16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

AUDIOVISUAL SPEECH INVERSION BY SWITCHING DYNAMICAL MODELING GOVERNED BY A HIDDEN MARKOV PROCESS

A. Katsamanis[1], *G. Ananthakrishnan*[2], *G. Papandreou*[1], *P. Maragos*[1], *O. Engwall*[2]

[1]School of Electrical and Computer Engineering,
National Technical University of Athens,
Athens, Greece
{nkatsam,gpapan,maragos}@cs.ntua.gr

[2]Centre for Speech Technology,
Lindstedtsvägen 24 KTH (Royal Institute of Technology),
Stockholm, Sweden
{agopal,engwall}@kth.se

ABSTRACT

We propose a unified framework to recover articulation from audiovisual speech. The nonlinear audiovisual-to-articulatory mapping is modeled by means of a switching linear dynamical system. Switching is governed by a state sequence determined via a Hidden Markov Model alignment process. Mel Frequency Cepstral Coefficients are extracted from audio while visual analysis is performed using Active Appearance Models. The articulatory state is represented by the coordinates of points on important articulators, e.g., tongue and lips. To evaluate our inversion approach, instead of just using the conventional correlation coefficients and root mean squared errors, we introduce a novel evaluation scheme that is more specific to the inversion problem. Prediction errors in the positions of the articulators are weighted differently depending on their relevant importance in the production of the corresponding sound. The applied weights are determined by an articulatory classification analysis using Support Vector Machines with a radial basis function kernel. Experiments are conducted in the audiovisual-articulatory MOCHA database.

1. INTRODUCTION

Audiovisual speech inversion refers to the problem of recovering properties of the speech production system, namely aspects of the vocal tract shape and dynamics, given audiovisual speech information, i.e., the audio speech signal and visual information from the speaker's face. Apart from its theoretical importance, a solution to this problem could allow devising efficient representations of the audio and visual aspects of speech by means of the underlying vocal tract configuration. This can be beneficial to important applications such as speech synthesis [1], speech recognition [2], speech coding [3] and language tutoring [4]. In the current paper, we propose a scheme to add dynamical constraints to audiovisual speech inversion and we also introduce a novel method to evaluate inversion results.

1.1 Previous Work

Speech inversion has been traditionally considered as the determination of the vocal tract shape from the audio speech signal only [5, 6, 7, 8, 9]. For example, in [5] codebooks are optimized to recover vocal tract shapes from formants while the inversion scheme in [6] builds on neural network techniques to recover articulatory coordinates from audio Mel-scale filterbank coefficients. In [7] a Gaussian Mixture Model based mapping is proposed for inversion from Mel Frequency Cepstral Coefficients (MFCCs).

The same speech representation, i.e. MFCCs, is used in [8], where an adaptive extended Kalman filtering scheme is presented to pose phonological and dynamical constraints to the inversion process. In their work, speech is segmented into so-called coproduction units, roughly related to diphones, via a maximum-likelihood process. Each such unit is modeled by a dynamical system with a nonlinear observation equation, which is piecewise linearized based on the corresponding training acoustic-articulatory vector pairs. Clustering into linear regions is performed via a Self-Organising Maps (SOMs) analysis. Articulatory trajectories are determined by extended Kalman smoothing.

A stochastic piecewise-linear approximation of the audio-articulatory relation is also presented in [9]. Each phoneme is modeled by a context-dependent Hidden Markov Model (HMM) and a separate linear regression mapping is trained at each HMM state between the observed MFCCs and the corresponding articulatory parameters. Given the observed audio parameters an optimal state sequence is determined and the hidden articulatory trajectories are obtained by Maximum A Posteriori estimation.

An inherent shortcoming of audio-only inversion approaches is that the mapping from the acoustic to articulatory domains is one-to-many, in the sense that there is a large number of vocal tract configurations which can produce the same speech acoustics, and thus the inversion problem is significantly under-determined. Incorporation of the visual modality in the speech inversion process can significantly improve inversion accuracy. Important articulators such as the lips, jaw, teeth, and tongue are to a certain extent visible. Therefore, visual cues can significantly narrow the solution space and ameliorate the ill-posedness of the inversion process. Indeed, a number of studies have shown that the speaker's face and the motion of important vocal tract articulators such as the tongue are significantly correlated [10, 11, 12, 13].

Motivated by such observations, in [14] we present a unified framework to automatically extract visual features from the speaker's face, integrate them with audio features and exploit this bimodal information to recover articulation from audiovisual speech. Visual features are efficiently extracted from the face by means of Active Appearance Models (AAMs). In this way, we explicitly take into consideration both facial shape and appearance variations, which is the main advantage compared to transform-based approaches as the Independent Component Analysis scheme of [13]. The nonlinear mapping between audiovisual and articulatory features is approximated by a piecewise linear model, governed by a Markov switching process; switching between the segmental linear mappings is determined based on a state sequence identified via an HMM alignment process, similarly to [9]. Our approach is evaluated in the Qualisys-Movetrack audiovisual-articulatory database and promising results are demonstrated.

Quantitative evaluation of audio/audiovisual speech inversion is typically performed by estimating the error between the predicted and the measured/true articulatory parameters. However, rather different articulatory parameters tend to be able to produce almost the same audio/audiovisual parameters [6]. This is due to the fact that for certain phonemes, some articulatory features are more important than others and some are of little importance. For example, in the phoneme /p/, the closure of the lips is the most important articulatory feature, while the positions of the other articulators are irrelevant.

The audio/audiovisual-to-articulatory estimator hence needs to be more accurate about the positions of some articulators when uttering certain phonemes, while it could be allowed to make larger errors for estimating the positions of other articulators. The most commonly used non-regenerative evaluation measures in the cur-

rent literature, Pearson's correlation coefficient and mean RMS error [10, 11, 15] do not take these accuracy constraints into account. Hence, these measures do not necessarily demonstrate the quality of the inversion, since they do not differentiate between crucial and irrelevant errors.

In [16] it is suggested that an articulatory classifier could be used to give a better picture of how successful an inversion method is for each phoneme. The articulatory classifier relies on a prototype articulation for each phoneme to find the closest prototype in each frame. The evaluation is less exact than a correlation score or an error compared to estimated parameters, but it does give information about whether important articulatory features are correct, and if not, the type of error made. This method, however, does not give a single measurable quantity to find whether a particular technique is better than another, especially if their performances are comparable for certain phoneme classes, while they vary for other phoneme classes. It is hence a qualitative analysis tool for finding the strengths and weaknesses of an inversion method, rather than a quantitative measurement of the reliability of an inversion technique.

1.2 Proposed Method

In this context, our contribution in the current paper is essentially twofold.

Dynamical Articulatory Constraints Firstly, to better handle articulatory dynamics and pose continuity and smoothness constraints, we suggest an audiovisual speech inversion scheme based on switching linear dynamical modeling. Inspired by work in audio-only inversion [8, 9] and building on our previous work we introduce a combined HMM and Kalman filtering framework to predict the hidden articulatory state given the observed sequence of audiovisual cues.

Weighted Evaluation Secondly, to evaluate our approach we propose a novel technique based on a weighted root mean square (W-RMS) error, which uses Support Vector Machines (SVM) to estimate the importance, i.e. the weights, of different articulators in different phonemic contexts. It obtains the articulatory parameter weights for each phoneme, based on each parameter's importance in discriminating the phoneme from the rest of the phonemes in the language. An articulatory classification technique is used, and more importance is given to those articulatory features which help in the classification. In this way, it provides a single measurable quantity, which evaluates the performance of the inversion technique, while taking into account, the significance of the articulatory parameters for the production of each phoneme in the language.

Experiments are conducted in the audiovisual-articulatory MOCHA database and results are presented and discussed. Ours is essentially the first study to exploit the video recordings of this database.

2. SWITCHING LINEAR DYNAMIC AUDIOVISUAL-TO-ARTICULATORY MODELING

In the Bayesian framework, audiovisual-to-articulatory speech inversion at a specific time t may be viewed as the articulatory configuration $\mathbf{x}_t$ that maximizes the posterior probability of the articulatory parameters given the available audiovisual information up to time t, i.e., $\mathbf{Y}_t = \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$:

$$p(\mathbf{x}_t|\mathbf{Y}_t) = \frac{p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{Y}_{t-1})}{p(\mathbf{y}_t|\mathbf{Y}_{t-1})}. \quad (1)$$

We have assumed that the observation $\mathbf{y}_t$ at time t is dependent only upon the current configuration $\mathbf{x}_t$. We may further have:

$$p(\mathbf{x}_t|\mathbf{Y}_{t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{Y}_{t-1})d\mathbf{x}_{t-1} \quad (2)$$

by marginalizing out the previous state $\mathbf{x}_{t-1}$. The parameter vector $\mathbf{x}_t$ (n elements) provides a proper representation of the vocal tract. This representation could be either direct, including space coordinates of real articulators, or indirect, describing a suitable articulatory model for example. The audiovisual parameter vector $\mathbf{y}_t$ (m elements), comprising acoustic and visual parameters $\mathbf{y}_t^{\mathrm{a}}$ and $\mathbf{y}_t^{\mathrm{v}}$, should ideally contain all the vocal-tract related information that can be extracted from the acoustic signal on the one hand and the speaker's face on the other. Formant values, linear spectral pairs or MFCCs have been applied as acoustic parameterization. For the face, space coordinates of key-points, e.g. around the mouth, could be used or alternatively parameters based on a more sophisticated face model (e.g., AAM). If we assume that:

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{w}_t \quad (3)$$

$$\mathbf{y}_t = C\mathbf{x}_t + \mathbf{v}_t \quad (4)$$

where $\mathbf{w} \sim N(0, Q)$ and $\mathbf{v} \sim N(0, R)$ independent noise processes and further $\mathbf{x}_0 \sim N(\boldsymbol{\mu}_0, V_0)$, then the maximum a posteriori solution to this problem is given by the Kalman filter [17].

Intuitively, in the case of continuous speech, we expect the linear approximation of Eq. (4) to only be valid as an observation equation for limited time intervals corresponding to a specific phoneme, or even a part of the phoneme, i.e., transition or steady state. The same holds for the state space equation governing the articulatory dynamics. It is thus expected that using different, phoneme-specific (or inter-phoneme specific as in [8]) observation and state-space equations will be more appropriate than using a global linear dynamical system. The proposed switching linear dynamical system is:

$$\mathbf{x}_t = A_{1,c}\mathbf{x}_{t-1} + A_{2,c}\mathbf{x}_{t-2} + B_c\mathbf{u}_c + \mathbf{w}_t \quad (5)$$

$$\mathbf{y}_t = C_c\mathbf{x}_t + \mathbf{v}_t \quad (6)$$

Essentially, there is a separate linear dynamical system corresponding to each class c. For each such class, motivated by physiological considerations, the articulatory configuration is modeled as a second order autoregressive process, as in [8]. It is further considered that $B_c = I - (A_{1,c} + A_{2,c})$ so that the mean value of the articulatory configuration at each state would be u_c. Noise covariances Q_c, R_c are also dependent on the state c.

Inference and learning in this switching linear dynamical modeling framework can be handled via variational approximations as described in [18]. In our current work, for simplicity, we accept that segmentation of the audiovisual-articulatory data into separate classes can be separately determined and is roughly related to phonemic properties. Phoneme-dependent audiovisual Hidden Markov Models (HMMs) are used for this purpose, as in [14]. Each HMM state corresponds to a separate linear dynamical model as described by Eqs. (5) and (6). The HMMs are trainable in the conventional way, by maximum likelihood, given the sequence of phoneme-labeled training audiovisual data. After a forced state alignment procedure of the audiovisual data, the occupation probabilities at each state/class c are estimated and so the training data corresponding to each linear dynamical model can be gathered. The state-space equation of each dynamical model is identified by maximum-likelihood given the training articulatory vectors. The parameters of the observation equations are determined by means of reduced-rank Canonical Correlation Analysis, as in [19].

In this setting, inversion requires finding the optimal state sequence given the observations (sequences of audio, visual or audiovisual features), effectively determining switching between the separate models. For each state-aligned observation vector, the corresponding articulatory vector is estimated via the state-specific linear dynamical system by Kalman filtering.

3. WEIGHTED EVALUATION

One part of estimating the importance of different articulators for a particular phoneme, is finding out which of the articulators help in distinguishing the phoneme from the rest of the phonemes in the language. Keeping this in mind, it is possible to construct a

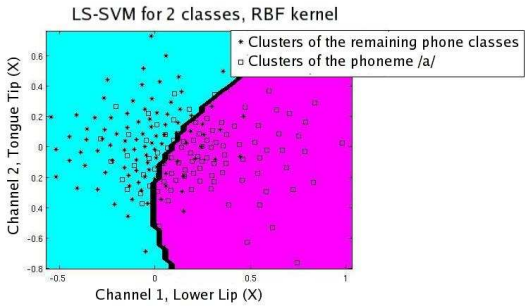

Figure 1: Non-linear separation boundary for sampled patterns from phoneme '/α/' against sampled patterns from the remaining phoneme classes. In this case, Channel 1 is the x-coordinates of the lower lip and Channel 2 is the x-coordinates of the tongue tip

problem, to find the weights to be given to the articulators, in order to provide the best classification of the phonemes. It can be looked at as a parameter selection problem, where you select the best parameters that help the classification. There are several methods for dimension/parameter reduction in the literature, but parameter weighting using SVM is elegant and simple, and additionally, provides individual weights for every parameter. There are several methods of feature selection using SVM, as discussed in [20]. For the current work we are using the SVM-Projection Recurrent Feature Elimination (SVM-Projection RFE) algorithm.

The SVM finds a hyperplane which separates the articulatory space into two classes, separating a phoneme from all others, while allowing for maximum possible error. The hyperplane can be in a higher dimension than the data, as used in Radial Basis Functions (RBF) or polynomial kernels. This means that the separation may not be linear in the dimension of the data. Thus, if the best separating hyperplane is obtained between a phoneme and the rest of the phonemes in the language, it will orient itself so as to make the maximum angle with the most discriminating dimension. By sorting the angles made by the hyperplane with each of the articulatory dimensions, we know which articulator is the most important for the particular phone.

The algorithm is described in detail in [20].

SVM-Projection RFE algorithm: For every phone class c versus the rest of the phoneme classes,

1. Train the SVM to get the separating hyperplane $g(Y)$. Here, Y are the data points from the two classes.

2. Compute the gradient $\bigtriangledown g(Y) \forall Y \in SV$ (support vectors)

$$\bigtriangledown g(Y) = \sum_{i \in SV} \alpha_i y_i \bigtriangledown_Y K(Y_i, Y) \tag{7}$$

   Here, K is the Kernel function and α are the coefficients of the Lagrange multipliers from which the hyperplane is constructed.

3. Compute ρ_i

$$\rho_i = \frac{|\bigtriangledown g(Y_i)|}{\| \bigtriangledown g(Y_i) \|^2}, \forall i \in SV \tag{8}$$

4. Compute the weights for the phoneme class w

$$w = \sum_{i \in SV} \rho_i \tag{9}$$

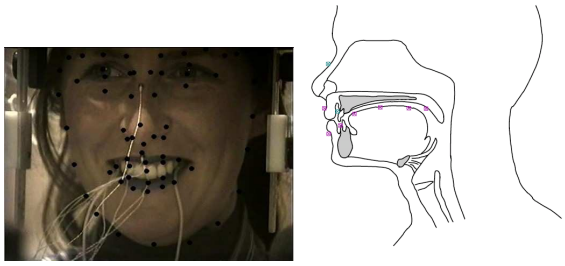

Figure 2: On the left, a sample image of the MOCHA fsew0 speaker's face. The black dots are landmarks that have been automatically localized by Active Appearance Modeling. On the right, a figure showing the placement of the EMA coils that have been tracked for MOCHA, i.e., on the lips, upper and lower incisors, on the tongue tip, dorsum, and blade and on the velum. The coils on the upper incisor and on the nose bridge are used to compensate for head movement.

One problem is that the number of instances (frames) for a single phoneme is much smaller than that of the remaining phonemes and the SVM is always biased towards the class with higher number of patterns. Another problem is the presence of outliers, which could cause the orientation of the hyperplane to change quite considerably. To avoid these two problems, the data in both classes (one phoneme versus the rest) are clustered using K-means clustering. The SVM is applied on the K cluster-centroids from each class. Thus, outliers are filtered out and the number of patterns from each class remains constant. The MATLAB package LS-SVM [21] was used with an RBF kernel for this work. A typical separating hyperplane in a dimension higher than the data is shown in figure 1.

In this way, we get a phoneme-weight matrix, W, with elements W_k^n for $1 \leq k \leq C$, where C is the number of channels, and $1 \leq n \leq P$, where P is the number of phoneme classes in the language.

For the true articulatory parameters Y and estimated parameters $\widehat{Y}$, we can obtain the W-RMS error, E_{wrms}, for N testing samples as follows

$$E_{wrms} = \frac{\sum_{k=1}^{P} \sqrt{\sum_{i \in k} (Y_i - \widehat{Y}_i)^T D_k (Y_i - \widehat{Y}_i)}}{N} \tag{10}$$

where

$$D_k = diag(W_k^{1 \leq n \leq C}) \tag{11}$$

Matrix D can be called the weighting matrix. The most informative features can probably correspond to a nonlinear combination, with contributions of more than one articulatory feature. This could be linearly approximated by a non-diagonal weighting matrix D. However, in this study we ignore the correlation among features and use a diagonal weighting matrix D.

4. EXPERIMENTS AND DISCUSSION

Audiovisual speech inversion experiments were performed in the audiovisual-articulatory MOCHA database.

4.1 Database Description

The MOCHA database [22] is a data-rich and widely used publicly available articulatory dataset, featuring audio, EMA (Electromagnetic Articulography) and EPG (Electropalatography) measurements of speakers uttering 460 British TIMIT utterances. It has been collected mainly for research in speech recognition exploiting speech production knowledge. EMA recordings are at 500 Hz and have been downsampled to 60 Hz. For the purpose of our experiments we have also obtained the video footage of the speaker's face that was recorded during the original data acquisition process and had been so far unused. Ours is thus the first study to exploit the

Table 1: Weights obtained for some typical vowels

Phoneme	/α/	/I/	/U/
Lower incisor (x)	0.02	0.07	0.05
Upper lip (x)	0.09	0.08	0.07
Lower lip (x)	0.07	0.00	0.13
Tongue tip (x)	0.04	0.02	0.02
Tongue blade (x)	0.01	0.08	*0.13*
Tongue dorsum (x)	0.00	0.10	0.13
Velum (x)	0.01	0.11	0.04
Lower incisor (y)	0.16	0.01	0.04
Upper lip (y)	0.00	0.04	0.06
Lower lip (y)	*0.22*	0.13	0.10
Tongue tip (y)	0.14	0.02	0.03
Tongue blade (y)	0.11	*0.17*	0.07
Tongue dorsum (y)	0.05	0.05	0.07
Velum (y)	0.04	0.08	0.04

Table 2: Weights obtained for some typical consonants

Phoneme	/p/	/t/	/k/
Lower incisor (x)	0.05	0.09	0.05
Upper lip (x)	0.02	0.03	0.00
Lower lip (x)	0.04	0.07	0.03
Tongue tip (x)	0.03	0.04	0.01
Tongue blade (x)	0.04	0.02	0.02
Tongue dorsum (x)	0.02	0.10	0.06
Velum (x)	0.11	0.10	0.10
Lower incisor (y)	0.00	0.10	0.02
Upper lip (y)	*0.34*	0.06	0.07
Lower lip (y)	0.07	0.02	0.07
Tongue tip (y)	0.04	*0.17*	0.02
Tongue blade (y)	0.09	0.02	0.13
Tongue dorsum (y)	0.05	0.04	*0.34*
Velum (y)	0.08	0.09	0.07

visual aspect of the MOCHA data. Currently, we have access only to the video recordings of the female subject 'fsew0', Fig. 2.

A practical issue we faced with the MOCHA corpus was the lack of labeling for the video data. We successfully resolved this problem by exploiting the already existing transcriptions for the audio data and automatically matching the segmented audio data with audio tracks extracted from the unprocessed raw video files. The extracted visual features were upsampled to 60 Hz to match the EMA frame rate.

4.2 Audio and Visual Front-Ends

To represent the speech signal we use 16 Mel Frequency Cepstral Coefficients (A). They are extracted from 35-ms pre-emphasized (coefficient: 0.97) and Hamming windowed frames of the signal, at 60 Hz, to match the frame rate at which the visual and EMA data are recorded. The 0-th MFCC coefficient is excluded. For the face, after active apperance modeling as described in [14], we have utilized 15 features representing shape and 26 representing appearance, i.e. 41 AAM parameters in total, corresponding to 95% of the training set variance.

4.3 Weighted Evaluation Scheme

By applying the scheme described in Section 3, we have estimated the weights of the articulatory parameters in the MOCHA database. Each data point in the SVM weight estimation algorithm is located in the 14-dimensional articulatory space, corresponding to the x and y coordinates of the 7 points on the articulators that are tracked by EMA in MOCHA. The x direction corresponds to the horizontal position, while the y direction corresponds to the vertical position.

The weights obtained from the SVM-Projection RFE algorithm give a number of interesting insights. As can be seen from Tables 1 and 2, the critical articulatory channels are the ones with maximum weights. Most often, this critical channel is in accordance with our intuition. For example, for the phonemes /α/, the vertical position of the lower lip is most critical, while for phoneme /t/, it is the tongue tip that is most crucial. However, a few obtained weights don't seem intuitive. For example, the horizontal position of the tongue blade seems to be the critical channel for the phoneme /U/ along with the horizontal positions of the lower lip and tongue dorsum. Similarly, the horizontal position of the velum and the vertical position of the lower lip seem quite important for the phoneme /t/ from the obtained weights.

4.4 Introducing Articulatory Dynamical Constraints

For our experiments, we have set aside 10% of our data for testing and used the rest for training. Our goal has been to evaluate the proposed approach in comparison with audiovisual inversion using a global linear dynamical system or using the HMM framework proposed in [14]. For reference, we also present the inversion results using the audio or visual only observations. The HMMs currently

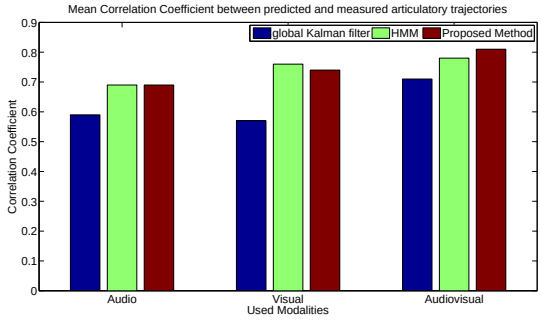

Figure 3: Audio, visual and audiovisual inversion evaluation using the mean correlation coefficient between the measured and the predicted articulatory trajectories. Three cases are given for comparison, i.e., using a global linear dynamical system, using only HMMs or the proposed switching linear dynamical model

used for the experiments are phoneme-based 1-state HMMs. Models with more states could not be sufficiently trained in our current setup and so their performance slightly deteriorated. In total, 46 HMMs are trained, one for each phoneme that appears in the MOCHA database and two more for breath and silence. Two non-emitting states are also incorporated in each model, at the beginning and at the end respectively, so that the transitions between models can also be taken into consideration [23]. For the testing utterances, their phonetic content is considered to be known and forced state alignment is performed by applying the Viterbi algorithm. Mean correlation coefficient and RMS difference are estimated between the predicted and measured trajectories of articulatory coordinates. As is shown in Fig. 3, accounting for articulatory dynamics in the proposed switching scheme is beneficial to inversion. This is also demonstrated in Table 3 where the corresponding mean RMS errors are given. This time, the RMS error in the case when a global linear dynamical system (LDS) is given for comparison. The weighted versions of the RMS errors, especially for the audiovisual case, give further evidence for the good quality of the inversion achieved. Predicted versus reference trajectories of the y-coordinates of the tongue tip and the lower incisor are given in Fig. 4 for a single MOCHA utterance.

5. CONCLUSIONS AND FUTURE WORK

We have presented a framework to introduce dynamical constraints to audiovisual speech inversion. The effectiveness of a switching linear dynamical modeling scheme is investigated and promising results are acquired. Switching is achieved via an audiovisual HMM state alignment process. Simplified learning and inference are discussed. We further describe a novel evaluation technique to prop-

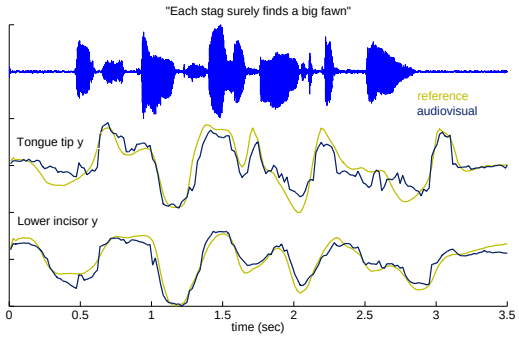

Figure 4: Predicted trajectories of the lower incisor and tongue tip y-coordinates as found using the proposed audiovisual inversion scheme. The corresponding measured trajectories are also given in light colour for reference.

Table 3: Weighted and unweighted root mean square errors in (mm) for three different inversion techniques, i.e., using a global linear dynamical system (LDS), using HMMs or using the proposed switching linear dynamical modeling (SLDS) approach. The audio, visual and audiovisual cases are given.

	Root Mean Square Error					
	Unweighted			Weighted		
	LDS	HMM	SLDS	LDS	HMM	SLDS
Audio	2.15	1.76	1.78	2.17	1.66	1.66
Visual	2.29	1.56	1.62	2.32	1.49	1.54
Audiovisual	1.89	1.53	1.43	1.88	1.47	1.36

erly weigh inversion errors depending on their importance in the production of a specific phoneme. For the introduction of the dynamical constraints we are currently looking into ways to better train the models using limited datasets, which is the case when data is segmented into multiple phoneme or state-of-phoneme classes, for each of which a linear dynamical system has to be trained. In parallel, a detailed analysis of the proposed evaluation scheme is in progress to better assess its importance for speech inversion.

Acknowledgements

We are grateful to Korin Richmond from CSTR in the University of Edinburgh for providing us the video recordings of the MOCHA database. Our work was supported partially by grant ΠΕΝΕΔ-2003-ΕΔ866 [co-financed by E.U.-European Social Fund (75%) and the Greek Ministry of Development-GSRT (25%)] and partially by European Community FP6 FET ASPI (contract no. 021324).

REFERENCES

[1] M. Sondhi, "Articulatory modeling: a possible role in concatenative text-to-speech synthesis," in *IEEE Workshop on Speech Synthesis, Santa Monica, USA*, 2002.

[2] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, February 2007.

[3] J. Schroeter and M. M. Sondhi, *Speech coding based on physiological models of speech production*. NewYork: Marcel Dekker Inc, 1992, ch. 8.

[4] O. Engwall, O. Bälter, A.-M. Öster, and H. Sidenbladh-Kjellström, "Designing the user interface of the computer-based speech training system ARTUR based on early user tests," *Journal of Behaviour and Information Technology*, vol. 25, no. 4, pp. 353–365, 2006.

[5] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *Journal of Acoustical Society of America*, vol. 118, no. 1, pp. 444–460, 2005.

[6] 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.

[7] 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.

[8] S. Dusan and L. Deng, "Acoustic-to-articulatory inversion using dynamical and phonological constraints," in *Proceedings of Seminar on Speech Production*, 2000, pp. 237–240.

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

[10] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, pp. 23–43, 1998.

[11] J. Jiang, A. Alwan, P. A. Keating, E. T. Auer Jr., and L. E. Bernstein, "On the relationship between face movements, tongue movements, and speech acoustics," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1174–1188, 2002.

[12] O. Engwall, "Introducing visual cues in acoustic-to-articulatory inversion," in *Interspeech*, 2005, pp. 3205–3208.

[13] H. Kjellström, O. Engwall, and O. Bälter, "Reconstructing tongue movements from audio and video," in *Interspeech*, 2006, pp. 2238–2241.

[14] 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. Int'l Conf. Acoustics, Speech, and Signal Processing*, 2008.

[15] J. Beskow, O. Engwall, and B. Granström, "Resynthesis of facial and intraoral motion from simultaneous measurements," in *Proc. of the 15th ICPhS*, 2003, pp. 431–434.

[16] O. Engwall, "Evaluation of speech inversion using an articulatory classifier," in *Proc. of the Seventh International Seminar on Speech Production*, 2006, pp. 431–434.

[17] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Prentice Hall, 1979.

[18] Z. Ghahramani and G. E. Hinton, "Variational learning for switching state-space models," *Neural Computation*, vol. 12, no. 4, pp. 831–864, 2000.

[19] A. Katsamanis, G. Papandreou, and P. Maragos, "Audiovisual-to-articulatory speech inversion using HMMs," in *Proceedings of International Workshop on Multimedia Signal Processing (MMSP)*, 2007.

[20] S. Youn, Eun, "Feature selection in support vector machines," Master's thesis, The Graduate School of the University of Florida, 2002.

[21] J. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

[22] A. Wrench and W. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," in *In Proc. 5th Seminar on Speech Production, Kloster Seeon, Bavaria*, 2000, pp. 305–308. [Online]. Available: http://www.cstr.ed.ac.uk/artic

[23] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland., *The HTK Book*. Cambridge University, 1997.