How can acoustic-to-articulatory maps be constrained?

Yves Laprie, Petros Maragos, Jean Schoentgen

HOW CAN ACOUSTIC-TO-ARTICULATORY MAPS BE CONSTRAINED?

Yves Laprie, Petros Maragos, and Jean Schoentgen

LORIA/CNRS
Nancy, France
Yves.Laprie@loria.fr

School of E.C.E
National Technical
University of Athens
Athens, Greece
maragos@cs.ntua.gr

L.I.S.T.
Université Libre de Bruxelles &
Fund for Scientific Research
Brussels, Belgium
jschoent@ulb.ac.be

ABSTRACT

The objective of the presentation is to examine issues in constraining acoustic-to-articulatory maps by means of facial data and other a priori knowledge regarding speech production. Constraints that are considered are the insertion of data on lip opening, spread and protrusion, as well as other facial data together with constraints on the vocal tract length. A priori knowledge that has been taken into account concerns the deformation and speed of deformation of the vocal tract as well as phonetic rules regarding vowel-typical tract shapes. Inverse maps that have been tested are formant-to-area and formant-to-parametric sagittal profile maps as well as audio/visual-to-electromagnetic coil trajectory maps. The results obtained while mapping audio-only data compared to audio combined with other data are discussed.

1. INTRODUCTION

Strong evidence exists in favor of the assumption that human speakers and listeners exploit the multimodality of speech, visual articulatory cues in particular. Indeed, the ability to observe articulators directly, such as the jaw and lips, improves the intelligibility of speech. Also, one knows from neurophysiology that a close link exists between articulatory and acoustic cognitive representations of speech, with the sensorimotor control of speech production being represented in so-called mirror neurons.

Even so, audiovisual-to-articulatory inversion is an unsolved problem at present. One major difficulty is the lack of a one-to-one mapping between the acoustic and articulatory domains. Therefore, several distinct vocal tract shapes may produce the same speech spectrum. In a fashion, the inverse map is under-determined. Accurately performing inverse mapping requests data that are not available either in quantity or kind. One important issue is therefore the addition of constraints to eliminate implausible solutions, constraints which must at the same time be restrictive and phonetically realistic. Restrictions may be implemented in the inverse map itself or applied a posteriori for selecting plausible solutions among several possible ones.

In practice, the origin of these restrictions is observed articulatory data of the vocal tract that enable developing speech production models with a view to acoustic-to-articulatory inversion, and/or visual data of the speaker's face that inform on visible articulators, in analogy with human speakers/listeners.

Broadly speaking, two general frameworks for acoustic-to-articulatory mapping exist. The first is the analysis by synthesis framework. In that, inverse mapping consists in searching parameters of a model that account best for the observed signal. This is achieved via a local or global optimization.

The synthesis stage involves an articulatory synthesizer that generates vocal tract shapes and an acoustical model that simulates wave propagation. Often, area function models are formed by concatenating elementary acoustic tubes that are conical or cylindrical. The area function is the cross-section of the vocal tract as a function of the distance from the glottis. Area function models are components of articulatory models or are used autonomously. When used autonomously, their advantage as well as disadvantage is their large number of degrees of freedom, which enable outputting a wide range of sounds, but without the guarantee that the area function shape is anatomically valid. Articulatory models, on the contrary, enable producing vocal tract shapes that are likely to agree with human anatomy. Their parameters (about ten) either describe the position of articulators [1, 2] or specify the weight of principal components (empirical deformation modes) derived from X-ray or IRM images of the vocal tract [3, 4].

One may keep in mind that the flexibility of an articulatory model comprises two facets. The first concerns the ease with which they may be adapted to the morphology of a given speaker (e.g. total length of the vocal tract, lengths of the mouth and pharynx cavities, or contour of the fixed vocal tract wall). The second aspect is their range of possible articulatory gestures. Flexibility per se is not a guarantee that articulatory models are able to approximate any observed vocal tract shape. Even so, articulatory models may generate vocal tract shapes, which are not relevant from a phonetic point of view (see Fig. 1).

An additional source of discrepancies between inferred and observed vocal tract shapes is the simulation of the acoustic wave propagation, which does not represent the acoustical behavior of the real vocal tract exactly, whatever the acoustic model that is used or the precision of the physical constants in the model.

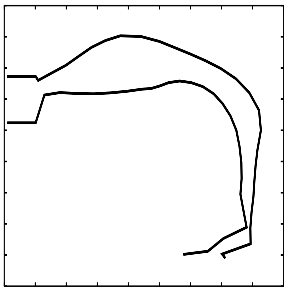


Figure 1: Phonetically invalid mid-sagittal profile generated by means of Maeda's articulatory model; the phonetic score (*ref. section 3*) , the position w.r.t. to the glottis of the main constriction and its cross-section in cm^2 are given.

The analysis stage in the analysis-by-synthesis framework consists in searching for the geometrical and/or articulatory parameters that generate speech spectra that are as close as possible to the observed spectra. Owing to the many-to-one nature of the acoustic-to-articulatory mapping this search needs to be controlled efficiently so as to keep only solutions that are phonetically relevant. An added risk is that, when attempting to decrease the distance between observed and synthetic speech signals, the optimization algorithm misuses the intrinsic flexibility of the articulatory/acoustic simulation to compensate for the mismatch between the model and the speaker. The analysis stage may thus give rise to unexpected (and phonetically unlikely) compensations between different artic-

ulators. The discovery and adequate insertion into synthesis models of constraints is therefore a key challenge in the analysis-by-synthesis approach to acoustic-to-articulatory mapping.

The second general framework for audiovisual-to-articulatory mapping is statistical. Statistical approaches do not involve a synthesis stage because they learn to map directly observed acoustic onto observed anatomical data. The maps are obtained from corpora of articulatory data, in the form of stochastic models, i.e. hidden Markov models [5], Bayesian networks [6], or artificial neural nets [7, 8]. As opposed to analysis-by-synthesis methods, no underlying physical or anatomical models are used to constrain the recovery of articulatory parameters and decrease the space of acceptable solutions. Instead, the size and extent of the phonetic coverage of the training data strongly influence the efficiency and accuracy of the mapping. One manner for inserting constraints is augmenting acoustic input data with data regarding the speaker's face.

This article presents several directions of research in the incorporation of constraints, which have been recently evaluated in the framework of ASPI, which is an European FET project.

2. CONSTRAINING TRACT LENGTH, CROSS-SECTIONS AND TRACT KINEMATICS IN THE FRAMEWORK OF FORMANT-TO-AREA MAPPING

2.1 Constraints on the tract kinematics

Formant-to-area mapping designates the inference of the cross-sections of an area function model from the first few observed formant frequencies. In that framework, fixing constraints that enable selecting a single solution among infinitely possible ones is mandatory. The reason is that the number of variable cross-sections must necessarily be larger than the number of observed formant frequencies in models that enable controlling independently a number of eigenfrequencies that is equal to the number of given formant frequencies [9]. Constraints that have been used frequently are the request that the vocal tract shape remains as close as possible to a reference shape, which usually has been chosen to be a quasi-neutral one or, alternatively, that the cross-sections evolve as slowly as possible. In the framework of the present project, two additional constraints have been evaluated, which have been the minimization of the acceleration and jerk of deformation of the vocal tract.

However, the use of these constraints and their impact on the inversion results are not independent of the algorithmic implementation of the inverse mapper as such. Here, the mathematical framework has been the local linearisation of the nonlinear link between formant frequencies and tract cross-sections and tract length. This local relation is formulated via the Jacobian matrix that links small cross-section increments to small formant frequency increments. The Jacobian matrix can be estimated numerically by feebly increasing the area function parameters and recording the corresponding small increases of the eigenfrequencies.

Because the Jacobian matrix is not square, it has been pseudo-inverted via singular value decomposition [10], which breaks up the Jacobian matrix into a product of a diagonal matrix and two orthogonal matrices that are invertible. The pseudo-inverse of the Jacobian matrix is obtained by zeroing those elements in the inverse of the diagonal matrix that in the original diagonal matrix have been smaller than a threshold, and multiplying with the inverses of the two other matrices.

The general solution consists in a special solution plus a linear combination of matrix columns that form the vector base of the general solution, which obtains all the area function parameter increments that agree with the observed formant frequency increments at that step in time.

The role of the constraints is to enable selecting a single solution among the infinitely many possible ones, so that the area function parameters can be updated and the Jacobian matrix can be estimated at the next time step. The special solution gives by construction the smallest possible parameter increments that agree with the observed frequency increments. It therefore implements the slowest deformation constraint because the time steps are equal.

On the contrary, parameters that satisfy the smallest deformation or smallest acceleration and jerk constraints are obtained via a least-squares formulation that enables computing the weights of the basis vectors that must be added to the special solution so that the general solution satisfies these criteria.

It is however the case that criteria other than the slowest deformation constraint do not mathematically guarantee that the parameter increments remain small. They may therefore conflict with the locally linear approximation of the link between eigen-frequencies and tract parameters. One therefore multiplies the excess increments by a constant < 1 to guarantee the validity of this linear approximation. This obligation hampers observing the effects of these constraints on the anatomical plausibility of the inferred tract shapes. This also suggests that the stepwise locally-linear approximation of the tract parameter-to-formant link intrinsically favors some solutions to the inverse mapping problem over others simply by requesting that the mapped parameter increments must be small so that the approximation remains locally valid.

2.2 Confinement of the labial cross-section and tract length

The constraints that are discussed in this section are optional. They concern the introduction of a priori knowledge in the map or the insertion of data that have been measured via other channels than the acoustic one. The mathematical formulation has involved a change of variables that replaces tract parameters that are free to vary with parameters that are confined to an interval, which may be very small. The hyperbolic tangent has been used to transform free into confined parameters. The tangent parameters have been chosen so that its asymptotes have been equal to the lower and upper boundaries one would like the tract parameters to confine to. This change of variables is equivalent to multiplying the Jacobian matrix by a diagonal matrix that involves the partial derivatives of the confined parameters with respect to the unconfined ones.

One drawback of this formalism is observed when a confined parameter comes near its upper or lower boundary. Then, owing to the change of variables, that parameter is incremented less and less when tracking the observed increments of the formant frequencies. This is a desirable property when the parameter moves towards its boundary, because one expects it to get blocked at the boundary. But, this is an annoying property when the parameter moves away from the boundary, because then one would like it to contribute fully to the observed formant increments.

This is not a disadvantage when the formalism is used to insert external constraints, because then one would like to confine a parameter into a narrow interval anyway. It may be a disadvantage when the formalism is used to confine a parameter to a wider interval inside of which one would expect it move freely. But in that case, the formalism is optional anyway. As an alternative, a hysteresis-like free $\rightarrow$ confined transform may implemented that behaves differently according to whether the parameter moves towards or away from its boundary.

Default constraints have been that the lower and upper boundaries of the cross-sections, other than those that are near the glottis, are $0cm^2$ and $12cm^2$ and the default length of the tract has been within $16-18cm$. In addition, the same formalism has been used to confine the labial cross-section to within $\pm 1\%$ of the observed cross-section and the tract length to the observed length.

2.3 Results

The effects of the constraints have been expressed numerically via the inter-correlation between mapped and observed area functions. Observed area functions and formant frequencies have been obtained from published data. The total number of speakers has been 8. They have sustained a total 78 vowels.

Results are the following. Minimal speed, acceleration and jerk of deformation constraints have yielded mapped tract shapes that are quasi-identical. As discussed above, the reason has been the request that the Jacobian matrix must be a valid local approximation of the area $\rightarrow$ eigenfrequency map.

The same constraints have given rise to statistically significantly larger similarities between observed and mapped tract cross-sections than the request that the deformation with respect to a reference shape be minimal. One concludes that vowel production is not subject to the constraint that the vocal tract stays as near as possible to the neutral vocal tract.

Tract shapes of French rounded back vowels have been the most difficult to map. Correlations between observed and mapped cross-sections have been statistically significantly lower for rounded back vowels than for front vowels. Inserting labial or length constraints has not qualitatively improved correlations between observed and mapped area functions.

Fixing the average tract length to the observed tract length of the speaker, has increased the correlation between observed and mapped area functions statistically significantly. Idem, when the labial cross-sections have been set equal to the observed labial cross-section of each vowel and speaker. Numerically speaking, improvements in correlation have been less than 10%, however. Labial constraints and model complexity have been observed to interact for reasons that are discussed in a presentation that has been submitted to the same congress.

3. CONSTRAINING ARTICULATORY PARAMETERS IN THE FRAMEWORK OF FORMANT-TO-ARTICULATORY MAPPING

Formant-to-articulatory mapping here designates the inference of the parameters of a mid-sagittal articulatory model from formant frequencies. The mapping is performed via table-lookup. Constraints on articulatory parameters can be directly derived from data regarding the speaker's face, or based on speaker-independent knowledge.

3.1 Constraints from visible articulators

A view of the face provides information regarding the jaw and lip opening, as well as lip protrusion and spreading, i.e. possibly four articulatory parameters. Out of the 7 parameters of Maeda's sagittal profile model, up to 3 can be obtained from facial data, i.e. jaw opening, lip opening and protrusion. When the acoustic input data are the first 3 formant frequencies, this means that up to 6 cues may by available from which 7 unknowns must be computed. The impact of articulatory constraints derived from facial data may therefore be substantial.

The use of facial data raises issues, which are the obtaining of the values of the parameters of the articulatory model from the facial data and the insertion of these into the inverse map. The focus here is on the first, which is the estimation of "visible" articulatory parameters, for instance those of Maeda's articulatory model, from facial data. In [11], the facial data have been $3D$ data recovered via stereovision. A first problem is that not all required articulatory parameters can be derived from $3D$ data directly. For Maeda's sagittal profile model that has been obtained from X-ray images, the jaw opening had been determined by measuring the distance between lower and upper incisors. This distance cannot be measured on images of a speaker's face. Markers painted onto the chin have been used instead. The chin is not a perfectly rigid structure, however, and the chin marker movement may have depended on jaw as well as lower lip movement. In addition, the estimation of the lip parameters has been performed by removing the contribution of the jaw movement first. Reference markers (on the forehead or nose) have been used to compensate for head movement.

A second problem is merging two articulatory models: one which describes the face of the speaker whose speech must be inverted, and a second which is the articulatory sagittal profile model into which observed audio-visual data have to be mapped. Usually, the latter has been obtained for another speaker and has involved other types of data, i.e. MRI or X-ray images. In [11], two solutions have been explored.

The first solution has consisted in building an articulatory model for the face only, which uses the same three "visible" parameters (jaw opening, lip opening and protrusion) as Maeda's sagittal profile model. Its construction has been based on the same guided factor analysis as used by Maeda. Once the facial model had been available, "visible" articulatory parameters could be estimated from $3D$ facial data via the known linear link between parameters and geometric modes, and inserted into the sagittal profile model. However, one observes that the two models do not agree exactly, because they have been obtained for different speakers and different types of data.

The second solution has consisted in turning measured facial cues directly into articulatory parameters of the mid-sagittal profile model, which has been adapted before to the speakers providing the audio-visual data. The remaining discrepancies have been taken into account via a transform that equalises the averages of the $3D$ facial cues and the averages of equivalent (but not identical) cues obtained from the original X-ray images. This transform thus has enabled turning observed facial $3D$ cues into the X-ray data-based measures, from which the articulatory model has been obtained originally. This procedure has involved one model only and has been observed to give better results.

Once visible articulatory parameters have been obtained, they have been used either to select relevant articulatory regions that are explored in the code-book to find inverse solutions, or to remove a posteriori unlikely inverse solutions that have been obtained from the speech signal only.

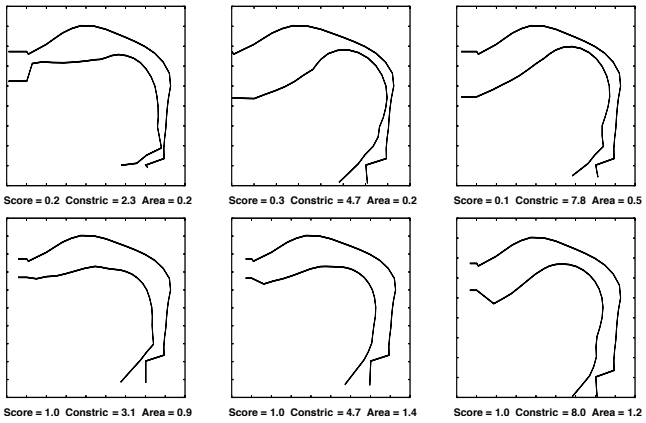
3.2 Phonetic constraints

By construction, factor analysis-based articulatory models implicitly involve constraints on the acceptable shapes of the vocal tract, because they only permit realistic deformations of the individual geometric modes that make up the model. However, even if all modes of deformation are acceptable individually, this does not guarantee that their combination generates vocal tract shapes a human speaker may articulate.

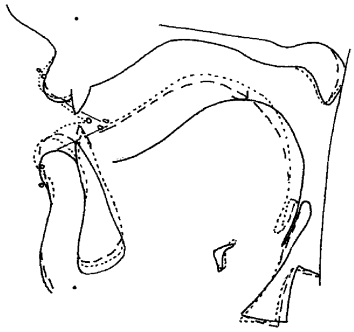
Phoneticians have developed articulatory descriptions [12, 13] of speech sounds. These have been exploited to derive restrictions on synergies between articulators in human speakers [14]. The same constraints can be used to isolate acceptable domains in a model articulatory space for a given domain in the acoustic space centred on a vowel quality. The objective has thus been to build a combined tessellation of the acoustic and articulatory spaces [15]. French vowels have been used to build the tessellation, because they provide a good coverage of the acoustic and articulatory spaces (other languages with a sufficiently dense vowel system could have been used). These constraints present the advantage of being speaker-independent.

Three types of constraints have been defined for vowels: mouth opening, lip protrusion and tongue dorsum position. In practice, these constraints have been used in the form of a phonetic score calculated for every inverse solution. The score has been a function of the distance from the mapped articulatory array to the articulatory domain that has been expected considering the inputted formant frequencies and a priori phonetic knowledge. The score has decreased from 1, when the mapped articulatory array belongs to the allowed articulatory domain, to 0, when the constraints have been violated.

The constraints have been evaluated for vowels for which the speech signal together with X-ray data have been available [15]. The results show that these constraints favour the recovery of vocal tract shapes that agree with the original X-ray data. In addition, the consistency of the phonetic constraints has been investigated by studying the phonetic scores involved when mismatching constraints, i.e. when using a constraint assigned to one vowel to invert acoustic cues corresponding to a vowel that has been different from an articulatory point of view. Results have shown that the relevant constraint, i.e. the one that should have been used, is the one that yields the highest phonetic score.



(a)



(b)

Figure 2: **2a***: Mid-sagittal profiles of the vocal tract for vowel [a]. For each profile, the phonetic score, the position of the main constriction (Constric) w.r.t. the glottis and the constriction area in* cm^2 *are given.* **2b***: X-ray mid-sagittal profiles for [a]: [aba] (solid line), [maf] (dashed line), [vwal] (dotted line)*

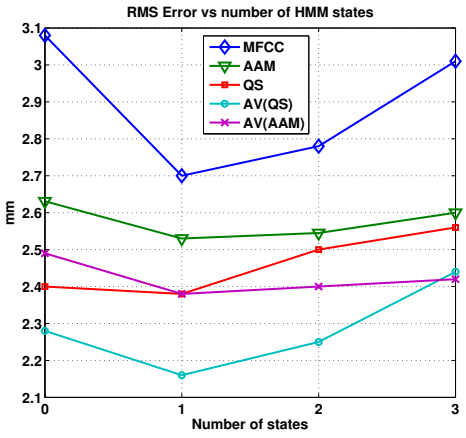


Figure 3: Root-mean-square distance (*RMS*) between original and mapped articulatory trajectories for increasing numbers of hidden Markov model states using facial data only (*AAM* or *QS*), audio-only (*MFCC*) and combined data (*AV*(*AAM*), *AV*(*QS*)).

4. AUDIOVISUAL-TO-ARTICULATORY INVERSION USING HIDDEN MARKOV MODELS

To exploit both audio and visual cues for speech inversion, a statistical framework, combining ideas from multistream hidden Markov models and canonical correlation analysis, has been applied. The visual modality has been represented by either coordinates of markers glued on the speaker's face and tracked during data acquisition, or by visual features extracted by means of active appearance modelling of the face. While the former representation has been accurate, the latter has been more useful in practice, because it can be derived automatically using only the frontal view of the face, without any special acquisition set-up. To cope with limited training data, reduced-rank linear regression models have been estimated when necessary, using canonical correlation analysis. More details may be found in [16, 17].

To evaluate this audiovisual approach to inversion, the Qualisys-Movetrack dataset has been used. The data have been made available by *KTH* and have been described in detail in [18]. The dataset has contained simultaneous measurements of the audio signal, tongue movement and facial motion during speech. Apart from the audio signal which has been sampled at $16kHz$ and the video which has been at $30fps$, each frame of the dataset (at the rate of $60fps$) has contained the $3D$ coordinates of 25 reflectors glued on a speaker's face, as well the $2D$ mid-sagittal plane coordinates of 6 electromagnetic articulography coils (reference coils included) glued on the speaker's tongue, teeth and lips. The data set has comprised in total around 65000 (audio-visual, articulatory) data pairs. These correspond to one repetition of 138 symmetric vowel-consonant-vowel sequences and 178 short everyday Swedish sentences. All data have been aligned in time and phoneme-level transcriptions have been included as well. Further experiments on the MOCHA database [7] lead to similar conclusions and are reported in [16, 17].

Experiments in audiovisual speech inversion have involved the following data. The speech signal has been represented via 16 Mel frequency cepstral coefficients (labelled *MFCC*); alternative acoustic representations such as Linear Spectral Pairs (LSPs) which also perform well for audio-only inversion could alternatively be applied. The MFCC features have been extracted from $35ms$ preemphasized and Hamming-windowed frames of the speech signal, at $60Hz$, to match the frame rate at which the visual and electromagnetic articulography data have been recorded.

For the face, active appearance modelling [19] (labelled *AAM*) has yielded 7 features representing shape and 17 representing texture variability. As an alternative, for comparison and also to show the full potential of using facial information for inversion, all the $3D$ coordinates of the face markers have been used as they have been provided in the database, i.e. a total of 75 features (labelled *QS*).

On the articulatory side, we have used the $2D$ coordinates of the 3 tongue coils (tip, blade, dorsum) and the coil on the lower incisor. The data have been centered by mean subtraction.

The complex audiovisual-to-articulatory mapping is approximated by an adaptive piece-wise linear model. Model switching is governed by a hidden Markov model (HMM) which captures articulatory dynamic information. We have explored recovering articulatory trajectories either from acoustic (labelled *MFCC*) or facial data (labelled *AAM* and *QS*) alone or from both combined (labelled $AV-AAM$ or $AV-QS$). We have randomly selected 90% of the QSMT utterances for training and used the remaining 10% for testing.

To evaluate the results, both the root-mean-square difference and the correlation coefficient between the original and estimated articulatory trajectories have been estimated. Results are summarized in Figs. 3 and 4. The correlation coefficients and the root-mean-square error are shown for increasing numbers of hidden Markov model states. In general, introduction of the visual AAM features ($AV-AAM$) has been beneficial compared to the audio-only case (*MFCC*). The AV-AAM performance only slightly lags the performance of the ground-truth facial $3D$ features fused with audio ($AV-QS$), which have overall performed best, as expected.

At a different level, we have explored many alternative HMM

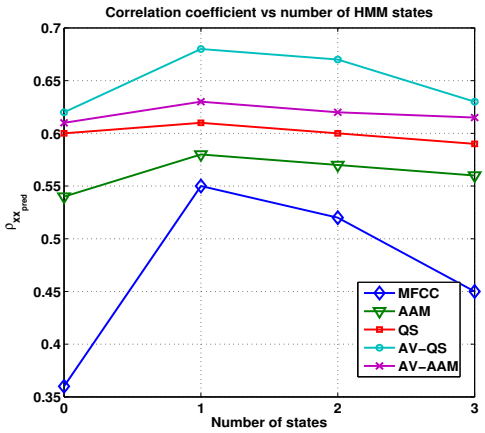


Figure 4: Correlation coefficient (ρ) between original and mapped articulatory trajectories for increasing numbers of hidden Markov model states using facial data only (*AAM* or *QS*), audio-only (*MFCC*) and combined data ($AV-AAM$, $AV-QS$). Zero states correspond to the case of a global linear model.

Features	Level	Type	States	RMS (mm)	$\rho_{x\hat{x}}$
Audio	P	HMM	2	2.56	0.60
QS	P	HMM	2	2.30	0.65
QS	V	HMM	3	2.24	0.66
A-QS	P	HMM	2	2.16	0.69
A-QS	P-P	HMM+LF	2-2	2.02	0.71
A-QS	P-V	HMM+LF	2-2	1.99	0.72
A-QS	**P**	**MS-HMM**	**2**	**1.95**	**0.74**

Table 1: root mean square (*RMS*) error and correlation coefficient (ρ) between inferred and observed articulatory trajectories using several hidden Markov model-based schemes. Audio features (*A*), 3*D* facial marker coordinates tracked via Qualisys (*QS*) or both have been used. The models may be either at the phoneme (*P*) or at the viseme (*V*) level and either single hidden Markov models are used, or in a late fusion (*LF*) configuration or as multistream ($MS-HMM$).

architectures for audio-visual fusion under our scheme. The experiments reported in Table 1 have also been performed on the QSMT dataset, while the visual information has been represented by the 3*D* Qualisys features (*QS*). Audio and visual information dynamics have been fused in three different ways, namely via simple hidden Markov models trained on concatenated feature vectors, single per-modality hidden Markov models with late fusion ($HMM+LF$), or finally, via multistream hidden Markov models ($MS-HMM$). For the late fusion scheme two variants are given in Table 1 differing from each other in whether the visual stream has been modeled as a sequence of phonemes (*P*) or visemes (*V*). Interestingly, the visemes demonstrate improved performance, both in the single modality case and in fusion.

Conclusion

Data augmentation for inverse mapping, i.e. using visual next to acoustic cues, is best suited to statistical methods because these are flexible and not data-specific. For inverse mapping-by-synthesis, the link between modelling for speech production and inversion purposes is closer, however, thus allowing a priori knowledge to be taken into account more easily.

Acknowledgements

The authors acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Sixth Framework Programme for Research of the European Commission, under FET-Open contract no. 021324. P.M. wishes to thank A. Katsamanis and G. Papandreou for research collaboration on HMM-based audiovisual speech inversion.

REFERENCES

[1] P. Mermelstein, "Articulatory model for the study of speech production," *JASA*, vol. 53, pp. 1070–1082, 1973.

[2] P. Birkholz and D. Jackèl, "A three-dimensional model of the vocal tract for speech synthesis," in *Proc. ICPhS*, 2003, pp. 2597–2600.

[3] S. Maeda, "Un modèle articulatoire de la langue avec des composantes linéaires," in *Actes 10èmes Journées d'Etude sur la Parole*, Grenoble, Mai 1979, pp. 152–162.

[4] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional articulatory modeling of the tongue, lips and face, based on mri and video images," *Journal of Phonetics*, vol. 30, no. 3, pp. 533–553, 2002.

[5] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an hmm-based speech production model," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 175–185, Mars 2004.

[6] T.A. Stephenson, H. Bourlard, S. Bengio, and A.C. Morris, "Automatic speech recognition using dynamic bayesian networks with both acoustic and articulatory variables," in *Proc. ICSLP*, Beijing, Oct. 2000.

[7] K. Richmond, "Mixture density networks, human articulatory data and acoustic-to-articulatory inversion of continuous speech," in *Workshop on Innovation in Speech Processing, Institute of Acoustics*, 2001, pp. 259–276.

[8] A. Toutios and K. Margaritis, "Mapping the speech signal onto electromagnetic articulography trajectories using support vector regression," in *Proc. Int. Conf. Text, Speech and Dialogue*, 2005.

[9] P. Mermelstein, "Determination of the vocal-tract shape from measured formant frequencies," *J. Acoust. Soc. Am*, vol. 41, pp. 1283–1294, 1967.

[10] R. C. Aster, B. Borchers, and C. H. Thurber, *Parameter Estimation and Inverse Problems*, Elsevier, Amsterdam, 2005.

[11] B. Potard and Y. Laprie, "Adapting visual data to a linear articulatory model," in *7th International Seminar on Speech Production*, Ubatuba, Brasil, Dec. 2006, pp. 485–492.

[12] P. Ladefoged, *A Course in Phonetics, 4th edition*, Heinle, 2001.

[13] A. Marchal, *Les sons et la parole*, Guérin, Montréal, 1980.

[14] V. Robert, B. Wrobel-Dautcourt, Y. Laprie, and A. Bonneau, "Strategies of labial coarticulation," in *Proc. Interspeech*, 2005.

[15] B. Potard, Y. Laprie, and S. Ouni, "Incorporation of phonetic constraints in acoustic-to-articulatory inversion," *to appear in JASA*, 2008.

[16] A. Katsamanis, G. Papandreou, and P. Maragos, "Audiovisual-to-articulatory speech inversion using active appearance models for the face and hidden markov models for the dynamics," in *Proc. ICASSP*, 2008, pp. 2237–2240.

[17] A. Katsamanis, G. Papandreou, and P. Maragos, "Face active appearance modeling and speech acoustic information to recover articulation," *IEEE Tr. on Acoustics, Speech, and Lang. Proc.*, 2008, under review.

[18] Olov Engwall and Jonas Beskow, "Resynthesis of 3d tongue movements from facial data," in *EUROSPEECH*, 2003.

[19] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.