Analysis of Head Gesture and Prosody Patterns for Prosody-Driven Head-Gesture Animation

M. E. Sargın, *Student Member, IEEE,* Y. Yemez, E. Erzin, *Senior Member, IEEE,* and A. M. Tekalp, *Fellow, IEEE*

***Abstract*— We propose a new two-stage framework for joint analysis of head gesture and speech prosody patterns of a speaker towards automatic realistic synthesis of head gestures from speech prosody. In the first stage analysis, we perform Hidden Markov Model (HMM) based unsupervised temporal segmentation of head gesture and speech prosody features separately to determine elementary head gesture and speech prosody patterns, respectively, for a particular speaker. In the second stage, joint analysis of correlations between these elementary head gesture and prosody patterns is performed using Multi-Stream HMMs to determine an audio-visual mapping model. The resulting audio-visual mapping model is then employed to synthesize natural head gestures from arbitrary input test speech given a head model for the speaker. In the synthesis stage, the audio-visual mapping model is used to predict a sequence of gesture patterns from the prosody pattern sequence computed for the input test speech. The Euler angles associated with each gesture pattern are then applied to animate the speaker head model. Objective and subjective evaluations indicate that the proposed synthesis by analysis scheme provides natural looking head gestures for the speaker with any input test speech, as well as in "prosody transplant" and "gesture transplant" scenarios.**

***Index Terms*— Multimedia computing, speech analysis, video signal processing, animation**

I. Introduction

State of the art visual speaker animation methods are capable of generating synchronized lip movements automatically from speech content; however, they lack automatic synthesis of speaker gestures from speech. Head and face gestures are usually added manually by artists, which is costly and often look unrealistic. Hence, learning the correlation between gesture and speech patterns of a speaker towards automatic realistic synthesis of speaker gestures from speech remains as a challenging research problem.

There exists significant literature on speaker lip animation, that is, rendering lip movements synchronized with the speech signal [1]. Since lip movement is physiologically tightly coupled with acoustic speech, it is relatively an easy task to find a mapping between the phonemes of speech and the visemes of lip movement. Many schemes exist to find such audio-to-visual mappings among which the HMM (Hidden Markov Model)-based techniques are the most common as they yield smooth animations exploiting temporal dynamics of speech [2]–[9]. Some of these works also incorporate synthesis of facial expressions along with the lip movements to make animated faces look more natural [3], [6], [8], [9]. The common strategy in these techniques is to train a joint HMM structure with extracted visual and audio feature vectors and then to use the trained HMM structure to generate speech-driven facial expressions and lip movements.

Despite exhibiting variations from person to person and in time, head and body gestures are also correlated with speech. For example, it has been observed that manual gestures are correlated with prosody [10], [11] and verbal content of the speech [12], whereas head gestures are mostly correlated with the prosody [11], [13], [14]. Although correlations between speech and head/body gestures have been investigated in several works, there are only a limited number of publications addressing speech-driven head and body gesture synthesis. In [15], audio streams from training videos are first segmented using pitch contour information. The same boundaries are also applied to the corresponding video streams for segmenting head motions. The co-occurring audio and head motion segments are stored as pairs in a database. Later, a new test audio stream is segmented, and an optimal head gesture sequence is determined from the database using dynamic programming to create synthetic head motions. A similar methodology is followed in [16], where audio/head motion feature pairs extracted from training videos are stored into a database indexed by audio features. Later, audio features extracted from a new test input speech are used to search for K-nearest neighbors. The optimum nearest neighbor combination, found by dynamic programming, is used to synthesize corresponding head motions. In [17], we presented a preliminary demonstration of natural looking head and arm gesture synthesis from speech using a manually determined audio-visual mapping from speech to head and arm motions.

The aim of this paper is to present a framework for joint analysis of head gesture and speech prosody patterns towards automatic generation of the audio-visual mapping from speech prosody to head gestures. Although the same framework can also be applied to analysis of co-occurring arm gesture and speech patterns, this is beyond the scope of the current paper. There are some open challenges involved in the joint analysis of head gestures and prosody towards prosody-driven head gesture synthesis: First, unlike phonemes and visemes in speech articulation, there does not exist a well-established set of elementary prosody and gesture patterns for gesture synthesis. Second, synchronicity of gesture and prosody patterns may exhibit variations. For instance, a speaker can move her/his head before the corresponding prosodic utterance with a variable time lag. Moreover, gestural patterns may span time intervals of different duration with respect to their prosodic counterparts. Third, prosody and gesture patterns are speaker dependent, and may exhibit variations in time even for the same speaker. Previously reported works [15]–[17] do not address any of these challenges; for instance, the asynchrony problem is either ignored or handled by manual alignment. In this work, we address these challenges by first processing the head gesture and prosody features separately by a parallel HMM structure to learn and model the gestural and prosodic elements (elementary patterns), respectively, over training data for a particular speaker. We then employ a multi-stream parallel HMM structure to find the jointly recurring gesture-prosody patterns and the corresponding audio-to-visual mapping.

HMM-based segmentation techniques are commonly employed in modeling multi-stream correlations; for example, for speech-driven lip animation in [7]–[9] and for audio-visual event detection in [18]. We can classify HMM based modeling techniques as supervised and unsupervised. Speech and lip motion correlation modeling can be thought of as a supervised analysis/segmentation problem, since phonemes and visemes constitute well-established elementary units for these modalities. Hence, speech-driven lip animation task is often equivalent to find a mapping between the phonemes of speech and the

M. E. Sargın, Y. Yemez, E. Erzin and A. M. Tekalp are with the College of Engineering, Koç University, Istanbul, Turkey. (e-mail: msargin@ku.edu.tr; yyemez@ku.edu.tr; eerzin@ku.edu.tr; mtekalp@ku.edu.tr).

This work has been supported by the European FP6 Network of Excellence SIMILAR (http://www.similar.cc).

visemes of lip movement. On the other hand, we shall consider the audio-visual gesture modeling/mapping as an unsupervised segmentation problem, where the recurrent joint events are not well defined and to be extracted from the joint feature streams.

The organization of this paper is as follows: In Section II, we first provide an overview of the proposed HMM-based analysis-synthesis framework, and then describe the computation of head gesture and speech prosody features. Robust and accurate tracking of the speaker head motions is an integral part of the overall system; hence, it is described in detail. Section III presents the proposed two-stage unsupervised analysis procedure to identify and model jointly recurring head gesture and prosody patterns. Section IV explains HMM-based synthesis of head gesture parameters from input test speech. In Section V, we describe the experiments conducted, and present objective and subjective evaluation of the prosody-driven head gesture synthesis results. Finally, Section VI provides conclusions.

II. Overview of the Proposed System and Feature Extraction

A block diagram of the proposed system for prosody-driven head gesture animation, which consists of analysis and synthesis parts, is depicted in Fig. 1. The analysis part includes two feature extraction modules and two-stages of analysis. Feature extraction modules compute the head gesture features $\boldsymbol{f}^g$ and speech prosody features $\boldsymbol{f}^p$, respectively, from training stereo video sequences of a speaker. At the first stage analysis, individual feature streams are used to train separate parallel HMM structures, which provide probabilistic models for temporal recurrent patterns in the corresponding modalities, respectively. The segments corresponding to these patterns are detected and labeled over the training video streams, where pattern labels for prosody and gesture are denoted by $\boldsymbol{l}^p$ and $\boldsymbol{l}^g$, respectively. At the second stage, the labels of temporally segmented gesture and prosody streams are used together to train a discrete multi-stream parallel HMM to identify jointly recurring patterns. The resulting joint HMM structure models the correlation between speech prosody and head gestures. The synthesis part makes use of the joint HMM to predict the gesture labels from the prosody labels computed for a test input speech using the prosody HMM obtained by the first stage analysis. The corresponding gesture features, i.e., head motion parameters, are synthesized using the gesture HMM obtained at the first stage analysis and finally animated on a 3D head model. The details of the two stages of the analysis, shown by Stage-I and Stage-II blocks in Fig. 1, are presented in Section III, whereas the gesture synthesis part is described in detail in Section IV. In the remainder of this section, we describe our methodology for extraction of head gesture and speech prosody features.

A. Extraction of Head Gesture Features

We define the head gesture feature vector, $\boldsymbol{f}_k^g$, for frame k to include the Euler angles associated with the 3D head rotation and their first differences,

$$\boldsymbol{f}_k^g = [\theta_k, \phi_k, \psi_k, \Delta\theta_k, \Delta\phi_k, \Delta\psi_k]^T \tag{1}$$

where θ_k, ϕ_k and ψ_k are the Euler angles of rotation, with respect to a reference frame k_r, around the x, y and z axes, respectively, and $\Delta\theta_k$, $\Delta\phi_k$, $\Delta\psi_k$ denote their respective first differences. The reference frame k_r can be selected as the first frame in which the subject's head is assumed to be at neutral position.

1) 3D Point Tracking: For video recording, we use a rectified stereo camera system with two identical cameras, and assume that the intrinsic camera parameters are known *a priori*. For each frame k, we initially detect a rectangular head region from one of the stereo views (e.g., the right or the left but not both) using a boosted Haar based cascade classifier structure, which was initially proposed by Viola and Jones [19] and later improved by Lienhart and Maydt [20]. The detected rectangular head region is used to initialize the search window within which facial pixels are segmented based on a Gaussian skin color distribution model computed over a training set of sampled skin colors. An ellipse $\mathcal{E}_k$ is then fitted to the facial skin region.

Let P_{k_r} denote the set of image points within the ellipse $\mathcal{E}_{k_r}$ of the reference frame k_r so that $P_{k_r} = \{\boldsymbol{p}_{k_r,1}, \boldsymbol{p}_{k_r,2}, \ldots, \boldsymbol{p}_{k_r,N}\}$ and $\boldsymbol{p}_{k_r,n} = [x_n, y_n]^T$. For each frame k, we employ the hierarchical Lukas-Kanade technique [21] to find the optical flow vectors, $\{\boldsymbol{v}_{k,1}, \boldsymbol{v}_{k,2}, \ldots, \boldsymbol{v}_{k,N}\}$, from frame k_r to frame k. The set P_k of the corresponding image points in frame k is then obtained by $\boldsymbol{p}_{k,n} = \boldsymbol{p}_{k_r,n} + \boldsymbol{v}_{k,n}$, $n = 0, 1, ..., N$.

In order to find the 3D world coordinates of the image points in each set P_k, we compute the disparity vectors at these points using bandpass images and a cross correlation measure based on the sum of absolute differences [22]. The disparity vectors are also validated using several criteria [23]. Given the disparity vectors for each frame and the intrinsic parameters of the rectified stereo camera system, the 3D world coordinates of the 2D points from both sets P_{k_r} and P_k are calculated by the well-known triangulation technique. Let $\boldsymbol{W}_k$ denote the $3 \times M$ matrix formed by the 3D world coordinates of the points associated with P_k, so that $\boldsymbol{W}_k = [\boldsymbol{w}_{k,1}, \boldsymbol{w}_{k,2}, \ldots, \boldsymbol{w}_{k,M}]$ and $\boldsymbol{w}_{k,m} = [X_m, Y_m, Z_m]^T$. While forming the matrix $\boldsymbol{W}_k$, we exclude those points in P_k that fall outside the ellipse $\mathcal{E}_k$ due to possible erroneous optical flow vectors. The excluded points are outliers which may corrupt the 3D motion capture process. Hence the dimension M of the matrices $\boldsymbol{W}_k$ and $\boldsymbol{W}_{k_r}$ are re-determined at each frame k according to the number of points that fall within the detected ellipse $\mathcal{E}_k$.

2) Computation of the Euler Angles: Let $\boldsymbol{R}_k$ and $\boldsymbol{t}_k$ denote the rotation matrix and the translation vector, respectively, of the rigid head motion from frame k_r to k. Then, $\boldsymbol{W}_k$ and $\boldsymbol{W}_{k_r}$ are related by

$$\boldsymbol{W}_k = \begin{bmatrix} \boldsymbol{R}_k & \boldsymbol{t}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{W}_{k_r} \\ \boldsymbol{1}^T \end{bmatrix}. \tag{2}$$

The rotation matrix $\boldsymbol{R}_k$ and translation vector $\boldsymbol{t}_k$ are estimated by a unitary constraint optimization technique as explained in the Appendix. Once estimated, the rotation matrix $\boldsymbol{R}_k$ can be decomposed into three matrices:

$$\boldsymbol{R}_k = [r_{ij}^k] = \boldsymbol{R}_x(\theta_k)\boldsymbol{R}_y(\phi_k)\boldsymbol{R}_z(\psi_k) \tag{3}$$

where $\boldsymbol{R}_x(\theta_k)$, $\boldsymbol{R}_y(\phi_k)$ and $\boldsymbol{R}_z(\psi_k)$ are the matrices that specify rotations around x, y and z axes, respectively [24], [25]. The Euler angle vector $\boldsymbol{e}_k = [\theta_k, \phi_k, \psi_k]^T$ which maps $\boldsymbol{W}_{k_r}$ to $\boldsymbol{W}_k$, is finally extracted from this decomposition by

$$\boldsymbol{e}_k = \left[\arctan(-r_{23}^k/r_{33}^k),\ \arcsin(r_{13}^k),\ \arctan(-r_{12}^k/r_{11}^k)\right]^T. \tag{4}$$

In cases where the head rotation between the current frame k and reference frame k_r is larger than a threshold angle (e.g., if $|\theta_k| > 25^\circ$ or $|\phi_k| > 25^\circ$ or $|\psi_k| > 25^\circ$), the optical flow vectors, hence the 3D point correspondences between two frames, may become unreliable. In such cases, we switch to incremental motion estimation, where the reference frame for frame k is set to frame $k-1$. Thus, we recompute optical flow vectors with respect to frame $k-1$; hence, the new 3D point correspondences and the resulting incremental Euler angle vector $\boldsymbol{\delta}_{k-1}$, which defines the rotation between frames k and $k-1$ are computed. Then, the Euler angle vector with respect to the reference frame k_r is given by

$$\boldsymbol{e}_k = \boldsymbol{e}_{k-1} + \boldsymbol{\delta}_{k-1} \tag{5}$$

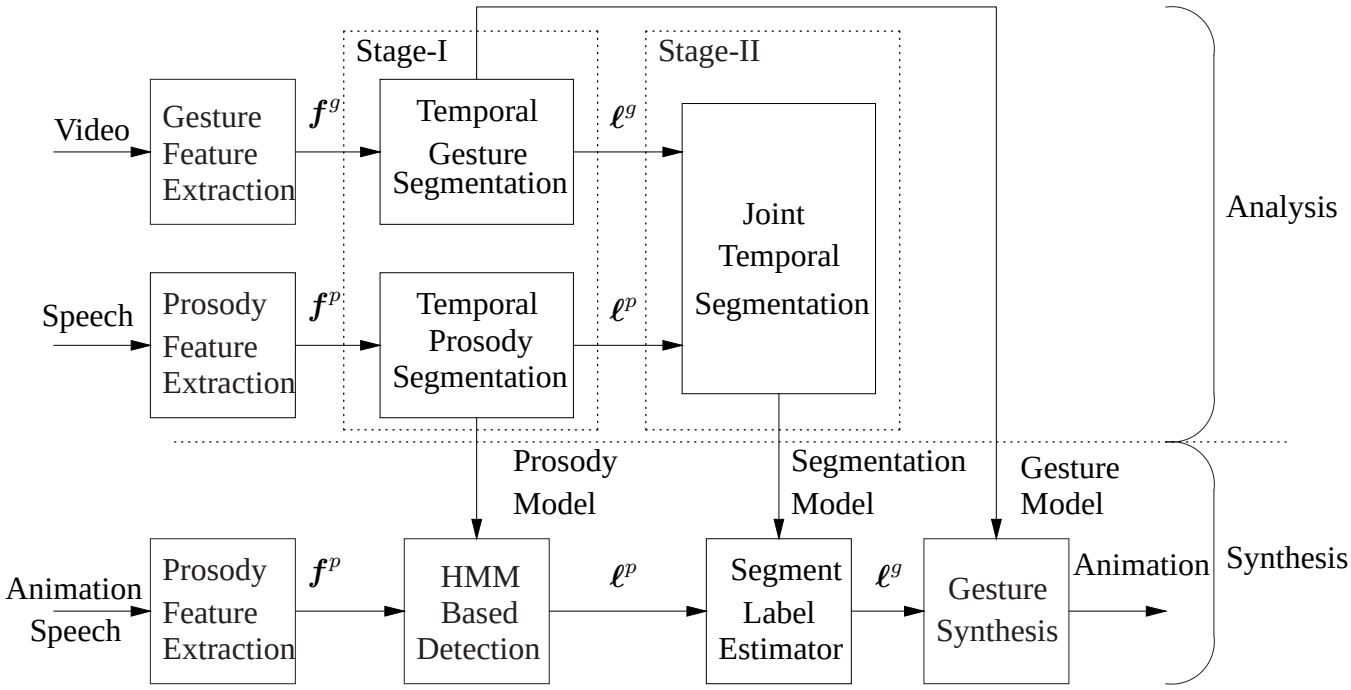

Fig. 1. Overview of the proposed synthesis-by-analysis system.

3) Smoothing of the Feature Vector by Kalman Filtering: We finally employ a Kalman filter for post smoothing of the computed (estimated) Euler angles, which are input as observations $\boldsymbol{z}_k$ to the Kalman filter. The measurement noise $\boldsymbol{r}_k$ models the estimation errors in the Euler angles. The head gesture feature vector, $\boldsymbol{f}_k$ (the superscript g is omitted for ease of notation), consisting of the Euler angles and their first differences, is selected as the state vector. The state-space representation of the Kalman filter is given by

$$\begin{aligned} \boldsymbol{f}_{k+1} &= \boldsymbol{F}\boldsymbol{f}_k + \boldsymbol{G}\boldsymbol{u}_k \\ \boldsymbol{z}_k &= \boldsymbol{H}\boldsymbol{f}_k + \boldsymbol{r}_k \end{aligned} \tag{6}$$

where

$$\boldsymbol{f}_k = \begin{bmatrix} \boldsymbol{e}_k \\ \Delta\boldsymbol{e}_k \end{bmatrix}, \quad \boldsymbol{F} = \begin{bmatrix} \boldsymbol{I}_{3\times3} & \boldsymbol{I}_{3\times3} \\ \boldsymbol{0}_{3\times3} & \boldsymbol{I}_{3\times3} \end{bmatrix} \\ \boldsymbol{G} = \boldsymbol{I}_{6\times6}, \quad \boldsymbol{H} = \begin{bmatrix} \boldsymbol{I}_{3\times3} \\ \boldsymbol{0}_{3\times3} \end{bmatrix} \tag{7}$$

The 3×1 vector $\Delta\boldsymbol{e}_k$ denotes the first differences of the Euler angles. The model noise $\boldsymbol{u}_k$ and measurement noise $\boldsymbol{r}_k$ are assumed to be uncorrelated, zero-mean white Gaussian processes. The output of the Kalman filter gives the final feature vector for the head gestures.

B. Extraction of Prosody Features

The prosodic speech events can be described by the temporal variations of loudness/intensity and pitch as well as pauses between phrases, phoneme durations, timing, and rhythm. Among these, the most expressive one is the pitch, which is the rate of vocal-fold cycling. In this study, pitch frequency, V, and speech intensity, I, are considered as prosody features.

The pitch contour is extracted at a rate of 100 Hz from the speech signal using the autocorrelation method as described in [26]. The mean of all pitch contours over all active utterances is removed to emphasize local variations [27], and then the resulting mean-removed contours are low-pass filtered to reduce discontinuities. The regions between utterances without a valid pitch are filled with zero mean unit variance Gaussian noise. The intensity features are also extracted over the active utterances. The squared sound intensities are weighted with a 32 ms Kaiser-20 window, and the speech signal intensity is calculated as the sum of these weighted samples. The 32 ms window is shifted by 10 ms for each frame to extract intensity values at 100 Hz frame rate. The intensity features are also mean removed over active utterances and between-utterance regions are filled with zero mean unit variance Gaussian noise. The first order derivative, ΔV_k, of the post-processed pitch frequency at frame k is calculated using the following regression formula:

$$\Delta V_k = \frac{\sum_{i=1}^{2} i(V_{k+i} - V_{k-i})}{2\sum_{i=1}^{2} i^2}. \tag{8}$$

Finally, the pitch frequency, its derivative and the intensity are concatenated to form the 3 dimensional prosody feature vector $\boldsymbol{f}_k^p$ at frame k:

$$\boldsymbol{f}_k^p = [V_k \ \Delta V_k \ I_k]^T \tag{9}$$

III. Head Gesture-Prosody Pattern Analysis

In this section, we propose a two stage HMM-based unsupervised analysis framework, where the first stage aims to separately extract elementary gesture and prosody patterns for a speaker, and the second stage determines a correlation model between these head gesture and prosody patterns. In the first stage analysis, recurring elementary gesture and prosody patterns are determined separately by unsupervised temporal clustering of individual gesture and prosody feature streams, respectively. The extracted elementary prosody and gesture patterns are analogous to phonemes and visemes in the speech and lip motion modeling. However, the elementary gesture and prosody patterns are not well established as in the case of phonemes and visemes, since the nature and strength of head gesture and prosody patterns may vary from person to person and in time. Hence, the need for unsupervised stage I analysis in order to extract these patterns for each speaker. Furthermore, the joint recurring nature of these patterns are also not well established as in the case of phoneme-viseme association; hence, the need for stage II analysis

for joint modeling of correlations between head gesture and prosody patterns. In order to find a mapping between prosody and gesture patterns, unsupervised temporal segmentation of joint gesture and prosody pattern labels is performed, which defines the correlation between gesture and prosody pattern streams and relates co-occurring head gesture and prosody patterns.

We note that if a multi-stream HMM structure were directly employed for joint analysis of gesture and prosody feature streams, as commonly used for event detection [18], instead of the proposed two-stage analysis, the resulting joint gesture-prosody feature segments would not necessarily correspond to *independent* meaningful elementary gesture and prosody patterns. As a result, the synthesized gesture sequence might contain poorly defined gestural elements, which would degrade the quality of prosody-driven head gesture animation.

A. Stage-I: Extraction of Elementary Head Gesture and Prosody Patterns

The first stage analysis defines recurrent elementary head gesture and prosody patterns separately using unsupervised temporal clustering over individual feature streams. The gesture and prosody feature streams $\boldsymbol{F}^g$ and $\boldsymbol{F}^p$ are separately used to train two HMM structures Λ_g and Λ_p, which capture recurrent head gesture segments $\boldsymbol{\varepsilon}^g$ and prosody segments $\boldsymbol{\varepsilon}^p$. For ease of notation, we use a generic notation to represent the HMM structure which is identical for the gesture and prosody streams. The HMM structure Λ, which is used for unsupervised temporal segmentation, has M parallel branches and N states as shown in Fig. 2. In the HMM structure Λ, observation probability densities are modeled by a single Gaussian with diagonal covariance for both gesture and prosody streams. The states labeled as s_s and s_e are non emitting start and end states of the parallel HMM structure. Fig. 2 clearly illustrates that the parallel HMM Λ is composed of M parallel left-to-right HMMs, $\{\lambda_1, \lambda_2, \ldots, \lambda_M\}$, where each λ_m is composed of N states, $\{s_{m,1}, s_{m,2}, \ldots, s_{m,N}\}$. The state transition matrix A_{λ_m} of each λ_m is associated with a sub-diagonal matrix of A_Λ. The feature stream is a sequence of feature vectors, $\boldsymbol{F} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_T\}$, where $\boldsymbol{f}_t$ denotes the feature vector at frame t. Unsupervised temporal segmentation using HMM model Λ yields L number of segments $\boldsymbol{\varepsilon} = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L\}$. The l-th temporal segment is associated with the following sequence of feature vectors,

$$\varepsilon_l = \{\boldsymbol{f}_{t_l}, \boldsymbol{f}_{t_l+1}, \ldots, \boldsymbol{f}_{t_{l+1}-1}\} \quad l = 1, 2, \ldots, L \tag{10}$$

where $\boldsymbol{f}_{t_1}$ is the first feature vector $\boldsymbol{f}_1$ and $\boldsymbol{f}_{t_{L+1}-1}$ is the last feature vector $\boldsymbol{f}_T$.

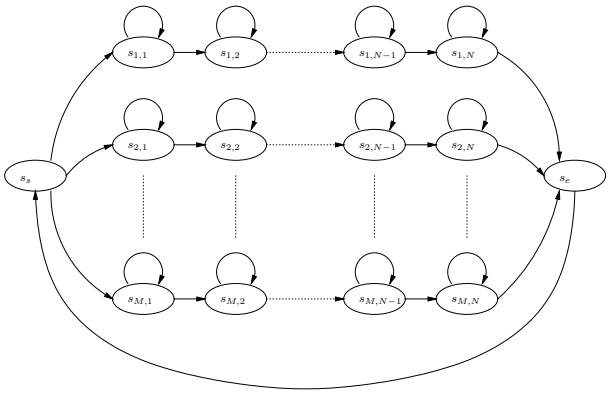

Fig. 2. Parallel HMM structure

The segmentation of the feature stream is performed using Viterbi decoding to maximize the probability of model match, which is the probability of feature sequence $\boldsymbol{F}$ given the trained parallel HMM Λ,

$$\begin{aligned} \mathrm{P}(\boldsymbol{F}|\Lambda) &= \max_{t_l, m_l} \prod_{l=1}^{L} \mathrm{P}(\{\boldsymbol{f}_{t_l}, \boldsymbol{f}_{t_l+1}, \ldots, \boldsymbol{f}_{t_{l+1}-1}\}|\lambda_{m_l}) \\ &= \max_{\varepsilon_l, m_l} \prod_{l=1}^{L} \mathrm{P}(\varepsilon_l|\lambda_{m_l}) \end{aligned} \tag{11}$$

where ε_l is the l-th temporal segment, which is modeled by the m_l-th branch of the parallel HMM Λ. One can show that λ_{m_l} is the best match for the feature sequence ε_l, that is,

$$m_l = \underset{m}{\operatorname{argmax}}\, \mathrm{P}(\varepsilon_l|\lambda_m) \tag{12}$$

Since, the temporal segment ε_l from frame t_l to $(t_{l+1} - 1)$ is associated with segment label m_l, we define the sequence of frame labels based on this association as,

$$\ell_t = m_l \qquad \text{for } t = t_l, t_l + 1, \ldots, t_{l+1} - 1 \tag{13}$$

where ℓ_t is the label of the t-th frame and we have a label sequence $\boldsymbol{\ell} = \{\ell_1, \ell_2, \ldots, \ell_T\}$ corresponding to the feature sequence $\boldsymbol{F}$. The first stage analysis extracts the frame label sequences $\boldsymbol{\ell}^g$ and $\boldsymbol{\ell}^p$ given the head gesture and prosody feature streams $\boldsymbol{F}^g$ and $\boldsymbol{F}^p$. While mapping the gesture and prosody features to discrete frame labels, the mismatch between the frame rates of gesture and prosody is eliminated by downsampling the frame rate of prosody label stream to the rate of gesture label stream.

The parallel HMM structure has two important parameters to set before training the model Λ. The first parameter is the number of states in each branch, N. It should be selected by considering the minimum duration of temporal patterns. Selecting a small N may hamper modeling long term statistics for each branch of the parallel HMM. The extreme case $N = 1$ reduces to K-Means unsupervised clustering. We select the number of states in each branch of the head gesture HMM Λ_g as $N_{\Lambda_g} = 10$, corresponding to the minimum gesture pattern duration of 10 frames ($\frac{1}{3}$ sec assuming 30 video frames/sec). Note that the gesture patterns can be longer than 10 frames since the HMM structure allows self-state transitions. On the other hand, the prosody patterns are expected to follow smooth pitch frequency movements over several syllables. Considering the average syllable durations and smoothness of the pitch contours, we set $N_{\Lambda_p} = 5$ in each branch of the prosody HMM model Λ_p.

The second parameter is the number of temporal patterns, M. Since the number of head gesture and prosody patterns is speaker dependent, we propose selection of M by using two fitness measures. The first fitness measure α, which is inversely related to in-class variance, is defined as the frame average of the log-probability of model match,

$$\alpha = \frac{1}{T} \log(\mathrm{P}(\boldsymbol{F}|\Lambda)) \tag{14}$$

The α measure is expected to saturate with increasing number of parallel branches in Λ, since the training database is expected to contain limited number of temporal patterns. However, small variations within temporal patterns are also expected, hence the number of branches M can be more than the actual number of temporal patterns in the training corpus. Consequently, the second fitness measure, which is the average statistical separation between two similar temporal patterns, increases with the decreasing number of temporal patterns. The second fitness measure β is considered as the average statistical separation between two similar temporal patterns, and it is defined as

$$\beta = \frac{1}{T} \sum_{l=1}^{L} \log\left(\frac{\mathrm{P}(\varepsilon_l|\lambda_{m_l})}{\mathrm{P}(\varepsilon_l|\lambda_{m_l^*})}\right), \tag{15}$$

where $\lambda_{m_l^*}$ is the second best match for the temporal segment ε_l, that is,

$$m_l^* = \underset{\forall m \neq m_l}{\operatorname{argmax}} \, \mathrm{P}(\varepsilon_l|\lambda_m) \tag{16}$$

In general, the α measure increases with the number of patterns M, while β measure decreases. Hence, a good value for M can be selected such that β is high enough, while α reaches a certain value.

B. Stage-II: Joint Modeling of Prosody-Gesture Patterns

In the second stage, unsupervised segmentation of the joint gesture-prosody label stream is performed to detect recurrent joint label patterns. Note that this task is similar to the task of stage I, except in the second stage we have a multi-stream discrete observation sequence. For this task, the parallel HMM structure in Fig. 2 is used with discrete multi-stream HMM branches. In multi-stream HMMs, all streams share the same state transition structure however emission probabilities are determined independently for each stream.

The joint gesture-prosody frame label stream, denoted by $\boldsymbol{\ell}^{gp}$, is defined such that for every frame k, $\ell_k^{gp} = [\ell_k^g, \ \ell_k^p]^T$. We represent the discrete multi-stream parallel HMM structure by Γ_{gp} and its m-th branch by γ_m^{gp}. The discrete HMM Γ_{gp} is trained over the joint gesture-prosody label stream. Each branch γ_m^{gp}, associated with a joint gesture-prosody temporal label pattern, is then described by

$$\gamma_m^{gp} = (\boldsymbol{A}_{\gamma_m^{gp}}, [\boldsymbol{B}_{\gamma_m^g} \boldsymbol{B}_{\gamma_m^p}], \boldsymbol{\Pi}_{\gamma_m^{gp}}) \tag{17}$$

where $\boldsymbol{A}_{\gamma_m^{gp}}$ denotes a state transition matrix, $\boldsymbol{B}_{\gamma_m^g}$ and $\boldsymbol{B}_{\gamma_m^p}$ are discrete observation probability distributions for gesture and prosody label streams, and $\boldsymbol{\Pi}_{\gamma_m^{gp}}$ is an initial state probability matrix. The distributions $\boldsymbol{B}_{\gamma_m^g}$ and $\boldsymbol{B}_{\gamma_m^p}$ define the probability of observing a gesture-prosody label at state s and frame k, given by

$$P(\ell_k^{gp}|s) = P(\ell_k^g|s)^{\kappa_g} P(\ell_k^p|s)^{\kappa_p} \tag{18}$$

where the exponents, κ_g and κ_p, are the stream weights, which may be set to unity.

For the purpose of synthesis, each multi-stream discrete HMM branch, γ_m^{gp}, can be split into two individual single-stream discrete HMM models $\gamma_m^g = (\boldsymbol{A}_{\gamma_m^{gp}}, \boldsymbol{B}_{\gamma_m^g}, \boldsymbol{\Pi}_{\gamma_m^{gp}})$ and $\gamma_m^p = (\boldsymbol{A}_{\gamma_m^{gp}}, \boldsymbol{B}_{\gamma_m^p}, \boldsymbol{\Pi}_{\gamma_m^{gp}})$, respectively for gesture and prosody streams. These single stream HMM models share the same state transition and initial state probability matrices but their discrete observation probability distributions are different. The individual observation distributions are then given by $P(\ell_k^g|s)$ and $P(\ell_k^p|s)$ for gesture and prosody models, respectively.

Unsupervised temporal segmentation of joint label streams is demonstrated by the following example, which also illustrates how the asynchrony between gestures and prosody is handled in our scheme.

Example: Let us have two label streams $\boldsymbol{\ell}^a$ and $\boldsymbol{\ell}^b$, where each label can assume values 1, 2, or 3. When temporal segmentation of the joint label stream is performed using the HMM structure Γ_{gp} with $M = 2$ patterns and $N = 3$ number of states for each pattern, we obtain the result shown in Fig. 3. One can observe that the recurrent joint label patterns are captured and the asynchrony between individual label streams is modelled by the first and the last states of the HMM branches.

The number of states $N_{\Gamma_{gp}}$ for each branch of Γ_{gp} should be selected according to the number of head gesture and prosody patterns determined by the stage I analysis, since Γ_{gp} models the recurrent joint gesture-prosody label pairs. Similarly, the number of branches $M_{\Gamma_{gp}}$ in Γ_{gp} should be selected by considering the two fitness measures α and β as defined in (14) and (15). The selection of $N_{\Gamma_{gp}}$ and $M_{\Gamma_{gp}}$ is further discussed in SectionV.

IV. Prosody-Driven Gesture Synthesis

In this section, we address prosody-driven gesture synthesis using the proposed gesture-prosody pattern model. A detailed block diagram of the proposed prosody-driven gesture synthesis system is shown in Fig. 4. The system takes speech as input and produces a sequence of head gesture features, i.e., Euler angle vectors, which are naturally correlated with the input speech. The details of the sub-blocks are described in the following.

1) Prosody Feature Extraction: The prosody features, $\boldsymbol{F}^p$, are extracted from the input speech signal as described in Section II-B.

2) Prosody Feature Segmentation: Temporal segmentation of prosody feature sequence $\boldsymbol{F}^p$ is performed using the HMM model Λ_p, which is trained in the stage I analysis in Section III-A. During the temporal segmentation, the conditional probability $P(\boldsymbol{F}^p|\Lambda_p)$ is maximized using Viterbi decoding to extract the temporal prosody segment sequence, $\boldsymbol{\varepsilon}^p$, and the sequence of prosody frame labels, $\boldsymbol{\ell}^p$.

3) Gesture Segment Label Estimation: The aim of this step is to predict the sequence of gesture frame labels, $\boldsymbol{\ell}^g$, given the prosody frame labels $\boldsymbol{\ell}^p$. To this effect, temporal segmentation of the prosody frame labels, $\boldsymbol{\ell}^p$ is performed using the HMM model Γ_p, which is extracted by splitting the jointly trained gesture-prosody HMM model Γ_{gp}. As a result of this temporal prosody label segmentation, a state sequence $\boldsymbol{s}^p = \{s_1^p, s_2^p, \ldots, s_K^p\}$ associated with $\boldsymbol{\ell}^p = \{\ell_1^p, \ell_2^p, \ldots, \ell_K^p\}$ is extracted. Then, the gesture frame label sequence $\boldsymbol{\ell}^g$ is predicted by maximizing the probability of observing gesture label on the state sequence path $\boldsymbol{s}^p$ over the gesture HMM model Γ_g, such that,

$$\ell_k^g = \arg\max_m P(m|s_k^p, \Gamma_g) \tag{19}$$

where k is the frame index, m runs over all possible M gesture patterns and the conditional probability $P(m|s_k^p, \Gamma^g)$ is defined by the discrete observation probability distribution $\boldsymbol{B}_{\gamma_m^g}$.

4) Generation of Euler Angles: This step computes the gesture segment sequence $\boldsymbol{\varepsilon}^g$, consisting of the Euler angle features, given the gesture frame label sequence $\boldsymbol{\ell}^g$. First, we find the segment frame boundaries, $\{t_l\}_{l=1}^L$, by merging the same gesture frame labels in the sequence $\boldsymbol{\ell}^g$. Then, the Euler angle features for the l-th segment, $\varepsilon_l^g = \{\boldsymbol{f}_{t_l}^g, \boldsymbol{f}_{t_l+1}^g, \ldots, \boldsymbol{f}_{t_{l+1}-1}^g\}$, are generated from the HMM $\lambda_{\ell_{t_l}}^g$, which is the ℓ_{t_l}-th branch of the parallel HMM model Λ_g (computed in stage I).

Note that the segment duration for the l-th segment is extended as $d_l = (t_{l+1} + \Delta - (t_l - \Delta))$ frames, where Δ is the number of overlapping frames at the segment boundaries to smooth segment-to-segment transitions. The state sequence $\boldsymbol{s}_l^g$ or equivalently the state occupancy durations for the l-th segment is calculated using the diagonal terms of the d_l-step state transition matrix of the HMM $\lambda_{\ell_{t_l}}^g$. Having the state sequence $\boldsymbol{s}_l^g$ and the continuous observation probability $P(\boldsymbol{f}^g|\boldsymbol{s}_l^g)$, which are modeled using a Gaussian distribution, the Euler angle features are generated along the state sequence associated with the distribution $P(\boldsymbol{f}^g|\boldsymbol{s}_l^g)$. The segment boundaries have $2\Delta + 1$ number of frame overlaps, where the overlapped and averaged features generate smoother segment-to-segment transitions.

5) Smoothing of Euler Angles: As the final step of the gesture synthesis, the Euler angles are smoothed using median filtering followed by a Gaussian low pass filter to remove motion jerkiness. The median filtering is performed over 11 visual frames and the Gaussian smoothing is performed over 15 visual frames. Fig. 5 depicts the samples generated from the HMM, and outputs of the median and Gaussian filters. The figure clearly shows that the median filter removes jitters within a state and the Gaussian low pass filter smooths the state-to-state transitions.

There are two main advantages of using HMMs for gesture synthesis. The first is the random variations in the synthesized gesture

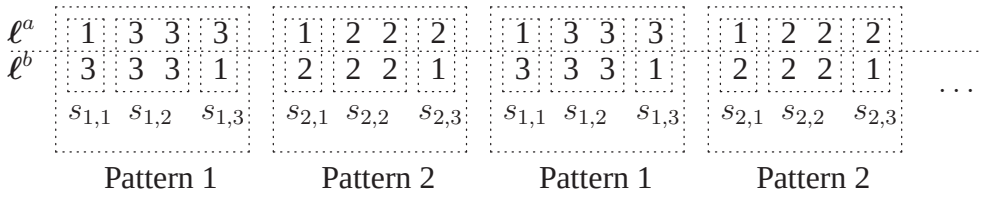

Fig. 3. Example for unsupervised joint label segmentation.

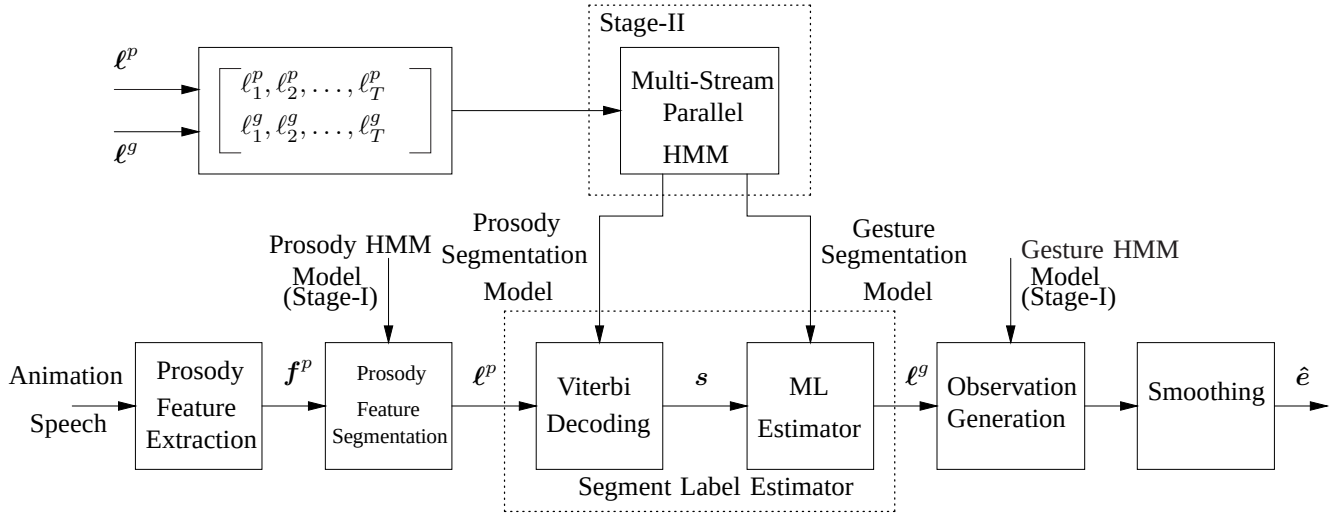

Fig. 4. The proposed prosody-driven gesture synthesis system.

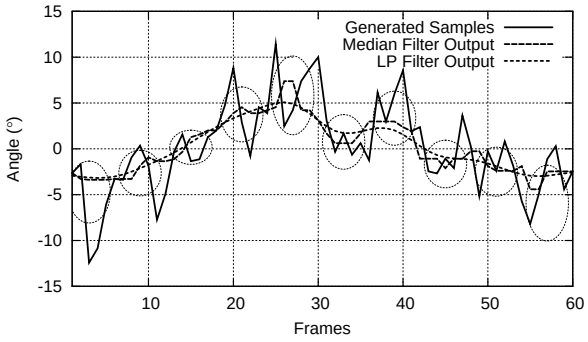

Fig. 5. The effect of filtering in the synthesis of Euler angle θ. The dashed circles represent the states of a single branch HMM model. The vertical position and size of each circle are adjusted considering the mean and variance of the Euler angles associated with each state.

patterns for each segment. This variation yields more natural looking synthesis results than using a fixed gesture dictionary, since humans produce slightly varying gestures at different occasions for the same semantics. The second advantage is generating gestures with varying durations in accordance with prosody of the speaker.

V. Evaluation and Results

In this section, we present experimental results and evaluation of the proposed system. Section V-A describes the audio-visual database, which is used in the experimental evaluation to generate objective and subjective results. The evaluation of the gesture-prosody pattern analysis is presented in Section V-B, and the objective and subjective performance results for synthesis are presented in Section V-C. Speaker dependency of the prosody-driven head gesture synthesis system is evaluated in Section V-D.

A. Database and Experimental Setup

We have conducted experiments using the MVGL-MASAL gesture-speech database. The database includes four recordings of two subjects telling stories in Turkish. The subjects are instructed to tell stories to children audience. All gestures are spontaneous within this context. Each story lasts approximately 7 minutes. The audio-visual data is synchronously captured from the stereo camera and the sound card. The stereo video includes only upper body gestures with 30 frames per second whereas the audio is recorded with 16 kHz sampling rate and 16 bits per sample. The detailed specification of the stereo camera can be found on [28]. The performance of the proposed analysis and synthesis system is evaluated in detail on the recordings of the first speaker, whereas the recordings of the second speaker are used to investigate the speaker dependency problem. For the first speaker, the database is partitioned into two parts such that three stories are used for training of the models and one story is used for testing. For objective evaluation of the synthesis, the Euler angles extracted from the test sequence are considered as the ground truth for the synthesized head motion.

B. Analysis Results

The head gesture and prosody correlation analysis includes unsupervised temporal segmentation of the individual feature streams as well as the joint gesture-prosody label stream. The objective and subjective evaluation of these tasks are presented in the following.

Segmentation of Head Gesture Patterns: The parallel HMM Λ_g is trained with features extracted from the training video using Expectation-Maximization (EM) algorithm. The resulting HMM structure provides a probabilistic cluster model for unsupervised segmentation of head gestures into recurring elementary patterns.

The number of branches, or equivalently the number of gesture patterns, M_{Λ_g} is a critical model parameter. In order to set M_{Λ_g}, the

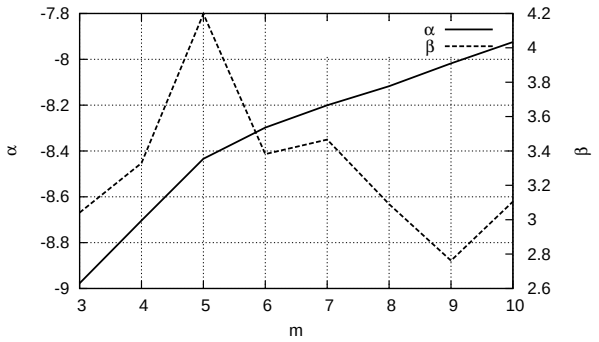

Fig. 6. The α and β fitness measures for varying number of head gesture patterns.

two fitness measures α and β, as respectively defined in (14) and (15), are calculated for varying number of gesture patterns and plotted in Fig. 6. The α value, which measures the probability of model match, increases with increasing number of patterns as expected. Note that β measures the statistical separation between patterns. A good value for M_{Λ_g} is such that β is high enough, while α reaches a certain value. Therefore, M_{Λ_g} can be selected in the vicinity where α and β curves (normalized with their minimum and maximum values) intersect. In Fig. 6, β reaches the maximum at $M_{\Lambda_g} = 5$, which is near the intersection point. Hence, we set the number of gesture patterns M_{Λ_g} to 5.

Consequently, when the training head gesture sequence is segmented using Λ_g, the segments belonging to the same gestural patterns are observed to be visually alike. The mean Euler angle vectors and the typical thumbnails for the five gesture patterns are depicted in Fig. 7.

Segmentation of Prosody Patterns: The speech prosody feature sequence is extracted from the audio part of the training database. As defined in stage I, the HMM model Λ_p is trained with prosodic features to obtain unsupervised temporal segmentation of the audio stream.

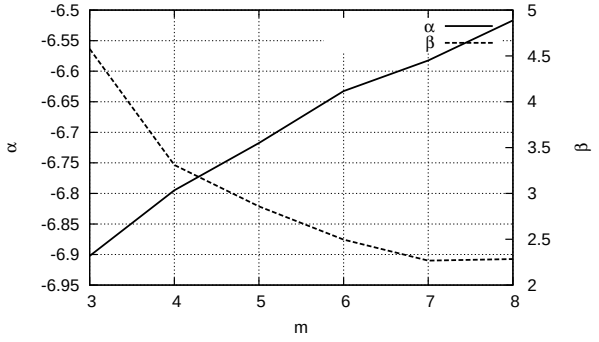

Fig. 8. The α and β fitness measures for varying number of prosody patterns.

The two fitness measures α and β are calculated for varying number of prosody patterns using HMM model Λ_p and plotted in Fig. 8. The α value, which measures the probability of model match, increases and the β value, which measures statistical separation between patterns, decreases with increasing number of patterns as expected. The number of prosody patterns M_{Λ_p} can thus be set to a value in the vicinity where α and β curves intersect. Hence, we select M_{Λ_p} as 5 in our experiments.

The means and standard deviations of the normalized intensity and pitch frequency trajectories for the five prosody patterns are depicted in Fig. 9. Note that the first pitch trajectory (upper-left) is associated with the no-pitch segments that we filled with zero mean and unit variance Gaussian noise. The noise filling is necessary for successful modeling of those segments with continuous density HMMs. The other four prosody patterns can be classified using the prosodic transcription conventions introduced by the American English Tones and Break Indices (ToBI) standard [29]. The two prosody patterns on the upper right are both falling boundary tones (L%); the pattern on the lower left is a falling boundary tone, which makes a peak before the last syllable (HL%), and the pattern on the lower right is a rising-falling boundary tone, which rises within the last syllable (LHL%). We should note that these prosody patterns are obtained using unsupervised clustering over the training database, and they do not define a complete prosodic transcription convention for Turkish.

Segmentation of Joint Gesture-Prosody Patterns: In the first stage analysis, we obtain two independent HMM structures, Λ_g and Λ_p, respectively for recurrent head gesture and prosody patterns. We then extract two independent and parallel streams of head gesture and prosody pattern labels via temporal segmentation using these HMM structures. In the second stage, the discrete multi-stream HMM structure Γ_{gp} is trained using EM over the joint gesture-prosody pattern label stream to perform unsupervised segmentation. The number of states for each branch of Γ_{gp} is selected as $N_{\Gamma_{gp}} = 4$ to model possible label pair transitions. These four states model four different gesture-prosody label pair combinations within a joint gesture-prosody label pattern. Note that the extreme case, $N_{\Gamma_{gp}} = 1$, can only model a single co-occurrence pattern of gesture-prosody labels.

The two fitness measures α and β for Γ_{gp}, and also the number of gesture patterns in Λ_g, are considered for selection of the number of joint gesture-prosody label patterns $M_{\Gamma_{gp}}$. The number of joint patterns $M_{\Gamma_{gp}}$ is expected to be larger than or equal to the number of gesture patterns M_{Λ_g}, since in a robust synthesis process all the gesture patterns need to be generated for some temporal prosody label pattern. Hence, for the selection of $M_{\Gamma_{gp}}$, we present the two fitness measures α and β together with the normalized Euclidean distance measure ϵ_n as defined in (20) for varying number of joint gesture-prosody label patterns in Fig. 10. The parameter $M_{\Gamma_{gp}}$ is selected as 6, since this value, which is near the intersection of α and β curves, is greater than M_{Λ_g}, and the distance ϵ_n has a minimum at $M_{\Gamma_{gp}} = 6$.

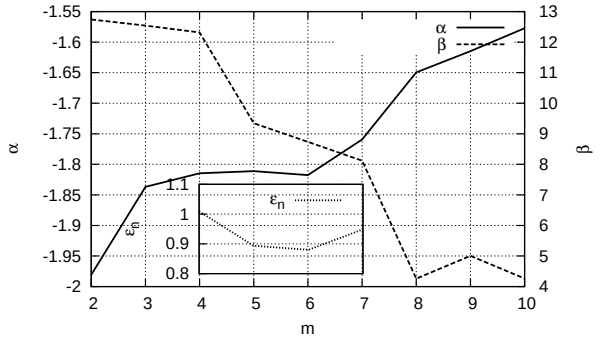

Fig. 10. The α and β fitness measures and the normalized Euclidean distance measure ϵ_n for varying number of joint gesture-prosody label patterns.

Observation probability distributions of the joint multi-stream HMM are plotted in Fig. 11. It can be seen that each branch of the HMM structure Γ_g models a temporal sequence of identical elementary gesture patterns. That is, in each of the six classes, a distribution of prosody patterns co-occurs with a single elementary gesture pattern. Note that this association between temporal prosody label patterns and a single gesture pattern is very beneficial to

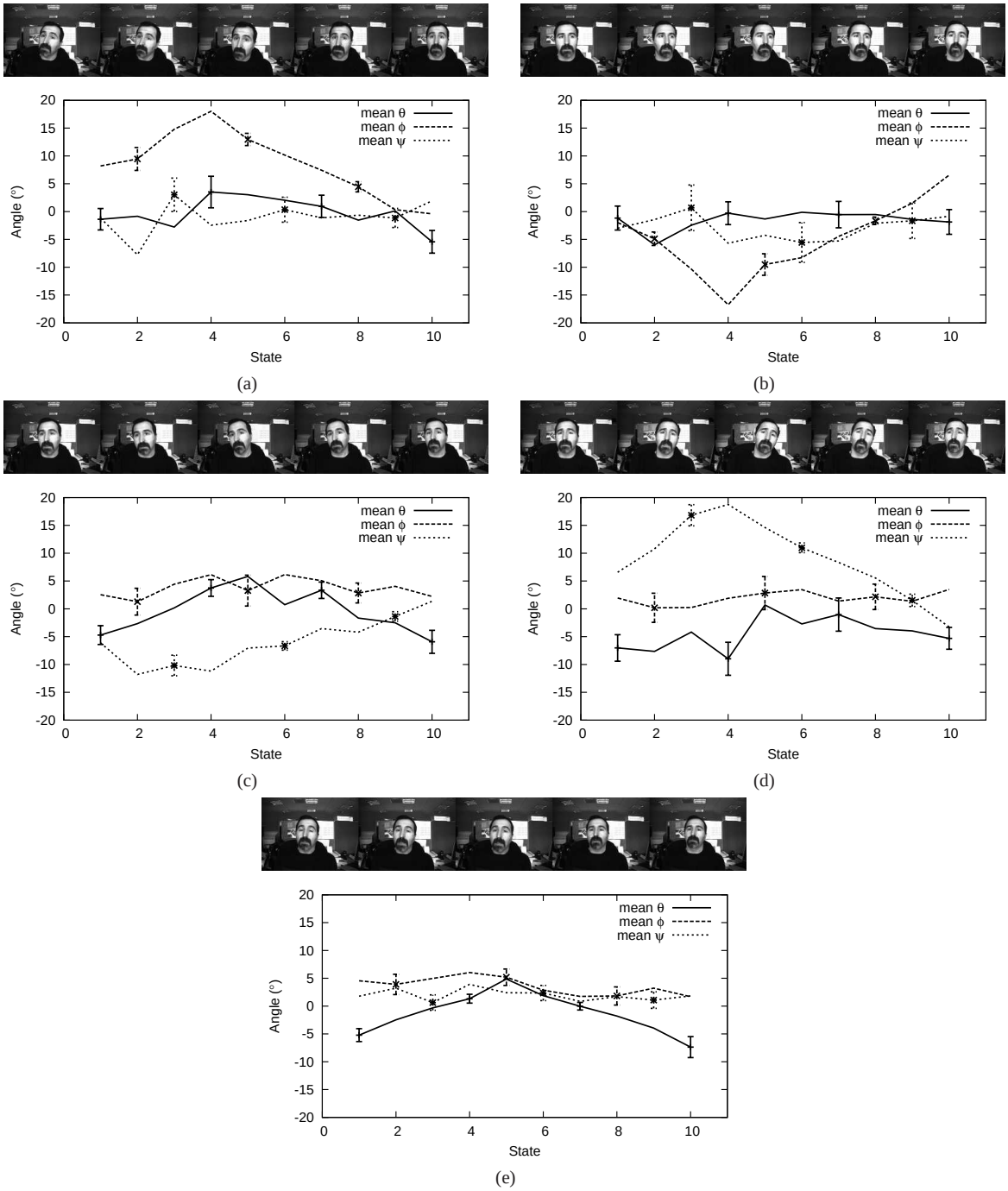

Fig. 7. The mean Euler angles with standard deviations and typical thumbnails for the five gesture patterns: (a) Turn Left, (b) Turn Right, (c) Tilt Left, (d) Tilt Right, and (e) Nod

obtain smooth prosody-driven head gesture animations. Furthermore, boundaries of the prosody patterns within the co-occurring gesture pattern are determined by the state transition probabilities of Γ_p, and hence, the asynchrony problem is handled through the learned statistics of the joint multi-stream HMM Γ_{gp}.

C. Synthesis Results

Prosody-driven head gesture synthesis generates an Euler angle sequence, which is naturally correlated to a given test speech signal. The details of the synthesis process is given in Section IV. In this section, we present objective and subjective evaluations of the prosody-driven head gesture synthesis process. The evaluations are performed over the test database, which is defined in Section V-A.

The objective evaluations compare the difference between original and synthesized Euler angles. Furthermore, A-B comparison type subjective evaluations are performed using the talking head avatar of *Momentum Inc.* [30], where the Euler angles that we deliver are used to drive head gestures/motion of the speech-driven talking head animation. The subjective tests are used to measure opinions on the naturalness of the synthesized head gestures using the speech-driven talking head animations.

We have adopted the Input-Output Hidden Markov Model

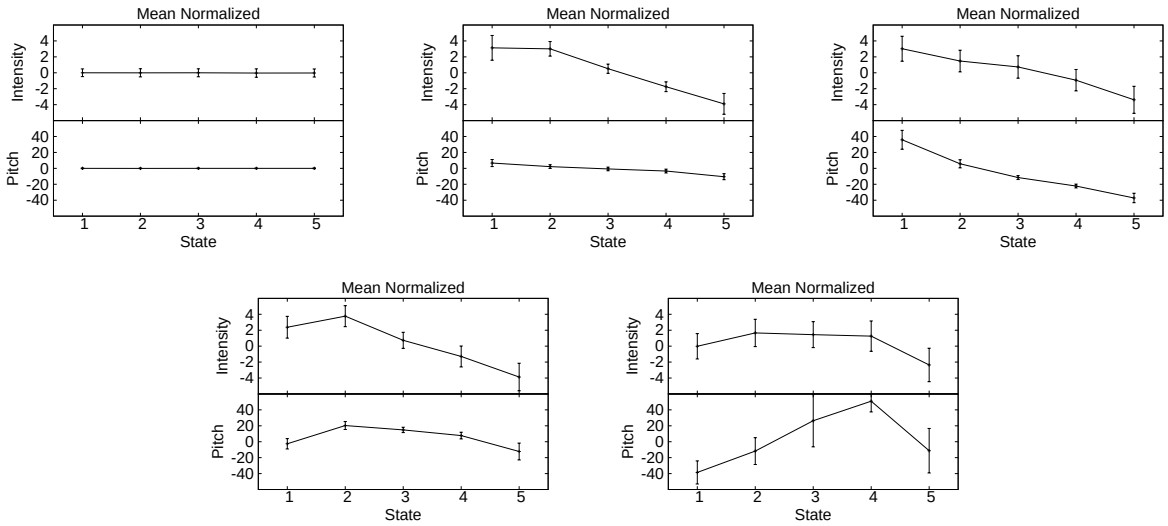

Fig. 9. The means and standard deviations of the normalized intensity (dB) and pitch frequency (Hz) trajectories for the five prosody patterns.

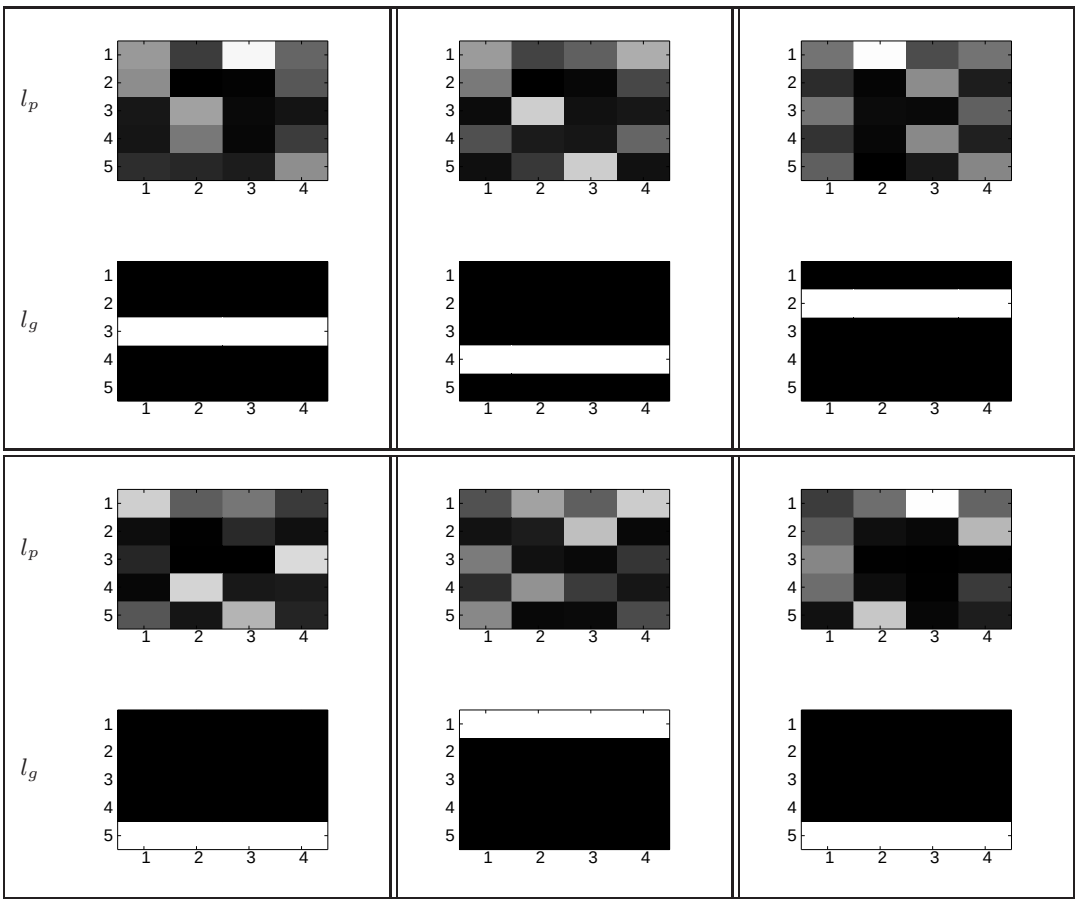

Fig. 11. Observation probability distributions of the joint multi-stream HMM. The discrete prosody and gesture labels are on the y-axis, states are on the x-axis. Dark and white regions represent low and high probability values, respectively. Note that in each of the six classes, a distribution of prosody patterns co-occurs with a single elementary gesture pattern.

(IOHMM) structure [8], [31] as a possible alternative scheme for the joint analysis of gesture and prosody label streams at the second-stage. In that case, the IOHMM structure replaces the HMM Γ_{gp} to predict gesture segment labels from prosody labels. The states in the IOHMM are fully connected and the number of states is selected to be the same as the number of states in the Γ_{gp} model, which is 24. The IOHMM implementation of the Torch Machine Learning Library [32] is used in our experiments.

Objective Results: The objective evaluations compare the distance between original and synthesized Euler angles. In our evaluations we have used three different distance measures. Let the original and synthesized Euler angles at frame k are represented with $\boldsymbol{e}_k$ and $\hat{\boldsymbol{e}}_k$,

TABLE I
THE MEAN AND STANDARD DEVIATION OF THE DISTANCE MEASURES BETWEEN THE ORIGINAL AND THE TWO SETS OF SYNTHESIZED EULER ANGLES, FROM THE PROPOSED Γ_{gp} AND IOHMM MODELS.

Model	Γ_{gp}	IOHMM
$[\mu_{\epsilon_n}, \sigma_{\epsilon_n}]$	$[0.817896, 0.010981]$	$[0.890652, 0.012861]$
$[\mu_{\epsilon_m}, \sigma_{\epsilon_m}]$	$[1.946043, 0.020871]$	$[2.290755, 0.073021]$
$[\mu_{\epsilon_e}, \sigma_{\epsilon_e}]$	$[13.694374, 0.158944]$	$[16.427827, 0.525493]$

respectively. The first distance measure ϵ_n is a normalized Euclidean distance measure, which penalize Euler angles in wrong directions [6],

$$\epsilon_n = \frac{\sum_{k=1}^{K}(\hat{\boldsymbol{e}}_k - \boldsymbol{e}_k)^T(\hat{\boldsymbol{e}}_k - \boldsymbol{e}_k)}{\sum_{k=1}^{K}(\hat{\boldsymbol{e}}_k + \boldsymbol{e}_k)^T(\hat{\boldsymbol{e}}_k + \boldsymbol{e}_k)} \tag{20}$$

The second measure ϵ_m is the Mahalanobis distance, which is the Euclidean distance weighted with the inverse covariance matrix, $\boldsymbol{\Sigma}^{-1}$, of the original Euler angles $\boldsymbol{e}_k$,

$$\epsilon_m = \frac{1}{K}\sum_{k=1}^{K}\sqrt{(\hat{\boldsymbol{e}}_k - \boldsymbol{e}_k)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{e}}_k - \boldsymbol{e}_k)} \tag{21}$$

The third distance measure is the standard Euclidian distance, $\epsilon_e = \frac{1}{K}\sum_{k=1}^{K}\sqrt{(\hat{\boldsymbol{e}}_k - \boldsymbol{e}_k)^T(\hat{\boldsymbol{e}}_k - \boldsymbol{e}_k)}$.

The original Euler angles are extracted from the visual part of the test database to be used as the ground truth in the objective evaluations. Two sets of synthesized Euler angles are generated using the audio part of the test database. The first set is generated with the proposed head gesture synthesis system based on the Γ_{gp} model. The second set is generated by replacing the second stage joint gesture-prosody correlation model Γ_{gp} by IOHMM. The error measure statistics for the three distance measures ϵ_n, ϵ_m and ϵ_e between the original and synthesized Euler angles are collected over synthesis trials repeated a hundred times. The mean and standard deviation of the distance measures are given in Table I. Note that all three distance measures favor the proposed joint gesture-prosody correlation model Γ_{gp}.

TABLE II
THE SUBJECTIVE A-B COMPARISON RESULTS

A-B pair	Preference Score
Original - Γ_{gp}	-0.23
Original - IOHMM	-0.83
Γ_{gp} - IOHMM	-0.56
Identical pairs	0.04

Subjective Results: Subjective A-B comparisons are performed using the speech-driven talking head animations to measure opinions on the naturalness of the synthesized head gestures. The subjects are asked to evaluate the naturalness of the speech-driven synthesized head gestures for an A-B test pair on a scale of $(-2,\ -1,\ 0,\ 1,\ 2)$, where the scale corresponds to (A much better, A better, no preference, B better, B much better).

The whole test database is manually partitioned into meaningful 15 segments, where each segment is approximately 12 seconds. For each evaluation 8 segments out of 15 are randomly selected. Three sets of A-B comparison pairs, each including these 8 segments, are considered for the speech-driven talking head animations using the original and two sets of synthesized Euler angles. Furthermore, three random startup A-B test pairs and another three test pairs with identical synthesis algorithms are also included to the subjective test set. Hence, the total number of A-B pairs in a test is 30. Apart from the three random start-up A-B pairs, all the pairs are randomized across conditions and pairwise. The subjective tests are performed over 15 subjects. The average preference scores for the three comparison sets are presented in Table II. Note that the scores of the three random start-up pairs are ignored in calculating the final preference scores. As expected, the subjective A-B comparisons indicate a preference for the talking head animations with the original Euler angles. On the other hand, animations synthesized with the proposed joint gesture-prosody correlation model Γ_{gp} are preferred over animations generated using the IOHMM correlation model with an average preference score of -0.56. Also note that the preference for the animations with the original Euler angles is stronger in the case of IOHMM driven animations as compared to the proposed Γ_{gp} driven animations. This is expected, since the output and transition probabilities in the IOHMM structure are conditional directly on the input sequence, whereas in the joint multi-stream HMM, the output gesture patterns are affected by the states only and not directly by the input. Hence, use of parallel multi-stream HMM in the second stage is more robust to any noise in the input stream.

Samples of the audio-visual sequences for the prosody-driven talking head animations are available online [33]. These samples are selected to demonstrate three possible related applications. The first one is the speaker dependent prosody-driven gesture synthesis application, where gesture-prosody correlation model of a speaker is used to animate the same speaker with her/his speech. The second application is head gesture transplant, where gesture-prosody correlation model of speaker *A* is used to animate speaker *B* from speaker *A*'s speech. Furthermore, the prosody transplant is considered as the third application, where gesture-prosody correlation model of speaker *A* is used to animate speaker *A* from speaker *B*'s speech. In the demonstration of the prosody transplant we used speech input from audio-book recordings in English, where the gesture-prosody correlation model is performed over the story telling recordings in Turkish. Although one should expect differences in prosody patterns across different languages, the naturalness of the animations is observed to be acceptable. We also note that the talking speed of these two speakers are different, where the native Turkish speaker has a faster rate than the native English speaker. As expected from the proposed correlation model, we observe slower head gesture animations for the native English speaker.

D. Speaker Dependency

The proposed analysis method is capable of providing personalized elementary head gesture and prosody patterns and a personalized prosody to gesture mapping model. To demonstrate this, we have repeated the experiments with a second speaker. The second speaker is also instructed to tell the same four stories to an audience of children. The system is then trained using the recordings of the speaker. At the end of the two-stage analysis for modeling gesture-prosody correlation, we have observed that the resulting elementary patterns for both prosody and head gestures significantly differ from those of the first speaker.

In order to set the number of gesture patterns, M_{Λ_g}, and prosody patterns, M_{Λ_p}, the two fitness measures α and β are calculated for varying number of gesture and prosody patterns, respectively, which are plotted in Fig. 12 and Fig. 13. In both plots, the probability of model match, α, increases and the statistical separation between patterns, β, decreases with increasing number of patterns. The numbers of gesture and prosody patterns are selected in the vicinity where α and β curves intersect, as $M_{\Lambda_g} = 4$ and $M_{\Lambda_p} = 5$. The mean Euler angle vectors and typical thumbnails for the four gesture patterns are plotted in Fig. 14. Similarly, the means and standard deviations of the normalized intensity and pitch frequency trajectories for the

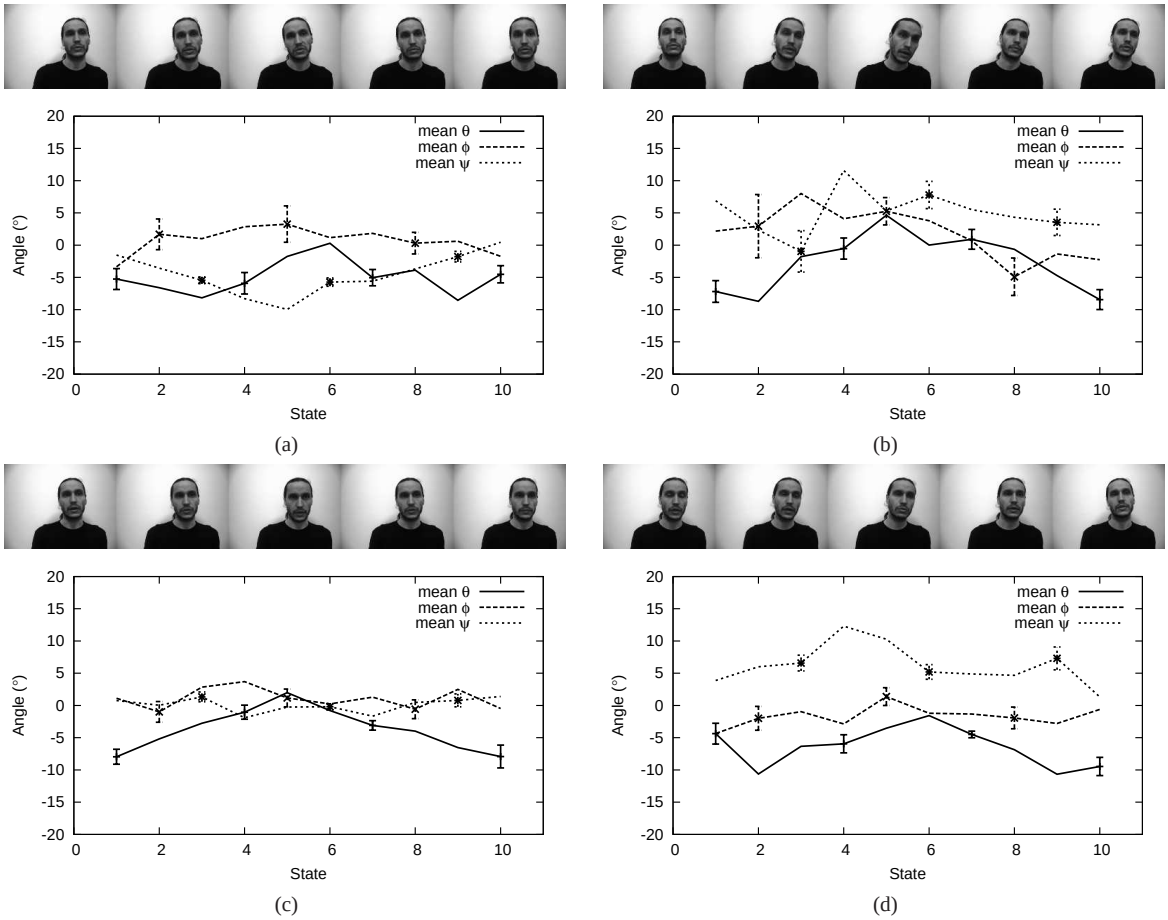

Fig. 14. The mean Euler angles with standard deviations and typical thumbnails from the second speaker for the four gesture patterns: (a) Tilt Left, (b) Nod with Tilt Right, (c) Nod, (d) Tilt Right.

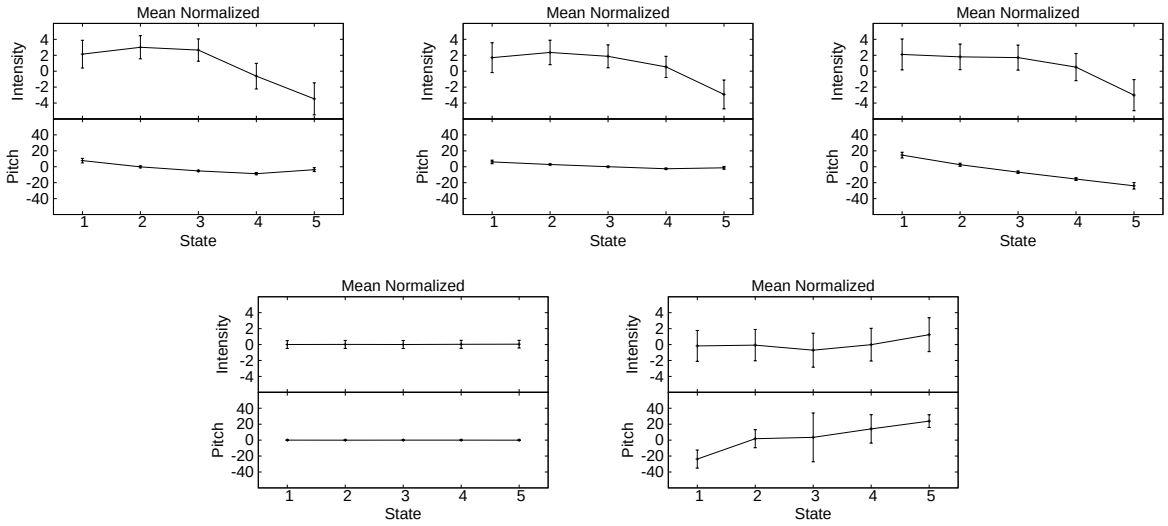

Fig. 15. The means and standard deviations of the normalized intensity (dB) and pitch frequency (Hz) trajectories for the five prosody patterns of the second speaker.

five prosody patterns are plotted in Fig. 15. Note that the elementary gesture patterns for the second speaker is distinctively different than the ones for the first speaker (see Fig. 7 for comparison). Sample video streams of the typical elementary gesture patterns are available online in [33]. The elementary prosody patterns also differ for the second speaker. Three of the prosody patterns are falling boundary tones (L%), and the other one is a rising boundary tone (H%) for the second speaker.

At the second stage analysis, the joint gesture-prosody pattern label stream is segmented in an unsupervised manner using the discrete

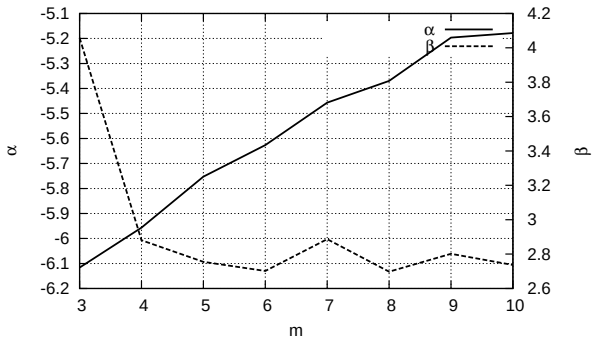

Fig. 12. The α and β fitness measures for varying number of head gesture patterns of the second speaker.

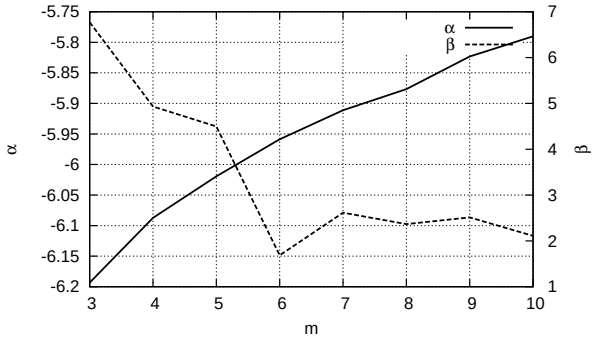

Fig. 13. The α and β fitness measures for varying number of prosody patterns of the second speaker.

multi-stream HMM structure Γ_{gp}. The two fitness measures for Γ_{gp} are plotted in Fig. 16. In the unsupervised segmentation, the number of joint gesture-prosody patterns is set to $M_{\Gamma_{gp}} = 5$. As for the demonstration of synthesis results, to better emphasize speaker dependency, we have used the same audio-book recordings in English and the same face model to derive the head gesture animations for the two different speakers. A sample animation video is available online in [33], where the video stream resulting from the second speaker's gesture-prosody correlation model, is presented in parallel with the video stream generated from the first speaker's model for visual evaluation of the speaker dependency performance.

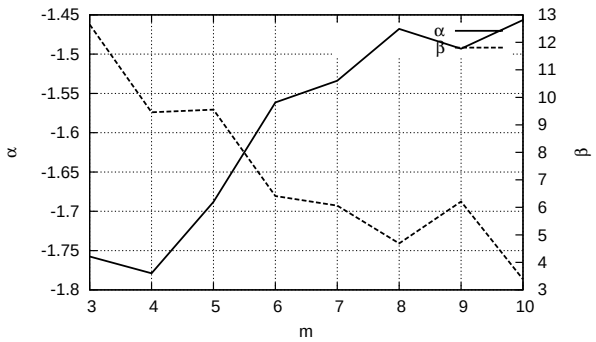

Fig. 16. The α and β fitness measures and the normalized Euclidean distance measure ϵ_n for varying number of joint gesture-prosody label patterns for the second speaker.

VI. Conclusions

We proposed a new two-stage joint head gesture and speech prosody analysis framework. In the first stage of the analysis, elementary gesture and prosody patterns are extracted using unsupervised segmentation for a speaker, and in the second stage, a correlation model between head gesture and prosody patterns is developed. The proposed two-stage analysis framework offers the following advantages: i) Meaningful elementary gesture and prosody patterns are defined for a speaker at the first stage. ii) A mapping between these elementary prosody and head gesture patterns is obtained with unsupervised segmentation of the joint gesture-prosody label stream. iii) The HMM-based analysis and synthesis yields flexibility in modeling structural and durational variations within gestural and prosodic patterns. iv) Automatic generation of the elementary gesture patterns produces natural looking prosody-driven head gesture synthesis.

In addition to successful demonstration of speaker dependent speech-driven head gesture synthesis system, different applications, such as head gesture transplant and prosody transplant, are also demonstrated. After extracting a gesture-prosody correlation model for speaker *A*, head gesture transplant animates speaker *B* from speaker *A*'s speech, and prosody transplant animates speaker *A* from speaker *B*'s speech. In the prosody transplant demonstration, gesture-prosody correlation model is trained with audio-visual recordings in Turkish, and prosody-driven gesture synthesis is performed with speech input recordings in English. The naturalness of the prosody transplant is found to be acceptable. Also in this demonstration, we observe slower head gesture animations for the native English speaker whose talking speed is slower.

The proposed HMM based two-stage head gesture and speech prosody analysis system can be utilized to model the correlation between any other loosely correlated modalities, such as facial expressions and speech prosody, arm gestures and speech semantics, etc. Furthermore, the proposed speaker dependent speech-driven head gesture synthesis system can be tailored to model speaker's emotion and mood. We also note that prosody patterns obtained using the proposed stage I analysis over a multi-speaker phonetically rich Turkish (or any other language) training database, can be used to define a complete ToBI-like prosodic transcription convention for Turkish (or any other language) intonation.

VII. Acknowledgments

The authors would like to thank *Momentum Inc.* for making the talking head avatar available and for their collaboration to build the MVGL-MASAL gesture-speech database.

Appendix
Rigid Motion Parameter Estimation by Constrained Optimization

This appendix summarizes the method used for estimating the rotation matrix, $\boldsymbol{R}$, and translation vector, $\boldsymbol{t}$, that describe the rigid motion between the world point coordinate matrices $\boldsymbol{W}_k$ and $\boldsymbol{W}_{k_r}$ (see Section II-A).

Let $\boldsymbol{m}_k$ denote the mean of the column vectors in the matrix $\boldsymbol{W}_k$ such that

$$\boldsymbol{m}_k = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{w}_k^i \tag{22}$$

and $\boldsymbol{m}_{k_r}$ be defined similarly. Then, the translation $\boldsymbol{t}$ between $\boldsymbol{W}_k$ and $\boldsymbol{W}_{k_r}$ is given by

$$\boldsymbol{t} = \boldsymbol{m}_k - \boldsymbol{m}_{k_r} \tag{23}$$

Furthermore, let $\boldsymbol{W}'_{k_r}$ and $\boldsymbol{W}'_k$ represent the mean-removed coordinate matrices such that

$$\boldsymbol{W}'_k = \boldsymbol{W}_k - \boldsymbol{m}_k \boldsymbol{1}^T, \quad \text{and} \quad \boldsymbol{W}'_{k_r} = \boldsymbol{W}_{k_r} - \boldsymbol{m}_{k_r} \boldsymbol{1}^T \tag{24}$$

Then, the rotation matrix $\boldsymbol{R}$ can be found by minimizing the cost function

$$f(\boldsymbol{R}) = \|\boldsymbol{E}\|_F^2 = \mathrm{tr}(\boldsymbol{E}\boldsymbol{E}^T) \tag{25}$$

where $\|\cdot\|_F$ and $\mathrm{tr}(\cdot)$ denote the Frobenius-norm and the matrix trace, respectively, and

$$\boldsymbol{E} = \boldsymbol{W}'_k - \boldsymbol{R}\boldsymbol{W}'_{k_r} \tag{26}$$

The minimization of the cost function $f(\boldsymbol{R}), f : \mathbb{R}^{3\times3} \rightarrow \mathbb{R}$, is a non-linear optimization problem, under the unitary constraint $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}$, which can be solved by the algorithm proposed in [34], where Manton describes a modified Newton method for optimization on the complex Stiefel manifold which defines the space related with the unitary constraint.

We simplified this method to minimize the cost function $f(\boldsymbol{R})$ for a square and real matrix $\boldsymbol{R}$ subject to the constraint $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}$ as follows:

1) Choose initial $\boldsymbol{R}$ such that $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}$.
   For small rotations, $\boldsymbol{R}$ can be approximated in terms of a parameter vector $\boldsymbol{u} = [u_x, u_y, u_z]^T$ such that [35]

$$\boldsymbol{R} \approx \boldsymbol{I} + \boldsymbol{S} = \boldsymbol{I} + \begin{bmatrix} 0 & -u_z & u_y \\ u_z & 0 & -u_x \\ -u_y & u_x & 0 \end{bmatrix} \tag{27}$$

   Equating the residual defined in (26) to zero, we obtain the following equation to solve for $\boldsymbol{S}$:

$$\boldsymbol{W}'_k - \boldsymbol{W}'_{k_r} = \boldsymbol{S}\boldsymbol{W}'_{k_r} \tag{28}$$

   which can be expressed in terms of $\boldsymbol{u}$ as

$$\begin{aligned} \mathrm{vec}(\boldsymbol{W}'_k - \boldsymbol{W}'_{k_r}) = \boldsymbol{K}\boldsymbol{u} = \begin{bmatrix} \boldsymbol{K}_1 \\ \vdots \\ \boldsymbol{K}_N \end{bmatrix} \boldsymbol{u} \\ \boldsymbol{K}_n = \begin{bmatrix} 0 & Z_n & -Y_n \\ -Z_n & 0 & X_n \\ -X_n & Y_n & 0 \end{bmatrix} \end{aligned} \tag{29}$$

   where operator $\mathrm{vec}(.)$ obtains a column vector by stacking the columns of the operand matrix and each 3×3 sub-matrix $\boldsymbol{K}_n$ is constructed using the nth point (X_n, Y_n, Z_n) from $\boldsymbol{W}'_{k_r}$. The least squares solution of (29) can then be used to find $\boldsymbol{u}$ and to construct $\boldsymbol{S}$. The initial guess for $\boldsymbol{R}$ can finally be obtained by projection onto the unitary space $\boldsymbol{R} = \pi(\boldsymbol{I} + \boldsymbol{S})$ (described in step 5 below).
2) Compute the derivative $\boldsymbol{D}_R$ and the Hessian $\boldsymbol{H}_R$ of f given by

$$\boldsymbol{D}_R = -2\boldsymbol{E}\boldsymbol{W}'^T_{k_r} \tag{30}$$

$$\boldsymbol{H}_R = -2((\boldsymbol{W}'_{k_r}\boldsymbol{W}'^T_{k_r}) \otimes \boldsymbol{I}_{3\times3}) \tag{31}$$

   where $\otimes$ denotes Kronecker product.
3) If $\sqrt{tr(\boldsymbol{D}_R^T\boldsymbol{D}_R - \boldsymbol{R}^T\boldsymbol{D}_R\boldsymbol{R}^T\boldsymbol{D}_R)} < \epsilon$, then stop.
4) Compute the Newton step size $\boldsymbol{Z} := \boldsymbol{Z}^{(cp)}$.
   The Newton step size is defined as the value of $\boldsymbol{Z}$, $\boldsymbol{Z} \in \mathbb{R}^{3\times3}$, confined to the tangent space V, at which the quadratic approximation $g(\boldsymbol{Z})$ has its critical point:

$$g(\boldsymbol{Z}) \approx f(\boldsymbol{R}) + \mathrm{tr}(\boldsymbol{Z}^T\boldsymbol{D}) + (1/2)\,\mathrm{vec}(\boldsymbol{Z})^T\boldsymbol{H}\,\mathrm{vec}(\boldsymbol{Z}) \tag{32}$$

   where

$$\boldsymbol{D} = \boldsymbol{D}_R, \quad \boldsymbol{H} = \boldsymbol{H}_R - (1/2)[(\boldsymbol{R}^T\boldsymbol{D}_R + \boldsymbol{D}_R^T\boldsymbol{R})^T \otimes \boldsymbol{I}] \tag{33}$$

   The tangent space V is defined as a subset of $\mathbb{R}^{3\times3}$ such that $\boldsymbol{Z} = \boldsymbol{R}\boldsymbol{A}$ where $\boldsymbol{A}$ is skew-symmetric. The critical point $\boldsymbol{Z}^{(cp)} \in V$, i.e. the Newton step size, satisfies the following linear constraint:

$$tr(\boldsymbol{Z}^T D) + \left[\mathrm{vec}(\boldsymbol{Z})^T\boldsymbol{H}\right]\mathrm{vec}(\boldsymbol{Z}^{(cp)}) = 0 \tag{34}$$

   By writing $\boldsymbol{Z}$ as $\boldsymbol{Z} = \sum_{i=1}^3 \alpha_i \boldsymbol{R}\boldsymbol{A}_i$, where $\boldsymbol{A}_i \quad (i = 1, 2, 3)$ is an arbitrary basis for skew-symmetric matrix, the critical point $\boldsymbol{Z}^{(cp)}$ can be found by solving the following linear equation for $\alpha_i \quad (i = 1, 2, 3)$

$$\frac{\partial g(\boldsymbol{Z})}{\partial \alpha_i} = \mathrm{tr}((\boldsymbol{R}\boldsymbol{A}_i)^T\boldsymbol{D}) + \mathrm{vec}(\boldsymbol{R}\boldsymbol{A}_i)^T\boldsymbol{H}\,\mathrm{vec}(\boldsymbol{Z}) = 0 \tag{35}$$

   Note that the above equation can be put into a matrix form:

$$\boldsymbol{K}^T\boldsymbol{H}\boldsymbol{K}\boldsymbol{\alpha} = \boldsymbol{K}^T\,\mathrm{vec}(\boldsymbol{D}) \tag{36}$$

   where

$$\begin{aligned} \boldsymbol{K} = (\boldsymbol{I} \otimes \boldsymbol{R})[\mathrm{vec}(\boldsymbol{A}_1), \mathrm{vec}(\boldsymbol{A}_2), \mathrm{vec}(\boldsymbol{A}_3)] \\ \boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3]^T \end{aligned} \tag{37}$$

5) Set $\boldsymbol{R}' := \pi(\boldsymbol{R} + \boldsymbol{Z})$.
   The projection $\pi(\boldsymbol{R})$, $\pi : \mathbb{R}^{3\times3} \rightarrow St$, onto the Stiefel manifold, $St = \{\boldsymbol{R} \in \boldsymbol{R}^{3\times3} : \boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}\}$, is defined as

$$\pi(\boldsymbol{R}) = \underset{\boldsymbol{Q}\in St}{\mathrm{argmin}} \|\boldsymbol{R} - \boldsymbol{Q}\|^2. \tag{38}$$

   If the singular value decomposition of $\boldsymbol{R}$ is $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, then the projection is simply given by [34]

$$\pi(\boldsymbol{R}) = \boldsymbol{U}\boldsymbol{I}_{3\times3}\boldsymbol{V}^T \tag{39}$$

6) If $f(\boldsymbol{R}) \leq f(\boldsymbol{R}')$ then abort.
7) Set $\boldsymbol{R} := \boldsymbol{R}'$. Go to Step 2.

REFERENCES

[1] T. Chen, "Audiovisual speech processing," *IEEE Signal Processing Mag.*, vol. 18, pp. 9–21, 2001.

[2] S. Morishima, K. Aizawa, and H. Harashima, "An intelligent facial image coding driven by speech and phoneme," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP '89)*, pp. 1795–1798, 1989.

[3] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," *Proc. ACM SIGGRAPH '97*, pp. 353–360, 1997.

[4] F. Huang and T. Chen, "Real-time lip-synch face animation driven by human voice," *Multimedia Signal Processing, 1998 IEEE Second Workshop on*, pp. 352–357, 1998.

[5] E. Yamamoto, S. Nakamura, and K. ShiKano, "Lip movement synthesis from speech based on hidden markov models," *Speech Communication*, pp. 105–115, 1998.

[6] M. Brand, "Voice puppetry," *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pp. 21–28, 1999.

[7] P. S. Aleksic and A. K. Katsaggelos, "Speech-to-video synthesis using facial animation parameters," *IEEE Trans. on Circuits and Systems for Video Tech.*, vol. 14, no. 5, pp. 682–692, 2004.

[8] Y. Li and H.-Y. Shum, "Learning dynamic audio-visual mapping with inputoutput hidden markov models," *IEEE Trans. on Multimedia*, vol. 8, no. 3, pp. 542–549, 2006.

[9] J. Xue, J. Borgstrom, J. Jiang, L. Bernstein, and A. Alwan, "Acoustically-driven talking face synthesis using dynamic bayesian networks," in *Proc. of the Int. Conf. on Multimedia and Expo 2006 (ICME 2006)*, 2006, pp. 1165–1168.

[10] L. Valbonesi, R. Ansari, D. McNeill, F. Quek, S. Duncan, K. E. McCullough, and R. Bryll, "Multimodal signal analysis of prosody and hand motion: Temporal correlation of speech and gestures," *Proc. of the European Signal Processing Conference 2002 (EUSIPCO'02)*, vol. 1, pp. 75–78, 2002.

[11] K. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility: Head movement improves auditory speech perception," in *PSYCHOLOGICAL SCIENCE*, vol. 15, no. 2, 2004, pp. 133–137.

[12] F. Quek, D. McNeill, R. Ansari, X. Ma, R. Bryll, S. Duncan, and K. McCullough, "Gesture cues for conversational interaction in monocular video," *ICCV99 Wksp on RATFGRTS*, pp. 64–69, 1999.

[13] T. Kuratate, K. G. Munhall, P. E. Rubin, E. Vatikiotis-Bateson, and H. Yehia, "Audio-visual synthesis of talking faces from speech production correlates," in *Sixth European Conference on Speech Communication and Technology (EUROSPEECH'99)*, 1999, pp. 1279–1282.

[14] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: facial movements accompanying speech," *Proc. of IEEE Int. Conf. Automatic Face and Gesture Recognition*, pp. 381–386, 2002.

[15] E. Chuang and C. Bregler, "Mood swings: expressive speech animation," *ACM Transactions on Graphics*, vol. 24, no. 2, pp. 331–347, 2005.

[16] Z. Deng, C. Busso, S. Narayanan, and U. Neumann, "Audio-based head motion synthesis for avatar-based telepresence systems," in *ACM SIGMM 2004 Workshop on Effective Telepresence (ETP'04)*, 2004, pp. 24–30.

[17] M. E. Sargin, F. Ofli, Y. Yasinnik, O. Aran, A. Karpov, S. Wilson, E. Erzin, Y. Yemez, and A. M. Tekalp, "Gesture-speech correlation analysis and speech driven gesture synthesis," in *Proc. of the Int. Conf. on Multimedia and Expo 2006 (ICME 2006)*, 2006.

[18] M. Naphade and T. Huang, "Discovering recurrent events in video using unsupervised methods," in *Proc. of the Int. Conf. on Image Processing 2002 (ICIP 2002)*, 2002, pp. II: 13–16.

[19] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *IEEE CVPR*, pp. 511–518, 2001.

[20] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," *Proc. of the Int. Conf. on Image Processing 2002 (ICIP'02)*, vol. 1, pp. 900–903, 2002.

[21] J. Y. Bouguet, "Pyramidal implementation of the lucas kanade feature trackerdescription of the algorithm," *Intel Corporation, Microprocessor Research Labs, OpenCVDocuments*, 1999.

[22] M. Brown, D. Burschka, and G. Hager, "Advances in computational stereo," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 8, pp. 993–1008, 2003.

[23] P. Fua, "Combining stereo and monocular information to compute dense depth maps that preserve depth discontinuities," *12th. International Joint Conference on Artificial Intelligence*, pp. 1292–1298.

[24] D. Varshalovich, A. Moskalev, and V. Khersonskii, *Description of Rotation in Terms of the Euler Angles*. Quantum Theory of Angular Momentum. World Scientific, 1988.

[25] K. Shoemake, "Animating rotation with quaternion curves," *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pp. 245–254, 1985.

[26] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Proc. of the Inst. of Phonetic Sciences*, vol. 17, pp. 97–110, 1993.

[27] S. Ananthakrishnan and S. Narayanan, "An Automatic Prosody Recognizer using a Coupled Multi-Stream Acoustic Model and a Syntactic-Prosodic Language Model," *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 1, 2005.

[28] Point Grey Research Inc. http://www.ptgrey.com/.

[29] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "Tobi: A standard for labeling english prosody," in *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, 1992, pp. 867–870.

[30] Momentum Inc. Speech-Driven Talking Head Avatar is available at http://www.momentum-dmt.com/.

[31] Y. Bengio and P. Frasconi, "Input-output HMMs for sequence processing," *Neural Networks, IEEE Transactions on*, vol. 7, no. 5, pp. 1231–1249, 1996.

[32] R. Collobert, S. Bengio, and J. Mariethoz, "Torch: a modular machine learning software library," *IDIAP Research Report*, vol. 2, p. 46, 2002.

[33] Prosody-Driven Head Gesture Animation demonstrations are available at http://mvgl.ku.edu.tr/prosodygesture/.

[34] J. H. Manton, "Optimisation algorithms exploiting unitary constraints," *IEEE Transactions on Signal Processing*, vol. 50, no. 3, pp. 635–650, March 2002.

[35] D. Demirdjian and T. Darrell, "Motion estimation from disparity images," in *Proc. of the Eighth IEEE Int. Conf. on Computer Vision*, vol. 1, 2001, pp. 213–218.

Mehmet Emre Sargın (S'04) received the B.Sc. degree in Electrical and Electronics Engineering from Middle East Technical University, Ankara, Turkey, in 2004, and the M.Sc. degree in Electrical and Computer Engineering from Koc University, Istanbul, Turkey, in 2006. He is currently pursuing the Ph.D. degree in Electrical and Computer Engineering at University of California, Santa Barbara, CA. His research interests include computer vision, pattern recognition, machine learning and bio-image informatics. He is the second prize winner in the ICASSP'07 student paper contest.

Yücel Yemez (M'03) received the B.S. degree from Middle East Technical University, Ankara, Turkey, in 1989, and the M.Sc. and Ph.D. degrees from Boğaziçi University, İstanbul, Turkey, respectively in 1992 and 1997, all in electrical engineering. From 1997 to 2000, he was a postdoctoral researcher in the Image and Signal Processing Department of Télécom Paris (Ecole Nationale Supérieure des Télécommunications). Currently he is an assistant professor of the Computer Engineering Department at Koç University, İstanbul, Turkey. His current research is focused on various fields of computer vision and 3D computer graphics.

Engin Erzin (S'88-M'96-SM'06) received his Ph.D. degree, M.Sc. degree, and B.Sc. degree from the Bilkent University, Ankara, Turkey, in 1995, 1992 and 1990, respectively, all in Electrical Engineering. During 1995-1996, he was a postdoctoral fellow in Signal Compression Laboratory, University of California, Santa Barbara. He joined Lucent Technologies in September 1996, and he was with the Consumer Products for one year as a Member of Technical Staff of the Global Wireless Products Group. From 1997 to 2001, he was with the Speech and Audio Technology Group of the Network Wireless Systems. Since January 2001, he is with the Electrical&Electronics Engineering and Computer Engineering Departments of Koç University, Istanbul, Turkey. His research interests include speech signal processing, pattern recognition and adaptive signal processing.

A. Murat Tekalp (S'80-M'84-SM'91-F'03) received the M.S. and Ph.D. degrees in electrical, computer, and systems engineering from Rensselaer Polytechnic Institute (RPI), Troy, NY, in 1982 and 1984, respectively. He has been with Eastman Kodak Company, Rochester, NY, from December 1984 to June 1987, and with the University of Rochester from July 1987 to June 2005, where he was promoted to Distinguished University Professor. Since June 2001, he has been a Professor at Ko University, Istanbul, Turkey. His research interests are in the area of digital image and video processing, including video compression and streaming, motion-compensated video filtering for high-resolution, video segmentation, content-based video analysis and summarization, 3DTV/video processing and compression, multi-camera surveillance video processing, and protection of digital content. He authored the book Digital Video Processing (Englewood Cliffs, NJ: Prentice- Hall, 1995) and holds seven U.S. patents. His group contributed technology to the ISO/IEC MPEG-4 and MPEG-7 standards.

Dr. Tekalp was named Distinguished Lecturer by the IEEE Signal Processing Society in 1998, and awarded a Fulbright Senior Scholarship in 1999. He received the TUBITAK Science Award (highest scientific award in Turkey) in 2004. He chaired the IEEE Signal Processing Society Technical Committee on Image and Multidimensional Signal Processing (January 1996December 1997). He served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING (1990 to 1992) and the IEEE TRANSAACTIONS ON IMAGE PROCESSING (1994 to 1996), and the Kluwer journal Multidimensional Systems and Signal Processing (1994 to 2002). He was an Area Editor for the Academic Press Journal Graphical Models and Image Processing (1995 to 1998). He was also on the Editorial Board of the Academic Press journal Visual Communication and Image Representation (1995 to 2002). He was appointed as the Special Sessions Chair for the 1995 IEEE International Conference on Image Processing, the Technical Program Co-Chair for IEEE ICASSP 2000 in Istanbul, the General Chair of IEEE International Conference on Image Processing (ICIP) in Rochester in 2002, and Technical Program Co-Chair of EUSIPCO 2005 in Antalya, Turkey. He is the Founder and First Chairman of the Rochester Chapter of the IEEE Signal Processing Society. He was elected as the Chair of the Rochester Section of IEEE for 1994 to 1995. At present, he is the Editor-in-Chief of the EURASIP journal Signal Processing: Image Communication (Elsevier). He is serving as the Chairman of the Electronics and Informatics Group of the Turkish Science and Technology Foundation (TUBITAK) and as an independent expert to review projects for the European Commission.