Estimation of Articulatory Trajectories Based on Gaussian Mixture Model (GMM) with Audio-Visual Information Fusion and Dynamic Kalman Smoothing

İ. Yücel Özbek, *Student Member, IEEE,* Mark Hasegawa-Johnson, *Member, IEEE,* and Mübeccel Demirekler, *Member, IEEE*

***Abstract*—This paper presents a detailed framework for Gaussian mixture model (GMM) based articulatory inversion equipped with special post-processing smoothers, and with the capability to perform audio-visual information fusion. The effects of different acoustic features on the GMM inversion performance are investigated and it is shown that the integration of various types of acoustic (and visual) features improves the performance of the articulatory inversion process. Dynamic Kalman smoothers are proposed to adapt the cutoff frequency of the smoother to data and noise characteristics; Kalman smoothers also enable the incorporation of auxiliary information such as phonetic transcriptions to improve articulatory estimation. Two types of dynamic Kalman smoothers are introduced: global Kalman (GK) and phoneme-based Kalman (PBK). The same dynamic model is used for all phonemes in the GK smoother; it is shown that GK improves the performance of articulatory inversion better than the conventional low-pass (LP) smoother. However, the PBK smoother, which uses one dynamic model for each phoneme, gives significantly better results than the GK smoother. Different methodologies to fuse the audio and visual information are examined. A novel modified late fusion algorithm, designed to consider the observability degree of the articulators, is shown to give better results than either the early or the late fusion methods. Extensive experimental studies are conducted with the MOCHA database to illustrate the performance gains obtained by the proposed algorithms. The average RMS error and correlation coefficient between the true (measured) and the estimated articulatory trajectories are 1.227 mm and 0.868 using audiovisual information fusion and GK smoothing, and 1.199 mm and 0.876 using audiovisual information fusion together with PBK smoothing based on a phonetic transcription of the utterance.**

***Index Terms*—Audiovisual to articulatory inversion, audiovisual fusion, Gaussian mixture model, Kalman smoother, maximum likelihood trajectory estimation**

I. Introduction

In his book *The Acoustic Theory of Speech Production*, Gunnar Fant published a series of nomograms: plots of formant frequency as a function of articulatory constriction location [1]. His nomograms captured the imagination of the scientific world by depicting the relationship between articulation, $\mathcal{X}$, and acoustics, $\mathcal{Z}$ in the form of a mathematical function, $\mathcal{Z} = f(\mathcal{X})$. With Fant's nomograms available, many researchers believed that the problems of speech technology would soon be solved: speech synthesis seemed to be no harder than implementation of the function $\mathcal{Z} = f(\mathcal{X})$, and speech recognition no harder than implementation of the inverse function, $\mathcal{X} = g(\mathcal{Z})$. Although the nomogram is no longer considered to be a summary of all speech technologies, estimation of the function $\mathcal{X} = g(\mathcal{Z})$ (the so-called *articulatory inverse* function) is still a goal of speech technology research, partly because it is an interesting challenge, but also, in part, because recent results continue to demonstrate that the accuracy of automatic speech-to-text transcription using both acoustic and articulatory measurements is higher than the accuracy of transcription using only acoustics (Markov et al. [2] and Stephenson et al. [3]).

The science of articulatory inversion has become more precise, since the development of electromagnetic articulography (EMA) made it relatively easy to record simultaneous articulatory and acoustic measurements (Perkell et al. [4]), and especially since the publication of a relatively large single-talker EMA database as part of the MOCHA database [5]. Using EMA data, recent studies have demonstrated acoustic-to-articulatory inversion using hidden Markov models (HMMs) (Hiroya et al. [6] and Zhang et al. [7]), neural networks (NN) (Richmond [8], [9]), and Gaussian mixture models (GMM) (Toda et al. [10]), using a wide variety of acoustic feature vectors (Qin and Carreira-Perpiñán [11]). The publication of visual features for the female speaker of the MOCHA database (Katsamanis et al. [12]) has allowed experiments to test audiovisual-to-articulatory inversion (Katsamanis et al. [13]). This paper builds on the methods published in these recent works. This paper develops an articulatory inverse function based on GMM, using audio, video, and audiovisual input features. The first goal of this paper is to extend the previous research on articulatory inversion by using a wider variety of acoustic features and audiovisual fusion strategies. The conducted experiments demonstrate that the incorporation of various types of acoustic features improves the system's performance; moreover the integration of MFCC and formant-related acoustic features (e.g., formant frequencies and formant energies or LSF) is particularly useful in articulatory inversion.

İ. Yücel Özbek and Mübeccel Demirekler are with Electrical and Electronics Eng. Dept., Middle East Technical University, 06531, Ankara, Turkey.

Mark Hasegawa-Johnson is with Electrical and Computer Eng. Dept., University of Illinois, Urbana IL, USA.

Manuscript received October 7, 2009.

Although it demonstrated the potential utility of articulatory inversion, the nomogram also demonstrated that articulatory inversion is, strictly speaking, an ill-posed problem: the function $\mathcal{Z} = f(\mathcal{X})$ is non-monotonic, therefore it has no inverse. Schroeder [14] and Mermelstein [15] demonstrated that there is a one-to-one map between the shape of a tube, on one hand, and the set of pole and zero frequencies of its driving point impedance, on the other; since the formant frequencies carry information only about the poles of the driving-point impedance, they concluded, the problem of acoustic-to-articulatory inversion is underspecified by a factor of two. They proposed two potential solutions: (1) measure both the poles and zeros of the driving-point impedance, e.g., using external stimulation, or (2) halve the degrees of freedom of the problem, by assuming any particular distribution of losses within the vocal tract. Wakita [16] demonstrated a special case of the latter solution: he demonstrated that if all losses are at the lips, then the vocal tract shape is provided by the reflection-line implementation of linear predictive coding (LPC). Atal, Chang, Mathews and Tukey [17] proposed a third solution: dynamic programming. They proposed that the formant frequency vector at each time step should be inverted in a one-to-many fashion, to find a list of matching articulatory vectors; dynamic programming is then used to find the most likely temporal sequence of articulations matching the observed acoustics. The majority of articulatory inversion methods published since [17] follow its lead in the use of some kind of model of articulatory dynamics to resolve the ambiguity of the one-to-many inverse mapping, though the methods used to model articulatory dynamics have varied widely, from implicit methods such as a low-pass filter (Toda et al. [10] and Richmond [8]) to explicit models such as an HMM (Hiroya et al. [6]).

The second goal of this paper is to demonstrate a better smoother for GMM-based articulatory inversion. In contrast to the currently typical low-pass filter based smoothers, we propose a dynamic model-based smoother that uses both mean and covariance provided by the GMM inverter, whereas approaches based on low-pass filtering use only the mean. We show that the proposed smoothing method significantly improves the performance of articulatory inversion. Furthermore, the proposed smoother can be easily adapted to make use of any available auxiliary information, (e.g., of a phonetic transcription) as it is based on a Bayesian-normalized generative model of the observed articulatory and acoustic data. We demonstrate that the use of this information dramatically reduces the error rate of the articulatory inversion.

Finally, the third main goal of our paper is to introduce several efficient and effective information fusion procedures that combine acoustic and visual information. In addition to the so-called early and late fusion types described in the literature, we propose a novel audio-visual fusion methodology (called "modified late fusion algorithm") by considering the visual observability of each articulator. We show that this fusion algorithm gives better performance than early-fusion or late-fusion algorithms that ignore articulator observability.

The rest of the paper is structured as follows. We start in Sec. II by describing the set of audio and video features considered in this article. We emphasize the advantage of using various combinations of audio features in articulatory inversion. Sec. III, for the sake of completeness, describes GMM-based articulatory inversion, more or less exactly as it was described in [10]. The dynamic smoothing process based on global- and phoneme-based Kalman smoothers that we propose is detailed in IV. Sec. V describes three different candidate methods for audiovisual fusion. The results of an extensive set of experiments are given in Sec. VI. Discussion and conclusion that are drawn from these results are given in Sec. VII.

II. Acoustic and Visual Features

Early studies of articulatory inversion (e.g. Mermelstein [15] and Wakita [16]) often focused on the relative utility of different types of acoustic features; Qin and Carreira-Perpiñán [11], among others, have compared different acoustic feature sets in a modern probabilistic paradigm. Video features have also been demonstrated to be useful for articulatory inversion (Katsamanis et al. [13]). This part of the paper examines the extraction of different types of acoustic and visual features for GMM-based articulatory inversion.

A. Audio Features

The papers of Fant [1], Mermelstein [15] and others suggest that good estimates of the formant frequencies may go a long way toward accurate articulatory inversion. Recent studies have approximated the formant frequencies using information from linear predictive coding (LPC), including the line spectral frequencies (LSFs). Qin and Carreira-Perpiñán [11], in particular, suggested that acoustic features directly related to the formants (e.g., LPC and LSF) are more useful for articulatory inversion than acoustic features based solely on the DFT spectrum. One of the goals of this paper is to demonstrate that better formant estimates produce better articulatory estimates. This paper will compare several of the same acoustic feature sets that were tested in [11], including Mel-frequency cepstral coefficients (MFCC), LPC, LSF. This paper will also test two additional feature sets (a vector of formant-related features and log area ratios (LAR) of the vocal-tract tube model) and various combinations of feature types. We have previously demonstrated that formant frequencies and energies are useful for articulatory inversion (Özbek et al. [18]); a goal of this paper is to validate our previous results by comparing the formant vector to a wider variety of standard acoustic feature vectors. MFCCs have often been reported to give optimal results in acoustic-to-articulatory inversion experiments ([10], but see [11]). In this study, 13 MFCCs (The 0th coefficient is excluded) are generated from 29 triangular band-pass filters, uniformly spaced on a Mel-frequency scale between 0 Hz and 8000 Hz. $M \triangleq [M_1, \ldots, M_{13}]$ denotes the 13-dimensional MFCC feature vector; $DM \triangleq [M, M_\Delta, M_{\Delta\Delta}]$ denotes a combination of the MFCCs and their velocity M_Δ, and acceleration $M_{\Delta\Delta}$ components. The LPC coefficients are adequate for articulatory inversion during vowels and glides, if the vocal tract is assumed to be a lossless tube with a lossy termination (Wakita [16]); even under more realistic

assumptions, the LPCs are a useful description of the vowel spectrum [11]. LPCs are estimated by the autocorrelation method and the order of the LPC filter is chosen to be 18. $L \triangleq [L_1, \ldots, L_{18}]$ denotes an 18-dimensional LPC feature vector; $DL \triangleq [L, L_\Delta, L_{\Delta\Delta}]$ denotes the combination of LPCs and their velocity L_Δ and acceleration $L_{\Delta\Delta}$ components. Log area ratio (LAR) coefficients are derived from LPC. In LAR analysis, the vocal tract is modeled by cascading uniform tubes (Wakita's model [16]). It is assumed that the first tube (the glottis) is closed (area = 0), and the last tube (just past the lips) has an infinite area. The LAR coefficients are formed by the log area ratio of cross-section areas of consecutive tubes. Let A_i be the ith LAR coefficient; it is calculated as

$$A_i \triangleq \log \frac{G_i}{G_{i+1}} = \log \frac{1-\rho_i}{1+\rho_i}$$

where G_i and G_{i+1} are the ith and $(i+1)$th cross-sectional area respectively, and ρ_i is the corresponding partial correlation coefficient derived from the Levinson-Durbin recursion. LAR features are denoted $A \triangleq [A_1, \ldots, A_{18}]$; DA is the combined feature set, including velocity and acceleration. The LSFs are the poles and zeros of the driving point impedance of Wakita's vocal tract model (Hasegawa-Johnson [19]). The LSFs are widely used in speech coding, because of their high intra-vector and inter-vector predictability (Kondoz [20]). LSF coefficients tend to cluster around spectral peaks, especially around formant frequencies, and are therefore closely related to articulation (Itakura [21]). LSF features are denoted $S \triangleq [S_1, \ldots, S_{18}]$, and DS is the combined feature set, including velocity and acceleration. Formant frequencies are another type of acoustic features extracted from LPC features. Formants are the resonant frequencies of the vocal tract and they change according to movement of the articulators (Özkan et al. [22]). Therefore, this paper considers the formant frequencies as effective acoustic features to be utilized for articulatory inversion. Extraction of four formant trajectories is handled via the formant tracker described in our previous work (Özbek et al. [18], [23]). The energy associated with each formant frequency is also extracted, using the algorithm described by Özbek et al. [18]. First, a filter with time-varying center frequency is generated corresponding to each formant. Each of the four filters has a Gaussian frequency response, centered at the corresponding formant frequency (and therefore time-varying), and with a time-invariant bandwidth set equal to the average bandwidth of the corresponding formant. Second, a short-time Fourier transform (STFT) is computed. Third, the energy level associated with the ith formant, E_i, is computed by multiplying the magnitude STFT, $|X(f)|$, by the frequency response of the ith time-varying filter, $\mathcal{W}_i(f)$, and summing over all frequencies

$$E_i \triangleq \ln \left(\sum_{f=0}^{F_s/2} \mathcal{W}_i(f)|X(f)| \right) \tag{1}$$

where F_s denotes the sampling frequency of the speech signal. In this study, $FE \triangleq [F_1, \ldots, F_4, E_1, \ldots, E_4]$ denotes an 8-dimensional vector containing the frequencies and energies of the first four formants. $DFE \triangleq [FE, FE_\Delta, FE_{\Delta\Delta}]$ denotes the combination of FE and their velocity (FE_Δ) and acceleration ($FE_{\Delta\Delta}$) components.

B. Visual Features

In addition to acoustic features, visual features are useful in articulatory inversion. Lips, jaw, teeth, and some parts of the tongue are visible articulators, and visual information from these articulators can be extracted from a camera. The visual features used in this paper consist of 12 shape and 27 texture features, extracted from camera images of the talkers face by Katsamanis et al. [12] using Active Appearance Models. [13]. In this study, $V \triangleq [V_1, \ldots, V_{39}]$ denotes the 39-dimensional visual feature vector, and $DV \triangleq [V, V_\Delta, V_{\Delta\Delta}]$ denotes the combination of visual features and their velocity (V_Δ) and acceleration ($V_{\Delta\Delta}$) components.

III. Gaussian Mixture Model Based Articulatory Inversion

One of the best known acoustic-to-articulatory mapping methods is nonlinear regression based on GMMs (Toda et al. [10]). The basic idea of the method is as follows. Let $\mathcal{Z}$, $\mathcal{X}$ denote random vectors from acoustic (and/or visual) and articulatory spaces. Articulatory inversion methods look for an inverse mapping $\mathbf{g}(\cdot)$ defined as

$$\mathcal{X} = \mathbf{g}(\mathcal{Z})$$

to estimate articulatory vectors from given acoustic (and/or visual) data. The mapping between the articulatory and the acoustic (and/or visual) spaces is quite nonlinear and analytically unknown. Therefore, it is not a trivial problem to estimate an inverse mapping directly. In a probabilistic framework, the inverse mapping function $\mathbf{g}(\cdot)$ can be approximated if enough realization data pairs (z_k, x_k) of ($\mathcal{Z}$, $\mathcal{X}$) are available. Let $\hat{\mathbf{g}}(\cdot)$ be an estimate of the true inverse mapping $\mathbf{g}(\cdot)$ defined as

$$\hat{\mathcal{X}} \triangleq \hat{\mathbf{g}}(\mathcal{Z}).$$

In the minimum mean square error (MMSE) sense, the optimal approximate mapping $\hat{\mathbf{g}}_{MMSE}(\cdot)$ can be estimated by minimizing the error variance $\mathcal{J}(\hat{\mathbf{g}}(\mathcal{Z}))$

$$\hat{\mathbf{g}}_{MMSE}(\mathcal{Z}) = \arg\min_{\hat{g}(\cdot)} \mathcal{J}(\hat{\mathbf{g}}(\mathcal{Z})) \tag{2}$$

where $\mathcal{J}(\hat{\mathbf{g}}(\mathcal{Z})) \triangleq \mathrm{E}\left[(\mathcal{X} - \hat{\mathbf{g}}(\mathcal{Z}))^T(\mathcal{X} - \hat{\mathbf{g}}(\mathcal{Z}))\right]$. Taking the derivatives of $\mathcal{J}(\hat{\mathbf{g}}(\mathcal{Z}))$ with respect to $\hat{\mathbf{g}}(\mathcal{Z})$ and equating to zero, the approximate mapping can be found as

$$\hat{x}_{MMSE} \triangleq \hat{\mathbf{g}}_{MMSE}(z_k) = \mathrm{E}(\mathcal{X}|\mathcal{Z} = z_k). \tag{3}$$

In other words, the best thing that we can do in the MMSE sense is to represent the inverse mapping $\mathbf{g}(\cdot)$ with the conditional expectation of articulatory data given audiovisual data. In order to find a mathematically tractable approximate mapping we need further assumptions. If it is assumed that $\mathcal{X}$, $\mathcal{Z}$ are jointly distributed according to a Gaussian mixture model (GMM), then the joint distribution can be written as

$$f_{\mathcal{X},\mathcal{Z}}(x_k, z_k) \triangleq \sum_{i=1}^{\mathcal{K}} \pi_i \mathcal{N}(x_k, z_k; \mu^i, \Sigma^i) \tag{4}$$

where $\mathcal{K}$ is the number of mixture components and π_i is the ith mixture weight satisfying $\sum_{i=1}^{\mathcal{K}} \pi_i = 1$. μ^i and Σ^i denote the mean and the covariance of the GMM components and are defined as

$$\mu^i \triangleq \begin{bmatrix} \mu^i_{\mathcal{X}} \\ \mu^i_{\mathcal{Z}} \end{bmatrix}, \Sigma^i \triangleq \begin{bmatrix} \Sigma^i_{\mathcal{X}\mathcal{X}} & \Sigma^i_{\mathcal{X}\mathcal{Z}} \\ \Sigma^i_{\mathcal{Z}\mathcal{X}} & \Sigma^i_{\mathcal{Z}\mathcal{Z}} \end{bmatrix}. \tag{5}$$

The conditional distribution $f_{\mathcal{X}|\mathcal{Z}}(x_k|z_k)$ can be calculated from the joint distribution $f_{\mathcal{X},\mathcal{Z}}(x_k, z_k)$ using the Bayesian rule

$f_{\mathcal{X},\mathcal{Z}}(x_k, z_k)$ using the Bayesian rule

$$f_{\mathcal{X}|\mathcal{Z}}(x_k|z_k) \triangleq \frac{f_{\mathcal{X},\mathcal{Z}}(x_k, z_k)}{\int f_{\mathcal{X},\mathcal{Z}}(x_k, z_k)\, dx}.$$

The resulting conditional distribution is again a GMM, given as

$$f_{\mathcal{X}|\mathcal{Z}}(x_k|z_k) \triangleq \sum_{i=1}^{\mathcal{K}} \beta^i(z_k)\mathcal{N}(x_k; \mu^{i,k}_{\mathcal{X}|\mathcal{Z}}, \Sigma^i_{\mathcal{X}|\mathcal{Z}}) \tag{6}$$

where $\beta^i(z_k) \triangleq \mathcal{P}(c = i|\mathcal{Z})$ is the posterior probability of the ith mixture component, and $c \in (1, ..., \mathcal{K})$ is the mixture indicator. The parameters $\mu^{i,k}_{\mathcal{X}|\mathcal{Z}}$ and $\Sigma^i_{\mathcal{X}|\mathcal{Z}}$ are the conditional mean and covariance, respectively, of $\mathcal{X}$ given mixture indicator $c = i$ and observation $\mathcal{Z}$, and are defined as

$$\mu^{i,k}_{\mathcal{X}|\mathcal{Z}} = \mu^i_{\mathcal{X}} + \Sigma^i_{\mathcal{X}\mathcal{Z}}(\Sigma^i_{\mathcal{Z}\mathcal{Z}})^{-1}(z_k - \mu^i_{\mathcal{Z}}) \tag{7}$$

$$\Sigma^i_{\mathcal{X}|\mathcal{Z}} = \Sigma^i_{\mathcal{X}\mathcal{X}} - \Sigma^i_{\mathcal{X}\mathcal{Z}}(\Sigma^i_{\mathcal{Z}\mathcal{Z}})^{-1}\Sigma^i_{\mathcal{Z}\mathcal{X}} \tag{8}$$

$$\beta^i(z_k) = \frac{\pi_i \mathcal{N}(z_k; \mu^i_{\mathcal{Z}}, \Sigma^i_{\mathcal{Z}\mathcal{Z}})}{\sum_{i=1}^{\mathcal{K}} \pi_i \mathcal{N}(z; \mu^i_{\mathcal{Z}}, \Sigma^i_{\mathcal{Z}\mathcal{Z}})} \tag{9}$$

Under these assumptions, the MMSE articulatory estimate $\hat{x}_k$ and corresponding error covariance Σ_k can be found as follows

$$\begin{aligned} \hat{x}_k &\triangleq \mathrm{E}(\mathcal{X}|\mathcal{Z} = z_k) \\ &= \sum_{i=1}^{\mathcal{K}} \beta^i(z_k) \left[\mu^i_{\mathcal{X}} + \Sigma^i_{\mathcal{X}\mathcal{Z}}(\Sigma^i_{\mathcal{Z}\mathcal{Z}})^{-1}(z_k - \mu^i_{\mathcal{Z}})\right] \end{aligned} \tag{10}$$

$$\begin{aligned} \Sigma_k &\triangleq \mathrm{Cov}(\mathcal{X}|\mathcal{Z} = z_k) \\ &= \sum_{i=1}^{\mathcal{K}} \beta^i(z_k) \left[\Sigma^i_{\mathcal{X}|\mathcal{Z}} + (\mu^i_{\mathcal{X}|\mathcal{Z}} - \hat{x}_k)(\mu^i_{\mathcal{X}|\mathcal{Z}} - \hat{x}_k)^T\right]. \end{aligned} \tag{11}$$

The parameter set of the GMM, $\Theta_{GMM} = \{\pi_i, \mu^i, \Sigma^i\}_{i=1}^{\mathcal{K}}$, may be estimated using the expectation maximization (EM) algorithm, as described in [24].

IV. Dynamic Smoothing of Articulatory Trajectories

The movements of the vocal organs (such as lip, jaw, and tongue) are slowly varying and quite smooth. That means the positions of articulators change continuously. However, since the various articulatory inversion methods given in the literature do not produce continuously changing trajectories, they need a smoothing stage. For this purpose, Hiroya et al. [6], Richmond [9], Zhang et al. [7], and Toda et al. [10] use the maximum likelihood trajectory estimation (MLTE) method, which employs the time derivatives (e.g., velocity and acceleration) of articulatory features. It has been shown that the use of these auxiliary features improves the smoothness of the articulatory trajectories. Zero-phase low-pass filters (FIR or IIR) can also be used to smooth the output of the articulatory estimators (Richmond [8], Toda et al. [10] and Toutis et al. [25]). Zero-phase filtering can be performed by first filtering forward in time, then filtering backward (in time) using the same filter. For each articulator, the cut-off frequencies of the low-pass filters are optimized by trial and error. In Toda et al. [10], it is reported that applying low-pass smoothing to the trajectories obtained from the MLTE method slightly improves the performance of the method. Zero-phase filtering has the advantage of simplicity, but the disadvantage of relative inflexibility. In particular, it is not possible to make use of relevant information provided by the GMM, such as the error covariance of estimated articulatory trajectories, using zero-phase filtering. It is also not possible to make use of extrinsic information sources, e.g., information about phoneme boundary times, should such information be available. This section proposes a general statistical smoothing method based on Kalman smoothing. The proposed method is capable of utilizing statistical output quantities such as the estimated error covariance, and extrinsic information such as phoneme boundaries if they are available. The proposed smoothing method could, in principle, be applied to any articulatory inversion method given in the literature, including methods based on HMM ([13], [6]), SVM ([25], [26]), or MDN models [8].

Since the articulatory trajectories are physical quantities and their motion varies slowly, they can be modeled as the output of a dynamic system. The dynamics of the articulators and their relation with the observed GMM articulatory estimates, $\hat{g}_{MMSE}(z)$, can be approximated by a linear Gauss-Markov model, for which the optimal Kalman smoother can be formulated. The Kalman smoother can be written as a Bayesian-normalized generative model, therefore it is relatively straightforward to integrate auxiliary information that will improve its inference power; as an example, this paper demonstrates the use of a phonetic transcription as auxiliary information to improve the performance of the articulatory inversion.

A. Smoothing Problem Formulation

In this subsection, we will denote the output of the first stage articulatory inversion algorithm as y_n and the corresponding true articulatory state as x_n. The trajectories of the variables are shown by subscripts like $x_{1:N} \triangleq \{x_1, \ldots, x_N\}$. We assume that the dynamics of x_n, and the relationship between x_n and y_n are governed by nonlinear relations that can be approximated by the following piece-wise linear Gaussian dynamic system Deng [27].

$$x_{k+1} = F_1(s_{k+1})x_k + F_2(s_{k+1})x_{k-1} + \varepsilon(s_{k+1}) + \vartheta_k(s_{k+1}), \tag{12}$$

$$y_k = x_k + d(s_k) + v_k(s_k), \tag{13}$$

where $s_k \in \{1, \ldots, \mathcal{S}\}$ is the regime variable representing the model index and $\mathcal{S}$ is the total number of regimes. Each regime is a unit with linear Gauss-Markov articulatory dynamics (e.g.,

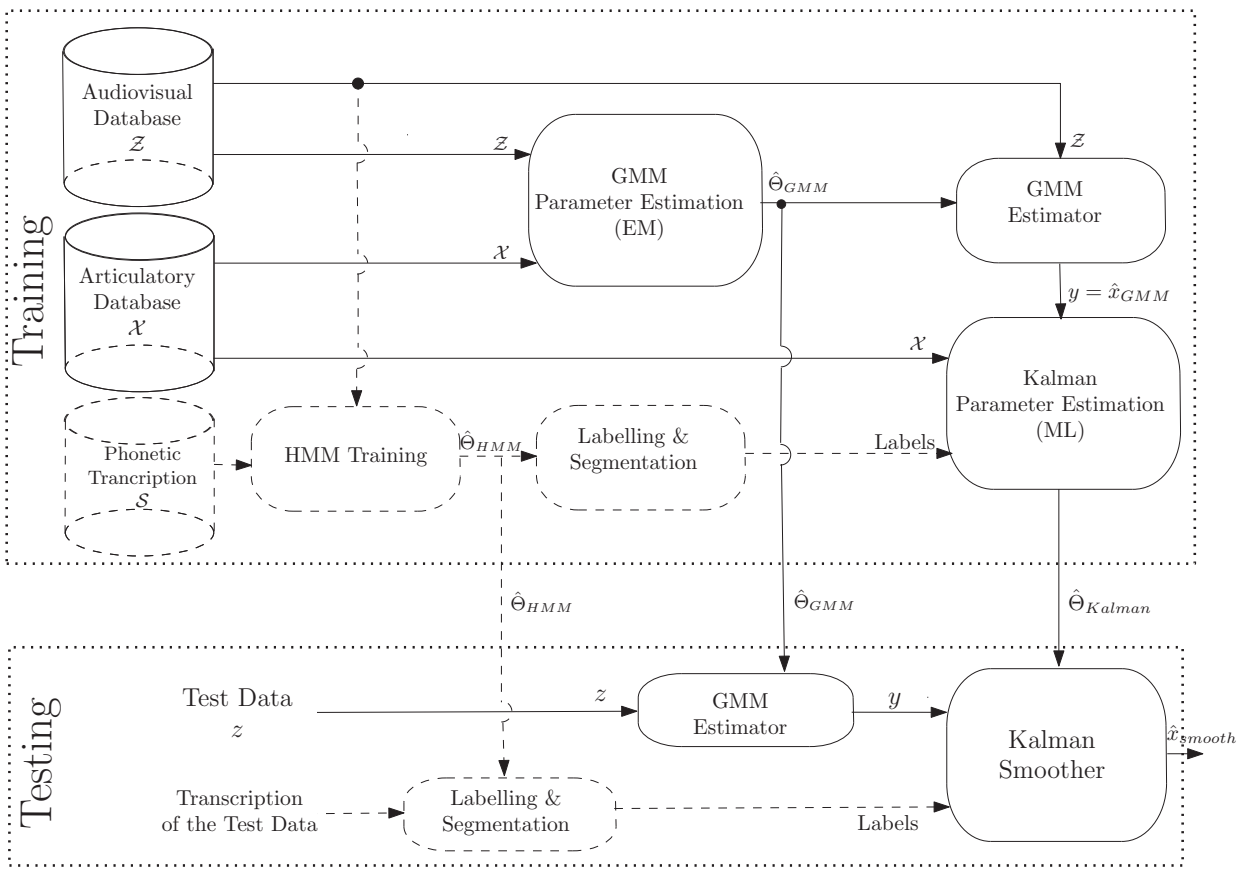

Fig. 1. The general block diagram of the smoothed GMM inversion (both training and testing phases) proposed in the paper. The dashed blocks and lines represent the operations concerning the auxiliary information of phonetic transcriptions. If such information is unavailable, they should be ignored.

we assume that phonemes have this property) and characterized by an s_k-dependent parameter set Θ_{s_k}. The second-order Markov dynamic system model of (12) and (13) can be converted into a first-order model by defining the augmented state vector $\mathbf{x}_k$ which is composed of the current and the previous articulatory state vectors, i.e., $\mathbf{x}_k \triangleq \left[x_k^T, x_{k-1}^T\right]^T$. Then, the first order model is given by

$$\mathbf{x}_{k+1} = F(s_{k+1})\mathbf{x}_k + u(s_{k+1}) + w_k(s_{k+1}), \quad (14)$$

$$y_k = H\mathbf{x}_k + d(s_k) + v_k(s_k), \quad (15)$$

where

- $\mathbf{x}_k \in \mathbb{R}^{2n_x}$ denotes the augmented state vector related to articulatory data and n_x is the dimension of articulatory state vector x_k,
- $y_k \in \mathbb{R}^{n_x}$ denotes the observation vector related to the output of GMM estimator with dimension of n_x,
- the augmented initial state $\mathbf{x}_0$ has the regime-dependent distribution $\mathcal{N}(\mathbf{x}_0; \bar{\mathbf{x}}(s), \Sigma(s))$. $\bar{\mathbf{x}}(s)$ and $\Sigma(s)$ are the initial mean and covariance of the state respectively,
- $F(s) \in \mathbb{R}^{2n_x \times 2n_x}$ is the regime-dependent state transition matrix given as

$$F(s) \triangleq \begin{bmatrix} F_1(s) & F_2(s) \\ I_{n_x \times n_x} & 0_{n_x \times n_x} \end{bmatrix}.$$

- $u(s) \in \mathbb{R}^{2n_x}$ is the regime-dependent bias vector given as $u(s) \triangleq [\varepsilon^T(s), 0_{1\times n_x}]^T$,
- $w_k(s) \triangleq [\vartheta_k^T(s), 0_{1\times n_x}]^T$, where $\vartheta_k(s) \sim \mathcal{N}(\vartheta(s); 0_{n_x\times n_x}, Q(s))$ and $Q(s) \in \mathbb{R}^{n_x \times n_x}$ is the covariance matrix of process noise,
- $H \in \mathbb{R}^{2n_x \times 2n_x}$ is the observation matrix defined as $H \triangleq [I_{n_x\times n_x}, 0_{n_x\times n_x}]$,
- $d(s) \in \mathbb{R}^{n_x}$ is the regime-dependent measurement bias,
- $v_k(s) \sim \mathcal{N}(v(s); 0_{n_x\times n_x}, R(s))$ is the regime-dependent Gaussian observation noise, and $R(s) \in \mathbb{R}^{n_x \times n_x}$ is the covariance matrix of the observation noise.

$I_{n_x\times n_x}$ and $0_{n_x\times n_x}$ are the identity and zero matrices with dimension $n_x \times n_x$. The vector $0_{1\times n_x}$ is the zero vector with dimension n_x. The regime dependent parameter set of the model can be defined as

$$\Theta_s = \{\bar{\mathbf{x}}_s, \Sigma_s, F(s), u(s), d(s), Q(s), R(s)\}.$$

In this work, it is assumed that the regime variables are known for each time instant n, and that if they are non-trivial, that they encode extrinsic information such as a phonetic transcription. In such a scenario, the MMSE estimate $\hat{x}_{k|N}$ of the true articulatory state, which is defined as

$$\hat{x}_{k|N} \triangleq E[x_k | y_{1:N}, s_{1:N}],$$

is given by a Kalman smoother if the dynamic models parameter set is known. Therefore, the process of smoothing articulatory trajectories involves two separate tasks: learning the parameter set, and inferring the state.

B. Learning the Parameter Set

The parameter set $\Theta_s = \{\bar{\mathbf{x}}_s, \Sigma_s, F(s), u(s), d(s), Q(s), R(s)\}$ for regime s is estimated using a training data set. We consider here a database composed of

- $\mathcal{M}$ acoustic trajectories shown as $z_{1:N_d}^{1:\mathcal{M}} \triangleq \{z_{1:N_d}^d\}_{d=1}^{\mathcal{M}}$, where N_d is the length of the dth trajectory;
- $\mathcal{M}$ corresponding true articulatory trajectories shown as $x_{1:N_d}^{1:\mathcal{M}} \triangleq \{x_{1:N_d}^d\}_{d=1}^{\mathcal{M}}$; and
- $\mathcal{M}$ corresponding regime trajectories shown as $s_{1:N_d}^{1:\mathcal{M}} \triangleq \{s_{1:N_d}^d\}_{d=1}^{\mathcal{M}}$.

Note that the availability of the regime variables does not cause a loss of generality because the case where the regime variables are not available is equivalent to the case that $\mathcal{S} = 1$ and $s_n^d = 1$ for all n and d.

The primary database described above is processed as follows to obtain a secondary database, which is then used to train the parameters of the dynamic model.

- An acoustic to articulatory GMM is trained using the EM algorithm on the primary database.
- All $\mathcal{M}$ acoustic trajectories $z_{1:N_d}^{1:\mathcal{M}}$ are input to the trained GMM, and the corresponding estimated articulatory trajectories ($y_{1:N_d}^{1:\mathcal{M}} \triangleq \{y_{1:N_d}^d\}_{d=1}^{\mathcal{M}}$) are obtained.
- All regime trajectories are partitioned into fragments of constant regime, i.e., each regime trajectory $s_{1:N_d}^d$ is divided into sub-regime trajectories $\{s_{n_s^j:n_e^j}^d\}_{j=1}^{m_d}$ such that when $s_{n_s^j:n_e^j}^d$ are concatenated, $s_{1:N_d}^d$ is obtained and all elements of $s_{n_s^j:n_e^j}^d$ are equal.
- All $\mathcal{M}$ estimated articulatory trajectories $y_{1:N_d}^{1:\mathcal{M}}$ and true articulatory trajectories $x_{1:N_d}^{1:\mathcal{M}}$ are partitioned according to their corresponding regime trajectories. Without loss of generality, all the partitioned sub-sequences are re-indexed and grouped according to the regime variables to obtain $\mathcal{M}_s$ estimated articulatory trajectories $y_{1:N_d}^{1:\mathcal{M}_s} \triangleq \{y_{1:N_d}^d\}_{d=1}^{\mathcal{M}_s}$ and true articulatory trajectories $x_{1:N_d}^{1:\mathcal{M}_s} \triangleq \{x_{1:N_d}^d\}_{d=1}^{\mathcal{M}_s}$ for each $s = 1, \ldots, \mathcal{S}$.
- The augmented states $\mathbf{x}_{2:N_d}^{1:\mathcal{M}_s} \triangleq \{\mathbf{x}_{2:N_d}^d\}_{d=1}^{\mathcal{M}_s}$ are formed from $x_{1:N_d}^{1:\mathcal{M}_s}$. For the sake of simplicity, we refer to the combination of $\mathbf{x}_{2:N_d}^{1:\mathcal{M}_s}$ and $y_{2:N_d}^{1:\mathcal{M}_s}$ as $\mathcal{D}_s$.

The database operations described above are illustrated schematically in the training part of Fig. 1, which shows both training and testing parts of the GMM inversion and smoothing operations used in this paper. Notice that the dashed blocks and lines in the figure represent the operations that depend on extrinsic phonetic transcriptions; if phonetic transcriptions are unavailable, the operations shown in dashed blocks and lines are omitted.

Suppose now that the training data set $\mathcal{D}_s$ is given for the sth regime. The unknown parameter set Θ_s can be estimated by maximizing the logarithm of the likelihood $\mathcal{L}(\Theta_s)$ of the training data set $\mathcal{D}_s$, i.e.,

$$\hat{\Theta}_s = \arg\max_{\Theta_s} \mathcal{L}(\Theta_s) \tag{16}$$

where $\mathcal{L}(\Theta_s) \triangleq \ln p(\mathbf{x}_{2:N_d}^{1:\mathcal{M}_s}, y_{2:N_d}^{1:\mathcal{M}_s}|\Theta_s)$. Under the assumption of Markov dynamics, the joint log-likelihood can be written as

$$\begin{aligned}\mathcal{L}(\Theta_s) =& \sum_{d=1}^{\mathcal{M}_s} \ln p(\mathbf{x}_{2:N_d}^d, y_{2:N_d}^d|\Theta_s) \qquad (17)\\ =& \sum_{d=1}^{\mathcal{M}_s} \ln p(\mathbf{x}_2^d|\Theta_s) + \sum_{d=1}^{\mathcal{M}_s}\sum_{k=3}^{N_d} \ln p(\mathbf{x}_k^d|\mathbf{x}_{k-1}^d, \Theta_s)\\ &+ \sum_{d=1}^{\mathcal{M}_s}\sum_{k=2}^{N_d} \ln p(y_k^d|\mathbf{x}_k^d, \Theta_s) \qquad (18)\end{aligned}$$

where

$$\begin{aligned}\mathbf{x}_2|\Theta_s &\sim \mathcal{N}(\mathbf{x}_s; \bar{\mathbf{x}}(s), \Sigma(s)), \qquad (19)\\ \mathbf{x}_k|\mathbf{x}_{k-1}, \Theta_s &\sim \mathcal{N}(\mathbf{x}_k; F(s)\mathbf{x}_{k-1} + u(s), Q(s)),\\ y_k|\mathbf{x}_k, \Theta_s &\sim \mathcal{N}(y_k; \mathbf{x}_k + d(s), R(s)).\end{aligned}$$

Taking derivatives of (17) for each unknown parameter, and setting the derivatives equal to zero, estimation formulas for parameter set Θ_s can be found as follows.

Algorithm 1 (Maximum Likelihood Parameter Estimation)**.** *Define the following summations:*

$$\bar{\mathbf{x}}_c \triangleq \sum_{d=1}^{\mathcal{M}_s}\sum_{k=3}^{N_d} \mathbf{x}_k^d, \quad \bar{\mathbf{x}}_p \triangleq \sum_{d=1}^{\mathcal{M}_s}\sum_{k=3}^{N_d} \mathbf{x}_{k-1}^d, \quad \bar{y}_c \triangleq \sum_{d=1}^{\mathcal{M}_s}\sum_{k=2}^{N_d} y_k.$$

The estimated parameters for regime s are given as

$$\begin{aligned}\hat{F}(s) =& \left(\sum_{d=1}^{\mathcal{M}_s}\sum_{k=3}^{N_d}[\mathbf{x}_k^d(\mathbf{x}_{k-1}^d)^T] - \frac{\bar{\mathbf{x}}_c\bar{\mathbf{x}}_p^T}{\sum_{d=1}^{\mathcal{M}_s}(N_d-2)}\right)\\ &\times\left(\sum_{d=1}^{\mathcal{M}_s}\sum_{k=3}^{N_d}[\mathbf{x}_{k-1}^d(\mathbf{x}_{k-1}^d)^T] - \frac{\bar{\mathbf{x}}_p\bar{\mathbf{x}}_p^T}{\sum_{d=1}^{\mathcal{M}_s}(N_d-2)}\right)^{-1}\\ \hat{u}(s) =& \frac{1}{\sum_{d=1}^{\mathcal{M}_s}(N_d-2)}[\bar{\mathbf{x}}_c - \hat{F}(s)\bar{\mathbf{x}}_p] \qquad (20)\\ \hat{d}(s) =& \frac{1}{\sum_{d=1}^{\mathcal{M}_s}(N_d-1)}[\bar{y}_c - \bar{\mathbf{x}}_c] \qquad (21)\\ \hat{Q}(s) =& \frac{1}{\sum_{d=1}^{\mathcal{M}_s}(N_d-1)}\sum_{d=1}^{\mathcal{M}_s}\sum_{k=3}^{N_d}[\mathbf{x}_k^d - \hat{F}(s)\mathbf{x}_{k-1}^d - \hat{u}(s)]\\ &\times[\mathbf{x}_k^d - \hat{F}(s)\mathbf{x}_{k-1}^d - \hat{u}(s)]^T \qquad (22)\\ \hat{R}(s) =& \frac{1}{\sum_{d=1}^{\mathcal{M}_s}(N_d-1)}\sum_{d=1}^{\mathcal{M}_s}\sum_{k=2}^{N_d}[y_k^d - \mathbf{x}_k^d - \hat{d}(s)]\\ &\times[y_k^d - \mathbf{x}_k^d - \hat{d}(s)]^T \qquad (23)\\ \hat{\bar{\mathbf{x}}}(s) =& \frac{1}{\mathcal{M}_s}\sum_{d=1}^{M_s}\mathbf{x}_2^d \qquad (24)\\ \hat{\Sigma}(s) =& \frac{1}{\mathcal{M}_s}\sum_{d=1}^{\mathcal{M}_s}[\mathbf{x}_2^d - \hat{\bar{\mathbf{x}}}(s)][\mathbf{x}_2^d - \hat{\bar{\mathbf{x}}}(s)]^T \qquad (25)\end{aligned}$$

C. Inference (State Estimation)

After the parameter learning stage, the smoothed state $\hat{\mathbf{x}}_{k|N}$ can be estimated by a Kalman smoother. For this purpose, first a Kalman filter estimates the whole filtered state $\hat{\mathbf{x}}_{k|k}$ in the forward direction, and then in the backward direction, the Kalman smoother estimates the smoothed state $\hat{\mathbf{x}}_{k|N}$ as described in (Simon [28]). The fundamental algorithms for Kalman smoothing are not repeated here because of space considerations.

Remark 1. *When there is extrinsic information available that can be used as a sequence of regime variables (phonetic transcriptions in our case), the quantity R_s calculated in* (23) *is an appropriate estimate of the dynamic systems measurement covariance. However, when this extra information is absent, R_s becomes too coarse in general. In this case, the covariance* (11) *provided by the GMM based inversion method is better suited for use as measurement covariance.* □

V. Acoustic and Visual Information Fusion

The fusion of audio and visual features improves the performance of articulatory inversion. The fusion of audio and visual information is examined in Katsamanis et al. [13] with an HMM based audiovisual inversion method and in Kjellström et al. [29] with both linear and nonlinear (Artificial Neural Network (ANN) based) audiovisual inversion. This part of the paper examines audiovisual information fusion for GMM based inversion. The fusion of audio and visual features is performed in three ways: early, late, and modified-late fusion. The early- and late-fusion strategies are modeled after the methods of Katsamanis et al [13]. The modified-late fusion strategy is a type of late fusion algorithm based on observability characteristics of articulatory state components.

A. Early Fusion

In early fusion (feature or centralized fusion), the audio features (MFCC, LPC,...) and visual features are augmented to form a large feature vector, and the GMM regression based inversion is conducted using these combined features. Mathematically speaking, the acoustic measurement vector z in (10) is replaced by z_e defined as

$$z_e \triangleq \left[\; z_a^T, z_v^T \; \right]^T \tag{26}$$

where z_a and z_v are vectors of audio and visual features respectively. The early-fusion MMSE articulatory estimate $\hat{x}_e$ and its covariance Σ_e are found by using (7) and (8) respectively. In this work, the early-fusion method is also used to examine various combinations of acoustic-only feature vectors (MFCC, LPC, LAR, etc.).

B. Late Fusion

Late fusion (distributed fusion) combines separate estimation results which are based on audio and visual measurements. In this method, there are two different GMMs, one for audio and the other for visual data. For each time frame, their estimates are averaged with matrix weights to obtain the fused estimate. Matrix weights are generated from the local estimate covariances. A summary of the fusion algorithm is as follows. Let $\hat{x}_a \triangleq \mathrm{E}[\mathcal{X}|\mathcal{Z}_a]$ and $\hat{x}_v \triangleq \mathrm{E}[\mathcal{X}|\mathcal{Z}_v]$ be the estimated articulatory trajectories using audio and visual measurements respectively, and $\Sigma_a \triangleq \mathrm{Cov}(\mathcal{X}|\mathcal{Z}_a)$ and $\Sigma_v \triangleq \mathrm{Cov}(\mathcal{X}|\mathcal{Z}_v)$ be their corresponding covariance matrices, which are estimated using (10), (11) . The late fusion estimate $\hat{x}_l$ can be calculated for each time frame as

$$\hat{x}_l = W_a \hat{x}_a + W_v \hat{x}_v. \tag{27}$$

The weighting matrices W_a and W_v are calculated by minimizing the error covariance of $\hat{x}_l$. Considering the fact that each estimate $\hat{x}_a$ and $\hat{x}_v$ uses a different feature set, it can be assumed that estimation errors corresponding to these estimates are independent. In that case, for the unbiased estimation, the weights W_a and W_v are found as follows

$$W_a = \Sigma_v(\Sigma_a + \Sigma_v)^{-1} \text{ and } W_v = \Sigma_a(\Sigma_a + \Sigma_v)^{-1}.$$

Hence, the late fusion estimate $\hat{x}_l$ and its covariance Σ_l can be written as

$$\hat{x}_l = \Sigma_v(\Sigma_a + \Sigma_v)^{-1}\hat{x}_a + \Sigma_a(\Sigma_a + \Sigma_v)^{-1}\hat{x}_v, \tag{28}$$

$$\Sigma_l = \Sigma_a(\Sigma_a + \Sigma_v)^{-1}\Sigma_v. \tag{29}$$

The smoothing of the output of late fusion can be handled via Kalman smoother using $\hat{x}_l$ and Σ_l as an observation vector and measurement covariance, respectively (Sec. IV). It is possible to combine the late fusion and the smoothing process into a single Kalman smoother that uses a state space model with the same state equation as (14), and with a modified observation equation given as

$$y_k^a = H_a \mathbf{x}_k + d_a(s) + v_k^a(s), \tag{30}$$

$$y_k^v = H_v \mathbf{x}_k + d_v(s) + v_k^v(s), \tag{31}$$

where y_k^a and y_k^v are the observation vectors which represent the output estimates $\hat{x}_a$ and $\hat{x}_v$ of the Audio-GMM and Visual-GMM respectively at time instant k. Considering independent audio and visual measurements again, the combination of the observation equations (30) and (31) yields the following augmented measurement model.

$$y_k^e = H_l \mathbf{x}_k + q(s) + \nu_k(s) \tag{32}$$

where

$$y_k^e \triangleq \begin{bmatrix} y_k^a \\ y_k^v \end{bmatrix}, \quad H_l \triangleq \begin{bmatrix} H_a \\ H_v \end{bmatrix}, \quad q(s) \triangleq \begin{bmatrix} d_a(s) \\ d_v(s) \end{bmatrix}, \tag{33}$$

$$\nu_k(s) \triangleq \begin{bmatrix} v_k^a(s) \\ v_k^v(s) \end{bmatrix}, \quad R(s) \triangleq \begin{bmatrix} R_a(s) & 0 \\ 0 & R_v(s) \end{bmatrix}. \tag{34}$$

The measurement matrices H_a and H_v are given as $H_a = H_v \triangleq [I_{n_x \times n_x}, 0_{n_x \times n_x}]$. The audio and visual GMM covariances Σ_a and Σ_v are used for the measurement covariances R_a and R_v respectively. The Kalman smoother can be run for the augmented model above to find the smoothed late fusion results. This combined scenario is depicted in Fig. 2.

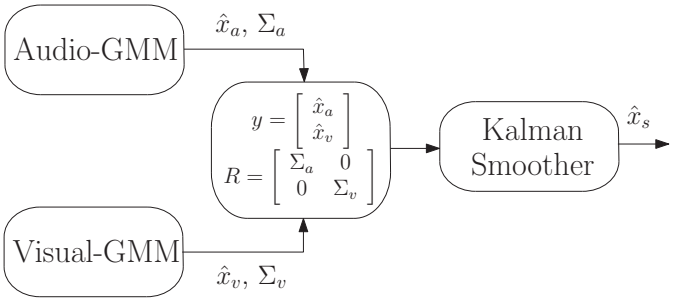

Fig. 2. Combined late fusion and smoothing process as a single smoother.

C. Modified Late Fusion

We propose a modified late fusion algorithm, modified to explicitly represent differences in the visual observability of different articulatory state components, and to correspondingly weight their contributions to articulatory estimation. The movement of visually apparent articulators such as lips (upper, lower lip) and jaw (lower incisor) can be captured by the camera quite accurately. On the other hand, the movement of the tongue tip, body, dorsum are only imperfectly captured on camera, and the movement of the velar articulator is not visible on an ordinary video record. Information about the observability of different articulators is relevant to the task of audiovisual fusion.We choose to modify the observation equations (30) and (31) of the late fusion process in order to explicitly represent the relative visual observability of the different articulators. The late fusion algorithm uses the observation matrices $H_a = H_v \triangleq [I_{n_x \times n_x}, 0_{n_x \times n_x}]$. In the modified late fusion method, these observation matrices are generalized as follows

$$H_a^m \triangleq \left[\begin{array}{cc} C_a & 0_{n_x \times n_x} \end{array} \right] \quad H_v^m \triangleq \left[\begin{array}{cc} C_v & 0_{n_x \times n_x} \end{array} \right]. \quad (35)$$

The acoustic $C_a \in \mathbb{R}^{n_x \times n_x}$ and visual $C_v \in \mathbb{R}^{n_x \times n_x}$ observation matrices are chosen to be diagonal. Each diagonal element of the observation matrix is selected from the interval [0,1] that represents the observability degree of the corresponding state component. The diagonal element 1 denotes a fully observable state while 0 denotes a completely unobservable state. In the experimental studies we used the values 0.9, 0.6, and 0.1 for "highly observable", "moderately observable" and "poorly observable" components of the matrix C_v.

VI. Experimental Studies

A. Experiments

In this work, we use the MOCHA database [5]. The acoustic data and EMA trajectories of one female talker (fsew0) are used; these data include 460 sentences. Audio and visual features were computed using a 36 ms window with 18 ms shift. A pre-emphasis filter (with $\alpha = 0.97$) is used in the extraction of all audio features. The summary of audiovisual feature types used in this paper is given in Table I. The articulatory data are EMA trajectories recorded at 500 Hz sampling frequency, which are the X and Y coordinates of the lower incisor, upper lip, lower lip, tongue tip, tongue body, tongue dorsum and velum. EMA trajectories are normalized by the methods suggested in Richmond [8] and down-sampled to match the 18 ms shift rate. All the model parameters of the GMM, HMM and Kalman smoother models are tested using 10-fold cross-validation. For each fold, nine-tenths of the data (414 sentences) are used for training and one-tenth (46 sentences) for testing. Cross-validation performance measures (RMS error and correlation coefficient) are computed as the average of all 10 folds. Phonetic segmentation and labeling are performed using HMM-based forced alignment. In this work, three-state, left-to-right HMMs are used for each phoneme. A total of 46 HMMs are trained for the MOCHA database: 44 for the phonemes, and 2 for breath and silence. The experiments can be classified into four classes.

- The first class of experiments uses only one type of feature vector as given in Table I.
- The second class of experiments uses a combination of two or three different types of acoustic features given in Table I. Note that there is a huge number of possible feature combinations.
  Hence, a small subset of the possible combinations is tested, based on the performance of individual feature types in the first experiment, and only the results of the selected feature combinations are given. The selected combinations and their performances are shown in Table II and in Fig. 4.
- The third class of experiments compares the proposed smoothing methods with the maximum likelihood trajectory estimation (MLTE) method.
- The fourth class of experiments fuses the acoustic and visual features. This set of experiments examines the performance of the fusion of acoustic and visual estimation results. Both the late and early fusion methods of Katsamanis et al. [13] and Kjellström et al. [29] are employed, and are compared to our proposed modified late fusion method.

Experiments are also distinguished based on the smoothing method that they use. Two Kalman smoothers are tested: a global Kalman smoother (GK) that uses a model generated by all of the available data, and a set of 46 phoneme-based Kalman (PBK) smoothers, in which each dynamic system model is trained to represent only the data that belong to a certain phonetic class. In a phoneme-based Kalman smoother, it is assumed that the phonetic transcription of test data is available and the test data can be segmented and labeled via HMM forced alignment.

B. Performance and Significance Test Measures

The performance of the algorithms is measured using three performance measures: RMS error, normalized RMS error, and correlation coefficient, all of which are as described in Richmond [8] and Katsamanis et al. [13]

- RMS error:

$$E_{RMS}^i \triangleq \sqrt{\frac{1}{\mathcal{I}} \sum_{k=1}^{\mathcal{I}} (x_k^i - \hat{x}_k^i)^2}, \quad i = 1, \ldots, n_x \quad (36)$$

  where x_k^i and $\hat{x}_k^i$ are true and estimated positions, respectively, of the ith articulator at the kth frame. $\mathcal{I}$ and n_x

TABLE I
AUDIO-VISUAL FEATURE TYPES USED IN THIS STUDY.

FE	Four formant frequencies and their energy levels $FE = [F_1, F_2, F_3, F_4, E_1, E_2, E_3, E_4]$
M	Mel-frequency cepstral coefficients ($MFCC$) $M = [M_1, \ldots, M_{13}]$
L	Linear Predictive Coding coefficients (LPC) $L = [L_1, \ldots, L_{18}]$
A	Log Area Ratio coefficients (LAR) $A = [A_1, \ldots, A_{18}]$
S	Line Spectral Frequencies coefficients (LSF) $S = [S_1, \ldots, S_{18}]$
V	Visual Active Appearance coefficients ($Visual$) $V = [V_1, \ldots, V_{39}]$
DX	Combination of X and its time derivatives; velocity and acceleration components $DX \triangleq [X, X_\Delta, X_{\Delta\Delta}]$ (X can be any feature type)

are the total number of frames in the database and the total number of articulators respectively.

- Normalized RMS error:

$$E^i_{NRMS} \triangleq \frac{E^i_{RMS}}{\sigma_i}, \quad i = 1, \ldots, n_x \tag{37}$$

where σ_i is the standard deviation of ith articulator x^i.

- Correlation coefficient:

$$\rho^i_{x,\hat{x}} \triangleq \frac{\sum_{k=1}^{\mathcal{I}}(x^i_k - \bar{x}^i_k)(\hat{x}^i_k - \bar{\hat{x}}^i_k)}{\sqrt{\sum_{k=1}^{\mathcal{I}}(x^i_k - \bar{x}^i_k)^2}\sqrt{\sum_{k=1}^{\mathcal{I}}(\hat{x}^i_k - \bar{\hat{x}}^i_k)^2}} \tag{38}$$

for $i = 1, \ldots, n_x$ where $\bar{x}^i$ and $\bar{\hat{x}}^i$ are the average positions of true and estimated ith articulator respectively.

- Significance Test: The following significance test is performed to compare two articulatory inversion methods (i.e., B_1 and B_2). Two hypotheses are compared: H_0 and H_1. The null hypothesis H_0 states that the performance of method B_1 is not better than that of B_2. The test hypothesis H_1 states that the performance of method B_1 is better than B_2.

$$\begin{aligned} H_0 &: e = J^{B_2} - J^{B_1} \leq 0 \\ H_1 &: e = J^{B_2} - J^{B_1} > 0 \end{aligned} \tag{39}$$

where J^{B_1} is the performance measure using the method B_1 and J^{B_2} is the performance measure obtained from the method B_2. We reject the null hypothesis if

$$z = \frac{\bar{e}}{\frac{\sigma_{\bar{e}}}{\sqrt{K}}} > t_0(\alpha) \tag{40}$$

where, $\bar{e} \triangleq \frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}} e_i$, $\sigma_{\bar{e}} \triangleq \sqrt{\frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}}(e_i - \bar{e})^2}$ and $t_0(\alpha)$ is the threshold based on the upper tail of the normal density with significance level α (for $\alpha = 0.01$, $t_0 = 2.33$). In order to validate the assumption of independent trials, each sentence is treated as a trial, rather than each frame; thus e_i is the average performance for the ith sentence and $\mathcal{M}$ is the total number of sentences. If the performance measure is chosen to be correlation coefficient, a similar significance test can be performed with some modifications.

C. Experimental Results

1) Experimental Results for Single Feature Set: The experimental results given in this section are for the articulatory inversion using only the acoustic data, or using only visual data. The comparison of acoustic features (MFCC, LPC, LAR, Visual, etc.) and different smoothing methods can be seen in Fig. 3. The first observation from these figures is that the MFCC feature vector gives better results than any other feature vector. This result is similar to those reported in Katsamanis et al. [13]. The LSF and FE are the second and third best features.

The second observation from these figures is that dynamic features improve the performance, as do low-pass or Kalman smoothers. All these reduce RMS error and improve the correlation coefficient for each feature set. The global Kalman (GK) smoother gives slightly better performance than the low-pass (LP) smoother. Moreover, the phone-based Kalman (PBK) smoother, which can be applied if an auxiliary phonetic transcription is available, gives the best results. As an example, using MFCC with only static features, the RMSE and correlation coefficient are about 1.695 mm and 0.69 respectively. If dynamic features are used in addition to static features, RMSE reduces to 1.6 mm and the correlation coefficient increases to 0.73. When the global Kalman (GK) smoother is applied to the articulatory trajectories, the RMSE and correlation coefficient become 1.428 mm and 0.81. The phone-based Kalman (PBK) smoother gives 1.389 mm RMSE and 0.823 correlation coefficient.

In addition, it is observed that LSF and LAR perform better than LPC, although they are derived deterministically from LPC. From the performance of the features, we can say that LSF expresses the autoregressive spectral information in a way that is more useful for articulatory inference than LPC or LAR. FE contains four formants, derived from LPC, and formant energies derived from LPC and the power spectrum; although the formant energies are correlated with LPC, they may contain additional information not provided by the LPC vector. Therefore the better performance of FE may be caused by the inclusion of energy in the feature vector, the recoding of LPC information into a more useful form, or some combination of these two factors.

It is also interesting to observe from Fig. 3 that the performance is highly improved by applying smoothing. The improvement is especially significant when the phoneme-based Kalman smoother is used for visual data. It seems perhaps that the articulators not directly measured by the camera are estimated in an almost open-loop fashion, based almost entirely on their dynamic behavior. We can say that the resultant system is *state observable*: unmeasured articulator positions are also estimated with some accuracy, so much so that visual performance is better than LPC. When we analyze the correlation coefficient given in Fig. 3.b we observe parallel results to RMS error, i.e., when RMS error is low, the correlation is high.

2) Experimental Results for Combined Acoustic Features: The experimental results given in this section are for the articulatory inversion using some combinations of the acoustic

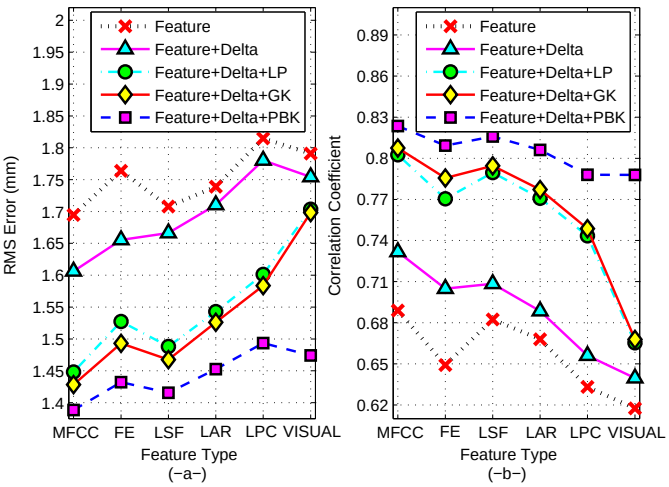

Fig. 3. RMS errors (a) and correlation coefficient (b) for the different audio/visual features and different smoothers.

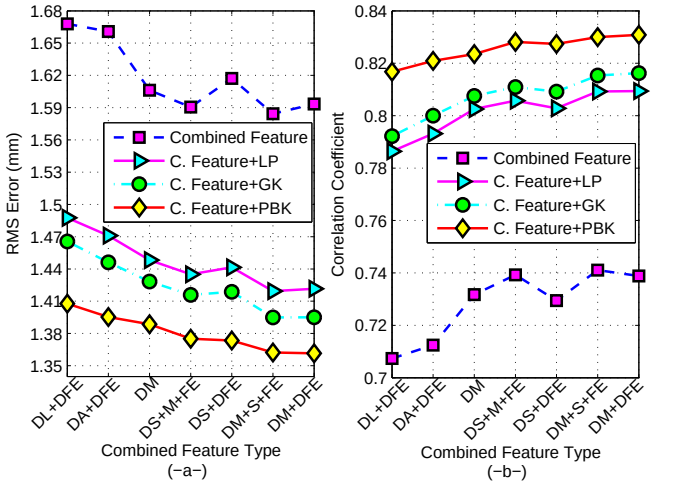

Fig. 4. RMS errors obtained using combinations of different acoustic features and different smoothers (a), and the corresponding correlation coefficient values (b).

features. By "combination" we mean the concatenation of the feature vectors, i.e., "early fusion." The combination patterns and performance results can be seen in Table II and Fig. 4. In Table II, RMS error given in each cell of the table corresponds to the performance obtained by combining the features in its corresponding row and column heading without using phonetic transcription. As an example, the RMS error for combination of DM and DFE gives 1.395 mm error, which can be seen on top of the last column. The best result for each row is highlighted. The first observation from this table is that the combination of formant-related features with almost any other feature vector gives better results than any other feature combination. The second observation is that the best combination is DM plus DFE, with the corresponding RMS error of 1.395 mm.

TABLE II
RMS ERRORS FOR VARIOUS COMBINATIONS OF ACOUSTIC FEATURES.

Features	M	FE	S	A	L	DFE
DM	–	1.415	1.397	1.415	1.422	1.395
DFE	1.418	–	1.423	1.439	1.447	–
DS	1.425	1.447	–	1.454	1.471	1.419
DA	1.452	1.474	1.467	–	1.507	1.446
DL	1.478	1.51	1.498	1.514	–	1.471
DM+S	–	1.398	–	1.42	1.433	–
DS+M	–	1.416	–	1.433	1.441	–

The RMS errors and correlation coefficients for highlighted feature combinations given in Table II are displayed with more detail in Fig. 4. In this figure, the performance of DM alone is also shown for the sake of comparison. DM is selected for comparison since it is the best single feature vector. According to this figure, some combinations of acoustic features improve the performance of the articulatory inversion. Second, as in the experiment with single-type feature vectors, global Kalman (GK) smoothing gives better results than a low-pass (LP) smoother, but the phone-based Kalman (PBK) smoother produces the best RMS error and correlation coefficient values.

The best combination of acoustic features is MFCC and formant-related features ($DM + DFE$). For this acoustic feature combination, the RMS error and correlation coefficients for global Kalman smoother are 1.395 mm and 0.816 respectively. In the case of phone-based Kalman smoother, RMS error and correlation coefficients are 1.362 mm and 0.83 respectively. Observing the results of single features, we have concluded that LPC contains irrelevant information and the features LSF and LAR are better than LPC although they are directly obtained from LPC. Because LSF and LAR are derived in a one-to-one mapping from LPC, the LPC-articulatory mapping does not suffer from greater non-uniqueness than LSF-articulatory mapping, but the LPC-articulatory mapping might be less smooth (indeed, this is one of the original motivations for the LSF representation (Itakura [21])), and is therefore learned less well by a GMM. On the other hand, MFCC had better performance than all the others, and the new experiments show that the combination of MFCC and formant-related features improves the performance. That means that MFCC and formants carry complementary information about the position of the articulators.

In this set of experiments, we also conducted two types of statistical significance tests. The first type of significance test examines whether or not articulatory inversion using the combined feature set $DM + DFE$ (MFCC and formant-related acoustic features) significantly outperforms articulatory inversion using only MFCC (DM) features.

Fig. 5 provides details regarding the utility of formant-related acoustic features in inversion and shows the significance test of the results. This figure shows that RMS error reduction using formant-related features in addition to MFCC features is significant at the $\alpha = 0.01$ level of significance for each articulator. The second type of significance test is performed to show whether or not articulatory inversion using a global Kalman smoother significantly outperforms articulatory inversion using only low-pass (LP) smoothing.

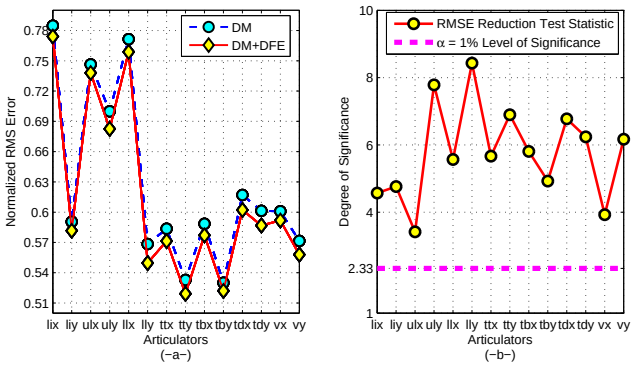

Fig. 5. Normalized RMS errors of using DM+DFE and DM features for each articulator (a). The corresponding significance test results for different articulators (b). The abbreviations li, ul,ll, tt,tb, td and v denote lower incisor, upper lip, lower lip, tongue tip, tongue body, tongue dorsum and velum, respectively. The suffixes x and y to the articulator abbreviations show the corresponding X and Y coordinates, respectively.

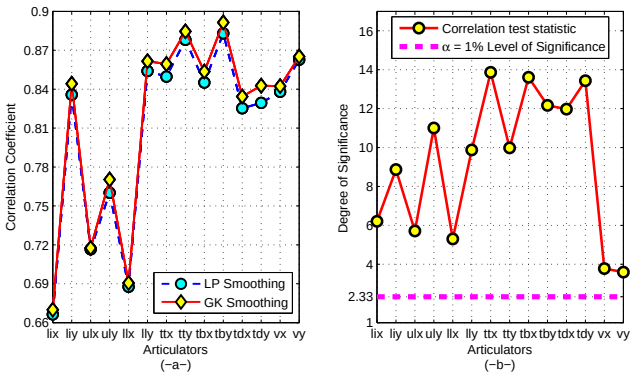

Fig. 6. (a) Correlation coefficients obtained using low-pass (LP) smoother and Kalman smoother; (b) The corresponding significance test results for different articulators. Abbreviations related to the names of the articulators are explained in Fig. 5.

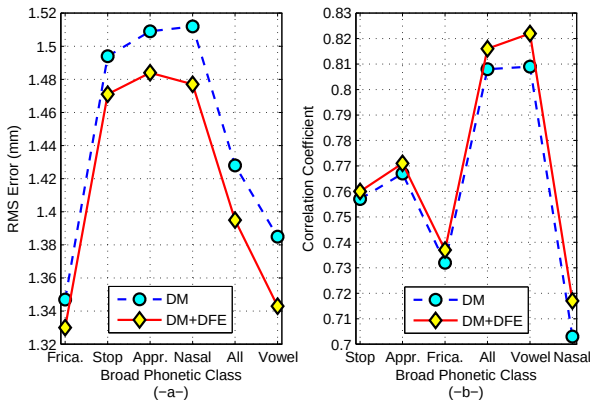

Fig. 7. (a) RMS errors of using $DM + DFE$ and DM for each broad phonetic class. (b) Corresponding correlation coefficients.

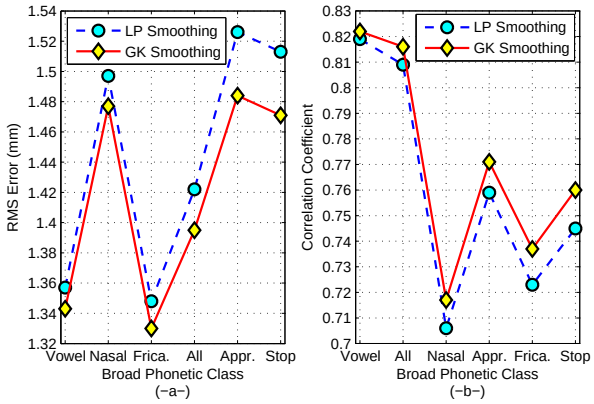

Fig. 8. (a) RMS errors obtained using global Kalman (GK) smoother and low-pass (LP) smoother for each broad phonetic class. (b) The corresponding correlation coefficients.

Fig. 6 shows that using a global Kalman smoother instead of a low-pass smoother gives statistically significant correlation improvements for each articulator. It is also interesting to analyze both types of improvement (augmenting feature vector DM by DFE and using Kalman smoother instead of LPF) for each broad phonetic class. Fig. 7 depicts the distribution of RMS error reduction and correlation improvement for each broad phonetic class using formant-related acoustic features and MFCC together. From this figure, it can be concluded that formant-related acoustic features are especially useful when the sounds are vowel and nasal.

Results of a similar experiment, conducted to examine the distribution of RMS error reduction and correlation improvement attributable to global Kalman vs. LP smoothing as functions of broad phonetic class, are shown in Fig. 8. Fig. 8 shows that the global Kalman smoother is especially useful for stop, fricative and approximant sounds.

3) Experimental results comparing the proposed smoothing methods with the MLTE method: In this section of the experiments, we compare the proposed smoothing methods with the maximum likelihood trajectory estimation (MLTE) method given in Richmond ([9], [30]), Toda et al. [10] and Hiroya et al. [31]. The GMM-based MLTE is introduced in Toda et al. [10]. In order to perform this comparison, we perform experiments similar to those reported in [10] to obtain MLTE-based trajectories

As described in Toda et al. [10], the feature vectors are generated by augmenting position and velocity components of the articulatory trajectories to the acoustic feature vector. These vectors are used in training of the GMM. Estimation of articulatory trajectories is done by the expectation maximization (EM) algorithm. The initial estimates are taken from MMSE-based estimation results. A more detailed explanation about MLTE used in this paper is given in the appendix for the sake of clarity.

The comparison of the proposed smoothing algorithms (GK and PBK smoothers) and MLTE for acoustic feature combination DM+DFE can be seen in Table III. The first observation from the table is that the results of MLTE (in short: ML) are better than the results of MMSE (in short: MS) based estimation. Furthermore, MS+LP (low-pass-smoothed MS trajectories) and ML have about similar performance. On the other hand, Richmond [9] and Toda et al. [10] report that ML gives better results than MS+LP. This difference may be

attributable to the selection of the acoustic features types. In this study we use the acoustic feature vector of one frame, but Richmond [9] and Toda et al. [10] use features which are obtained by the augmentation of the acoustic features of multiple frames. In Toda et al. [10], ML-based trajectories are further smoothed by applying a low-pass filter and the performance of the method is slightly improved. Therefore, we also applied the same smoothing procedure to compare the results. Smoothed ML is denoted by ML+LP. Comparison shows that ML+LP is slightly better than MS+LP. Moreover, Table III also shows that MS+GK gives lower RMS error and similar correlation coefficient compared with ML+LP. Similar to other experimental results given in this paper, MS+PBK filter gives the best results.

TABLE III
RMS ERRORS AND CORRELATION COEFFICIENTS FOR MMSE BASED ESTIMATION VS. MLTE METHOD

Per.	MS	ML	MS+LP	ML+LP	MS+GK	MS+PBK
rmse	1.593	1.437	1.422	1.413	1.395	1.362
corr.	0.738	0.806	0.809	0.814	0.816	0.831

4) Experimental Results for Audiovisual Fusion: The results for the early fusion inversion can be seen in Fig. 9. This figure shows that the combination of the visual features with MFCC and formant-related audio features (FE) gives better results than other combinations. The graphs also demonstrate the effectiveness of smoothing, especially when a good model, i.e., phone-based, is used in the smoother.

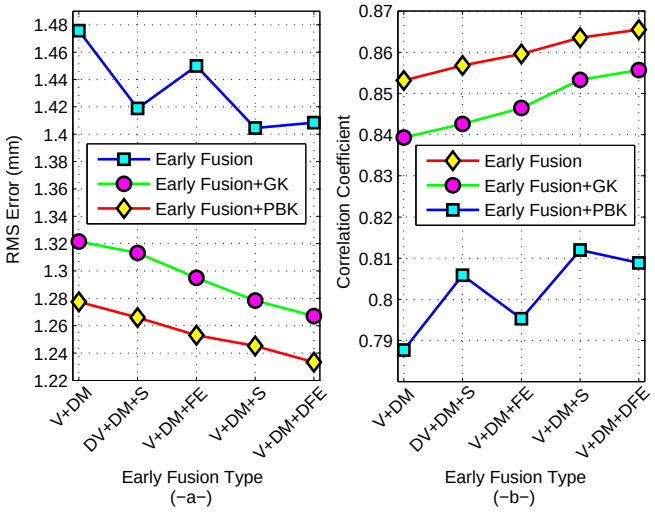

Fig. 9. (a) RMS errors for the different audio-visual features and smoothers using early audio-visual fusion, and (b) corresponding correlation coefficients.

The results for different fusion types can be seen in Fig. 10 for three sets of articulators: lips and jaw, tongue, and velum. A detailed examination of this figure suggests that inversion using only audio features, on average, gives better performance for tongue related and velar articulators, while inversion using only the visual gives better performance for lips and jaw (lower

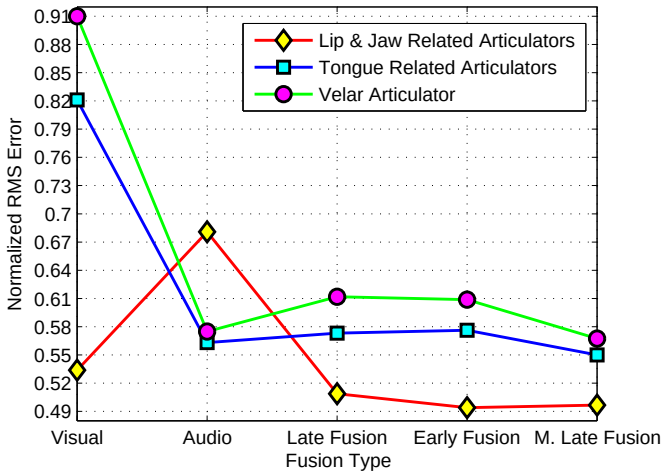

Fig. 10. Normalized RMS error with different sets of articulators for the various fusion types.

incisor). Therefore, the modified late fusion method uses the following C_a and C_v matrices (in equation (35)) in order to improve the inversion performance of the late fusion method.

$$C_a \triangleq [0.9 I_{14x14}]$$
$$C_v \triangleq \text{blkdiag}\,[0.9 I_{6x6}, 0.6 I_{6x6}, 0.1 I_{2x2}] \qquad (41)$$

The overall results for different fusion types are depicted in Fig. 11. This figure demonstrates that the modified late fusion is the best fusion type for GMM based audiovisual-articulatory inversion. Furthermore, the error reduction curve shows that for the modified late fusion scenario, the difference between the global Kalman and phone-based Kalman smoothers is reduced.

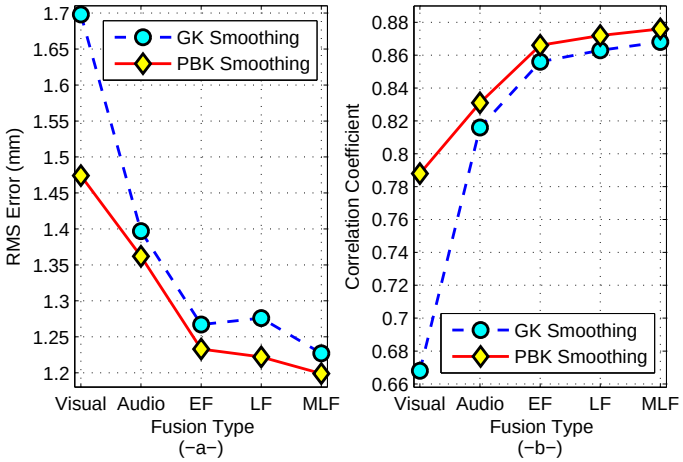

Fig. 11. (a) RMS errors for the various fusion types (blue lines and left axis), and (b) the corresponding correlation coefficients. The abbreviations EF, LF, MLF denote early, late and the modified late fusions.

VII. Discussion and Conclusion

This paper demonstrates the utility of vocal-tract-related acoustic features, and of selective audiovisual fusion, for

TABLE IV
BEST EXPERIMENTAL RESULTS FOR GMM BASED ARTICULATORY INVERSION.

Phonetic Transcription	Inversion Type	Feature Type	# of Mixture Component	Normalized RMSE	RMSE (mm)	Corr. Coeff.
Unavailable	Visual-only	DV	8	0.71	1.7	0.668
	Audio-only	DM+DFE	154	0.615	1.395	0.816
	Audiovisual	DV and DM+DFE	8+154	0.529	1.227	0.868
Available	Visual-only	DV	8	0.62	1.474	0.788
	Audio-only	DM+DFE	154	0.6	1.362	0.831
	Audiovisual	DV and DM+DFE	8+154	0.517	1.199	0.876

the purpose of audiovisual-to-articulatory inversion. Table IV summarizes the best results from all experiments. As reported by other authors, MFCCs are the best single feature set for articulatory inversion, but the performance of the MFCC-based inversion algorithm can be improved by appending a vector of line spectral frequencies, or even better, a vector of formant frequencies and energies. As reported by other authors, early integration of audio and video features is better than audio-only inversion. We also find that audiovisual fusion is especially useful for articulators that are usually visible (the lips and jaw). It is possible to obtain fairly good inversion results by using a modified late fusion scheme, in which video features contribute more to position estimation of the visible articulators (lips and jaw) and less to position estimation of the tongue and velum.

This paper has also demonstrated a probabilistic smoothing method, based on the Kalman smoother, for the purpose of computing a smoothed articulatory inverse. The Kalman smoother outperforms a simple low-pass smoother by a small margin even without auxiliary information, but the Kalman filters key advantage is that it can incorporate auxiliary information in a natural way. Given information about the phoneme transcription, for example, the Kalman smoother is able to reduce the RMS error of every inversion technique.

This paper also compares the proposed smoothing algorithms with the maximum likelihood trajectory estimation (MLTE) given in Richmond [9] and Toda et al. [10]. There are certain parallels between the proposed dynamic smoothers and the MLTE method. Both of them use past and future position values of the articulators to obtain the smoothed estimate of the current position values, and in this sense both Kalman smoothing and MLTE can be considered dynamic system models, but there is an important difference: the proposed GK and PBK smoothing algorithms are based on an explicit state space In other words, the relation between the current and past state values is explicitly defined by the state equation (14). Measurement noise and process noise are both explicitly represented, and their balance allows the algorithm to explicitly compensate for modeling error. This converts the smoothing process into a stochastic dynamic smoothing process. On the other hand, MLTE can be considered as an implicit dynamic smoothing algorithm. The relation between current and past state values is not defined explicitly, but the addition of velocity (and acceleration) of the state values into the estimation process smooths the state. The velocity (and acceleration) of the state are calculated deterministically and from this point of view MLTE may be considered as a deterministic dynamic smoothing method. It is likely for these reasons that the MLTE method benefits from a further post-processing smoothing stage (e.g., low-pass smoother).

Considering the observation/measurement usage, MLTE uses acoustic (and/or visual) features as observations together with the initially estimated trajectories to calculate the posterior probability of the state in the current time instant, and the smoothing process is conducted in an iterative manner. The proposed smoothing algorithm, on the other hand, uses the initial estimated trajectories (output of the GMM estimator) as observations (in equation (15)). According to these observations and the selected state space model, the smoothed articulatory trajectories are estimated by the Kalman smoother.

From a computational point of view, a direct solution of the problem by the MLTE has heavy computational complexity due to huge matrix inversions and multiplications process (although complexity is limited by the banded structure of the inverted matrices). However, Tokuda et al. [31] proposes an iterative solution for MLTE, in which computational complexity is reduced significantly. On the other hand, In GK and PBK smoothing methods are originally efficient iterative process in time and no such huge matrices in the process. Generalization of both GK and MLTE smoothing into a phoneme-based smoothing form is possible, e.g., Hiroya et al. [6] and Zhang et al. [7] demonstrated an HMM-based MLTE that is comparable in some ways to the PBK smoother introduced here.

The experimental results of the comparison of the proposed smoothing and MLTE methods show that inversion with global Kalman smoother (MS+GK) gives better performance than MLTE. Moreover, smoothing MLTE trajectories via low-pass filter improves the performance of the method; MS+GK has lower RMS error than LP-MLTE, but similar correlation coefficient. The phone-based Kalman (PBK) smoother gives the best performance among all experiments.

By using formant frequencies and energies in acoustic-to-articulatory inversion, Table IV and Fig. 4 demonstrate an RMS error of 1.395 mm and a correlation coefficient of 0.816 without using any phonemic information; to our knowledge, the best results reported in the literature for this task are 1.45 mm, 0.79 in Toda et al. [10] and 1.37 mm in Richmond [30]. By using audiovisual modified fusion, Table IV and Fig. 11 report an RMS error of 1.227 mm; to our knowledge, the best result reported in the literature for this task is 1.38 mm (Katsamanis et al. [13]). By providing a phonetic transcription as auxiliary information to the Kalman smoother, the RMS

error is further reduced to only 1.199 mm, with a correlation coefficient of 0.876; to our knowledge, these are among the best results reported in the literature for this task.

APPENDIX

For the sake of completeness, this appendix offers a brief explanation of the MLTE method used in this work. The MLTE method has two main stages: training a GMM using acoustic and articulatory trajectories with their velocity components, and estimating the articulatory trajectories via the EM algorithm. The following section explains these stages in some detail.

A. Training the GMM

In the MLTE method, first a GMM is trained. The training procedure is similar to the GMM training explained in Sec.III. The velocity (and may be the acceleration) components of the acoustic and articulatory trajectories are also used in the training process in addition to the position. Let $\mathscr{X} and \mathscr{Z}$ be random vectors of the acoustic (and/or visual) and the articulatory spaces respectively and let $\mathbb{x}_k$ and $\mathbb{z}_k$ be their realizations at time instant k, where $\mathbb{x}_k \triangleq \left[x_k^T, \Delta x_k^T\right]^T$ and $\mathbb{z}_k \triangleq \left[z_k^T, \Delta z_k^T\right]^T$. The velocity (delta) components Δz_k and Δx_k are defined as

$$\Delta x_k \triangleq \frac{1}{2}(x_{k+1} - x_{k-1}), \Delta z_k \triangleq \frac{1}{2}(z_{k+1} - z_{k-1}) \qquad (42)$$

Assuming that $\mathscr{X} and \mathscr{Z}$ are jointly distributed according to a GMM, their joint distribution can be written as follows

$$f_{\mathscr{X},\mathscr{Z}}(\mathbb{x}_k, \mathbb{z}_k) \triangleq \sum_{i=1}^{\mathcal{K}} \pi_i \mathcal{N}(\mathbb{x}_k, \mathbb{z}_k; \eta^i, \Psi^i). \qquad (43)$$

In this representation $\mathcal{K}$ is the number of mixture components and π_i is the weight of the ith mixture. The η^i and Ψ^i denote the mean and the covariance of the GMM components and their definitions are similar to the one given in (5) except that$\mathscr{X}$ and $\mathscr{Z}$ are used here instead of $\mathcal{X}$ and $\mathcal{Z}$ of (5). The parameter set of the GMM is defined below. $\Theta_{\mathscr{X}\mathscr{Z}} \triangleq \{\pi_i, \eta^i, \Psi^i\}_{i=1}^K$. These parameters are estimated using the expectation maximization (EM) algorithm.

The conditional distribution $f_{\mathscr{X}|\mathscr{Z}}(\mathbb{x}_k|\mathbb{z}_k)$ can also be written as a Gaussian mixture with same number of mixture components as follows.

$$f_{\mathscr{X}|\mathscr{Z}}(\mathbb{x}_k|\mathbb{z}_k) \triangleq \sum_{i=1}^{\mathcal{K}} \beta^i(\mathbb{z}_k) \mathcal{N}(\mathbb{x}_k; \eta^{i,k}_{\mathscr{X}|\mathscr{Z}}, \Psi^i_{\mathscr{X}|\mathscr{Z}}). \qquad (44)$$

The parameter set of the conditional distribution is $\Theta_{\mathscr{X}|\mathscr{Z}} \triangleq \{\beta^i(\mathbb{z}_k), \eta^{i,k}_{\mathscr{X}|\mathscr{Z}}, \Psi^i_{\mathscr{X}|\mathscr{Z}}\}_{i=1}^K$. Parameters can be computed using formulae similar to the ones given in (7-9). The estimated parameter sets $\Theta_{\mathscr{X}|\mathscr{Z}}$ and $\Theta_{\mathscr{X}\mathscr{Z}}$ are used in the estimation of the articulatory trajectories which is explained in the following section.

B. Trajectory Estimation

Let $\mathbb{Z} \in \mathbb{R}^{2Nn_z}$ and $\mathbb{X} \in \mathbb{R}^{2Nn_x}$ be the time sequences of the acoustic and the articulatory trajectories together with their velocity components respectively, i.e. $\mathbb{X} \triangleq [\mathbb{x}_1^T, \ldots, \mathbb{x}_N^T]^T$, $\mathbb{Z} \triangleq [\mathbb{z}_1^T, \ldots, \mathbb{z}_N^T]^T$. Let $S = [s_1, \ldots, s_N]$ be the sequence of underlying mixture component indicator for each vector. The MLTE method estimates articulatory trajectories using the following criterion

$$\hat{X} = \arg\max_X p(\mathbb{X}|\mathbb{Z}, \Theta_{\mathscr{X}\mathscr{Z}}) \qquad (45)$$

where $X \triangleq [x_1^T, \ldots, x_N^T]^T$ is the time sequence of articulatory trajectories without velocity components. Since mixture label sequence S is also unknown, there is no closed-form solution of the problem given in (45). Therefore an iterative solution, i.e., EM algorithm is required. Starting from an initial guess of the model parameters, $\hat{\mathbb{X}}^{(0)}$, the EM algorithm re-estimates articulatory trajectories $\hat{\mathbb{X}}^{(r)}$ in such a way that the expected complete-data log-likelihood function (auxiliary function) $Q(\mathbb{X}, \hat{\mathbb{X}}^{(r)}) = \mathrm{E}\left\{\ln p(\mathbb{X}, \mathbb{Z}, S)|\hat{\mathbb{X}}^{(r)}, \mathbb{Z}\right\}$ is maximized. The resulting likelihood function is guaranteed to be non-decreasing at each iteration. The auxiliary function $Q(\mathbb{X}, \hat{\mathbb{X}}^r)$ can be written as

$$\begin{aligned} Q(\mathbb{X}, \hat{\mathbb{X}}^r) &= \mathrm{E}\left\{\ln p(\mathbb{X}, \mathbb{Z}, S)|\hat{\mathbb{X}}^r, \mathbb{Z}\right\} \\ &= \sum_{k=1}^{N}\sum_{i=1}^{K} P(s_k = i|\hat{\mathbb{x}}_k^r, \mathbb{z}_k) \ln p(\mathbb{x}_k|s_k = i, \mathbb{z}_k) + C \end{aligned}$$

where, $p(\mathbb{x}_k|s_k = i, \mathbb{z}_k) \sim \mathcal{N}(\mathbb{x}_k; \eta^{i,k}_{\mathscr{X}|\mathscr{Z}}, \Psi^i_{\mathscr{X}|\mathscr{Z}})$ and C is the constant with respect to x_k. The EM algorithm has two alternating phases. In one phase the expected value of the complete-data log-likelihood is calculated where the expectation is taken over the hidden variables S. In the second phase the auxiliary function is maximized for articulatory trajectories. The summary of the EM algorithm is as follows.

EM Algorithm for rth iteration:

- Estep: Calculate the posterior probability of each mixture component

$$\begin{aligned} \xi_i(\hat{\mathbb{x}}_k^r, \mathbb{z}_k) &\triangleq P(s_k = i|\hat{\mathbb{x}}_k^r, \mathbb{z}_k) \\ &= \frac{\pi_i \mathcal{N}(\hat{\mathbb{x}}_k^r, \mathbb{z}_k; \eta^i, \Psi^i)}{\sum_{i=1}^{\mathcal{K}} \pi_i \mathcal{N}(\hat{\mathbb{x}}_k^r, \mathbb{z}_k; \eta^i, \Psi^i)} \end{aligned} \qquad (46)$$

- Mstep: Calculate the articulatory trajectories

$$\begin{aligned} \hat{X}^{(r+1)} &= \arg\max_X Q(\mathbb{X}, \hat{\mathbb{X}}^r) \\ &= \arg\max_X \left\{\sum_{k=1}^{N}\sum_{i=1}^{\mathcal{K}} \xi_i(\hat{\mathbb{x}}_k^r, \mathbb{z}_k) \ln p(\mathbb{x}_k|s_k = i, \mathbb{z}_k)\right\} \\ &= \arg\max_X \left\{\sum_{k=1}^{N} -\frac{1}{2}\mathbb{x}_k^T \bar{\Psi}_k \mathbb{x}_k + \mathbb{x}_k^T \bar{\mathbb{d}}_k\right\} \\ &= \arg\max_X \left\{-\frac{1}{2}\mathbb{X}^T \bar{\Psi} \mathbb{X} + \mathbb{X}^T \bar{\mathbb{D}}\right\} \\ &= \left(\mathbb{W}^T \bar{\Psi} \mathbb{W}\right)^{-1} \mathbb{W}^T \bar{\mathbb{D}} \end{aligned} \qquad (47)$$

where

$$\bar{\Psi} \triangleq \text{blkdiag}\left(\bar{\Psi}_1, \ldots, \bar{\Psi}_N\right), \bar{\Psi}_k \triangleq \sum_{i=1}^{\mathcal{K}} \xi_i(\hat{\mathbb{x}}_k^r, \mathbb{z}_k) \left(\Psi^i_{\mathscr{X}|\mathscr{Z}}\right)^{-1}$$

$$\bar{\mathbb{D}} \triangleq [\bar{\mathbb{d}}_1^T, \ldots, \bar{\mathbb{d}}_N^T]^T, \bar{\mathbb{d}}_k \triangleq \sum_{i=1}^{\mathcal{K}} \xi_i(\hat{\mathbb{x}}_k^r, \mathbb{z}_k) \left(\Psi^i_{\mathscr{X}|\mathscr{Z}}\right)^{-1} \eta^{i,k}_{\mathcal{X}|\mathcal{Z}}$$

The relation between $\mathbb{X}$ and X can be written as $\mathbb{X} = \mathbb{W}X$, where $\mathbb{W}$ is a transformation matrix whose entries are the coefficients used in the computation of the delta coefficients in (42). The structure of the transformation matrix can be seen in Toda et al. [10]. The iteration is performed until a certain stopping condition is satisfied. As mentioned above, the MLTE method needs an initial estimation of articulatory trajectories $\hat{\mathbb{X}}^{(0)}$. The initial estimates are taken from MMSE-based estimation results, as explained in Sec.III

ACKNOWLEDGMENTS

We would like to thank the Scientific and Technological Research Council of Turkey (TUBITAK) for its financial support. We are also grateful to Dr. Umut Orguner (LIU in Sweden) for helpful discussions of and support for this work.

REFERENCES

[1] G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton and Co., 1960.

[2] K. Markov, J. Dang, and S. Nakamura, "Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework," *Speech Commun.*, vol. 48, pp. 161–175, 2006.

[3] T. Stephenson, H. Bourlard, S. Bengio, and A. Morris, "Automatic speech recognition using dynamic Bayesian networks with both acoustic and articulatory variables," in *Proc. ICSLP*, 2000, pp. 951–954.

[4] J. S. Perkell, M. H. Cohen, M. A. Svirsky, and M. L. Matthies, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," *J. Acoust. Soc. Am.*, vol. 92, no. 6, pp. 3078–3096, 1992.

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

[6] 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, March 2004.

[7] L. Zhang and S. Renals, "Acoustic-articulatory modelling with the trajectory HMM," *IEEE Signal Processing Letters*, vol. 15, pp. 245–248, 2008.

[8] K. Richmond, "Estimating articulatory parameters from the speech signal," Ph.D. dissertation, The Center for Speech Technology Research, Edinburgh, UK, 2002.

[9] ——, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *Proc. Intrespeech*, 2006.

[10] 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.

[11] C. Qin and M. Á. Carreira-Perpiñán, "A comparison of acoustic features for articulatory inversion," in *Proc. Intrespeech*, 2007.

[12] A. Katsamanis, T. Roussos, G. Papandreou, and P. Maragos. Computer vision, speech communication & signal processing group. Downloaded data: AAM-based visual features for the female speaker fsew0 of the MOCHA database. [Online]. Available: http://cvsp.cs.ntua.gr/research/inversion/

[13] A. Katsamanis, G. Papandreou, and P. Maragos, "Face active appearance modeling and speech acoustic information to recover articulation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 3, pp. 411–422, 2009.

[14] M. Schroeder, "Determination of the geometry of the human vocal tract by acoustic measurements," *J. Acoust. Soc. Am.*, vol. 41, no. 4, pp. 1002–1010, 1967.

[15] P. Mermelstein, "Determination of the vocal-tract shape from measured formant frequencies," *J. Acoust. Soc. Am.*, vol. 41, no. 5, pp. 1283–1294, 1967.

[16] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. Audio Electroacoustics*, vol. 21, no. 5, pp. 417–427, 1973.

[17] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *J. Acoust. Soc. Am.*, vol. 63, no. 5, pp. 1535–1555, May 1978.

[18] İ. Y. Özbek, M. Hasegawa-Johnson, and M. Demirekler, "Formant trajectories for acoustic-to-articulatory inversion," in *Proc. Intrespeech*, 2009.

[19] M. Hasegawa-Johnson, "Line spectral frequencies are the poles and zeros of a discrete matched-impedance vocal tract model," *J. Acoust. Soc. Am.*, vol. 108, no. 1, pp. 457–460, 2000.

[20] A. M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communication Systems*. New York, NY: John Wiley and Sons, 1995.

[21] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Am.*, vol. 57, p. 535, 1975.

[22] E. Özkan, İ. Y. Özbek, and M. Demirekler, "Dynamic speech spectrum representation and tracking variable number of vocal tract resonance frequencies with time varying Dirichlet process mixture models." *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, pp. 1518–1532, November 2009.

[23] İ. Y. Özbek and M. Demirekler, "Vocal tract resonances tracking based on voiced and unvoiced speech classification using dynamic programming and fixed interval Kalman smoother," in *Proc. ICASSP*, 2008.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc*, 1977.

[25] A. Toutios and K. Margaritis, "Contribution to statistical acoustic-to-EMA mapping," in *Proc. EUSIPCO*, 2008.

[26] V. Mitra, İ. Y. Özbek, H. Nam, X. Zhou, and C. Espy-Wilson, "From acoustic to vocal tract time functions," in *Proc. ICASSP*, 2009.

[27] L. Deng, *Dynamic Speech Models: Theory, Algorithms, and Applications*. Morgan & Claypool Publishers, 2006.

[28] D. Simon, *Optimal State Estimation: Kalman, $\mathcal{H}_\infty$, and Nonlinear Approaches*. Wiley, 2006.

[29] H. Kjellström and O. Engwall, "Audiovisual-to-articulatory inversion," *Speech Commun.*, vol. 51, no. 3, pp. 195–209, 2009.

[30] K. Richmond, "Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion," in *Proc. NOLISP*, ser. Lecture Notes in Computer Science, vol. 4885. Springer-Verlag Berlin Heidelberg, Dec. 2007, pp. 263–272.

[31] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP*, 1995, pp. 660–663.