Face Active Appearance Modeling and Speech Acoustic Information to Recover Articulation

Athanassios Katsamanis, *Student Member, IEEE*, George Papandreou, *Student Member, IEEE*, and Petros Maragos, *Fellow, IEEE*

***Abstract*—We are interested in recovering aspects of vocal tract's geometry and dynamics from speech, a problem referred to as speech inversion. Traditional audio-only speech inversion techniques are inherently ill-posed since the same speech acoustics can be produced by multiple articulatory configurations. To alleviate the ill-posedness of the audio-only inversion process, we propose an inversion scheme which also exploits visual information from the speaker's face. The complex audiovisual-to-articulatory mapping is approximated by an adaptive piecewise linear model. Model switching is governed by a Markovian discrete process which captures articulatory dynamic information. Each constituent linear mapping is effectively estimated via canonical correlation analysis. In the described multimodal context, we investigate alternative fusion schemes which allow interaction between the audio and visual modalities at various synchronization levels. For facial analysis, we employ active appearance models (AAMs) and demonstrate fully automatic face tracking and visual feature extraction. Using the AAM features in conjunction with audio features such as Mel frequency cepstral coefficients (MFCCs) or line spectral frequencies (LSFs) leads to effective estimation of the trajectories followed by certain points of interest in the speech production system. We report experiments on the QSMT and MOCHA databases which contain audio, video, and electromagnetic articulography data recorded in parallel. The results show that exploiting both audio and visual modalities in a multistream hidden Markov model based scheme clearly improves performance relative to either audio or visual-only estimation.**

***Index Terms*—Active appearance models (AAMs) , audio-visual-to-articulatory speech inversion, canonical correlation analysis (CCA), multimodal fusion.**

I. Introduction

TREATING speech as essentially a multimodal process has led to interesting advances in speech technologies the recent years. For example, by properly exploiting visual cues from the speaker's face, speech recognition systems have gained robustness in noise [1]. The introduction of speaking faces or avatars in speech synthesis systems improves their naturalness and intelligibility [2]. In general, accounting for the visual aspect of speech in ways inspired by the human speech production [3] and perception mechanisms [4] can substantially benefit automatic speech processing and human–computer interfaces.

In this context, we are interested in recovering speech production properties, namely aspects of the vocal tract shape and dynamics, by exploiting not only the speech audio signal but also the speaker's moving face. The problem in its general form could be referred to as audiovisual-to-articulatory speech inversion. Apart from its theoretical importance, it could allow representing the audio and visual aspects of speech by the corresponding vocal tract configuration. This representation can be beneficial to important applications such as speech synthesis [5], speech recognition [6], speech coding [7], and language tutoring [8].

Speech inversion has been traditionally considered as the determination of the vocal tract shape from the audio speech signal only [5]. Recent audio-only inversion approaches are typically based on sophisticated machine learning techniques. For example, in [9], codebooks are optimized to recover vocal tract shapes from formants, while the inversion scheme of [10] builds on neural networks. In [11], a Gaussian mixture model (GMM)-based mapping is proposed for inversion from Mel frequency cepstral coefficients (MFCCs), while a hidden Markov model (HMM)-based audio-articulatory mapping is presented in [12]. Each phoneme is modeled by a context-dependent HMM and a separate linear regression mapping is trained at each HMM state between the observed MFCCs and the corresponding articulatory parameters. Similar approaches have been applied to the complementary problem of audio-to-lips inversion, i.e., lip synchronization driven by audio [13]–[15]. Lip and audio parameters are jointly modeled using phoneme Gaussian mixture HMMs in [16] while more sophisticated dynamic Bayesian networks incorporating articulatory information are used in [17].

An inherent shortcoming of audio-only inversion approaches is that the mapping from the acoustic to articulatory domains is one-to-many [9], 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 underdetermined. 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 alleviate 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 [3], [18]–[20]. In [3], the authors explore simple global linear

Manuscript received January 27, 2008; revised June 20, 2008. Current version published February 11, 2009. This work was supported in part by the European FP6 FET Research Project ASPI (IST-FP6-021324), in part by the European FP6 Network of Excellence MUSCLE (IST-FP6-507752), and in part by the Project ΠΕΝΕΔ-2003ΕΔ866, which is cofinanced by the E.U.-European Social Fund (80%) and the Greek Ministry of Development-GSRT (20%). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gerhard Rigoll.

The authors are with the School of Electrical and Computer Engineering, National Technical University of Athens, Athens 15773, Greece (e-mail: nkatsam@cs.ntua.gr; gpapan@cs.ntua.gr; maragos@cs.ntua.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2008.2008740

mappings to unveil associations between the behavior of facial data, acoustics, 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. The visual modality is represented by the 3-D coordinates of 12 or 18 infrared LEDs glued on the face and tracked by a special-purpose motion capture setup. The audio signal is recorded and the positions of electromagnetic sensors on the tongue, teeth, and lips are tracked concurrently for two speakers. The study concludes 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 [18] on similar data, i.e., 20 retro-reflectors glued on the face are tracked by analogous equipment, and again inversion is performed 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 Consonant–Vowel (CV) syllables and how they influence speech intelligibility. More recently, in [19] articulatory parameters are recovered from facial and audio data by global nonlinear regression techniques.

Despite their promising results, one may identify two main shortcomings in these approaches to audiovisual-to-articulatory inversion. Firstly, the visual modality is captured via complex acquisition setups and tracking systems which limits the applicability of these techniques in a laboratory setting. In more realistic scenarios, a single optical camera is expected to be recording the speaker's face, which is also expected to be free of any markers. Second, these studies have utilized a single global mapping. Although this can serve as a first approximation, a fixed global linear audiovisual mapping cannot sufficiently account for the underlying nonlinear and one-to-many relations between audiovisual features and articulatory positions. While more general fixed nonlinear mappings can be more effective, they are more difficult to train, especially when available data are limited, and they do not easily allow the incorporation of speech dynamics into the inversion process.

In this paper, extending our previous preliminary work [21], [22], we deal with both these issues. As far as facial analysis is concerned, we propose a computer vision approach to automatically extract visual features from just the frontal view of the face without needing any markers. Our visual front-end is based on active appearance models (AAMs) [23]. These are generative image models which facilitate effective and robust face modeling. Their main advantage compared to transform-based techniques, such as the independent component analysis scheme of [20], is that they explicitly take into consideration both facial shape and appearance variations. Model initialization is performed automatically using an Adaboost-based face detector [24]. Our AAM-based system allows reliable extraction of shape and appearance specific facial features which we subsequently use for articulatory inversion.

Further, to overcome the limitations of a fixed audiovisual-to-articulatory mapping and inspired by the audio-only inversion approaches of [12] and [25], we propose an adaptive inversion technique which switches between alternative class-specific (e.g., phoneme or viseme-specific) linear mappings. The underlying switching mechanism is governed by a hidden Markov process which allows imposition of constraints to the dynamic behavior of the articulatory parameters. Despite the simplicity of each individual linear mapping, the resulting piecewise approximation can successfully capture the complex audiovisual-articulatory interactions. At the same time, the constituent mappings can be estimated by efficient multivariate analysis methods. In particular, we discuss the use of canonical correlation analysis (CCA) which is well-suited for linear model estimation with the limited data corresponding to each specific class under our model. The proposed inversion scheme requires the determination of the Markov hidden state sequence for each utterance. For this purpose, we have investigated alternative state alignment techniques which combine audio and visual information at various synchronization levels [26]. In the case of synchronous fusion, the two modalities share a common state and are jointly aligned using state-synchronous multistream hidden Markov models (MS-HMMs), whereas in the case of fully asynchronous late fusion each modality has independent states and is separately aligned using individual HMMs. Given the determined hidden state sequence, inversion is performed by properly weighting the audio and visual information taking into consideration the reliability of each modality. We evaluate the proposed method on the MOCHA [27] (MultiCHannel Articulatory) and QSMT (Qualisys-Movetrack) [28] databases, which comprise simultaneously acquired audio, video, and electromagnetic articulography data. Our goal is to predict the trajectories of electromagnetically tracked coils which are glued on important articulators, e.g., tongue and teeth.

In Section II, we discuss linear modeling for inversion with particular emphasis on CCA-based linear model estimation. Our adaptive audiovisual-to-articulatory mapping scheme is discussed in Section III and various fusion alternatives are presented. Details of our visual front-end are given in Section IV, followed by presentation of our experimental setup and results in Section V.

II. Inversion by Linear Models

From a probabilistic point of view, the solution to audiovisual (AV) speech inversion may be seen as the articulatory configuration that maximizes the posterior probability of the articulatory characteristics $\mathbf{x}$ given the available AV information $\mathbf{y} = \{\mathbf{y}_{\mathrm{a}}, \mathbf{y}_{\mathrm{v}}\}$

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}. \tag{1}$$

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 column parameter vector $\mathbf{x}$ (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 audiovisual column parameter vector $\mathbf{y}$ (m elements), comprising acoustic and visual parameters $\mathbf{y}_{\mathrm{a}}$ and $\mathbf{y}_{\mathrm{v}}$, 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, line spectral frequencies (LSFs) 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, as the AAM of this work.

For the maximization, the distribution $p(\mathbf{y})$ is irrelevant since it does not depend on $\mathbf{x}$. The prior distribution $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \bar{\mathbf{x}}, \Sigma_x)$ is assumed to be Gaussian, with mean $\bar{\mathbf{x}}$ and covariance matrix Σ_x. 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 the linear mapping

$$\mathbf{y} - \bar{\mathbf{y}} = W(\mathbf{x} - \bar{\mathbf{x}}) + \boldsymbol{\epsilon}. \tag{2}$$

The error $\boldsymbol{\epsilon}$ of the approximation is regarded as zero-mean Gaussian with covariance Q, yielding $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}; \bar{\mathbf{y}} + W(\mathbf{x} - \bar{\mathbf{x}}), Q)$. The stochastic character of this approximation is justified by the fact that the acoustic and visual representations may not be fully determined by the vocal tract shape. For example, a spectral representation for the acoustics is also affected by the glottal source and a textural representation for the face might also be conditioned by a certain facial expression. Further, modeling and possible measurement uncertainty should also be taken into consideration. The maximum *a posteriori* solution is

$$\hat{\mathbf{x}} = \left(\Sigma_x^{-1} + W^T Q^{-1} W\right)^{-1} \left(\Sigma_x^{-1}\bar{\mathbf{x}} + W^T Q^{-1}(\mathbf{y} - \bar{\mathbf{y}} + W\bar{\mathbf{x}})\right). \tag{3}$$

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

A. 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 engineering; one can find a comprehensive introduction in [29]. It is well known that, when we completely know the underlying second-order statistics in the form of covariance matrices $R_{xx} = E[(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T]$, $R_{yy} = E[(\mathbf{y} - \bar{\mathbf{y}})(\mathbf{y} - \bar{\mathbf{y}})^T]$, and $R_{yx} = E[(\mathbf{y} - \bar{\mathbf{y}})(\mathbf{x} - \bar{\mathbf{x}})^T]$, then the optimal in the MSE sense choice for the $m \times n$ matrix W corresponds to the Wiener filter solution

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

and the covariance of the approximation error in (2) is

$$Q = R_{yy} - R_{yx} R_{xx}^{-1} R_{yx}^T. \tag{5}$$

Since the second-order statistics are in practice unknown *a priori*, we must contend ourselves with sample-based estimates thereof. If we have N samples $\mathbf{x}_t$ and $\mathbf{y}_t$, with $t = 1, \ldots, N$, then reasonable estimates for the mean and covariance of $\mathbf{x}$ are $\bar{\mathbf{x}} \approx 1/N \sum_{t=1}^{N} \mathbf{x}_t$ and $R_{xx} \approx 1/N \sum_{t=1}^{N} (\mathbf{x}_t - \bar{\mathbf{x}})(\mathbf{x}_t - \bar{\mathbf{x}})^T$, respectively, and similarly for $\bar{\mathbf{y}}$, 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 $\mathbf{x}$, m of $\mathbf{y}$, 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.

B. Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a multivariate statistical analysis technique for analyzing the covariability of two sets of variables, $\mathbf{x}$ and $\mathbf{y}$ [29, 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 $\mathbf{x}$ and $\mathbf{y}$ are maximally correlated, and therefore CCA is especially suited for regression tasks, such as articulatory inversion. In the case that $\mathbf{x}$ and $\mathbf{y}$ 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 $\mathbf{x}$ and $\mathbf{y}$ [30]. CCA is also related to linear discriminant analysis (LDA): similarly to LDA, CCA performs dimensionality reduction to $\mathbf{x}$ discriminatively; however, the target variable $\mathbf{y}$ in CCA is vector-valued and continuous, whereas in LDA is single-valued and discrete.

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

$$\rho(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^T R_{xy} \mathbf{b}}{\sqrt{\mathbf{a}^T R_{xx} \mathbf{a}}\sqrt{\mathbf{b}^T R_{yy} \mathbf{b}}}. \tag{6}$$

Having found the first such pair of *canonical correlation directions* $(\mathbf{a}_1, \mathbf{b}_1)$, along with the corresponding *canonical correlation coefficient* ρ_1, one continues iteratively to find another pair $(\mathbf{a}_2, \mathbf{b}_2)$ of vectors to maximize $\rho(\mathbf{a}, \mathbf{b})$, subject to $\mathbf{a}_1^T R_{xx} \mathbf{a}_2 = 0$ and $\mathbf{b}_1^T R_{yy} \mathbf{b}_2 = 0$; the analysis continues iteratively and one obtains up to $k = \operatorname{rank}(R_{xy}) \leq \min(m, n)$ direction pairs $(\mathbf{a}_i, \mathbf{b}_i)$ and CCA coefficients ρ_i, with $1 \geq \rho_1 \geq \cdots \geq \rho_k \geq 0$, which, in decreasing importance, capture the directions of co-variability of $\mathbf{x}$ and $\mathbf{y}$. For further information on CCA and algorithms for performing it, one is directed to [29].

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

$$W = R_{yx} R_{xx}^{-1} = R_{yy} B K A^T \tag{7}$$

where $A = [\mathbf{a}_1 \ldots \mathbf{a}_k]$ and $B = [\mathbf{b}_1 \ldots \mathbf{b}_k]$ have the canonical correlation directions as columns, and $K = \operatorname{diag}(\rho_1, \ldots, \rho_k)$ is a diagonal matrix of the ordered canonical correlation coefficients. One can prove [30] 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

$$W_r \triangleq R_{yy} B_r K_r A_r^T \tag{8}$$

with $A_r = [\mathbf{a}_1 \ldots \mathbf{a}_r]$ and $B_r = [\mathbf{b}_1 \ldots \mathbf{b}_r]$, and $K_r = \operatorname{diag}(\rho_1, \ldots, \rho_r)$, 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 [31]. 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.

III. Dynamics and Audiovisual Fusion

A. Dynamically Switched Mapping for Adaptive Inversion

This framework can be extended to handle the inversion of time-varying audiovisual parameter sequences acquired during continuous speech. The probabilities in (1) will now concern vector sequences. The main consideration is to find accurate observation and prior models that make the solution tractable. This is not straightforward given the complexity of the relationship between the acoustic and the articulatory spaces, which in general is nonlinear and one-to-many. The multimodal character of the time-evolving audiovisual information poses further challenges.

Intuitively, in the case of continuous speech, we expect the linear approximation of (2) to only be valid 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 articulatory prior model, i.e., the probability distribution of $\mathbf{x}$. We thus expect that using different, phoneme-specific (or inter-phoneme specific as in [25]) mappings and priors will be more effective than using a global linear approximation. This requires determining the switching process between these models, essentially leading to a piecewise linear approximation of the relation between the observed and the articulatory parameters.

Phoneme-dependent hidden Markov models (HMMs) may be used for this purpose [12]. Each state corresponds to a different prior model $p(\mathbf{x})$ for the articulatory parameters and observation model $p(\mathbf{y}|\mathbf{x})$ for the linear mapping between observed and articulatory features. More specifically, extending the analysis of Section II, the prior and conditional probability distributions at state c are considered to be

$$p_c(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \bar{\mathbf{x}}_c, \Sigma_{x,c}) \tag{9}$$

$$p_c(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}; \bar{\mathbf{y}}_c + W_c(\mathbf{x} - \bar{\mathbf{x}}_c), Q_c\right). \tag{10}$$

Then (e.g., see [32, Sec. 2.3.3]) the corresponding marginal distribution for $\mathbf{y}$ is

$$p_c(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \bar{\mathbf{y}}_c, \Sigma_{y,c}) \tag{11}$$

with $\Sigma_{y,c} = W_c \Sigma_{x,c} W_c^T + Q_c$, and the conditional distribution for $\mathbf{x}$ given $\mathbf{y}$ is

$$p_c(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}; \hat{\mathbf{x}}_c, \Sigma_{\hat{x},c}) \text{ with} \tag{12}$$

$$\hat{\mathbf{x}}_c = \Sigma_{\hat{x},c}\left(\Sigma_{x,c}^{-1}\bar{\mathbf{x}}_c + W_c^T Q_c^{-1}(\mathbf{y} - \bar{\mathbf{y}}_c + W_c \bar{\mathbf{x}}_c)\right) \tag{13}$$

$$\Sigma_{\hat{x},c}^{-1} = \Sigma_{x,c}^{-1} + W_c^T Q_c^{-1} W_c. \tag{14}$$

Note that (13) is the multiple-model generalization of the estimator in (3). In this setting, to determine the switching process between the separate models $\mathcal{M}_c = \{W_c, Q_c, \bar{\mathbf{x}}_c, \bar{\mathbf{y}}_c, \Sigma_{x,c}\}$ (one for each state), inversion requires finding the optimal state sequence $\mathbf{c}^*$ given the observations (sequences $\mathcal{Y}$ of audio, visual, or audiovisual features)

$$\mathbf{c}^* = \arg\max_{\mathbf{c}} P(\mathcal{Y}|\mathbf{c}). \tag{15}$$

Given (11), this can be achieved using the Viterbi algorithm, as with conventional HMMs [12]. For each state-aligned observation vector, the corresponding articulatory vector is then estimated using the state-specific estimator of (13). To impose continuity to the estimated articulatory trajectories, one may apply a postprocessing stage as in [12] using the derivatives of the observations and the articulatory parameters or utilize a more sophisticated prior state-space model in a combined HMM and Kalman filtering approach [33].

The HMM state prior and transition probabilities, as well as the state-specific means and variances $\{\bar{\mathbf{y}}_c, \Sigma_{y,c}\}$ corresponding to the observations $\mathbf{y}$ are trainable in the conventional way by likelihood maximization via the expectation-maximization (EM) algorithm [12]. Given the final occupation probabilities $\gamma_t(c)$, each being the probability of $\mathbf{y}_t$ being in state c at time t and estimated using the forward–backward procedure [34], we have

$$\bar{\mathbf{x}}_c = \frac{\sum_t \gamma_t(c)\mathbf{x}_t}{\sum_t \gamma_t(c)} \tag{16}$$

$$\Sigma_{x,c} = \frac{\sum_t \gamma_t(c)(\mathbf{x}_t - \bar{\mathbf{x}}_c)(\mathbf{x}_t - \bar{\mathbf{x}}_c)^T}{\sum_t \gamma_t(c)} \tag{17}$$

where $\mathbf{x}_t$ is the articulatory parameter vector at time t. To find W_c we have to solve the equations [12]

$$\sum_t \gamma_t(c)[(\mathbf{y}_t - \bar{\mathbf{y}}_c) - W_c(\mathbf{x}_t - \bar{\mathbf{x}}_c)](\mathbf{x}_t - \bar{\mathbf{x}}_c)^T = 0 \tag{18}$$

which are identical to the equations derived when solving the weighted least squares regression problem where $\mathbf{x}_t$ and $\mathbf{y}_t$ are weighted by $\gamma_t(c)^{1/2}$ [35]. We estimate W_c by CCA as described in Section II-B using exactly these weighted versions of the data. The optimal CCA model rank is determined via cross validation as further discussed in Section V. Finally, for Q_c we have

$$Q_c = \frac{\sum_t \gamma_t(c)\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^T}{\sum_t \gamma_t(c)} \text{ where } \boldsymbol{\epsilon}_t = \mathbf{y}_t - \bar{\mathbf{y}}_c - W_c(\mathbf{x}_t - \bar{\mathbf{x}}_c). \tag{19}$$

B. Audiovisual Fusion for Inversion

Identification of the hidden speech dynamics and recovery of the underlying articulatory properties can significantly benefit from the appropriate introduction of visual information in the proposed scheme. The audio and visual mapping switching processes can interact at various synchronization levels. We have investigated various audiovisual fusion alternatives.

1) Synchronous Case (Multistream HMMs): The fully synchronized scenario is based on the assumption that articulatory variations are simultaneously reflected on the two modalities. The shared dynamics are efficiently represented by means of multistream HMMs. Such models have been widely and successfully applied for audiovisual speech recognition [26], [36]. Joint state alignment is feasible via proper application of the Viterbi algorithm. Essentially, the audio and visual cues form two streams $\mathbf{y}_{\mathrm{a}}$ and $\mathbf{y}_{\mathrm{v}}$, thus allowing separate weighting at the scoring phase of the alignment process. In this way, the involvement of each stream in alignment is independently controllable, which is not the case for the simple HMMs. The modified class-score is

$$b(\mathbf{y}|c) \propto \mathcal{N}(\mathbf{y}_{\mathrm{a}}; \bar{\mathbf{y}}_{\mathrm{a},c}, \Sigma_{\mathrm{a},c})^{w_{\mathrm{a}}} \mathcal{N}(\mathbf{y}_{\mathrm{v}}; \bar{\mathbf{y}}_{\mathrm{v},c}, \Sigma_{\mathrm{v},c})^{w_{\mathrm{v}}} \tag{20}$$

where c is the common state for both streams and the weights w_{a} and w_{v} sum to one. Though this approach provides a straightforward way to integrate the two modalities, it can be quite restrictive as far as synchronization is concerned. More flexible hidden Markov model variants such as Product-HMMs [37] could partially alleviate this problem.

2) Asynchronous Case (Late Fusion): At the other extreme, the audio-articulatory and visual-articulatory dynamics can be modeled in a fully asynchronous way. They are assumed to be governed by separate switching processes and different HMMs are used for each stream. Integration of the complementary information is then achieved at a late stage, after both observation streams have been independently inverted to articulatory parameter trajectories. Taking advantage of the resulting flexibility, more representative and accurate stream models can be considered, e.g., viseme-based HMMs for the face and phoneme-based ones for speech acoustic information. Visemes correspond to groups of phonemes that are visually indistinguishable from each other and constitute more natural constituent units for visual speech [1]. For example, the viseme P corresponds to the group of phonemes $/p/$, $/b/$, $/m/$. This scheme is partially limited in the sense that it does not exploit interrelations between the streams to determine the underlying composite articulatory state sequence. However, it offers modeling flexibility and does not require any prior knowledge or assumption related to the synchronization of the involved modalities.

Given the composite hidden state $c = \{c_{\mathrm{a}}, c_{\mathrm{v}}\}$, i.e., the switching sequence that determines the applied piecewise audiovisual-to-articulatory mapping, the audio and visual streams contribute to the inversion process weighted by their relative reliability. This is achieved both in the synchronous, i.e., $c_{\mathrm{a}} = c_{\mathrm{v}}$, and in the asynchronous cases. Assuming independence of the single-stream measurement errors, the compound audiovisual articulatory configuration estimate is

$$\begin{aligned}\hat{\mathbf{x}}_c = \Sigma_{f,c}\big(&\Sigma_{x,c}^{-1}\bar{\mathbf{x}}_c + W_{\mathrm{a},c_a}^T Q_{\mathrm{a},c_a}^{-1}(\mathbf{y}_{\mathrm{a}} - \bar{\mathbf{y}}_{\mathrm{a},c} + W_{\mathrm{a},c_{\mathrm{a}}}\bar{\mathbf{x}}_c)\\ &+ W_{\mathrm{v},c_{\mathrm{v}}}^T Q_{\mathrm{v},c_{\mathrm{v}}}^{-1}(\mathbf{y}_{\mathrm{v}} - \bar{\mathbf{y}}_{\mathrm{v},c} + W_{\mathrm{v},c_{\mathrm{v}}}\bar{\mathbf{x}}_c)\big)\end{aligned} \tag{21}$$

where $\Sigma_{f,c}^{-1} = \Sigma_{x,c}^{-1} + W_{\mathrm{a},c_{\mathrm{a}}}^T Q_{\mathrm{a},c_{\mathrm{a}}}^{-1} W_{\mathrm{a},c_{\mathrm{a}}} + W_{\mathrm{v},c_{\mathrm{v}}}^T Q_{\mathrm{v},c_{\mathrm{v}}}^{-1} W_{\mathrm{v},c_{\mathrm{v}}}$ gives the uncertainty $\Sigma_{f,c}$ of the fused inversion estimate, comprising prior and observation model uncertainties. Linear models $W_{s,c}, Q_{s,c}$ in the form of (2) are estimated as described in Section III-A, each corresponding to a different stream s. The more accurate is a stream, i.e., with smaller error covariance $Q_{s,c}$, the more it influences the final estimate. Relaxing the independence assumption, in the case of synchronous fusion, we can account for correlations between the involved streams by using a composite audiovisual linear model per state $W_{\mathrm{av},c}, Q_{\mathrm{av},c}$. The predicted articulation becomes

$$\hat{\mathbf{x}}_c = \Sigma_{f,c}\left(\Sigma_{x,c}^{-1}\bar{\mathbf{x}}_c + W_{\mathrm{av},c}^T Q_{\mathrm{av},c}^{-1}(\mathbf{y} - \bar{\mathbf{y}}_c + W_{\mathrm{av},c}\bar{\mathbf{x}}_c)\right) \tag{22}$$

where in this case the prediction precision $\Sigma_{f,c}^{-1} = \Sigma_{x,c}^{-1} + W_{\mathrm{av},c}^T Q_{\mathrm{av},c}^{-1} W_{\mathrm{av},c}$ is derived from the prior and composite audiovisual observation modeling uncertainties.

IV. Facial Analysis With Active Appearance Models

We use *active appearance models* (AAMs) [23] of faces to accurately track the speaker's face and extract visual speech features from both its shape and texture. AAMs are generative models of object appearance and are proven particularly effective in modeling human faces for diverse applications, such as face recognition or tracking. In the AAM scheme, an object's shape is modeled as a wireframe mask defined by a set of L landmark points whose coordinates constitute a shape vector $\mathbf{s}$ of length $2L$. We allow for deviations from the mean shape $\mathbf{s}_0$ by letting $\mathbf{s}$ lie in a linear N_p-dimensional subspace, yielding

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{N_p} p_i \mathbf{s}_i. \tag{23}$$

The difference of the shape $\mathbf{s}$ from the mean shape $\mathbf{s}_0$ defines a warp $\mathbf{W}(\mathbf{p})$, which is applied to bring the face exemplar on the current image frame I into registration with the mean face template. After registration, the face color texture registered with the mean face can be modeled as a weighted sum of "eigenfaces" $\{A_i\}$, i.e.,

$$I(\mathbf{W}(\mathbf{p})) \approx A_0 + \sum_{i=1}^{N_\lambda} \lambda_i A_i \tag{24}$$

where A_0 is the mean texture of faces. Both eigenshape and eigenface bases are learned during a training phase, using a representative set of hand-labeled face images [23]. The training set shapes are first aligned and then a subsequent PCA yields the main modes of shape variation $\{\mathbf{s}_i\}$. Similarly, the leading principal components of the training set texture vectors constitute the eigenface set $\{A_i\}$. The first three of them extracted by such a procedure are depicted in Fig. 1.

Given a trained AAM, model fitting amounts to finding for each video frame I_t the parameters $\mathbf{q}_t \equiv \{\mathbf{p}_t, \boldsymbol{\lambda}_t\}$ which minimize the squared texture reconstruction error $I_t(\mathbf{W}(\mathbf{p}_t)) - A_0 - \sum_{i=1}^{N_\lambda} \lambda_{t,i} A_i$. We have used the efficient iterative algorithms described in [38] to solve this nonlinear least-squares problem. Due to the iterative nature of AAM fitting algorithms, the AAM shape mask must be initialized not too far from the face position for successful AAM matching. To automate the AAM mask initialization, we employ an Adaboost-based face detector [24] to get the face position in the first frame and initialize the AAM shape, as shown in Fig. 2(a). Then, for each subsequent frame, we use the converged AAM shape result from the previous frame for initializing the AAM fitting procedure.

In our experiments, we use a hierarchy of two AAMs. The first *Face-AAM*, see Fig. 2(b), spans the whole face and can reliably track the speaker in long video sequences. The second

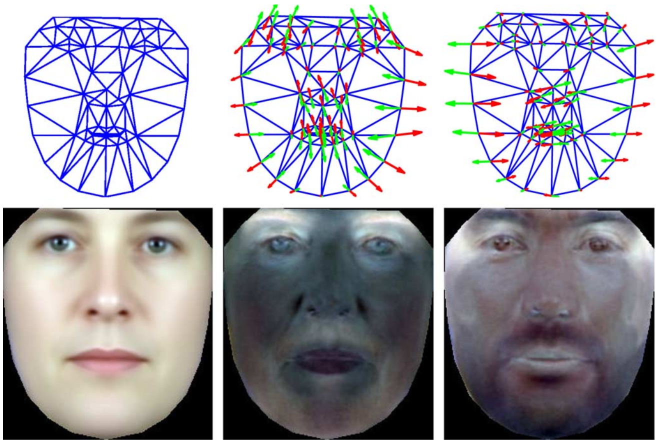
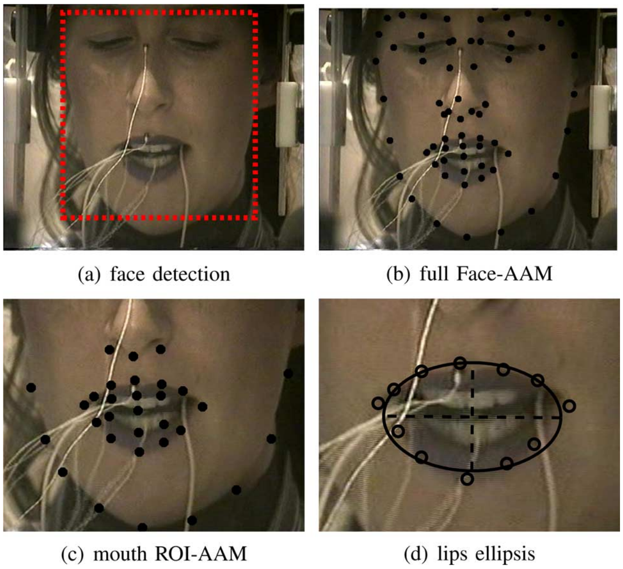


Fig. 1. Active appearance models. *Top*: Mean shape $\mathbf{s}_0$ and the first two eigenshapes $\mathbf{s}_1$ and $\mathbf{s}_2$. *Bottom*: Mean texture A_0 and the first two eigenfaces A_1 and A_2.



(a) face detection

(b) full Face-AAM

(c) mouth ROI-AAM

(d) lips ellipsis

Fig. 2. MOCHA speaker face analysis with our AAM-based visual front-end. (a) Automatic face detection result for AAM initialization. (b) Dots corresponding to the full *Face-AAM* landmarks, as localized by automatic AAM fitting. (c) Landmarks of the lip *ROI-AAM*. (d) Small circles are the subset of the ROI-AAM landmarks which outline the speaker's lips. The ellipsis shown best fits these lip points.

Region of Interest AAM (*ROI-AAM*), see Fig. 2(c), spans only the region-of-interest around the mouth and is thus more focused to the area most informative for visual speech. Since the ROI-AAM covers too small an area to allow for reliable tracking, it is used only for analyzing the shape and texture of the mouth area already localized by the Face-AAM. As final AAM visual feature vector for speech inversion we use the analysis parameters $\mathbf{q}_t$ of the ROI-AAM.

Having localized key facial points with the AAM tracker, we can further derive alternative measurements of the speaker's face which are simple to interpret geometrically. To demonstrate this, we fit for each video frame an elliptical curve on the AAM landmarks outlining the speaker's lips using the technique of [39], as shown in Fig. 2(d). The ellipsis's major and minor axes correspond to mouth width and opening, respectively.

V. Experiments and Discussion

In our experiments, we demonstrate that the proposed approach can effectively extract and exploit visual information from the speaker's face along with audio to recover articulation properties. Sequences of speech acoustic features and facial features are properly combined to recover the corresponding articulatory trajectories. These are trajectories of points on important articulators, e.g., tongue, teeth, and lips, and essentially provide a simple way to represent the vocal tract state during speech. To train our models we have used simultaneously acquired audio, video, and electromagnetic articulography (EMA) data. The latter comprise coordinates of coils on the articulators as these have been tracked by special purpose equipment. Part of the available data has been left out for evaluation.

A. Evaluation Criteria

The shape and dynamics of the predicted articulatory trajectories are compared with the measured ones using two quantitative criteria, i.e., the root-mean-squared (rms) error e_{rms} and the Pearson product-moment correlation coefficient $\rho_{x\hat{x}}$. The rms error indicates the overall difference between the estimated and measured trajectories, $\hat{x}_{1:T}$ and $x_{1:T}$, respectively. For an articulatory parameter i and for duration T of the corresponding trajectory, it is calculated by

$$e_{\mathrm{rms}}[i] = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{\mathbf{x}}_t[i] - \mathbf{x}_t[i])^2}, \quad i = 1, \ldots, n \qquad (25)$$

and provides a performance measure in the same units as the measured trajectories, i.e., in millimeters. However, to get an estimate that can better summarize the inversion performance for all articulators, we use the non-dimensional mean normalized rms error $\bar{e}_{\mathrm{nrms}}$. This is defined by

$$\bar{e}_{\mathrm{nrms}} = \frac{1}{n}\sum_{i=1}^{n}\frac{e_{\mathrm{rms}}[i]}{\sigma_i} \qquad (26)$$

and it allows to also account for the fact that the standard deviations $(\{\sigma_i\},\ i = 1, \ldots, n)$ of the different articulator parameters are not the same.

The mean correlation coefficient measures the degree of amplitude similarity and the synchrony of the trajectories and is defined as

$$\rho_{x\hat{x}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{t=1}^{T}(\mathbf{x}_t[i] - E[\mathbf{x}[i]])(\hat{\mathbf{x}}_t[i] - E[\hat{\mathbf{x}}[i]])}{\sqrt{\sum_{t=1}^{T}(\mathbf{x}_t[i] - E[\mathbf{x}[i]])^2}\sqrt{\sum_{t=1}^{T}(\hat{\mathbf{x}}_t[i] - E[\hat{\mathbf{x}}[i]])^2}}. \qquad (27)$$

These criteria are easy to estimate and they provide a way to quantify the speech inversion accuracy.

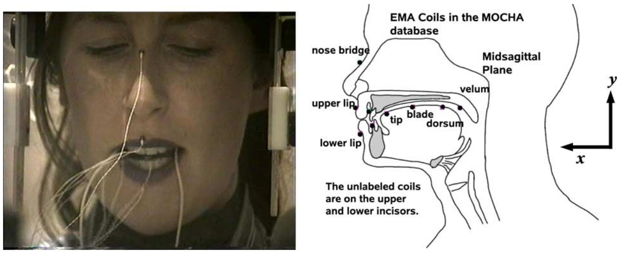

EMA Coils in the MOCHA database
nose bridge
Midsagittal Plane
velum
upper lip
tip
blade
dorsum
lower lip
The unlabeled coils are on the upper and lower incisors.
y
x

Fig. 3. On the left, a sample image of the MOCHA fsew0 speaker's face. On the right, a figure showing the placement of the electromagnetic articulography coils in MOCHA. The coils on the nose bridge and upper incisor are used for head movement correction.

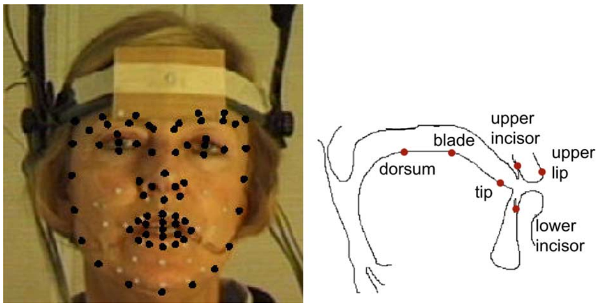


Fig. 4. Qualisys-movetrack database. *Left*: Landmarks on the speaker's face have been localized by active appearance modeling and are shown as black dots. White dots are markers glued on the face and tracked during data acquisition. *Right*: Dots correspond to coils on the speaker's tongue (dorsum, blade and tip from left to right), teeth and lips that have been tracked by electromagnetic articulography. The database also contains speech which is recorded concurrently.

B. Database Description

Experiments and evaluation have been performed on the MOCHA and QSMT databases, which contain both articulatory and concurrently acquired audiovisual data. The MOCHA database [27] is a data-rich and widely used publicly available articulatory dataset, which, among others, features audio recordings and concurrent articulatory, i.e., tongue, lip, and jaw, movement measurements by electromagnetic articulography. It has been collected mainly for research in speech recognition exploiting speech production knowledge and comprises recordings of speakers uttering 460 British TIMIT sentences. The EMA measurements are at 500 Hz and have been downsampled to 60 Hz to have a common reference with the Qualisys-Movetrack (QSMT) dataset. In total, seven EMA coils are tracked; they are glued on the upper and lower lips, on the lower incisor, on the velum and on the tongue tip, blade, and dorsum, as shown in Fig. 3. Two coils on the nose bridge and upper incisor are used to correct the measurements for head movement. For the purpose of our experiments, we have also obtained the video footage of one 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 visual aspect of the MOCHA data. Currently, we have access only to the video recordings of the female subject "fsew0," Fig. 3.

The QSMT dataset was made available by the speech group at the Speech, Music, and Hearing Department in KTH and is described in detail in [28]. It 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 16 kHz and the video which is sampled at 30 fps, each frame of the dataset (at the rate of 60 fps) contains the 3-D coordinates of 25 reflectors glued on the speaker's face (75-dimensional vector, tracked by a motion capture system), as well as the 2-D mid-sagittal plane coordinates of six EMA coils glued on the speaker's tongue, teeth, and lips (12-dimensional vector), comprising in total around 60 000 multimodal data frames. These correspond to one repetition of 135 symmetric VCV (Vowel–Consonant–Vowel) words, 37 CVC (Consonant–Vowel–Consonant) words, and 134 short everyday Swedish sentences. Apart from the video recordings, all other data are temporally aligned and transcribed at the phoneme-level. A sample image from the dataset along with the placement of the EMA coils are shown in Fig. 4.

The fact that the QSMT database includes the ground-truth coordinates of exterior face markers makes it particularly interesting for our purposes since this allows more easily evaluating the quality of fit of our AAM-based automatic tracking and visual feature extraction system.

A practical issue we faced with both QSMT and MOCHA corpora 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 transcribed 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. Further, synchronization issues were resolved by maximizing the correlation of each feature stream with the articulatory data. Correlation was measured by canonical correlation analysis as proposed in [40]. Significant global asynchrony, i.e., more than 120 ms, was detected and corrected only between the articulatory data and video in the QSMT dataset.

C. Experimental Setup

Experiments have been carried out on both datasets independently. Separate models were trained on each and evaluation was performed in parallel. For the MOCHA database, the chosen articulatory representation comprises 14 parameters, i.e., the coordinates of the seven EMA coils in the mid-sagittal plane, while eight parameters are used for QSMT corresponding to the mid-sagittal coordinates of the coils on the tongue (tip, blade, dorsum) and the lower incisor. To avoid possible bias in our results due to the limited amount of data, we follow a tenfold cross validation process. The data in each case are divided in ten distinct sets, nine of which are used for training and the rest for testing, in rotation.

For reference, we first investigate the performance of global linear models, as described in Section II, to invert audio, visual, or audiovisual speech observations and predict the trajectories of the hidden articulatory parameters. This also allows an initial evaluation of the advantages of linear models built using CCA. A simple method for rank selection has been devised and is described in Section V-D; acquired results verify that CCA-based

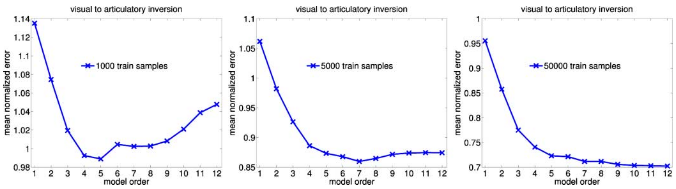
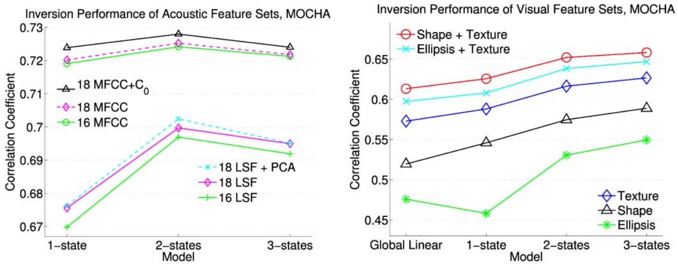


Fig. 5. Recovering articulation from face expression. Generalization error of the linear regression model versus model order for varying training set size. Reduced-rank CCA-based modeling can effectively cope with limited data.



(a) (b)

Fig. 6. MOCHA database: Single modality inversion performance in MOCHA. Alternative audio/visual only representations are compared with respect to the correlation coefficient. Left: Audio only inversion using MFCCs or LSFs. Right: Visual-only inversion using various AAM-based feature sets.

reduced-rank models can indeed outperform their full-rank variants especially with limited training data.

Next, by only considering single-modality-based observations, we systematically evaluate the inversion performance of different audio and visual feature sets. Phoneme-based audio models are built for MFCCs or LSFs, while viseme-based models are used for the AAM-based facial feature set variants. MFCCs with the zeroth coefficient included are shown to outperform alternative acoustic representations while on the visual side, inclusion of both AAM shape and texture features proved to be the most beneficial.

Fusion of the single modalities for audiovisual-to-articulatory inversion is then explored with the various alternative scenarios of Section III-B. Late fusion is in general found to give the best results, outperforming both single-stream HMMs and multistream HMMs, i.e., early and intermediate fusion respectively. Qualitatively interpreting the results in terms of how well certain phonemes are inverted and how accurate is the prediction of individual articulatory parameters appears to lead to reasonable conclusions.

D. CCA-Based Reduced-Rank Linear Model and Rank Selection

We demonstrate the improved performance of the reduced-rank linear mapping, relative to the conventional multivariate regression model. The goal of this experiment is to predict the QSMT 12-dimensional articulatory configuration vector $\mathbf{x}$ (we have used all QSMT EMA coordinates in this experiment) from the corresponding ground-truth 75 facial marker coordinates $\mathbf{y}_\mathrm{v}$ available in QSMT, 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 (4) and (8), respectively. Note that for this dataset $W = W_k$, with $k = 12$.

Fig. 5 depicts the prediction error of the model when computing articulation $\mathbf{x}$ from the face expression $\mathbf{y}_\mathrm{v}$ for varying order r; each plot 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 the big training set with $N = 50\,000$ samples, although then the full-order model performs best, reduced rank models with $r \geq 7$ perform almost as well. These results suggest integrating the CCA-based reduced-rank approach with the HMM-based system described in Section III, which incorporates individual regressors for each HMM-state, and thus the effective training data corresponding to each model are very few.

Automatic model rank selection is addressed via cross validation. To find the optimal rank, we divide the model training data into two sets and try to predict the smaller set using a model trained on the other one for various ranks. This is repeated for every validation fold. The rank giving the minimum squared error is chosen and the final model is trained using the full training set.

E. Single Modality Inversion Experiments

We discuss next our experiments to recover articulation from either acoustic or visual only information. As far as the audio speech signal is concerned, we have experimented with two basic acoustic parameterizations, i.e., MFCCs as given in [41] and LSFs [42]. Both feature sets have been shown to perform similarly in acoustic-to-articulatory inversion using neural networks [43]. In our case, they are extracted from 30-ms-long, pre-emphasized (coefficient: 0.97) and Hamming windowed frames of the signal, at 60 Hz, to match the frame rate at which the QSMT visual and EMA data are recorded. For the MFCCs, 24 filters are applied while for the LSFs the number of coefficients used matches the corresponding linear prediction coding (LPC) analysis order. We have investigated the importance of the total number of extracted features (from 12 to 22) as well as the importance of the inclusion of the zeroth MFCC coefficient. Since in our experiments we have used phoneme HMMs with diagonal observation covariance matrices $\Sigma_{y,c}$ we have also tried to assess the effect of principal component analysis (PCA) on the LSFs, which are not in general expected to have a nearly diagonal covariance, as is the case with the MFCCs. In Fig. 6(a) indicative results are given for the MOCHA database. In both databases, the conclusions are similar; MFCCs perform better than LSFs in our setup even when the performance of the latter is slightly improved by PCA. Further, the inclusion of the zeroth MFCC coefficient is advantageous while 18 is found to be a quite satisfactory choice for the number of cepstral coefficients to retain.

For the audio models, we found that two-state left-right phoneme-based HMMs perform the best in the described audio-to-articulatory inversion setup and probably biphone models could have performed even better, provided that sufficient training data were available [12]. In MOCHA, 46 models

TABLE I
VISEME CLASSES AS DETERMINED IN THE MOCHA DATABASE FOLLOWING A DATA-DRIVEN BOTTOM-UP CLUSTERING APPROACH. THE PHONETIC SYMBOLS AND CORRESPONDING EXAMPLES USED ARE AS FOUND IN THE MOCHA PHONETIC TRANSCRIPTIONS

Viseme	Phonemes	Word Examples from MOCHA
v_1	zh, sh, jh, ch	pleasure, educa**ti**on, **g**eological, porch
v_2	uh, uu, ow	lunch, too, found
v_3	i@, ii, iy	year, peel, bare**ly**,
v_4	a, ai, e, ei, eir, h	N**a**ncy, pie, **e**lderly, day, where, her
v_5	b, m, p	obtain, more, public
v_6	dh, th	the, wealth
v_7	f, v	for, live
v_8	k, breath, g, i, n, ng	cream, good, Acropol**i**s, hand, mango
v_9	d, s, sil, t, z	wild, seldom, to, is
v_{10}	@, l, y	w**a**s, lily, why
v_{11}	@@, aa	were, arm
v_{12}	o, oi, ou, u	often, enj**oy**, do, would
v_{13}	oo, w	all, why
v_{14}	r	rabbits

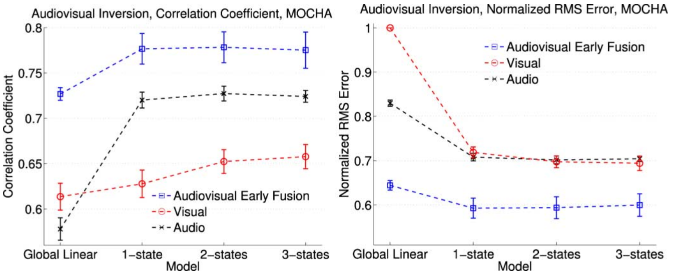


Fig. 7. MOCHA database: Correlation coefficient and normalized rms error between original and predicted articulatory trajectories for increasing number of HMM states using facial information only, via AAM shape and texture, audio only, via MFCCs, and both. The global linear model performance is also given for comparison.

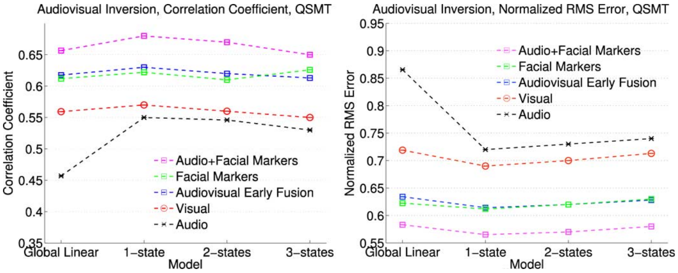


Fig. 8. QSMT database: Correlation coefficient and rms error between original and predicted articulatory trajectories for increasing number of HMM states using facial information only (via AAM features or ground truth facial markers), audio only (via MFCCs) and both. The global linear model performance is also given for comparison.

are trained in total, i.e., 44 for the phonemes and 2 for breath and silence, while in QSMT 52 models are trained for the 51 phonemes and silence that appear in the phonetic transcriptions of the data. For the visual-to-articulatory inversion, however, improved performance can be achieved if we use viseme-based HMMs instead. Though several viseme sets have been proposed for the TIMIT (MOCHA) phone-set, we considered that a data-driven determination of the viseme classes would be more appropriate [44]. Starting from individual phoneme-based clusters (as these are determined by phonetic transcriptions) of the visual data, we followed a bottom up clustering approach to finally define 14 visemes in MOCHA and 34 in QSMT. Viseme classes for the MOCHA database are given in Table I and clustering results appear to be quite intuitive in most cases.

To build the linear observation-articulatory mappings at each state, we have applied canonical correlation analysis as described in Section II-B and further detailed in Section V-D. In this process, insufficient available data may lead to improper canonical correlation coefficients, namely first coefficient equal to unity [45], or degenerate estimates of the model error covariance. To cope with this, we cluster the problematic model with the closest one (with respect to the Euclidean distance) and reestimate.

In this setup, to explore the performance of different visual representations, we have experimented with the nature of the AAM-based facial features used for inversion. The number of AAM shape features, 12 for MOCHA, 9 for QSMT, and AAM texture features, 27 for MOCHA and 24 for QSMT, corresponds to 95% of the observed variance in the facial data in each database. Shape alternatively is compactly described by the set of ellipsis-based geometric features derived from the AAM, as described in Section IV. Interestingly, this representation also appears to be effective, although not as effective as the original AAM shape feature vector. Inversion results for different scenarios are summarized in Fig. 6(b) for MOCHA. Overall we find that the concatenation of AAM shape and texture features performs the best.

F. Audiovisual-to-Articulatory Inversion Experiments

For audiovisual inversion, we first experiment with early fusion of the audio and visual feature sets. Corresponding feature vectors are concatenated at every time instance to form a single audiovisual feature vector and single stream phoneme HMMs are trained to determine linear model switching. Results are summarized in Figs. 7 and 8 for MOCHA and QSMT, respectively, for a global linear model and increasing number of HMM states. Error bars represent the standard deviation of the corresponding correlation coefficient or normalized error estimates, as these are given by cross validation. Visual and audio only inversion results are also included for comparison. In both datasets, integration of the two modalities is clearly beneficial. In QSMT, for which reference facial data are available, the audiovisual based inversion is almost as good as the inversion based on the fused audio and ground-truth facial markers. Further, measurement of the AAM features is much more practical since it does not require any special or inconvenient acquisition setup but only frontal view video of the speaker's face.

Results are improved when intermediate fusion is adopted, i.e., when multi- instead of single-stream HMMs are used. In Fig. 9, the best intermediate fusion results are shown for MOCHA, acquired when two-state multistream HMMs are used. The stream weights are applied for the determination of the optimal HMM state sequence via the Viterbi algorithm, as explained in Section III-B. Determination of this sequence

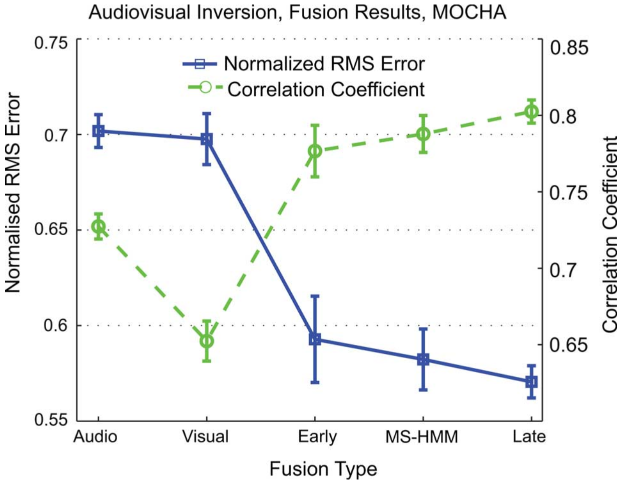


Fig. 9. MOCHA database: The best results for each inversion scenario are given, namely two-state audio HMMs, three-state visual HMMs, one-state single-stream audiovisual HMM, one-state audiovisual MS-HMM and the best performing 2+3-state audiovisual late fusion.

is actually an alignment and not a recognition process, as we consider that the phonemic content of each utterance is known. We have found that the performance is optimal in case the alignment is performed using only the audio features, that is by assigning a zero stream weight to the visual stream. This observation is in accordance with similar experience in audiovisual speech recognition for audio noise-free experiments [26]. 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.

Performance gets even better if we consider the two streams to be asynchronous and model them separately, i.e., using two-state phoneme HMMs for audio and three-state viseme HMMs for the visual modality. Final articulatory trajectories' estimates are then obtained by late fusion with (21), as described in Section III-B. This is actually the best performing scenario as is shown in Fig. 9. The effectiveness of the asynchronous model should be attributed to its flexibility in selecting for each modality the optimal HMM topology and hidden state representation (phoneme/viseme sets). Similar conclusions, only with larger uncertainty due to the small size of the dataset, can be drawn for the QSMT database as well.

For the MOCHA database, in Fig. 10 we show the articulators for which audiovisual inversion using late fusion is most successful. Both the rms error, to also give a feeling of the performance in physical units (mm), and its normalized version are depicted. As expected, prediction of lip movements is significantly improved, compared to the audio-only inversion case. In general, relative improvement is bigger for the recovery of y-coordinates, which is expected since a 2-D frontal view of the face is quite hard to give information on the x-coordinate movements which are only indirectly "seen". This observation can explain for example why the movement of "lower incisor x" is relatively not so accurately recovered. Interestingly, there are improvements in the prediction of the tongue movements as well.

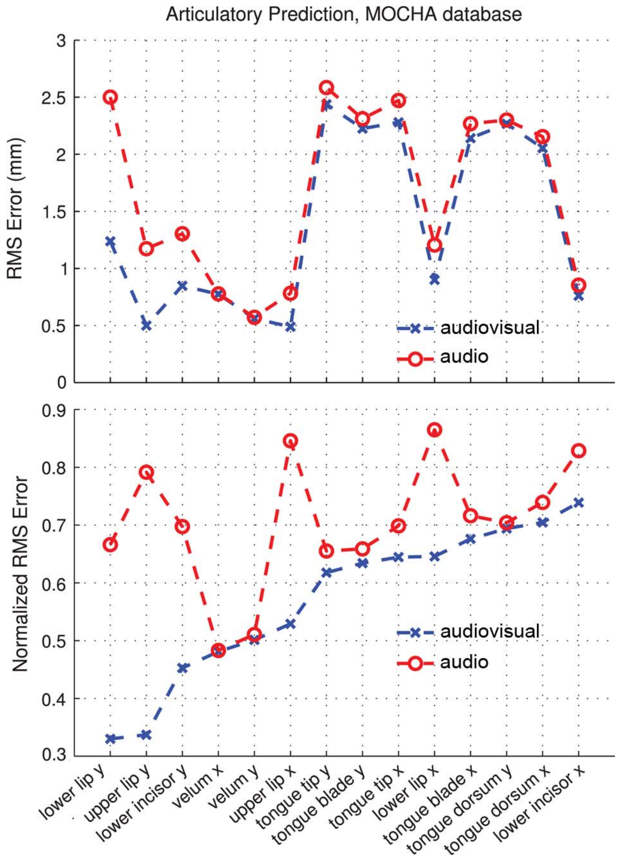


Fig. 10. MOCHA database: RMS prediction error and its normalized version for tracked points in the vocal tract, using audio-only or audiovisual information. The results correspond to the best setup for both observation scenarios. We use two-state HMMs for audio and we integrate them with three-state visual HMMs in late fusion.

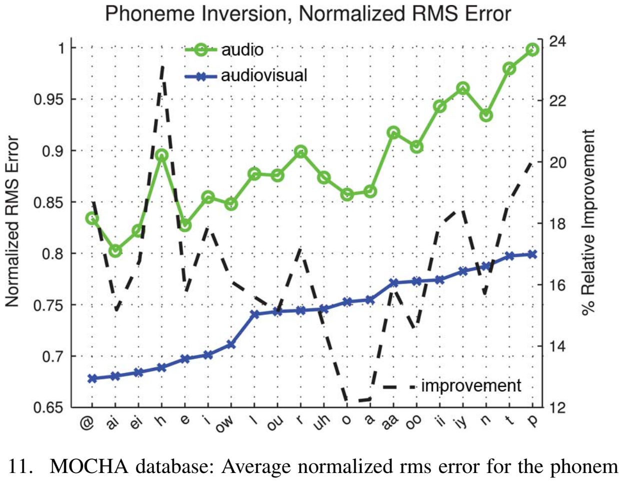


Fig. 11. MOCHA database: Average normalized rms error for the phonemes that have been inverted with minimum error. Results for both the audio-only and audiovisual inversion scenarios are depicted. Again, two-state HMMs are used for audio and 2 + 3-state HMMs in late fusion.

These observations could possibly justify the improvements in inversion when viewed in terms of phonemes as in Fig. 11. The rms error for the 20 best audiovisually inverted phonemes is depicted. The relative improvement is also given.

A qualitative example of the predicted trajectories for the midsagittal y-coordinates of the upper lip and tongue tip against the measured ones is shown in Fig. 12 for a MOCHA utterance.

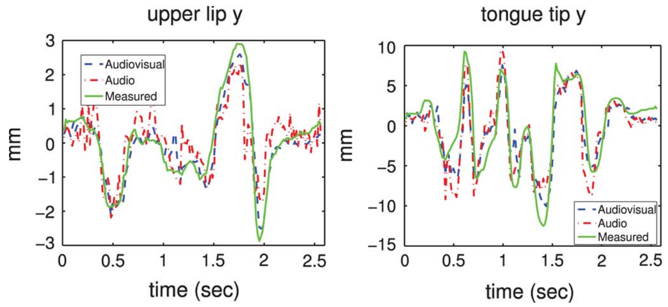


Fig. 12. Upper lip and tongue tip y-coordinates as measured with EMA and predicted from audio only and audiovisual observations of an example MOCHA utterance.

The audiovisual estimator more accurately follows the articulatory trajectories.

VI. Conclusion

We have presented a framework based on hidden Markov models and canonical correlation analysis to perform audiovisual-to-articulatory speech inversion. Experiments have been carried out in the MOCHA and QSMT databases to recover articulatory information from speech acoustics and visual information. Facial analysis is performed by means of active appearance modeling. In this way, it is possible to use visual information without a special motion capturing setup, that would require for example gluing markers on the speaker's face. Experiments regarding modeling and fusion schemes show that modeling the visual stream at the viseme level may improve performance and that the intermediate and late fusion schemes are better suited to audiovisual speech inversion than the early/feature fusion approach.

Acknowledgment

The authors would like to thank the speech group of KTH for providing the QSMT database and K. Richmond from the Centre for Speech Technology Research, University of Edinburgh for providing the video recordings of the MOCHA database.

References

[1] G. Potamianos, C. Neti, G. Gravier, and A. Garg, "Recent advances in the automatic recognition of audio-visual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, Sep. 2003.

[2] G. Bailly, M. Bérar, F. Elisei, and M. Odisio, "Audiovisual speech synthesis," *Int. J. Speech Technol.*, vol. 6, no. 4, pp. 331–346, Oct. 2003.

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

[4] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[5] J. Schroeter and M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, pp. 133–150, Jan. 1994.

[6] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 121, no. 2, pp. 723–742, Feb. 2007.

[7] J. Schroeter and M. M. Sondhi, "Speech coding based on physiological models of speech production," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1992.

[8] 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," *J. Behavior Inf. Technol.*, vol. 25, no. 4, pp. 353–365, 2006.

[9] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *J. Acoust. Soc. Amer.*, vol. 118, no. 1, pp. 444–460, 2005.

[10] K. Richmond, S. King, and P. Taylor, "Modelling the uncertainty in recovering articulation from acoustics," *Comput. Speech Lang.*, vol. 17, pp. 153–172, 2003.

[11] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Commun.*, vol. 50, pp. 215–227, 2008.

[12] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 175–185, Mar. 2004.

[13] T. Chen, "Audiovisual speech processing," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 9–21, Jan. 2001.

[14] E. Yamamoto, S. Nakamura, and K. Shikano, "Lip movement synthesis from speech based on hidden Markov models," *Speech Commun.*, vol. 26, pp. 105–115, 1998.

[15] G. Englebienne, T. Cootes, and M. Rattray, "A probabilistic model for generating realistic speech movements from speech," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007.

[16] K. Choi, Y. Luo, and J.-N. Hwang, "Hidden Markov model inversion for audio-to-visual conversion in an MPEG-4 facial animation system," *J. VLSI Signal Process.*, vol. 29, pp. 51–61, 2001.

[17] L. Xie and Z.-Q. Liu, "Realistic mouth-synching for speech-driven talking face using articulatory modeling," *IEEE Trans. Multimedia*, vol. 9, no. 3, pp. 500–510, Apr. 2007.

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

[19] O. Engwall, "Introducing visual cues in acoustic-to-articulatory inversion," in *Proc. Int. Conf. Spoken Lang. Process.*, 2005, pp. 3205–3208.

[20] H. Kjellström, O. Engwall, and O. Bälter, "Reconstructing tongue movements from audio and video," in *Proc. Int. Conf. Spoken Lang. Process.*, 2006, pp. 2238–2241.

[21] A. Katsamanis, G. Papandreou, and P. Maragos, "Audiovisual-to-articulatory speech inversion using HMMS," in *Proc. Int. Workshop Multimedia Signal Process. (MMSP)*, 2007, pp. 457–460.

[22] 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. Conf. Acoust. , Speech, Signal Process.*, 2008, pp. 2237–2240.

[23] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[24] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Int. Conf. Comp. Vision Pattern Recog.*, 2001, vol. I, pp. 511–518.

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

[26] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, Sep. 2000.

[27] A. Wrench and W. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," in *Proc. 5th Seminar Speech Production*, Kloster Seeon, Bavaria, 2000, pp. 305–308. .

[28] O. Engwall and J. Beskow, "Resynthesis of 3D tongue movements from facial data," in *Proc. Eur. Conf. Speech Commun. Technol.*, 2003, pp. 2261–2264.

[29] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*. New York: Academic, 1979.

[30] L. L. Scharf and J. K. Thomas, "Wiener filters in canonical coordinates for transform coding, filtering, and quantizing," *IEEE Trans. Speech Audio Process.*, vol. 46, no. 3, pp. 647–654, May 1998.

[31] L. Breiman and J. H. Friedman, "Predicting multivariate responses in multiple linear regression," *J. Roy. Statist. Soc. (B)*, vol. 59, no. 1, pp. 3–54, 1997.

[32] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[33] A. Katsamanis, G. Ananthakrishnan, G. Papandreou, P. Maragos, and O. Engwall, "Audiovisual speech inversion by switching dynamical modeling governed by a hidden Markov process," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2008, CD-ROM.

[34] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[35] W. DeSarbo and W. Cron, "A maximum likelihood methodology for clusterwise linear regression," *J. Classification*, vol. 5, pp. 249–282, 1988.

[36] *Speechreading by Humans and Machines*, D. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer, 1996.

[37] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 169–172.

[38] G. Papandreou and P. Maragos, "Adaptive and constrained algorithms for inverse compositional active appearance model fitting," in *Proc. IEEE Int. Conf. Comp. Vision and Patern Recog.*, 2008.

[39] A. Fitzgibbon, M. Pilu, and R. Fisher, "Direct least square fitting of ellipses," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 5, pp. 476–480, May 1999.

[40] M. E. Sargin, Y. Yemez, E. Erzin, and M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1396–1403, Nov. 2007.

[41] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book (for HTK version 3.2) Cambridge Univ. Eng. Dept., Tech. Rep, 2002.

[42] F. K. Soong and B.-H. Juang, "Line spectrum pair and speech data compression," in *Proc. Int. Conf. Acoust., Speech Signal Process*, 1984, vol. 9, pp. 37–40.

[43] C. Qin and M. Carreira-Perpinan, "A comparison of acoustic features for articulatory inversion," in *Proc. Int. Conf. Spoken Lang. Process.*, 2007, pp. 2469–2472.

[44] T. J. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 3, pp. 1082–1089, May 2006.

[45] M.-S. Tso, "Reduced-rank regression and canonical analysis," *J. R. Statist. Soc. (B)*, vol. 43, pp. 183–189, 1981.

Athanassios Katsamanis (S'03) received the Diploma in electrical and computer engineering (with highest honors) from the National Technical University of Athens, Athens, Greece, in 2003, where he is currently pursuing the Ph.D. degree .

He is currently a graduate Research Assistant with the National Technical University of Athens. From 2000 to 2002, he was an undergraduate Research Associate with the Greek Institute for Language and Speech Processing (ILSP), participating in projects in speech synthesis, signal processing education, and machine translation. During the summer of 2002, he worked on Cantonese speech recognition at the Hong Kong Polytechnic University, while in the summer of 2007 he visited Télécom Paris (ENST) working on speech production modeling. His research interests lie in the area of speech analysis and include speech production, synthesis, recognition, and multimodal processing. In these domains and in the frame of his Ph.D. degree and European research projects, since 2003 he has worked on multimodal speech inversion, aeroacoustics for articulatory speech synthesis, speaker adaptation for non-native speech recognition and multimodal fusion for audiovisual speech recognition.

George Papandreou (S'03) received the Diploma in electrical and computer engineering (with highest honors) from the National Technical University of Athens, Athens, Greece, in 2003, where he is currently working towards the Ph.D. degree.

Since 2003, he has been a Research Assistant at the National Technical University of Athens, participating in national and European research projects in the areas of computer vision and audiovisual speech analysis. During the summer of 2006, he visited Trinity College Dublin, Dublin, Ireland, working on image restoration. From 2001 to 2003, he was an undergraduate Research Associate with the Institute of Informatics and Telecommunications of the Greek National Center for Scientific Research "Demokritos," participating in projects on wireless Internet technologies. His research interests are in image analysis, computer vision, and multimodal processing. His published research in these areas includes work on image segmentation with multigrid geometric active contours (accompanied with an open-source software toolbox), image restoration for cultural heritage applications, human face image analysis, and multimodal fusion for audiovisual speech processing.

Petros Maragos (S'81–M'85–SM'91–F'95) received the Diploma in electrical engineering from the National Technical University of Athens in 1980 and the M.Sc.E.E. and Ph.D. degrees from the Georgia Institute of Technology (Georgia Tech), Atlanta, in 1982 and 1985, respectively.

In 1985, he joined the faculty of the Division of Applied Sciences, Harvard University, Cambridge, MA, where he worked for eight years as a Professor of electrical engineering. In 1993, he joined the faculty of the Electrical and Computer Engineering School, Georgia Tech. During parts of 1996 and 1998, he was on sabbatical and academic leave working as a Director of Research at the Institute for Language and Speech Processing, Athens. Since 1998, he has been working as a Professor at the NTUA School of Electrical and Computer Engineering. His research and teaching interests include signal processing, systems theory, pattern recognition, communications, and their applications to image processing and computer vision, speech and language processing, and multimedia.

Prof. Maragos has served as an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND SIGNAL PROCESSING and the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, and editorial board member for the journals *Signal Processing* and *Visual Communications and Image Representation*; General Chairman or Co-Chair of conferences or workshops (VCIP'92, ISMM'96, VLBV'01, MMSP'07); and member of IEEE DSP committees. His research has received several awards, including a 1987 NSF Presidential Young Investigator Award, the 1988 IEEE Signal Processing Society's Young Author Paper Award for the paper "Morphological Filters," the 1994 IEEE Signal Processing Senior Award, and the 1995 IEEE Baker Award for the paper "Energy Separation in Signal Modulations with Application to Speech Analysis," the 1996 Pattern Recognition Society's Honorable Mention Award for the paper "Min-Max Classifiers," and the 2007 EURASIP Technical Achievements Award for contributions to nonlinear signal processing and systems theory, image processing, and speech processing.