INTERSPEECH
2009 BRIGHTON

Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden Markov models

Atef Ben Youssef, Pierre Badin, Gérard Bailly, Panikos Heracleous

GIPSA-lab (Département Parole & Cognition / ICP), UMR 5216 CNRS – Universités de Grenoble,
961 rue de la Houille Blanche, D.U. - BP 46, F-38402 Saint Martin d'Hères cedex, France
Pierre.Badin@gipsa-lab.inpg.fr

Abstract

In order to recover the movements of usually hidden articulators such as tongue or velum, we have developed a data-based speech inversion method. HMMs are trained, in a multistream framework, from two synchronous streams: articulatory movements measured by EMA, and MFCC + energy from the speech signal. A speech recognition procedure based on the acoustic part of the HMMs delivers the chain of phonemes and together with their durations, information that is subsequently used by a trajectory formation procedure based on the articulatory part of the HMMs to synthesise the articulatory movements. The RMS reconstruction error ranged between 1.1 and 2. mm.

Index Terms: Speech inversion, augmented speech, automatic speech recognition, HTK, Electro-Magnetic Articulography (EMA), hidden Markov model (HMM), trajectory formation, HTS.

1. Introduction

There is strong evidence that human speakers/listeners exploit the articulatory origin of speech: the view of visible articulators, i.e. jaw and lips, improves speech intelligibility [1], speech imitation is faster when listeners perceive articulatory gestures [2], and the vision of hidden articulators still increases intelligibility [3]. More recently, brain studies have evidenced the recruitment of motor areas during speech perception, which supports the motor theory of speech perception [4]. Our laboratory is thus involved in the development of an *inversion* system that allows producing *augmented speech* from the speech 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.

Speech inversion is a long-standing problem, as testified by the famous work by Atal *et al.* [5] in the seventies. Speech inversion was traditionally based on analysis-by-synthesis, as implemented by [6], or by [7] 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 or marker tracking devices based on classical or infrared video.

2. State-of-the-art

Hiroya & Honda [8] have developed a method that determines articulatory movements from speech acoustics using a hidden Markov model (HMM)-based speech production model. After proper labelling of the training corpus, each allophone 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 maximising a posteriori estimation.

Toda *et al.* [9] modelled the joint probability density of an articulatory parameter and an acoustic parameter using a Gaussian Mixture Model (GMM) based on a parallel acoustic–articulatory speech database, in order to establish both an articulatory-to-acoustic mapping and an acoustic-to-articulatory inversion mapping without using phonetic information.

Kjellström & Engwall [10] implemented audiovisual-to-articulatory inversion using either simple multilinear regression or Artificial Neural Networks. Depending on the type of fusion (early or late) between the audio signal and the video signal (based on independent component images of the mouth region), they obtained RMS reconstruction errors for the tongue shape ranging from 2.5 to 3 mm.

Katsamanis *et al.* [11] approximated the audiovisual-to-articulatory mapping by an adaptive piecewise linear model. Model switching was governed by a Markovian discrete process which captures articulatory dynamic information. Each constituent linear mapping is effectively estimated via canonical correlation analysis. For facial analysis, active appearance models (AAMs) demonstrated fully automatic face tracking and visual feature extraction capabilities. Exploiting both audio and visual modalities in a multistream hidden Markov model based scheme, they found RMS errors ranging from 0.5 to 2.5 mm, depending on the articulator involved.

This article evaluates a method for acoustic-to-articulatory inversion based on jointly trained acoustic and articulatory phone HMM models that proceeds in two steps: a procedure of phoneme recognition of the uttered acoustic speech signal by means of the acoustic part of the phone HMMs, followed by a procedure of speech synthesis by articulatory trajectory formation using the articulatory part of the phone HMMs.

3. Articulatory and acoustic data

3.1. The corpus

Training phone HMMs necessitates an appropriate corpus of speech. For this preliminary study, a corpus already recorded was used [3]. 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 phone labels and phone boundaries were performed using the *Praat* software [12]. The centre of each phone was automatically chosen as the average between its beginning and end. Altogether the corpus contained 7352 phones, i.e. about 12 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 [3, 13]. 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

Before starting the modelling procedures, we explore the articulatory data by computing and displaying the dispersion ellipses of the six coils in the midsagittal plane for each phoneme corresponding to a standard deviation of one. 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 displays these ellipses for phoneme /t/, and shows for instance that the variability of the tongue tip coil is very low 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 this instant corresponds to the actual centre of the phone if the trajectories are not symmetrical.

3.4. Grouping phonemes in context classes

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 Euclidian distances of the coils coordinates between each pair of phone. Each allophone was represented by its mean over all the associated instances.

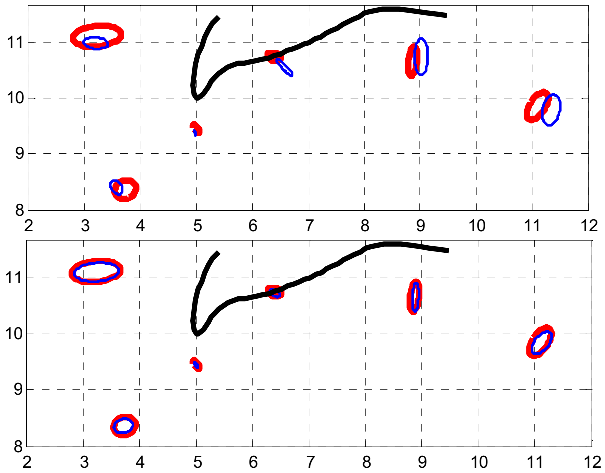


Figure 1. Dispersion ellipses of the original coordinates (thick lines) for phoneme /t/ for all contexts pooled. The reconstructed dispersion ellipses (thin lines), are also displayed for the no-ctx condition (top,) and for the L-ctx-R condition (bottom). The shape of the upper incisors and hard palate is displayed for reference purpose.

Using hierarchical clustering to generate dendrograms allowed to define six coherent classes for vocalic contexts ([a ɛ ɛ̃ | o ɔ ɑ̃ ɔ̃ | e i | u | ø œ œ̃ | y]) and eleven coherent classes for consonantal contexts ([p b m | ə _ | f v | ʁ | l | ʃ ʒ | t d s z n | j | ɥ | k g | w | __]). Using acoustic spectral distances did lead to classes less satisfactory from the point of view of phonetic knowledge.

4. Articulatory and acoustic HMM models

4.1. Feature vectors

Acoustic feature vectors consist of the 12 Mel-Frequency Cepstral Coefficients (MFCC) and of the logarithm of the energy, along with their 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 consist of the x and y coordinates of the six active coils. Their first time derivatives are also added. Note that the coils trajectories are down sampled to a frame rate of 100 Hz, synchronous with the MFCC + Log Energy frames.

4.2. Various context for the phonemes

Four different contextual schemes are tested: phonemes without context (no-ctx) (36 in the whole corpus), with left context (L-ctx) (392), with right context (ctx-R) (387) and with left and right contexts (L-ctx-R) (1376). For the determination of the contexts, the schwa and the short pause are supposed targetless, i.e. they are removed from the phonetic chain in order to take into account the next preceding or following target phoneme.

4.3. Articulatory and acoustic HMM models

Left-to-right with no skip, 3-state phone HMMs with one Gaussian per state and a diagonal covariance matrix are used. For training and test the HTK3.4 toolkit is used [14].

The training is performed using the Expectation Maximization (EM) algorithm based on the Maximum Likelihood (ML) criterion.

In order to ensure that acoustic and the articulatory HMMs have the same phone boundaries (and even same states boundaries within phone), the acoustic and articulatory features vectors are considered as two streams in the HTK multistream training procedure. Subsequently, the HMMs obtained are split into *articulatory HMMs* and *acoustic*

HMMs (this is compatible with the choice of diagonal covariance matrices, since no correlations between the acoustic and articulatory variables are taken into account in this case).

4.4. Language model

A bigram language model was trained over the complete corpus, for each type of context. Thus, the recognised phoneme sequences respect French phonotactics.

Due to the limited size of the training sets, some phonemes in context were missing. In order to overcome this problem, each missing L-ctx-R model inherits properties of the corresponding ctx-R model if it existed and by the no-ctx model if this latter model do not exist either.

5. Inversion procedures and evaluation

5.1. Phoneme recognition from acoustics

The acoustic-to-articulatory inversion is achieved in two steps. The first step performs phoneme recognition, based on the acoustic HMMs. No duration model is used. The result is a sequence of recognised phonemes, with their durations.

Two cases have been used to evaluate this procedure. The first evaluation uses the entire corpus for both training and recognition, and serves as maximum performance. The second evaluation, closer to reality, uses two third of the corpus for training (740 utterances, 4859 allophones) and the remaining third for testing (369 utterances, 2493 instances). The recognition results are given in the Table 1. The recognition performances are increased by the use of phonemes in context. Note however that, the good performance obtained for L-ctx-R when the whole corpus is used for training is due to overtraining: when the training set is reduced, the use of both left and right contexts decreases the recognition scores. Note however that the replacement of the missing HMMs, which aims to compensate for the too small size of the training sets (cf. 4.4), increases the recognition rate from 78.5 to 84.8 %.

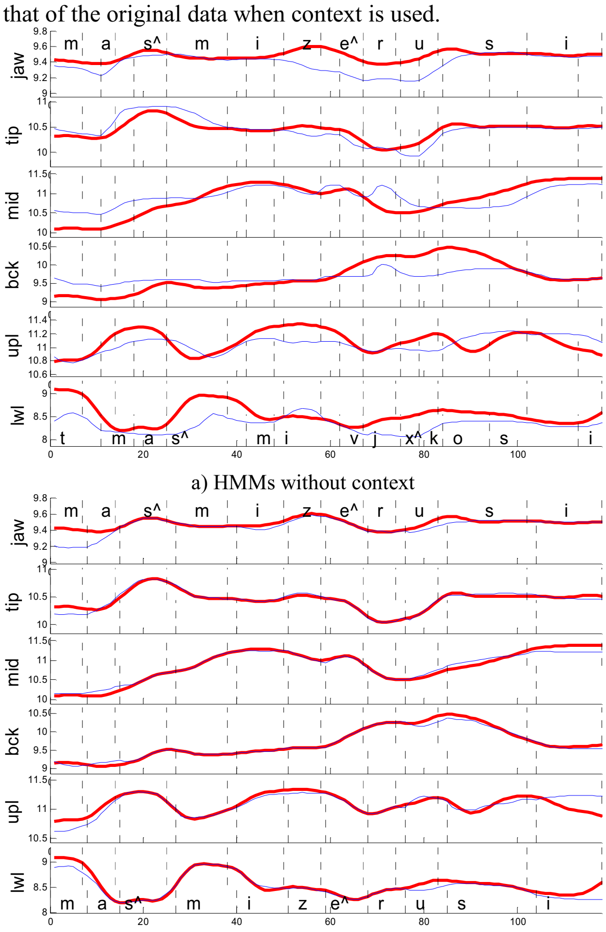
Table 1. *Recognition rates (Percent Correct, Accuracy) for the experiments with different types of contexts. The numbers of phones present in the training sets are displayed. The star * indicates the series of experiments for which missing HMMs were replaced by the closest model.*

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	
	89,7	70,4	97,6	91,3	98,1	93,0	**99,3**	97,9
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	
			89,1	72.3	**91,5**	77,2	**84,0**	75.69

5.2. Articulatory synthesis by trajectory formation

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 [15], the synthesis is performed as follows, using the software developed by the HTS group [16, 17]. A linear sequence of HMM states is built by concatenating the corresponding segmental 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 [17]. No maximisation of variance is performed.

Figure 2, that illustrates the measured and reconstructed time trajectories of the coil coordinates when training and testing sets are identical, for the no-ctx and L-ctx-R context conditions, shows that the use of both left and right contexts improves notably the inversion. Figure 1 confirms that the variability of the reconstructed coordinates is much closer to that of the original data when context is used.



a) HMMs without context

b) HMMs with left and right contexts

Figure 2. Example of measured (thick lines) and reconstructed (thin lines) articulatory trajectories of the coil coordinates, in the case of training on the whole corpus for the sentence [maʃmizɛʁusi] ("Ma chemise est roussie"). For each plot, the phone boundaries are shown by the upper vertical bars for the measured trajectories, and by the lower vertical bars for the reconstructed ones.

5.3. Evaluation

In order to evaluate this second step, or in fact the complete inversion chain, we computed the RMS errors between the measured coils coordinates and the corresponding reconstructed ones over all the frames of the test set. Table 2 gives these errors for the different cases. The case where the testing set is identical to the training set and that constitutes the upper boundary of possible level of errors lead to an RMS error of 1.05 mm, for phone HMMs with left and right contexts, of course higher than the errors for the phone HMMs with only one or no context. The error in the more realistic case where the training set is two third of the whole corpus amounts to 2.3 mm. Note that maximum errors are still very high (between 15 and 20 mm), mostly located at the beginning or the end of the sentences. In order to assess the contribution of the trajectory formation to RMS errors of the complete inversion method, we have synthesised these trajectories directly from the original labels, simulating a perfect acoustic recognition step, Table 2 shows the reduction of the RMS error. Note that the reduction is significant ($p < 0.03$) when the training set is two third of the whole corpus, and not significant ($p > 0.65$) when the training set is the whole corpus. We conclude that a significant part of the overall error is due to the trajectory formation step.

For comparison purposes, note that Hiroya & Honda [8] found, using 5% of sentences as test set, an average RMS error of 1.5 mm for the inversion from the speech acoustics and the phonemic information in an utterance, and 1.7 mm from the speech acoustics only. More recently, Toda *et al.* [9] obtained, using a 1/5 cross-validation test, RMS errors of 1.6 mm and 1.5 mm, for the female and male speaker of the MOCHA data bases respectively, when using an MMSE-based mapping for 32 GMMs mixtures components. In the case of an MLE-based mapping, they found RMS errors of 1.4 mm with 64 mixture components. They reported that Richmond *et al.* [18] obtained an RMS error of 1.6 mm for the same female speaker using a multilayer perceptron. Recall that Kjellström & Engwall [10] obtained RMS errors between 2.5 to 3 mm, while Katsamanis *et al.* [11] found RMS errors between 0.5 to 2.5 mm.

Considering the relatively small size of our corpus, and the relatively small size of our training test, our best result, i.e. 2. mm, compares well with these results. Note also that the upper bound reference of 1.1 mm for the close test is very promising.

Table 2. *RMS error and correlation coefficient 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. The ^ indicate that the synthesis is generated from the originals labels.*

Train - Test	no-ctx		L-ctx		ctx-R		L-ctx-R	
RMS(mm), Corr	RMS	Corr	RMS	Corr	RMS	Corr	RMS	Corr
1 - 1	2.26	0.72	1.62	0.82	1.62	0.83	**1.05**	**0.90**
2/3 - 1/3	2.32	0.70	2.15	0.71	**2.06**	**0.73**	2.31	0.69
2/3 - 1/3 *			2.07	0.72	**1.96**	**0.75**	2.08	0.73
1 - 1 ^	2.16	0.76	1.64	0.81	1.70	0.77	**1.11**	**0.89**
2/3 - 1/3 *^	2.21	0.75	1.87	0.75	1.86	0.77	**1.74**	**0.82**

6. Conclusions and perspectives

These preliminary results are encouraging, as they are a number of issues to explore to improve the performance of our inversion system. A larger corpus will be recorded, in order to increase the representativity of the HMMs, especially for the L-ctx-R context condition. Classical improvements such as increasing the number of Gaussians per state, using tied states for the articulatory stream, as well as tied mixtures will be explored. As the major aim is not phonemic recognition per se, but acoustic-to-articulatory inversion, it would be very interesting to strengthen the links between the recognition and articulatory synthesis stages: using recognised state durations instead of z-scoring at the synthesis stage, and using reconstruction error as optimisation criterion in the training stage for instance. Learning state-level phasing models between acoustics and articulatory boundaries as done by Govokhina *et al.* [19] is also planed. Finally, the use of additional visual information, for instance as AAM, would surely be beneficial.

7. Acknowledgements

We sincerely thank Ch. Savariaux and C. Vilain for helping us with the EMA recordings. This work has been partially supported by the French ANR-08-EMER-001-02 grant.

8. References

[1] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, pp. 212-215, 1954.

[2] C. A. Fowler, J. M. Brown, L. Sabadini, and J. Weihing, "Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks," *Journal of Memory & Language*, vol. 49, pp. 396-413, 2003.

[3] P. Badin, Y. Tarabalka, F. Elisei, and G. Bailly, "Can you "read tongue movements"?," presented at Proceedings of Interspeech 2008, Brisbane, Australia, 2008.

[4] J.-L. Schwartz, M. Sato, and L. Fadiga, "The common language of speech perception and action: a neurocognitive perspective," *Revue Française de Linguistique Appliquée - Communiquer par la parole: des processus complexes*, vol. XIII-2, pp. 9-22, 2008.

[5] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "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.

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

[7] S. Ouni and Y. Laprie, "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.

[8] S. Hiroya and M. Honda, "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.

[9] 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.

[10] H. Kjellström and O. Engwall, "Audiovisual-to-articulatory inversion," *Speech Communication*, vol. 51, pp. 195-209, 2009.

[11] A. Katsamanis, G. Papandreou, and P. Maragos, "Face Active Appearance Modeling and speech acoustic information to recover articulation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, pp. 411-422, 2009.

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

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

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

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

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

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

[18] K. Richmond, S. King, and P. Taylor, "Modelling the uncertainty in recovering articulation from acoustics," *Computer Speech and Language*, vol. 17, pp. 153-172, 2003.

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