Audiovisual-to-Articulatory Speech Inversion Using HMMs

Athanassios Katsamanis, George Papandreou and Petros Maragos
School of E.C.E., National Technical University of Athens, Athens 15733, Greece
Email: {nkatsam, gpapan, maragos}@cs.ntua.gr

***Abstract*— We address the problem of audiovisual speech inversion, namely recovering the vocal tract's geometry from auditory and visual speech cues. We approach the problem in a statistical framework, combining ideas from multistream Hidden Markov Models and canonical correlation analysis, and demonstrate effective estimation of the trajectories followed by certain points of interest in the speech production system. Our experiments show that exploiting both audio and visual modalities clearly improves performance relative to either audio-only or visual-only estimation. We report experiments on the QSMT database which contains audio, video, and electromagnetic articulography data recorded in parallel.**

I. Introduction

There has been a number of studies showing that there is important correlation between the speaker's face and the motion of important vocal tract articulators such as the tongue, [1]–[4]. Motivated by such findings we investigate a statistical framework to recover vocal tract related information by exploiting both the speech signal and visual cues from the speaker's face concurrently recorded.

In [4], the authors explore simple global linear mappings to unveil associations between the behavior of facial data and articulatory data during speech. They show that analysis can be facilitated by performing a dimensionality reduction process which determines the components that mostly influence the relation between the visual and articulatory spaces. Their experimental data consist of measurements of marker positions on the face and electromagnetic sensors in the vocal tract as well as the generated speech acoustics, for two speakers. They conclude that a high percentage (80%) of the variance observed in the vocal tract data can be recovered from the facial data. This conclusion is also verified in [3] on similar data and again by means of global multivariate linear regression. In the latter work, the authors mainly focus on the variations of the articulatory-visual relations for various CV (Consonant-Vowel) syllables and how they influence speech intelligibility. More recently, in [1], [2] articulatory parameters are recovered from facial and audio data either via relevance vector machines or a global linear mapping. These previous studies have shown that, although a global linear mapping is arguably a rough approximation of the underlying complex non-linear interaction between audio-visual features and articulatory positions, it can nonetheless serve as a first approximation, and also as a baseline system on which more advanced techniques have to improve.

On the other hand, to recover articulatory motion from acoustics only, various sophisticated approaches have been followed. In [5] it is found that Mixture Density Networks perform better than Multilinear Perceptrons in acoustic-to-articulatory inversion. The experiments were performed on the MOCHA database [5]. Note that this database also includes video recordings of the speaker's face that however have not been exploited in [5]. To estimate articulatory trajectories from Mel Frequency Cepstrum Coefficients (MFCCs) derived from the audio signal, a Hidden Markov Model(HMM)-Based Speech Production Model is proposed in [6]. This model allows the imposition of more elaborate constraints to the dynamic behavior of the articulatory parameters that are estimated for given speech acoustics. The HMM framework is reported to outperform other inversion approaches based on codebooks.

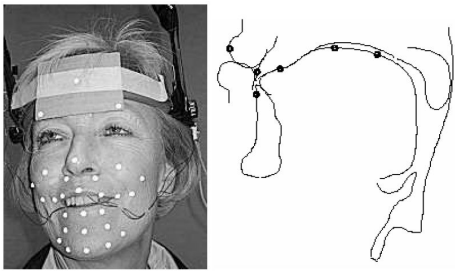

Fig. 1. Qualisys-Movetrack Data Acquisition setup. The positions of the face markers are tracked by the Qualisys system. In parallel, electromagnetic articulography is applied to track the coils placed on the tongue, lips and teeth. Speech is recorded concurrently (Figure from [7]).

In this context, our contribution is twofold. Firstly, we properly extend the framework in [6] in order to effectively fuse visual and audio cues to predict articulatory trajectories. For this purpose, we introduce multistream HMMs, which are commonly used in state-of-the-art audiovisual (AV) speech recognition systems [8]. Secondly, we give a viewpoint of multivariate regression and the related Wiener filter by means of Canonical Correlation Analysis (CCA). This naturally leads to optimal reduced-rank linear regression models, which are novel in the area of articulatory inversion and can potentially improve the predictive performance of the multivariate linear model. These reduced-rank approaches are particularly relevant in the case of models trained on only few data, such as the linear regressors embedded in the applied HMM-based system described in Sec. II, where each regressor corresponds e.g. to a single phoneme with only $O(100)$ occurrences in the training set, and thus reduced-rank or other regularization techniques are essential for obtaining regression models with reasonable generalization performance. Experiments are reported on the Qualisys-Movetrack (QSMT) database which has been collected and kindly provided by KTH [7].

II. Proposed Method

Linear Models for Speech Inversion From a probabilistic point of view, the solution to AV speech inversion may be seen as the articulatory configuration that maximizes the posterior probability of the articulatory characteristics given the available AV information:

$$p(x|y) = p(y|x)p(x)/p(y) \tag{1}$$

It would be intuitive to first consider the static case in which both the articulatory and the audiovisual characteristics do not vary with time. The parameter vector x (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 AV parameter vector y (m elements) should ideally contain all the vocal-tract related information that can be extracted from the acoustic signal on the one hand and

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.

For the maximization, the distribution $p(y)$ is irrelevant since it does not depend on x. Distribution $p(x) \sim N(x; \bar{x}, \sigma_x)$ is assumed to be Gaussian, for simplicity. The relationship between the AV and articulatory parameter vectors is in general expected to be nonlinear but could be to a first order stochastically approximated by a linear mapping (both x and y are centered by mean subtraction):

$$y = Wx + \epsilon \tag{2}$$

The error ϵ of the approximation is regarded as zero-mean Gaussian with covariance Q. The stochastic character of this approximation is justified by the fact that the acoustic and visual representations may not be directly related to the vocal-tract shape due to imperfect source cancellation and possible measurement uncertainty which should be taken into consideration.

The maximum a posteriori probability solution is:

$$\hat{x} = (\sigma_x^{-1} + W^T Q^{-1} W)^{-1} (\sigma_x^{-1} \bar{x} + W^T Q^{-1} y) \tag{3}$$

The estimated solution is a weighted mean of both the observation and the prior models. The weights are proportional to the relative reliability of the two summands.

Linear Mapping Estimation The linear mapping can be determined by means of multivariate linear analysis techniques. Such techniques constitute a class of well studied methods in statistics and other quantitative disciplines; one can find a comprehensive introduction in [9]. We can easily see that, when we completely know the underlying second-order statistics in the form of covariance matrices R_{xx}, R_{yy}, and R_{yx}, then the optimal in the MSE sense choice for the $m \times n$ matrix W corresponds to the Wiener filter

$$W = R_{yx} R_{xx}^{-1}, \tag{4}$$

and the covariance of the approximation error in (2) is $Q \triangleq E\{(y - \hat{y})(y - \hat{y})^T\} = R_{yy} - R_{yx} R_{xx}^{-1} R_{yx}^T$.

Since the second order statistics are in practice unknown a-priori, we must contend ourselves with sample-based estimates thereof; for example, if the $N \times n$ matrix X gathers N samples of x, then a reasonable estimate is $R_{xx} \approx \frac{1}{N} X^T X$, and similarly for R_{yy}, and R_{yx}. These estimates may not be reliable enough when the training set size N is small relatively to the feature dimensions n of x, m of y, and, consequently, when plugged into (4) to yield W, can lead to quite poor performance when we apply the linear regressor (2) to unknown data. We will see that CCA, among other benefits, provides a sound mechanism to select reduced-rank multivariate linear regression models which can outperform the conventional full-rank model in the small training set size case.

Canonical Correlation Analysis Canonical Correlation Analysis is a multivariate statistical analysis technique for analyzing the covariability of two sets of variables, x and y [9, Ch. 10]. Similarly to the better-known principal component analysis (PCA), CCA reduces the dimensionality of datasets, and thus produces more compact and parsimonious representations of them. However, unlike PCA, it is specifically designed so that the preserved subspaces of x and y are maximally correlated, and therefore CCA is especially suited for regression tasks, such as articulatory inversion. In the case that x and y are Gaussian, one can prove that the subspaces yielded by CCA are also optimal in the sense that they maximally retain the mutual information between x and y [10]. CCA is also related to Linear Discriminant Analysis (LDA): similarly to LDA, CCA performs dimensionality reduction to x discriminatively; however the target variable y in CCA is vector-valued and continuous, whereas in LDA is single-valued and discrete.

In CCA we seek directions, a (in the x space) and b (in the y space), so that the projections of the data on the corresponding directions are maximally correlated, i.e. one maximizes with respect to a and b the correlation coefficient between the projected data $a^T x$ and $b^T y$

$$\rho(a, b) = \frac{a^T R_{xy} b}{\sqrt{a^T R_{xx} a} \sqrt{b^T R_{yy} b}}. \tag{5}$$

Having found the first such pair of *canonical correlation directions* (a_1, b_1), along with the corresponding *canonical correlation coefficient* ρ_1, one continues iteratively to find another pair (a_2, b_2) of vectors to maximize $\rho(a, b)$, subject to $a_1^T R_{xx} a_2 = 0$ and $b_1^T R_{yy} b_2 = 0$; the analysis continues iteratively and one obtains up to $k = \text{rank}(R_{xy}) \leq \min(m, n)$ direction pairs (a_i, b_i) and CCA coefficients ρ_i, with $1 \geq \rho_1 \geq \ldots \geq \rho_k \geq 0$, which, in decreasing importance, capture the directions of co-variability of x and y. For further information on CCA and algorithms for performing it, one is directed to [9].

Interestingly, the Wiener filter regression matrix (4) of the multivariate regression model can be expressed most conveniently by means of CCA as

$$W = R_{yx} R_{xx}^{-1} = R_{yy} B P A^T, \tag{6}$$

where $A = [a_1 \ldots a_k]$ and $B = [b_1 \ldots b_k]$ have the canonical correlation directions as columns, and $K = \text{diag}(\rho_1, \ldots, \rho_k)$ is a diagonal matrix of the ordered canonical correlation coefficients. One can prove [10] that by retaining only the r first, $1 \leq r \leq k$, canonical correlation directions/coefficients, i.e. by using the *reduced-order* Wiener filter

$$W_r \triangleq R_{yy} B_r P_r A_r^T, \tag{7}$$

with $A_r = [a_1 \ldots a_r]$ and $B_r = [b_1 \ldots b_r]$, and $K_r = \text{diag}(\rho_1, \ldots, \rho_r)$, one can achieve optimal filtering in the class of order-r filters in the MSE sense. What is more important for us, when the training set is too small to accurately estimate the covariance matrices in hand, these reduced-rank linear predictors can exhibit improved prediction performance on unseen data in comparison to the full-rank model [11]. This is analogous to the improved performance of PCA-based models in well-studied pattern recognition tasks, such as face recognition, when only a subset of the principal directions are retained.

Determination of Articulatory Parameter Trajectories This framework can be extended to handle the inversion of time-varying AV parameter sequences. The probabilities in Eq. (1) will now concern vector sequences. The main consideration is to find accurate observation and prior models that would make the solution tractable. This is not straightforward given the complexity of the relationship between the acoustic and the articulatory space, which in general is nonlinear and one-to-many. Further, visual information should be properly exploited in order to somehow constrain inversion and reduce the number of possible solutions. Motivated by current research in AV speech recognition, we extend the work in [6] to multistream HMMs in order to better fuse the audio and visual modalities.

Intuitively, in the case of continuous speech, we expect the linear approximation of Eq. (2) to only be acceptable for limited time intervals corresponding to a specific phoneme, or at least a part of the phoneme. We also expect that using different, phoneme-specific mappings Of course, the mapping should change for different

phonemes. Hence, we would have a piecewise linear approximation for the observation model. As a prior model for the dynamics of the articulatory parameters, an HMMs is used. Articulator dynamics are in general expected to be phoneme-dependent and so we have one HMM for each phoneme and one articulatory-to-audiovisual mapping for each state. Further, as in audiovisual speech recognition [8] we assume that the audio and visual cues form two separate streams y_a and y_v correspondingly which are weighted differently when determining the HMM c output probability, $p(c|y) \propto N(y_a; m_{c,a}, \Sigma_{c,a})^{w_a} N(y_\nu; m_{c,\nu}\Sigma_{c,\nu})^{w_\nu}$. We accept that the weights w_a and w_ν should sum to one. The distribution of the articulatory parameters at each HMM state is Gaussian. A separate linear mapping $y = W_j x + \epsilon_j$ is considered at each state.

Speech inversion involves finding the optimal state sequence given the audiovisual data and then for each state-aligned analysis frame estimate the corresponding articulatory parameters as in Eq.3, exploiting the state-specific linear mapping. The state sequence is found by the Viterbi algorithm using the audiovisual data in two properly weighted streams. *HMM training* is performed in the conventional way by likelihood maximization [6]. Given the occupation probabilities at each state, the linear mappings between audiovisual and articulatory data are estimated by means of reduced-rank canonical correlation analysis.

III. Experimental Results and Discussion

Database Description For our experiments we have used the QSMT dataset made available by O. Engwall and described in detail in [7]. This dataset contains simultaneous measurements of the audio signal, tongue movements and facial motion during speech. In short, apart from the audio signal which is sampled at 16kHz, each frame of the dataset (at the rate of 60 fps) contains the 3D coordinates of 25 reflectors glued on a speaker's face (75-dimensional vector x), as well the 2D mid-sagittal plane coordinates of 6 EMA (Electromagnetic Articulography) coils glued on the speaker's tongue, teeth and lips (12-dimensional vector y), comprising in total around 65000 data pairs (x_t, y_t). These correspond to one repetition of 138 symmetric VCV (Vowel-Consonant-Vowel) words and 178 short everyday Swedish sentences. All data are temporally aligned and phoneme-level transcriptions are included as well. The data aqcuisition setup is shown in Fig. 1. The three points on the top of the face are used to compensate for head movement and the coils on the upper lip and upper incisor are used to align visual and EMA data.

Global CCA based reduced rank linear model experiment We present first an experiment that demonstrates the potential for improved performance of the reduced-rank linear mapping, relative to the conventional multivariate regression model. The goal of the experiment is to predict the tongue configuration x from the corresponding face expression y_ν by means of a globally linear model. We have split the dataset into training and testing parts; we estimate second-order statistics on the training set and compute from them either the linear regression matrix W or its reduced-order variants W_r, $r = 1, \ldots, 12$, from Eqs. (4) and (7), respectively. Note that for this dataset $W = W_k$, with $k = 12$.

The left column of Fig. 2 depicts the prediction error of the model when computing the tongue's articulation y from the face expression x for varying order r; each row in the figure corresponds to a different training set size $N = 1000, 5000, 50000$ samples. We observe that for small training set sizes, $N = 1000, 5000$, the reduced-order models W_r with $r = 5$ or 6 generalize better than the full-rank model with $W = W_{12}$. Even for the case of big training set with $N = 50000$ samples, although then the full-order model performs best, reduced rank models with $r \geq 7$ perform almost as well. Similar comments can be made about the reverse experiment in which we predict the face expression x from the tongue's articulation y, whose results are depicted in the right column of Fig. 2. These results are particularly relevant and encouraging for the integration of the CCA-based reduced-rank approach with the HMM-based system described in Subsec. II, which incorporates individual regressors for each HMM-state, and thus the effective training data corresponding to each model are very few.

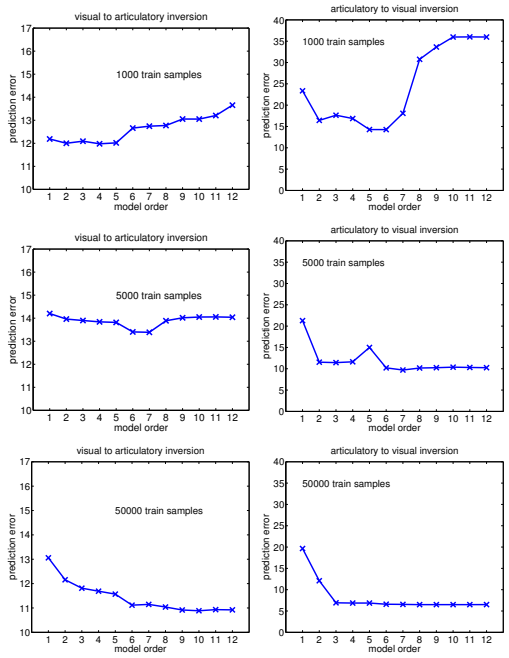

Fig. 2. Generalization error of the linear regression model *vs.* model order for varying training set size. *Left column*: Tongue position from face expression. *Right column*: Face expression from tongue position.

Audiovisual-to-articulatory inversion experiments Next we give our experiments in audiovisual speech inversion. To represent the speech signal we use 16 MFCCs. They are extracted from 35-ms preemphasized (coefficient: 0.97) and Hamming windowed frames of the signal, at 60Hz, to match the frame rate at which the visual and EMA data are recorded. The 0-th coefficient is excluded. From the face, all the 3D coordinates of the 25 reflectors are utilized. On the articulatory side, we use the 2D coordinates of the 3 coils on the tongue (tip, blade, dorsum) and the coil on the lower incisor. The data have been centered by mean subtraction. For training, we have randomly selected 90% of the utterances and testing is performed on the rest 10%. To evaluate the obtained results we have estimated both the RMS difference between the original x and the estimated $\hat{x}$ trajectories as well as the Pearson product-moment correlation coefficient, $\rho_{x\hat{x}} = tr(E[x\hat{x}^T])/\sqrt{tr(E[xx^T])tr(E[\hat{x}\hat{x}^T])}$.

We have built models to recover articulatory trajectories either from acoustic and facial data separately or from both combined. Results are summarized in Fig. 3. The correlation coefficient and the RMS error for the predicted trajectories are shown for increasing number of HMM states. One left-right HMM per phoneme and two separate for silence and breath have been trained. Initially, full order linear mappings are trained at each state. The results at zero states correspond to global linear models and are included for comparison.

In general, audiovisual-to-articulatory inversion outperforms either

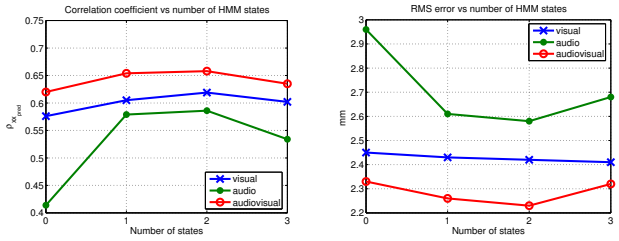

Fig. 3. Correlation coefficient and RMS Error between original and predicted articulatory trajectories for increasing number of HMM states using video only, audio only and both. Zero states correspond to the case of a global linear model.

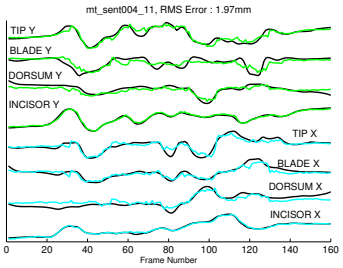

Fig. 4. Measured (black) and predicted (light color) articulatory trajectories.

single-modality inversion. Further, it should be noted that even the simplest time-dependent piecewise linear approximation achieved by only single-state phoneme HMMs is importantly more effective than the global linear model. This holds especially for the single audio modality and it could be justified by the fact that the mapping between acoustic and articulatory data is expected to be highly nonlinear. The observation that the single visual modality performs better than audio has been made in previous studies as well [4].

To focus on the audiovisual case, there are a few interesting issues that should be brought up. Though we have applied multi-stream HMMs the stream weights are not involved in the linear models trained at each HMM state. These are trained using the concatenated audiovisual feature vectors and the corresponding EMA feature vectors at the particular state. The stream weights are essentially applied only for the determination of the optimal HMM state sequence via the Viterbi algorithm. This process is actually an alignment and not a recognition process, as we consider that the phonemic content of each utterance is known. It is based only on the audiovisual and not on the articulatory data. For the results presented in Fig. 3 equal stream weights have been applied. We have found however that performance may be even better in case the alignment is performed using only the audio features, that is if we assign a zero weight to the visual stream. In this case we have found correlation coefficient equal to 0.69 for both the single and the 2-state HMMs with the former giving a slightly lower RMS prediction error, $2.16mm$ vs. $2.17mm$. This observation is in accordance with similar experience in audiovisual speech recognition for audio-noise free experiments [8]. The audio should be exclusively trusted for recognition when no noise is present. In our audiovisual-to-articulatory inversion setup it appears that in the absence of audio noise, the audio stream should be trusted for alignment but, given the optimal state alignment, the contribution of the visual modality in inversion is very important in any case.

Interestingly, we have also observed that by using reduced-order models at each HMM state by means of CCA we could in general achieve similar or even slightly better performance compared to the full-order models. To be more specific, the smallest RMS prediction error $2.14mm$ was achieved using 1-state HMMs with 6th-order linear models (the original order was 8). The 2-state HMMs with 6th-order linear models at each state performed identically. An example of predicted trajectories against the measured ones is shown in Fig.4 for a Swedish phrase. The corresponding RMS error is $1.97mm$.

IV. Conclusions and Future Work

We have presented a statistical framework based on multi-stream HMMs and CCA to perform audiovisual to articulatory speech inversion. Experiments have been carried out on the QSMT dataset to recover EMA coil movements from face motion and speech acoustics. The results demonstrate clear improvement compared to the simple global linear model mapping audiovisual to articulatory parameters, that had been used in earlier works in the area. Reducing the order of the linear model at each HMM state by CCA was beneficial as well but the full benefits are expected to be unveiled in a more detailed phoneme-based analysis that is currently in progress. In addition, modifications concerning continuity and more detailed imposition of dynamic constraints, e.g. related to coarticulation, would be as well interesting. For example, in [6], it is shown that further improvements may be expected if the used representations are enriched by parameter derivatives and accelerations and biphone instead of single-phone models are applied. Last but not least, in the proposed framework, we plan to elaborate on the fusion of the modalities at the state-level, and see how this differs from the linear model involving the concatenated audio-visual feature vector we have used here.

Acknowledgements This research was co-financed partially by E.U.-European Social Fund (75%) and the Greek Ministry of Development-GSRT (25%) under Grant ΠΕΝΕΔ-2003ΕΔ866, and partially by the European research project ASPI under Grant FP6-021324. We would also like to thank O. Engwall from KTH for providing us the QSMT database.

References

[1] H. Kjellstrom, O. Engwall, and O. Balter, "Reconstructing tongue movements from audio and video," in *Interspeech*, 2006, pp. 2238–2241.
[2] O. Engwall, "Introducing visual cues in acoustic-to-articulatory inversion," in *INTERSPEECH*, 2005, pp. 3205–3208.
[3] 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.
[4] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Sp. Comm.*, vol. 26, pp. 23–43, 1998.
[5] 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.
[6] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an hmm-based speech production model," *IEEE TSAP*, vol. 12, no. 2, pp. 175–185, March 2004.
[7] O. Engwall and J. Beskow, "Resynthesis of 3d tongue movements from facial data," in *EUROSPEECH*, 2003.
[8] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Tr. Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.
[9] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*. Acad. Press, 1979.
[10] L. L. Scharf and J. K. Thomas, "Wiener filters in canonical coordinates for transform coding, filtering, and quantizing," *IEEE TSAP*, vol. 46, no. 3, pp. 647–654, 1998.
[11] L. Breiman and J. H. Friedman, "Predicting multivariate responses in multiple linear regression," *Journal of the Royal Stat. Soc. (B)*, vol. 59, no. 1, pp. 3–54, 1997.