. Bailly (2000) Synthesis of French fricatives by audio-video to articulatory inversion. *Acta Acustica*, **86: 1**, pp. 136-146.

[Oun05] Ouni, S. and Y. Laprie (2005) Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *Journal of the Acoustical Society of America*, **118: 1**, pp. 444-460.

[Tam99] Tamura, M., S. Kondo, T. Masuko and T. Kobayashi (1999) Text-to-audio-visual speech synthesis based on parameter generation from HMM *EUROSPEECH'99* (Budapest, Hungary pp. 959-962.

[Tar07] Tarabalka, Y., P. Badin, F. Elisei and G. Bailly (2007) Can you "read tongue movements"? Evaluation of the contribution of tongue display to speech understanding *Conférence Internationale sur l'Accessibilité et les systèmes de suppléance aux personnes en situation de Handicaps (ASSISTH)* (Toulouse - France pp. 187-193.

[Tod08] Toda, T., A. W. Black and K. Tokuda (2008) Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication*, **50: 3**, pp. 215-227.

[Tra08] Tran, V.-A., G. Bailly, H. Loevenbruck and C. Jutten (2008) Improvement to a NAM captured whisper-to-speech system *Interspeech* (Brisbane, Australia pp. 1465-1468.

[You06] Young, S., G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland (2006) The HTK Book. Revised for HTK Version 3.4 December 2006.

[Zen04] Zen, H., K. Tokuda and T. Kitamura (2004) An introduction of trajectory model into HMM-based speech synthesis *Fifth ISCA ITRW on Speech Synthesis (SSW5)* (Pittsburgh, PA, USA pp. 191-196.

[Zha08] Zhang, L. and S. Renals (2008) Acoustic-articulatory modeling with the trajectory HMM. *IEEE Signal Processing Letters*, **15**, pp. 245-248.