Generating Human-like Behaviors using Joint, Speech-driven Models for Conversational Agents

Soroosh Mariooryad, *Student Member, IEEE* and Carlos Busso, *Member, IEEE*

***Abstract*—During human communication, every spoken message is intrinsically modulated within different verbal and nonverbal cues that are externalized through various aspects of speech and facial gestures. These communication channels are strongly interrelated, which suggests that generating human-like behavior requires a careful study of their relationship. Neglecting the mutual influence of different communicative channels in the modeling of natural behavior for a conversational agent may result in unrealistic behaviors that can affect the intended visual perception of the animation. This relationship exists both between audiovisual information and within different visual aspects. This paper explores the idea of using joint models to preserve the coupling not only between speech and facial expression, but also within facial gestures. As a case study, the paper focuses on building a speech-driven facial animation framework to generate natural head and eyebrow motions. We propose three *dynamic Bayesian networks* (DBNs), which make different assumptions about the coupling between speech, eyebrow and head motion. Synthesized animations are produced based on the MPEG-4 facial animation standard, using the audiovisual IEMOCAP database. The experimental results based on perceptual evaluations reveal that the proposed joint models (speech/eyebrow/head) outperform audiovisual models that are separately trained (speech/head and speech/eyebrow).**

***Index Terms*—Conversational Agent, Visual Prosody, Facial Animation, Dynamic Bayesian Network**

I. Introduction

SPOKEN language is a sophisticated, multimodal process in which several communicative channels are simultaneously manipulated to externalize verbal and nonverbal cues [1]. Along with the semantic content, we communicate emotions, desires and intentions which are encoded to produce a unifying message. Listeners perceive multiple cues that are communicated, decode each aspect of the message, and make inferences or representations of what is being said. This characterization of vocal communication, which is part of the Brunswikian lens model [2], implies that perception of spoken language is essentially multimodal. This claim is supported by the McGurk effect, which emphasizes the interaction between hearing and vision in the perception of speech [3]. Therefore, the relationships between different acoustic and visual modalities should be carefully considered to create believable *conversational agents* (CAs).

In human interaction, gestures and speech are intricately coordinated to express and emphasize ideas [4]–[6]. The tone and intensity of speech, facial expressions, and head motion are all combined in a nontrivial manner, as they unfold in natural human communication. These communicative channels are not only strongly connected, but also systematically synchronized along different scales (i.e., phonemes, words, phrases and sentences) [4].

The common approach in current CAs relies on sets of rules that are carefully coded after observing human-human interactions [4], [7]–[9]. One drawback of this approach is that the set of rules cannot easily model the rich and complex variability and timing of human behaviors. As users interact with the interface, they are likely to observe repetitive behaviors, reducing the credibility and perception of the system [10]. An alternative approach to synthesize human-like behaviors is the use of data-driven approaches [11]–[13]. Although the algorithms are usually more complex, they have the potential of capturing naturalistic variations of the behaviors [10]. One useful and accessible modality that can be used to drive facial behaviors is speech. Spoken language carries important information beyond the verbal message that a CA engine should capitalize on. A speech-driven CA can have an important role in many areas including audiovisual mobile interfaces for hearing impaired individuals, video conferences in virtual environment, animated feature films, immersive human computer interfaces, gaming and entertainment. In these areas, using large motion capture datasets to create the facial behaviors is expensive, and in many cases unfeasible, since it can be only applied to delicately planned scenarios. The use of speech represents an affordable solution to generating realistic avatars.

In our previous work, a statistical time series framework was used to generate head motion sequences that were temporally synchronized with speech [11], [14], [15]. If other facial gestures are separately generated using speech-driven algorithms, the resulting animation will fail to capture the complex relation within different facial gestures (e.g., gaze, eyebrow, head motion). Therefore, the virtual agent will be perceived unbelievable, even though each facial gesture is perfectly coupled with speech. The proposed study aims to overcome this challenge by using an integrative approach with joint models for head and eyebrow motions driven by speech. The underlying hypothesis is that the proposed integrative models will also capture the relation, timing and phase within different facial gestures, which cannot be easily achieved when the behaviors are separately generated by speech-driven models.

This study analyzes the effect of considering mutual influence between head and eyebrow motions in speech-driven animations. Three DBN models are introduced to incorporate

The authors are with the Multimodal Signal Processing (MSP) laboratory, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: sxm096221@utdallas.edu, busso@utdallas.edu).

different levels of statistical dependency between head and eyebrow motions in speech-driven animation. These joint models are referred to as jDBN in the text. These jDBNs are compared to similar separate models for head and eyebrow motions, with objective measure (canonical correlation analysis) and perceptual evaluations. The experiments show that jDBNs obtain better performance than the ones achieved with separate audiovisual models.

The rest of this paper is organized as follows. Section II describes previous studies on analyzing the relation between audio and visual modalities and common approaches utilized for designing CAs. Section III introduces the database and the framework for generating the animations. The proposed DBNs and the respective inference and learning algorithms are discussed in Section IV. Section V describes the implementation of the proposed facial animation framework. Section VI describes experimental setups for objective and subjective evaluations. It explores the implications of using joint models to generate human-like behaviors for CAs. Finally, Section VII gives the conclusions and future directions of this work.

II. Related Works

Gestures and speech are closely tied in meaning, time, function, development and dissolution [5]. Valbonesi et al. studied this idea, which was referred to in their study as the excitatory hypothesis [6]. They showed that more than 90% of the acoustic events, defined as maximum and minimum of the pitch and RMS energy, occur during hand gesture strokes. In some discourse contexts, about 75% of all clauses are accompanied by gestures [4]. Interestingly, McNeill reported that 90% of all gestures occur when the speaker is uttering something [5]. Our previous study showed that facial expression and speech also present high level of correlation [16]. Vatikiotis-Bateson et al. showed that facial expressions are directly connected with the articulatory production [17]. During spoken language production, the vocal-tract is shaped, and the face is manipulated to reach the articulatory targets, affecting regions quite further from the oral aperture [18]. The intrinsic dependency between gestures and speech suggests that the production of these verbal and nonverbal behaviors are combined at some point, producing a single internal encoding process, sharing the same semantic meaning in different communication channels [4]. These findings suggest that the control system to animate virtual, human-like characters needs to be closely related and synchronized with the information provided by the acoustic signal.

Facial expression and body movements that complement the verbal message are known as visual prosody. Head and eyebrow motions are two important modalities that are highly influenced by the prosodic structure of an utterance. Granström and House reported that eyebrow movement is an important factor to find prominent segments in a sentence in Swedish [19]. Even though there is no direct muscular connection, Cavé et al. indicated that rising and falling eyebrow co-occur with rising in the fundamental frequency [20]. Similarly, Flecha-Garca reported that raising eyebrows usually co-occur with pitch accents [21]. Graf et al. concluded that head and eyebrow motions are consistently correlated with the prosodic structure of speech [22]. The study of Munhall et al. revealed that head motion affects speech intelligibility [23]. From these studies it can be concluded that head and eyebrow motions are two fundamental aspects which should be considered for generating a convincing CA.

One of the common approaches to enrich virtual characters with human-like behaviors consists in designing rules that relate facial and head movements with conversational functions [4], [7]–[9]. Based on detailed analysis of human behaviors, Cassell et al. proposed a rule-based system to generate facial expressions, hand gestures and spoken intonation, which were properly synchronized according to dialogic rules [4]. They extended their work to create a human-like agent, called REA that was able to respond to discourse function using gestures [24]. Pelachaud et al. proposed another rule-based system to synthesize facial expressions and head movement from text responding to discourse function (e.g., conversational signal and punctuators) [25]. They later introduced another embodied CA, called Greta [26]. In addition to facial expressions, Greta can display rich verbal and nonverbal behaviors including gestures, gaze and head movements. DeCarlo et al. presented a coding-based platform for real-time facial animation [27]. The movements were driven by manual annotations of specific head motion gestures co-occurring with prominent words in the text. An important drawback of the rule-based systems is that they cannot easily capture the rich and complex variability observed in facial behaviors [10].

In data-driven approaches, the relation between speech and gestures, especially facial expressions, is analyzed as the results of articulatory processes [11], [12], [14]. Data-driven approaches need a mechanism to map acoustic information to corresponding sequence of visual information. Yehia et al. used linear regression to infer head motion from F0 [28]. Morishima generated sequences of visual information with vector quantization [29]. Rao et al. proposed *Gaussian mixture models* (GMMs), which does not take advantage of the temporal correlation of audiovisual information [30]. To address this limitation, *hidden Markov models* (HMMs) were proposed by Choi et al. [31]. Our previous work presented a data-driven approach to synthesize appropriate head movement by sampling from HMMs [14]. This study showed that using emotion-dependent HMMs to generate appropriate head motion can enhance the emotional content of the animation [11]. Zoric proposed a real-time HMM framework to recognize temporal segments which are likely to contain different facial gestures [32]. The system was driven by prosodic features. Xue et al. adapted DBNs which were previously used for audiovisual speech recognition [33], to the problem of speech-driven facial animation [34]. These DBNs model different statistical dependencies between audio and visual features. Cao et al. proposed a concatenative approach to generate facial expression and synchronized lip motion that considers the lexical and emotional content of the utterance [35].

The aforementioned studies have shown the potential of speech-driven methods to synthesize facial animations. Notice that when more than one facial aspect is considered, the common approach consists in generating each behavior separately.

For example, Yehia et al. proposed separate speech driven models to synthesize facial expression and head motions [28]. However, it is not clear whether the relationship within facial gestures is preserved when these gestures are separately generated from speech (e.g., relation between eyebrow and head motions). It is also not clear whether modeling these relations between facial aspects is important to synthesize human-like behaviors. This is the precise problem considered in this paper. Our results reveal the importance of considering joint statistics of different facial motions to produce natural behaviors driven by speech.

III. Methodology

The section describes the database considered in this study, and the facial and acoustic features used by the proposed framework.

A. IEMOCAP database

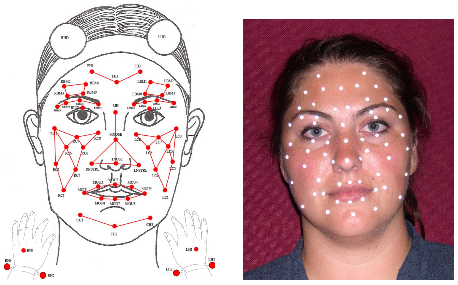
The study relies on the *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) database [36]. This corpus was designed to study expressive human communication. Ten actors were recorded in dyadic interaction in five sessions. In each session, two actors (one male and one female) were asked to perform three plays that were selected to elicit emotional manifestations. In addition, they were asked to improvise hypothetical scenarios such as losing baggage in the airport and getting married. The selected scripts and scenarios aroused different emotions, including happiness, anger, sadness and frustration, which facilitates the understanding and modeling of natural expressive behaviors. The recording includes motion capture technology to collect detailed motion information from the head, face and hands of one of the actors at a time. The markers layout used during the recording is depicted in Figure 1. In total, 53 markers were placed on the face following, in most of the cases, the location of *Feature Points* (FPs) defined in the MPEG-4 facial animation standard [37]. As discussed in section III-B, the placement of the facial markers facilitates the estimation of *facial animation parameters* (FAPs) to produce the animation.

A preprocess step is required to extract the head rotation from the marker data. First, the position of the markers are translated so that the nose marker becomes the center of the coordinate system. Then, a neutral head pose parameterized by a 53×3 matrix with the 3D position of each marker is selected as reference (M_{ref}). A similar matrix is estimated for each frame t (M_t). After calculating the *singular value decomposition* (SVD), UDV^T, of $M_t^T \cdot M_{ref}$, the rotation matrix R_t for each frame t is obtained by the product of VU^T [38]. R_t is used to compensate for rotation. The 3D Euclidean angles are estimated from this rotation matrix.

$$M_t^T \cdot M_{ref} = UDV^T \tag{1}$$

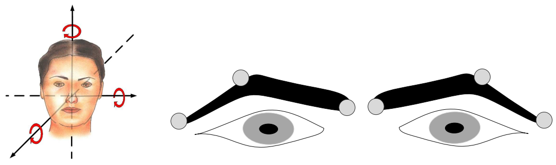
$$R_t = VU^T \tag{2}$$

The database provides synchronized speech, which is used to build audiovisual models for speech-driven animation.



(a) (b)

Fig. 1. IEMOCAP database. (a) Markers layout, (b) Actress with markers attached to her face.



(a) (b)

Fig. 2. (a) 3D head rotation, (b) Six *Feature Points* (FPs) used to generate eyebrow motion.

B. Facial Action Parameters

The proposed parameterization of facial behaviors is based on the MPEG-4 standard for facial animation, which provides specification for efficient coding of shape and animation of human faces [37]. The standard defines a set of 84 *feature points* (FPs) to describe a neutral face. These FPs are controlled by *facial action parameters* (FAPs). Therefore, providing a sequence of FAPs is sufficient to generate a desired animation. The standard defines 68 FAPs including 2 high-level FAPs (visemes and expressions) and 66 low-level FAPs. In most of the cases, an FAP affects a single FP, which simplify their mapping [39]. Generating a believable animation does not require to specify all FAPs. The proposed approach parameterizes the face using only 43 FAPs. As a case study, this paper focuses only on head and eyebrow motions. Three FAPs are dedicated to the Euler angles for head pose, which are depicted in Figure 2(a) (pitch, roll and yaw). Eyebrow motion is controlled by six FAPs, which correspond to the FPs described in Figure 2(b). The rest of the FAPs are responsible for other aspects of the talking face. For the sake of simplicity, these FAPs are derived from the original marker data as explained below.

The dynamics of the face is captured by the marker information. The marker information is mapped into FAP values with the use of linear mappings. It is straightforward to associate FAPs with facial markers, since the placement of most of the markers in the IEMOCAP database follows the location of the FPs defined in the MPEG-4 standard. Each FAP describes the displacement of corresponding FP from its neutral position. Calculating the displacements and, therefore, the FAPs requires a reference neutral pose. This neutral pose is approximated by the average value of the markers over a 1-sec window, in which the actors had a neutral pose. Figure 3 describes the linear transformation to map the range of a

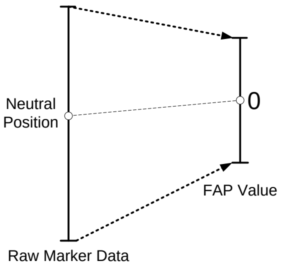


Fig. 3. Linear mapping between markers and corresponding FAPs. A linear function with respect to its neutral position is used to map each marker to its corresponding FAP value.

marker into the range of the corresponding FAP. The value of its neutral pose is mapped to zero. Notice that the range of the marker position is derived from the data. The mapping for unidirectional FAPs is even simpler, since they take only positive values. Although lip synchronization is not perfect, this approach generates good quality animations, as discussed in Section VI-C.

C. Acoustic Features

While there are many aspects of speech that can be used to generate virtual animations, this study only considers speech prosody, extracted with Praat [40]. Based on our previous work, the selected features are the fundamental frequency, RMS energy and their first and second derivatives [11], [16]. Unvoiced segments of speech produce discontinuity in pitch contour. To avoid propagating the discontinuities to the models (and eventually visual information), unvoiced segments are interpolated. The audio data is sampled at 48KHz. The motion capture data provides 120 samples per second. However, the fundamental frequency and energy are estimated every 16.67 milliseconds. Therefore, we upsampled the acoustic features to produce 120 samples per second.

IV. Proposed Dynamic Bayesian Networks

The proposed approach to jointly model speech and facial behaviors is based on *dynamic Bayesian networks* (DBNs). DBN is a powerful and flexible framework to capture temporal dependencies between time series sequences [41]. It models causal relationship between variables, and, therefore, it is an ideal scheme for exploratory data analysis and inferences. Likewise, it offers an efficient and principled approach to avoid overfitting the data. This framework has been applied successfully in several multimodal information processing tasks such as audiovisual speech recognition [33]. DBN not only serves as a modeling tool, but also has a flexible platform for estimating missing values of a system. These properties make DBN a good fit for the problem of speech-driven animation [34].

This paper proposes three DBNs to study the advantages of using joint speech-facial models for head and eyebrow motions (Figure 4(c)-(e)). These three DBNs provide different coupling strategies between speech and facial gestures. Before describing these models, we will introduce the basic structures that are proposed to separately synthesize head (Figure 4(a)) and eyebrow motion (Figure 4(b)) driven by speech. In the Figures 4(a)-(b), the node $Speech$ represents an observed continuous variable. The nodes $Eyebrow$ and $Head$ represent facial gestures which are only observed during training. These nodes are regarded as unobserved variables during synthesis. Since $Head$, $Eyebrow$ and $Speech$ are continuous variables, they are modeled with Gaussian distributions. All the circle nodes depicted in Figure 4, represent conditional Gaussian distributions with full covariance matrix.

The nodes H_h and H_e represent hidden discrete state variables describing the current state or configuration (i.e., similar to the hidden states in HMM). These hidden state variables serve as discrete codebooks of joint behaviors that constrain the possible configurations observed in the speech-head (or speech-eyebrow) space. Therefore, each of the n states of the variables H_h and H_e describes a characteristic coupling between the acoustic and facial gestures. For instance, high pitch values are usually accompanied by raised eyebrows [20]. The discrete hidden variables are expected to capture these relationships across modalities. Notice that we do not constrain any state transition in H_h and H_e (i.e., ergodic models). The causal relations between H_h and the nodes $Speech$ and $Head$ are explicitly imposed in the DBNs (see Figs. 4(a)). Since the variable H_h is hidden, the variables $Speech$ and $Head$ are dependent of each other. For simplicity, modeling the temporal acoustic-facial relationship is limited to the nodes H_h and H_e. Also, it is assumed that the transition probabilities follow the Markov property to obtain computationally tractable models. This assumption is commonly used in speech and facial processing tasks, since speech and facial gestures are inherently non-stationary signals. For inference, the acoustic features are entered into the network as evidences, which are propagated through the network. The expected values of head and eyebrow motions are used to generate the animation. The time unit of the DBNs corresponds to individual frames. Notice that we do not build independent models for different emotional states or phonetic units. This setting offers a number of advantages including avoiding preprocessing steps to infer the underlying phone sequence and emotional content. It also allows us to process continuous utterances frame by frame without introducing delay associated to longer time units.

This paper introduces an integrative approach with joint models to preserve the relation, timing and phase between facial gestures (i.e., eyebrow and head motion). Three DBNs are proposed, which are depicted in Figures 4(c)-(e). These configurations are chosen since they offer various degrees of coupling between the modalities. In Figure 4(c), two separate discrete state variables are considered for each facial gesture (H_h and H_e). These state nodes are related through the $Speech$ node. Once the sequence of acoustic features is known, Bayesian inference updates the marginal probabilities of H_h and H_e. Consequently, $Head$ and $Eyebrow$ become indirectly dependent on the $Speech$ node. Henceforth, this model is referred to as *jDBN-1*. This model provides the weakest link between facial modalities. The second DBN that is proposed has a single joint discrete state variable ($H_{h\&e}$) for both head and eyebrow (Figure 4(d)). Therefore, each state contains coupled information of both facial gestures. This model is referred to as *jDBN-2*. A single hidden state variable in this model provides a direct relation between $Head$ and $Eyebrow$. Hence, *jDBN-2* provides a stronger coupling in comparison to *jDBN-1*. However, the second-order statistics (correlation) between head and eyebrow in each state is ignored. The third DBN considers this correlation by using a single covariance matrix for head and eyebrow, as described in Figure 4(e). This model is referred to as *jDBN-3*.

The *Separate* DBNs, *jDBN-2* and *jDBN-3* are similar to conventional HMMs. However, the variables for facial gestures ($Head$ and $Eyebrow$) are not observed during the synthesis step. Similarly, *jDBN-1* can be regarded as a variation of *factorial HMM* (FHMM) [42]. Here, we adapt the inference and learning algorithms used for HMM and FHMM to train the proposed graphical models. The inference methods take advantage of the well-known forward-backward algorithm and are based on the derivation provided by Murphy [43].

A. Inference

Considering a DBN with a single discrete hidden state variable H with n states, the inference process consists in computing the marginal state probabilities given the observations, $\gamma_t(i) = P(H_t = i \mid y_{1:T})$. $y_{1:T}$ is the set of observed variables (y) from time $t = 1$ to $t = T$, and H_t is the value of the state variable at time t. The parameters of the model include the initial state probability distribution, $\pi(i) \triangleq P(H_1 = i)$, the state transition probabilities, $A_{ij} \triangleq P(H_t = j \mid H_{t-1} = i)$ and the observation probability model, $O_t(i) \triangleq P(y_t \mid H_t = i)$. The forward-backward algorithm recursively computes two variables $\alpha_t(i) \triangleq P(H_t = i \mid y_{1:t})$ and $\beta_t(i) \triangleq P(y_{t+1:T} \mid H_t = i)$ to obtain the marginal probabilities (γ_t) as follows:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum\limits_{j=1}^{n} \alpha_t(j)\beta_t(j)} \tag{3}$$

The initializations of α and β are given in Equations 4 and 5. Equations 6 and 7 give the recursions, which are derived from the definitions of $\alpha_t(i)$ and $\beta_t(i)$.

$$\alpha_1(i) = \frac{O_1(i)\pi(i)}{\sum\limits_{j=1}^{n} O_1(j)\pi(j)} \tag{4}$$

$$\beta_T = 1 \tag{5}$$

$$\alpha_t(i) = \frac{O_t(i) \sum\limits_{k=1}^{n} A_{ki}\alpha_{t-1}(k)}{\sum\limits_{j=1}^{n} O_t(j) \sum\limits_{k=1}^{n} A_{kj}\alpha_{t-1}(k)} \tag{6}$$

$$\beta_t(i) = \frac{\sum\limits_{k=1}^{n} O_{t+1}(k)A_{ik}\beta_{t+1}(k)}{\sum\limits_{j=1}^{n} \sum\limits_{k=1}^{n} O_{t+1}(k)A_{jk}\beta_{t+1}(k)} \tag{7}$$

Separate DBNS, *jDBN-2* and *jDBN-3* have only one hidden state variable (H_h, H_e, $H_{h\&e}$ and $H_{h\&e}$, respectively). Thus, the derivations provided in this section can be used for inference in these models. The observation probability model, $O_t(i)$, depends on the graphical structure of each model and also the set of observed variables in each step. Notice that inference is integral part of both training and synthesis. During the training step, both speech and facial gestures are observable variables. But, in the synthesis step, only speech is observed, which affects $O_t(i)$. Thus, full observation probability model ($O_t^F(i)$) is used during training, while partial observation probability ($O_t^P(i)$) is used during synthesis. For instance, $O_t^F(i)$ and $O_t^P(i)$ for *jDBN-2*, are given in Equations 8 and 9, respectively. Notice that the node H d-separates $Speech$ from $Head$ and $Eyebrow$. Thus, $O_t^P(i)$ is simplified. These probabilities for *Separate* DBNs (figs. 4(a)-(b)) and *jDBN-3* (Figure 4(e)) are easily adapted based on the graphical structure of each model.

$$\begin{aligned} O_t^F(i) =& P(Speech_t \mid H_{h\&e_t} = i) \cdot P(Head_t \mid H_{h\&e_t} = i) \\ & \cdot P(Eyebrow_t \mid H_{h\&e_t} = i) \end{aligned} \tag{8}$$

$$O_t^P(i) = P(Speech_t \mid H_{h\&e_t} = i) \tag{9}$$

The inference procedure for *jDBN-1* is adapted from the algorithm developed by Ghahramani and Jordan for FHMM [42], which is more efficient than converting the DBN into its equivalent HMM with a single state variable. Likewise, observation probabilities are similarly adapted to reflect the *jDBN-1* structure.

B. Learning

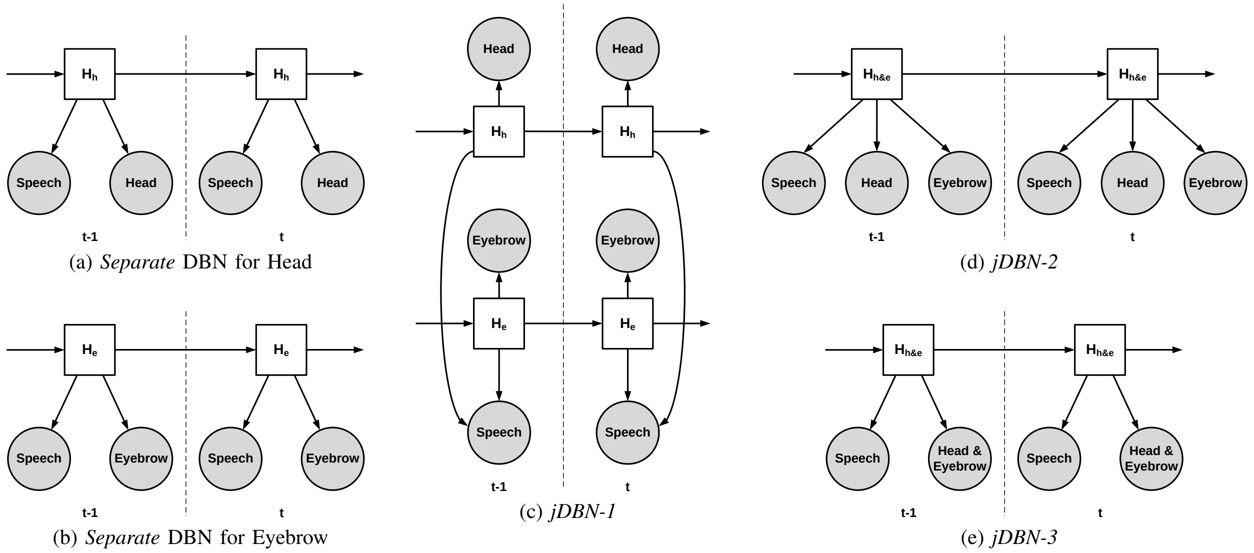
The parameters of the discussed graphical models are estimated via the *expectation maximization* (EM) algorithm. The algorithm computes posterior probabilities during the E step, which is the inference problem described above with respect to all observed variables (speech, head and eyebrow). These probabilities are used to update the model parameters to maximize the expected likelihood of the observations (the M step). The algorithm iterates between the two steps, until convergence. During the M step, Equation 10 updates the initial state probabilities, where L is the number of observation sequences. Equation 11 updates the state transition probabilities. $\xi_t(i,j) \triangleq P(H_t = i, H_{t+1} = j \mid y_{1:T})$ is the probability of being in state i at time t and being in state j at time $t+1$ given the observations. T_l is the length of the l^{th} observation sequence.

$$\hat{\pi}(i) = \frac{\sum\limits_{l=1}^{L} \gamma_1^l(i)}{L} \tag{10}$$

$$\hat{A}_{ij} = \frac{\sum\limits_{l=1}^{L} \sum\limits_{t=1}^{T_l-1} \xi_t^l(i,j)}{\sum\limits_{l=1}^{L} \sum\limits_{t=1}^{T_l-1} \gamma_t^l(i)} \tag{11}$$

$$\xi_t^l(i,j) = \gamma_t^l(i)A_{ij} \tag{12}$$

Similar to the inference step, the graphical structure of each model determines the update rule for the parameters of the observation probabilities. For example, Equations 13 and 14



(a) *Separate* DBN for Head

(b) *Separate* DBN for Eyebrow

(c) *jDBN-1*

(d) *jDBN-2*

(e) *jDBN-3*

Fig. 4. Proposed DBNs for speech-driven facial animation. (a), (b) Separate DBNs which are used as baseline, to generate head and eyebrow motions separately. (c), (d), (e) Joint DBNs to capture interaction of head and eyebrow. *jDBN-1* connects $Head$ and $Eyebrow$ through the $Speech$ node. *jDBN-2* provides a single state variable ($H_{h\&e}$). *jDBN-3* provides a single state space variable ($H_{h\&e}$) and joint covariance matrix for $Head$ and $Eyebrow$.

give the update rule for the mean and covariance for $Head$ in *jDBN-2*. In *jDBN-3*, $Head$ and $Eyebrow$ share a single covariance matrix.

$$\hat{\mu}^i_{Head} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T_l}\gamma_t^l(i)Head_t^l}{\sum_{l=1}^{L}\sum_{t=1}^{T_l}\gamma_t^l(i)} \tag{13}$$

$$\hat{\Sigma}^i_{Head} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T_l}\gamma_t^l(i)(Head_t^l - \mu^i_{Head})(Head_t^l - \mu^i_{Head})^T}{\sum_{l=1}^{L}\sum_{t=1}^{T_l}\gamma_t^l(i)} \tag{14}$$

The M step of FHMM, discussed in Ghahramani and Jordan [42], is adapted to update the parameters of *jDBN-1*.

C. Predicting Facial Gestures

When a sequence of speech features is entered as evidence to the graphical models, the posterior probability of state variables are updated (γ_t). We extract the expected value of the head and eyebrow modalities by summing up the weighted mean across the n states (soft decision). The resulting sequences are used to generate the facial behaviors, which are smoother than the ones achieved by estimating the mean of the most likely state using a dynamic programming algorithm (hard decision). For instance, Equation 15 generates a sequence of head motion, based on the input speech, where $\hat{Head}_t$ is the estimated head pose at time t. The eyebrow motion is predicted in a similar manner.

$$\hat{Head}_t = E[Head_t \mid Speech_{1:T}] = \sum_{i=1}^{n}\gamma_t(i)\mu^i_{Head} \tag{15}$$

V. System Architecture

Figure 5 illustrates the proposed architecture for speech-driven facial animation. Audiovisual models are built during training. During synthesis, acoustic features are entered into the network as evidence. Then, the expected values of facial gestures are used to generate the sequence of FAPs.

Discrete state variables characterize the state space of the audiovisual modalities. It is expected to observe discontinuities in the generated sequences, which will degrade the quality of the animation. Following a similar approach adopted in our previous work, a smoothing technique based on spherical cubic interpolation is used to eliminate discontinuities in head motion [14]. First, the generated sequence is downsampled to 10 frames per second. These frames are referred to as keyframes. The Euler angles are transformed into quaternion [44]. Then, we apply spherical cubic interpolation (squad) to generate the path between the keyframes. The squad function is based on the spherical linear interpolation, slerp. The definition of these functions are shown in equations 16 and 17,

$$slerp(q_1, q_2, \mu) = \frac{sin(1-\mu)\theta}{sin(\theta)}q_1 + \frac{sin\mu\theta}{sin\theta}q_2 \tag{16}$$

$$\begin{aligned} &squad(q_1, q_2, q_3, q_4, \mu) = \\ &\quad slerp(slerp(q_1, q_4, \mu), \quad slerp(q_2, q_3, \mu), 2\mu(1-\mu)) \end{aligned} \tag{17}$$

where q_is are quaternions, $cos\theta = q_1 \cdot q_2$ and μ is a parameter between 0 and 1 that determines the frame position of the interpolated quaternion. By setting specific values for μ, 120 frames per second are generated. Similarly, the eyebrow sequences are down sampled to 10 samples per second. Then, the data is extrapolated using spline to generate a continuous sequence.

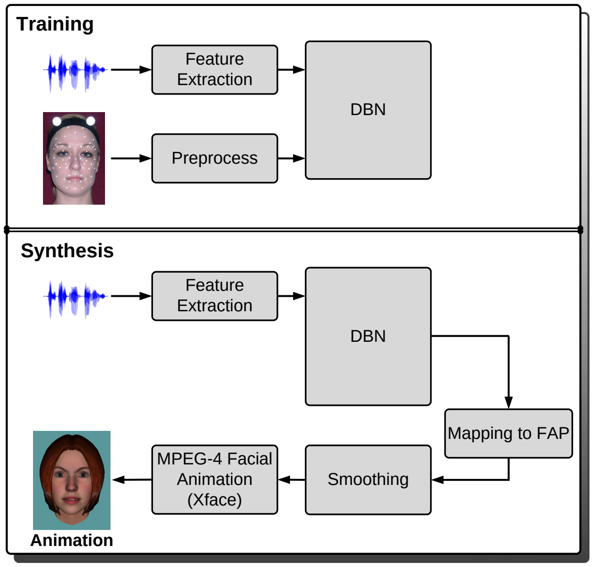


Fig. 5. Proposed system overview. Coupling between acoustic features and facial gestures are captured during training step. Inference process maps each sequence of acoustic feature to its visual counterpart. Finally, synchronized animation is generated with Xface.

The Xface package is used to render the virtual character with FAPs. Xface is an open source MPEG-4 compliant toolkit for 3D virtual talking head [45].

VI. Experiment Results

Objective and subjective metrics are used to evaluate the performance of the proposed audiovisual models. Objective metrics are used to select the optimal configuration for the DBNs (sec VI-A) and to assess whether the synthesized facial behaviors preserve the temporal relationship between the modalities (sec VI-B). Perceptual evaluations are conducted to assess the naturalness of the synthesized behaviors (sec VI-C).

The study relies on *canonical correlation analysis* (CCA) as an objective metric. CCA provides a scale-invariant optimal linear framework to measure the correlation between two streams of data with equal or different dimensions [46]. The basic principle in CCA is to project the features into a common space in which Pearson's correlation can be computed. A higher value of the first-order canonical correlation indicates that the dynamic temporal patterns observed in the streams of data are successfully modeled by the proposed network.

A. Parameter Optimization

Since the aim of the proposed models is to preserve the relationship between facial gestures, we estimated the CCA between the generated head and eyebrow sequences (CCA_{he}). This metric is adopted to select the optimum configuration for the models. As a reference, we estimated the CCA between the original head (3D) and eyebrow (6D) sequences. The average first-order canonical correlation is $r = 0.89$ ($std = 0.086$). This result supports the underlying hypothesis about the high correlation between eyebrow and head motion. If the CCA_{he} approaches this upper bound, we conclude that the synthesized facial behaviors preserve their temporal relationship.

CCA is used to evaluate the performance of different configurations for the proposed DBNs. Using an objective metric to select the optimum configuration allows us to reduce the set of animation sequences that need to be perceptually evaluated. The extensive search for the optimal configuration of the proposed models requires time consuming trainings and testing evaluations. Therefore, the parameter optimization step is performed only on a single subject (the actress in the first session). Notice that we use the data from all the ten subjects in the evaluations. There are 418 utterances from the selected actress which are used for training (75%) and testing (25%).

The configuration of the proposed models is defined by the number of discrete hidden states of H_e, H_h or $H_{h\&e}$ (see Figure 4). To find the optimal value for the discrete hidden states, different configurations have been tested. Figure 6(a) gives the performance in term of CCA_{he} for separate models with different number of hidden states (H_e and H_h). The figure shows that the CCA_{he} is maximized when the number of states for H_h and H_e are set equal to 6 and 12, respectively. The *jDBN-1* model also has separate state variables (H_e and H_h). Figure 6(b) shows that the optimum configuration for this model is 8 and 10 for H_e and H_h, respectively. The models *jDBN-2* and *jDBN-3* have only a single joint discrete hidden variable ($H_{h\&e}$). Figure 6(c) shows that the models maximize the CCA_{he} when the number of states is set to 14, for *jDBN-2*, and 12, for *jDBN-3*. The aforementioned configurations for each of the models are used in the rest of the experiments (see Tables I and III).

B. Objective Evaluation

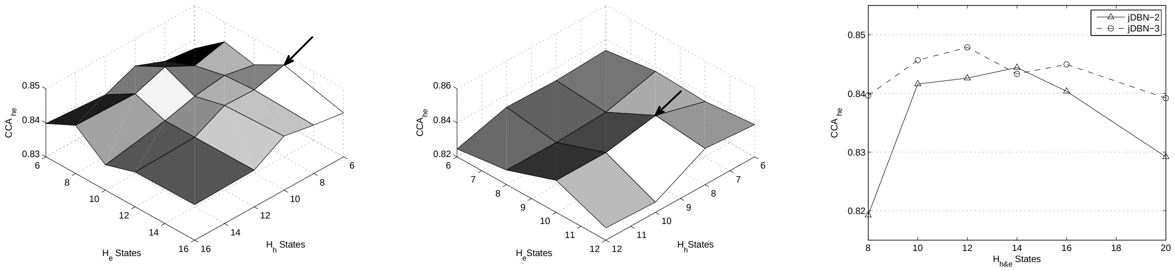
CCA is also used to compare the performance of the proposed models. In addition to CCA_{he}, the analysis includes the CCA between the synthesized and original sequences for head motion, CCA_h, the CCA between the synthesized and original sequences for eyebrow motion, CCA_e, the CCA between speech features and synthesized head sequences, CCA_{hs}, and the CCA between speech features and synthesized eyebrow sequences, CCA_{es}.

1) Speaker Dependent Experiment: Table I gives the results of the objective evaluation for each of the proposed models. This experiment is independently conducted for each of the ten speakers in the IEMOCAP database. The reported results correspond to the average values across all speakers. For each subject, the models were trained with 75% of the data and tested with the rest 25%. The partitions were randomly selected. This table also provides the number of parameters that are needed for each model with the selected configurations ($Param\#$).

Table I indicates that the coupling between the synthesized eyebrow and head motion sequences increases from *jDBN-1* to *jDBN-3* (see column CCA_{he}). *jDBN-3* achieved stronger coupling than the *Separate* models. The CCA between the synthesized and original head motions (CCA_h) follows a similar pattern. The *Separate* models have the highest CCA between speech and eyebrow motion (CCA_{es}). Notice that joint facial models still achieve high correlation levels between speech and facial behaviors. These correlations are higher than

TABLE I
OBJECTIVE EVALUATION OF THE PROPOSED MODELS (CCA_{he}=CCA BETWEEN SYNTHESIZED HEAD AND EYEBROW, CCA_{h}=CCA BETWEEN SYNTHESIZED AND ORIGINAL HEAD, CCA_{e}=CCA BETWEEN SYNTHESIZED AND ORIGINAL EYEBROW, CCA_{hs}=CCA BETWEEN SPEECH AND SYNTHESIZED HEAD, CCA_{es}=CCA BETWEEN SPEECH AND SYNTHESIZED EYEBROW). REPORTED VALUES FOR *Original* ARE COMPUTED FROM MARKER DATA.

$Model$	$Param\#$	$States\#$			CCA_{he}	CCA_{h}	CCA_{e}	CCA_{hs}	CCA_{es}
		$H_{h\&e}$	H_h	H_e	Mean(STD)	Mean(STD)	Mean(STD)	Mean(STD)	Mean(STD)
Separate DBNs	1042	-	6	12	0.84(0.120)	0.78(0.148)	0.79(0.131)	0.93(0.059)	0.91(0.074)
jDBN-1	2664	-	8	10	0.84(0.119)	0.77(0.151)	0.77(0.132)	0.91(0.067)	0.86(0.108)
jDBN-2	1077	14	-	-	0.84(0.119)	0.78(0.145)	0.79(0.129)	0.93(0.060)	0.90(0.076)
jDBN-3	1115	12	-	-	0.85(0.118)	0.79(0.142)	0.79(0.128)	0.93(0.093)	0.89(0.082)
Original	-	-	-	-	0.89(0.086)	-	-	0.78(0.130)	0.71(0.164)



(a) (b) (c)

Fig. 6. Optimizing number of hidden states for the proposed methods with respect to CCA_{he}. (a) Separate Models. (b) *jDBN-1*. (c) *jDBN-2*, *jDBN-3*

the values estimated from speech and the original sequences derived from the markers (see the last row in Table I). This table also shows that the correlation levels for *jDBN-2* and *jDBN-3* are equal or higher than the ones for *jDBN-1* for all of the proposed metrics. Also, the use of joint covariance for head and eyebrow in *jDBN-3* produces a slightly higher CCA_{he} than the one achieved by the *jDBN-1* and *jDBN-2*.

Table II gives the average processing time, in seconds, for the inference and learning procedures. The reported values for inference are the average time across all sequences. The learning time is the total time required to train the DBNs for each subject. This table indicates that multiple hidden state variables (*jDBN-1*) significantly increases the time complexity of the algorithms. While training the models is usually time consuming, Table II shows that the time of the inference algorithm is reasonable for real-time applications. Notice that we have not attempted to implement the approach in real time, since this is not the scope of the paper (the current implementation of the inference algorithm uses a forward-backward scheme, which is not causal). However, the inference algorithm can be replaced by a simple forward filtering approach, which will give the causal estimates of head and eyebrow motions. Notice that this algorithm requires less processing time compared to the forward-backward method.

TABLE II
PROCESSING TIME REQUIRED FOR INFERENCE AND LEARNING OF THE PROPOSED DBNS.

$Model$	$Time(sec)$	
	Inference	Learning
Separate DBNs (Head)	0.044	3120.34
Separate DBNs (Eyebrow)	0.082	6229.56
jDBN-1	1.124	63407.71
jDBN-2	0.102	10957.38
jDBN-3	0.086	6721.83

2) Speaker Independent Experiment: This section explores whether the proposed models generalize when speech from another subject is used to generate the sequences. We train the models for each subject using the same partition used in Section VI-B1. Then, we generate the eyebrow and head motion sequences using the speech from the other 9 subjects (mismatched conditions). Since we use data from multiple subjects, a key preprocessing step is to normalize the acoustic and facial features to reduce inter-speaker variability. We follow a normalization scheme motivated by our previous study [47]. The main idea of the approach is that the properties observed in neutral sentences across the subjects should match.

Table III gives the results of speaker independent experiment. The results follow similar patterns to the ones observed in speaker dependent experiments (compare Table I with Table III). We concluded that the proposed DBNs cope with speaker variability. Therefore, they can be employed to generate facial behaviors from speech of any individual.

C. Subjective Evaluation

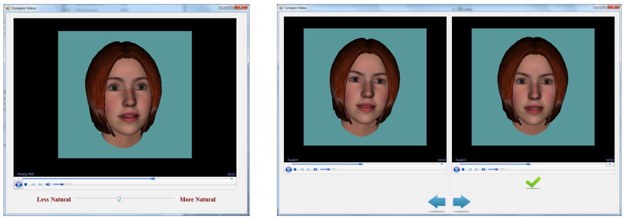
CCA provides one approach to assess the performance of the proposed models. A complementary approach based on perceptual evaluations can bring new insights about the effect of joint modeling of speech and facial gestures. The study includes perceptual evaluations to compare how natural the behaviors generated with the proposed DBNs are perceived.

The IEMOCAP database was recorded from five male and five female subjects. Presenting an animation with a male voice and a female avatar will affect the natural perception of the animation. To avoid this artifact, we could use a different

TABLE III

OBJECTIVE EVALUATION OF THE PROPOSED MODELS IN A SPEAKER INDEPENDENT EXPERIMENT (CCA_{he}=CCA BETWEEN SYNTHESIZED HEAD AND EYEBROW, CCA_h=CCA BETWEEN SYNTHESIZED AND ORIGINAL HEAD, CCA_e=CCA BETWEEN SYNTHESIZED AND ORIGINAL EYEBROW, CCA_{hs}=CCA BETWEEN SPEECH AND SYNTHESIZED HEAD, CCA_{es}=CCA BETWEEN SPEECH AND SYNTHESIZED EYEBROW). REPORTED VALUES FOR *Original* ARE COMPUTED FROM MARKER DATA.

$Model$	CCA_{he}	CCA_h	CCA_e	CCA_{hs}	CCA_{es}
	Mean(STD)	Mean(STD)	Mean(STD)	Mean(STD)	Mean(STD)
Separate DBNs	0.84(0.119)	0.78(0.145)	0.79(0.128)	0.93(0.061)	0.91(0.072)
jDBN-1	0.84(0.121)	0.77(0.149)	0.77(0.130)	0.90(0.070)	0.86(0.109)
jDBN-2	0.84(0.119)	0.78(0.147)	0.79(0.130)	0.93(0.061)	0.90(0.075)
jDBN-3	0.84(0.118)	0.78(0.150)	0.79(0.129)	0.93(0.062)	0.90(0.079)
Original	0.89(0.086)	-	-	0.78(0.130)	0.71(0.164)



(a) (b)

Fig. 7. Graphical user interfaces for perceptual evaluation. (a) Interface used to rate how natural the animations are perceived, (b) Interface used for comparing two animations.

facial model for male and female subjects. However, potential differences in the intrinsic quality of the facial models will introduce another variable into the analysis. For these reasons, this section reports results on the data recorded from the five female subjects. The evaluation also considers speaker dependent and speaker independent experiments.

1) Speaker Dependent Experiment: For the experiment reported in this section, the models were separately trained and tested with data from the same speaker (matching condition). For each subject, an emotionally-balanced subset of the test set was randomly selected for subjective evaluations. This subset includes five sentences that are labeled with the emotions sadness, happiness, anger, frustration and neutral states. In total, it gives 25 sentences for the five subjects. For each sentence, eight animations are generated: *Original*, *No Eyebrow*, *No Head*, *Separate*, *Only Head*, *jDBN-1*, *jDBN-2* and *jDBN-3*. The *Original* sequences are directly generated from the markers, including head and eyebrow motions. These sequences are used as reference. The *No Eyebrow* and *No Head* sequences are generated from the markers, but without eyebrow and head motions, respectively. These sequences are used to study the importance of eyebrow and head motions on the perceived quality of the animation. In the *Separate* sequences, the audiovisual models described in Figures 4(a)-(b) are used to separately synthesize eyebrow and head motions. The *Only Head* keeps the eyebrows still and generates the head motion with the separate model for head movement only (Figure 4(a)). Finally, the *jDBN-1*, *jDBN-2* and *jDBN-3* sequences convey the behaviors generated with the proposed integrative audiovisual DBNs (Figure 4(c)-(e)). Two separate evaluations are conducted to assess (i) how natural the animations are perceived, and (ii) the preferences of the evaluators between DBNs.

In the first experiment, 20 evaluators assessed how natural the talking head is perceived using a slider that it is mapped into the range 0 to 5 (less natural – more natural). Figure 7(a) displays the interface for performing this evaluation. Eight animations for each of the 25 utterances result in 200 animations per subject. The set of animations was presented to the subjects in two sessions with a break between them (100 animations per session). The presentation of the videos was randomized for each subject to avoid bias.

Figure 8 shows the result of the first evaluation. The average quality for the *Original* sequences is 3.4. Although some evaluators indicated that lips motion was not perfectly synchronized with speech, this value is similar to other perceptual evaluations with *Original* sequences [11]. Therefore, we conclude that the mapping from facial markers to FAPs proposed in Section III-B is effective to animate a CA. Figure 8 shows that removing eyebrow motion does not change the natural perception of the animation. However, the naturalness perception is significantly affected when head motion is not synthesized (from 3.4 to 2.32). This finding agrees with the perceptual evaluation conducted in our previous study [11].

Figure 8 provides a subjective metric to compare the animations generated with separate and joint models. One way *Analysis of Variance* (ANOVA) evaluation indicates that there are significant differences between the scores associated to different approaches ($F[7, 3992] = 46.85$, $p < 1e-63$). Multiple pairwise comparison tests show that the scores for *jDBN-3* and *Separate* are significantly different at p-value < 0.0185. We conclude that the differences in quality between *Separate* and *jDBN-3* models are statistically significant. The statistical tests also show that the scores of *jDBN-3* are not significantly different from the *Original* sequence (p-value > 0.2). This result indicates that the sequences generated with *jDBN-3* are perceived as natural as the ones generated with the *Original* sequences. However, the same test shows a significant drop in quality from *Original* to *Separate* models (p-value $< 1e-8$). These evaluations show the superiority of the proposed *jDBN-3* to generate facial animations. Notice that *Separate* and *Only Head* conditions received similar scores (see Figure 8). The differences were not statistically significant. We conclude that separate models do not improve the perceptual quality by independently including the speech driven facial behaviors. However, the perceptual quality significantly increases when eyebrow motion is jointly modeled with head motion (see *jDBN-3* in Figure 8).

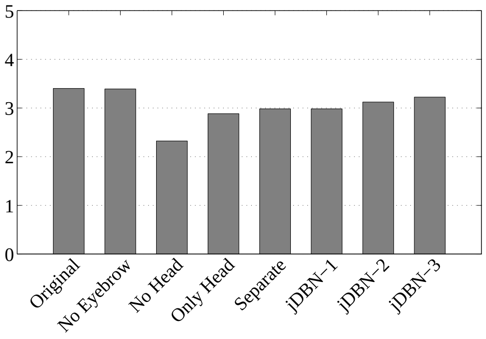


Fig. 8. Results for perceptual evaluation to assess how natural the animations are perceived. For each approach, the figure shows the average values across evaluator, where 0 corresponds to less natural, and five corresponds to more natural (see Figure 7(a)).

Another interesting result is that the perception of naturalness improves when stronger dependencies are included in the models (from *jDBN-1* to *jDBN-3*). According to Table II, *jDBN-3* requires less processing time for inference and learning. It also obtained the highest score among the joint DBNs (see Figure 8). Based on these observations, we selected *jDBN-3* as the most suitable joint model for this problem. This result is expected since *jDBN-3* has shared hidden state variable with full covariance for head and eyebrow. These capabilities are useful for preserving the coupling between facial gestures.

During the second experiment, two facial animations with the same speech content are presented, and 20 evaluators are asked to select the one that displays more realistic behaviors. Since two conditions for the same sentence are directly compared, the results of this experiment will identify the most attractive model to synthesize facial behaviors. To reduce the time of the evaluation, only animations generated with *jDBN-3* are compared with the ones generated with *Separate* and *Original* sequences. The user interface used for this subjective evaluation is shown in Figure 7(b). With 3 animations per sentence (*Original*, *Separate* and *jDBN-3*), there are 3 possible permutations. This generates 75 comparisons (25 sentences), which are presented to the evaluators in two sessions. The order of the presentation and the position of the videos in the GUI (left or right) were randomized for each evaluator.

Figure 9 gives the average results for the preference evaluation. Figure 9(a) shows that when the animations generated with *jDBN-3* and *Separate* models were compared, the raters selected the joint models 57.6% of the times. The 95% confidence interval for this proportion is $[54\%, 61.2\%]$. Since the value 50% is not included in the interval, we concluded that the number of subjects that preferred the animations generated with *jDBN-3* is statistically higher than the ones that chose the animations generated with *Separate* models. In fact, the 99.9$\%$ confidence interval for this comparison is [50.8$\%$, 64.4%] which is still statistically significant. These findings, which agree with the results shown in Figure 8, confirm our hypothesis that facial gestures need to be jointly modeled with speech to preserve their timing, phase and relation. When the animations with the synthesized sequences were compared with the ones conveying the original sequences, the raters showed a clear preference for the original sequences. Notice that the gap is reduced when joint models are used to synthesize the animations (see Figs. 9(b) and 9(c)).

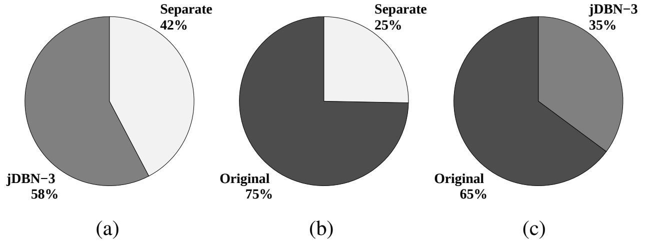


Fig. 9. Perceptual evaluation results (preference). (a) Comparing animations generated with *Separate* and *Original*. (b) Comparing animations generated with *jDBN-3* and *Original*. (c) Comparing animations generated with *Separate* and *jDBN-3* sequences.

2) Speaker Independent Experiment: This section gives the results of subjective evaluations with mismatch between the subject used for training and the subject used for testing (mismatched condition). For each of the five speakers, one of the utterances used in the previous experiment was randomly selected. Then, the head and eyebrow motions for the utterance are generated with the *jDBN-3* models trained with the other four speakers. Therefore, 20 animations are generated with mismatched conditions (5 subject × 1 sentence × 4 mismatched conditions). This set was randomly presented among the animations with matched conditions, used in the first part of the subjective evaluations (see Sec VI-C1). Figure 10 compares the average scores of *jDBN-3* with and without mismatched conditions. Notice that the number of animations with mismatch is four times the number of animations without mismatch. The ANOVA test indicates that the difference between the perceived naturalness of the animations with and without mismatched conditions is not statistically significant ($F(1, 498) = 0.32$, $p = 0.5705$). We believe that the normalization schemes for acoustic and facial features adopted in the paper (Section VI-B2) were effective in reducing speaker variability. Therefore, the generated sequences are perceived as natural as the ones generated without speaker mismatches. Notice that in contrast to other tasks such as speaker/speech recognition, there is not a single correct ground truth solution for the speech-driven facial animation problem (i.e., many different sequences can be perceived natural). Therefore, the behaviors learned from different speakers can be used to generate facial behavior for mismatched speakers. This is an important result, which indicates that the approach is suitable for the applications mentioned in Section I.

VII. Discussion and Conclusion

This study proposed DBNs to generate animations driven by speech. The DBNs aimed to jointly model the interrelations not only between speech and facial gestures, but also within facial gestures. In particular, three DBNs with different coupling strategies are proposed to synthesize eyebrow and head motions. The objective and subjective evaluations confirm that animations generated with the proposed joint models achieve

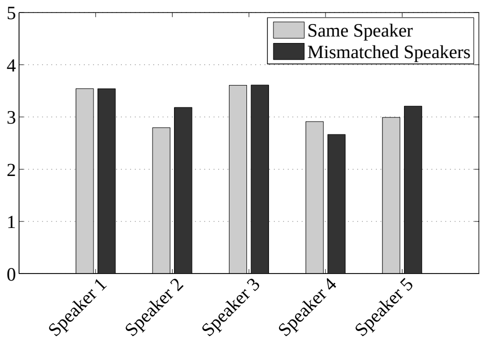


Fig. 10. Results for perceptual evaluation with mismatched conditions for training and testing of *jDBN-3*. For each subject, the figure shows the average values across evaluators, with and without mismatched condition.

more natural facial behaviors than the ones generated by separate audiovisual models. This result suggests that facial gestures need to be jointly modeled with speech to preserve their timing, phase and relation. Also, the robustness of the proposed models to generalize in mismatched conditions (train model from one subject and testing with another) is confirmed through objective and subjective evaluations.

Figure 8 shows that evaluators perceived the *jDBN-3* and *Original* animations with similar level of naturalness. When they were asked to select one of them, however, they preferred the original sequences (Figure 9). These results indicate that the proposed joint models can be improved. The proposed DBNs are only driven by acoustic prosodic features. Other features such as *Mel frequency cepstral coefficients* (MFCCs) can be integrated in the proposed model to generate other aspects of the face such as lip motions and eye gaze. Besides, there are important aspects in spoken language that directly influence the verbal and nonverbal behaviors displayed during human interaction [4] [5] (e.g., head nod for yes, and raising eyebrows for surprise). Therefore, by identifying the underlying semantic structure in the sentence, we can impose constraints in our integrative model to consider the relation between facial gestures and high-level linguistic functions such as dialog acts and emotions. However adding these aspects will increase the complexity of the network both in the number of parameters and the processing time for training and inference. We are exploring different approaches and simplifications to scale the models while keeping a tractable DBN. For example, a linear transformation can be estimated to decorrelate the acoustic and facial features. This transformation can be implemented using *principal component analysis* (PCA). Therefore, the full covariance matrix used in the proposed DBNs can be replaced by a diagonal matrix after applying this linear transformation. This approach will significantly reduce the number of parameters in the models. These are some of our future work to synthesize believable human behaviors driven by speech.

REFERENCES

[1] C. Busso and S. Narayanan, "Joint analysis of the emotional fingerprint in the face and speech: A single subject study," in *International Workshop on Multimedia Signal Processing (MMSP 2007)*, Chania, Crete, Greece, October 2007, pp. 43–47.

[2] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, April 2003.

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

[4] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Bechet, B. Douville, S. Prevost, and M. Stone, "Animated conversation: Rule-based generation of facial expression gesture and spoken intonation for multiple conversational agents," in *Computer Graphics (Proc. of ACM SIGGRAPH'94)*, Orlando, FL,USA, 1994, pp. 413–420.

[5] D. McNeill, *Hand and Mind: What gestures reveal about thought*. Chicago, IL, USA: The University of Chicago Press, 1992.

[6] L. Valbonesi, R. Ansari, D. McNeill, F. Quek, S. Duncan, K. McCullough, and R. Bryll, "Multimodal signal analysis of prosody and hand motion: Temporal correlation of speech and gestures," in *European Signal Processing Conference (EUSIPCO 02)*, Tolouse, France, September 2002, pp. 75–78.

[7] J. Cassell, T. Bickmore, L. Campbell, H. Vilhjálmsson, and H. Yan, "More than just a pretty face: Conversational protocols and the affordances of embodiment," *Knowledge-Based Systems*, vol. 14, pp. 55–64, March 2001.

[8] E. Marsi and F. van Rooden, "Expressing uncertainty with a talking head," in *Workshop on Multimodal Output Generation (MOG 2007)*, Aberdeen, Scotland, January 2007, pp. 105–116.

[9] I. Poggi and C. Pelachaud, "Performative facial expressions in animated faces," in *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, Eds. Cambridge, MA, USA: MIT Press, 2000, p. 154188.

[10] M. Foster, "Comparing rule-based and data-driven selection of facial displays," in *Workshop on Embodied Language Processing, Association for Computational Linguistics*, Prague, Czech Republic, June 2007, pp. 1–8.

[11] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1075–1086, March 2007.

[12] J. Cassell, Y. Nakano, T. Bickmore, C. Sidner, and C. Rich, "Non-verbal cues for discourse structure," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL2001)*, Toulouse, France, July 2001, pp. 114–123.

[13] M. Stone, D. DeCarlo, I. Oh, C. Rodriguez, A. Stere, A. Lees, and C. Bregler, "Speaking with hands: Creating animated conversational characters from recordings of human performance," *ACM Transactions on Graphics (TOG)*, vol. 23, pp. 506–513, August 2004.

[14] C. Busso, Z. Deng, U. Neumann, and S. Narayanan, "Natural head motion synthesis driven by acoustic prosodic features," *Computer Animation and Virtual Worlds*, vol. 16, no. 3-4, pp. 283–290, July 2005.

[15] ——, "Learning expressive human-like head motion sequences from speech," in *Data-Driven 3D Facial Animations*, Z. Deng and U. Neumann, Eds. Surrey, United Kingdom: Springer-Verlag London Ltd, 2007, pp. 113–131.

[16] C. Busso and S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: a single subject study," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2331–2347, November 2007.

[17] E. Vatikiotis-Bateson, K. Munhall, Y. Kasahara, F. Garcia, and H. Yehia, "Characterizing audiovisual information during speech," in *Fourth International Conference on Spoken Language Processing (ICSLP 96)*, vol. 3, Philadelphia, PA, USA, October 1996, pp. 1485–1488.

[18] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1-2, pp. 23–43, 1998.

[19] B. Granström and D. House, "Audiovisual representation of prosody in expressive speech communication," *Speech Communication*, vol. 46, no. 3-4, pp. 473–484, July 2005.

[20] C. Cavé, I. Guaïtella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser, "About the relationship between eyebrow movements and F0 variations," in *International Conference on Spoken Language (ICSLP)*, vol. 4, Philadelphia, PA, USA, October 1996, pp. 2175–2178.

[21] M. L. Flecha-García, "Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English," *Speech Communication*, vol. 52, pp. 542–554, June 2010.

[22] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: Facial movements accompanying speech," in *Proc. of IEEE International Conference on Automatic Faces and Gesture Recognition*, Washington, D.C., USA, May 2002, pp. 396–401.

[23] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, “Visual prosody and speech intelligibility: Head movement improves auditory speech perception,” *Psychological Science*, vol. 15, no. 2, pp. 133–137, February 2004.

[24] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjalmsson, and H. Yan, “Embodiment in conversational interfaces: Rea,” in *International Conference on Human Factors in Computing Systems (CHI-99)*, Pittsburgh, PA, USA, May 1999, pp. 520–527.

[25] C. Pelachaud, N. Badler, and M. Steedman, “Generating facial expressions for speech,” *Cognitive Science*, vol. 20, no. 1, pp. 1–46, January 1996.

[26] E. Bevacqua, M. Mancini, R. Niewiadomski, and C. Pelachaud, “An expressive ECA showing complex emotions,” in *Proceedings of the Artificial Intelligence and Simulation of Behaviour (AISB 2007) Annual Convention*, Newcastle, UK, April 2007, pp. 208–216.

[27] D. DeCarlo, C. Revilla, M. Stone, and J. Venditti, “Making discourse visible: coding and animating conversational facial displays,” in *Computer Animation (CA 2002)*, Geneva, Switzerland, June 2002, pp. 11–16.

[28] H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, “Facial animation and head motion driven by speech acoustics,” in *5th Seminar on Speech Production: Models and Data*, Kloster Seeon, Bavaria, Germany, May 2000, pp. 265–268.

[29] S. Morishima and H. Harashima, “A media conversion from speech to facial image for intelligent man-machine interface,” *IEEE Journal on Selected Areas in Communications*, vol. 9, pp. 594–600, May 1991.

[30] R. Rao, T. Chen, and R. Mersereau, “Audio-to-visual conversion for multimedia communication,” *IEEE Transactions on Industrial Electronics*, vol. 45, pp. 15–22, February 1998.

[31] K. Choi, Y. Luo, and J. Hwang, “Hidden Markov model inversion for audio-to-visual conversion in an MPEG-4 facial animation system,” *The Journal of VLSI Signal Processing*, vol. 29, pp. 51–61, August 2001.

[32] G. Zoric, “Hybrid approach to real-time speech driven facial gesturing of virtual characters,” Ph.D. dissertation, University of Zagreb, July 2010.

[33] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, “Dynamic Bayesian networks for audio-visual speech recognition,” *EURASIP Journal on Applied Signal Processing*, vol. 2002, pp. 1274–1288, January 2002.

[34] J. Xue, J. Borgstrom, J. Jiang, L. Bernstein, and A. Alwan, “Acoustically-driven talking face synthesis using dynamic Bayesian networks,” in *IEEE International Conference on Multimedia and Expo (ICME 2006)*, Toronto, ON, Canada, July 2006, pp. 1165–1168.

[35] Y. Cao, W. Tien, P. Faloutsos, and F. Pighin, “Expressive speech-driven facial animation,” *ACM Transactions on Graphics*, vol. 24, pp. 1283–1302, October 2005.

[36] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[37] I. Pandzic and R. Forchheimer, *MPEG-4 Facial Animation - The standard, implementations and applications*. John Wiley & Sons, November 2002.

[38] K. Arun, T. Huang, and S. Blostein, “Least-squares fitting of two 3-d point sets,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 5, pp. 698–700, September 1987.

[39] N. Sarris, N. Grammalidis, and M. Strintzis, “FAP extraction using three-dimensional motion estimation,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, pp. 865–876, October 2002.

[40] P. Boersma and D. Weeninck, “Praat, a system for doing phonetics by computer,” Institute of Phonetic Sciences of the University of Amsterdam, Amsterdam, Netherlands, Technical Report 132, 1996, http://www.praat.org.

[41] F. Jensen and T. Nielsen, *Bayesian networks and decision graphs*. New York, NY, USA: Springer Verlag, November 2010.

[42] Z. Ghahramani and M. I. Jordan, “Factorial hidden Markov models,” *Machine Learning*, vol. 29, pp. 245–273, November 1997.

[43] K. Murphy, “Dynamic Bayesian networks: Representation, inference and learning,” Ph.D. dissertation, University of California Berkely, Fall 2002.

[44] D. Eberly, *3D Game Engine Design: A Practical Approach to Real-Time Computer Graphics*. San Francisco, CA, USA: Morgan Kaufmann Publishers, 2000.

[45] K. Balci, “Xface: MPEG-4 based open source toolkit for 3D facial animation,” in *Conference on Advanced Visual Interfaces (AVI 2004)*, Gallipoli, Italy, May 2004, pp. 399–402.

[46] C. Dehon, P. Filzmoser, and C. Croux, “Robust methods for canonical correlation analysis,” in *Data Analysis, Classification, and Related Methods*, Springer-Verlag, Berlin, 2000, pp. 321–326.

[47] C. Busso, S. Lee, and S. Narayanan, “Analysis of emotionally salient aspects of fundamental frequency for emotion detection,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 582–596, May 2009.

Soroosh Mariooryad (S’2012) received his B.S degree (2007) with high honors in computer engineering from Ferdowsi University of Mashhad, and his M.S degree (2010) in computer engineering (artificial intelligence) from Sharif University of Technology (SUT), Tehran, Iran. He is currently pursuing his Ph.D. degree in Electrical Engineering at the University of Texas at Dallas (UTD), Richardson, Texas, USA. From 2008 to 2010, he was a member of the Speech Processing Lab (SPL) at SUT. In 2010, he joined as a research assistant the Multimodal Signal Processing (MSP) laboratory at UTD. His research interests are in speech and video signal processing, probabilistic graphical models and multimodal interfaces. His current research includes modeling and analyzing human non-verbal behaviors, with applications to speech-driven facial animations and emotion recognition. He has also worked on statistical speech enhancement and fingerprint recognition.

Carlos Busso (S’02-M’09) is an Assistant Professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He received his B.S (2000) and M.S (2003) degrees with high honors in electrical engineering from University of Chile, Santiago, Chile, and his Ph.D (2008) in electrical engineering from University of Southern California (USC), Los Angeles, USA. Before joining UTD, he was a Postdoctoral Research Associate at the Signal Analysis and Interpretation Laboratory (SAIL), USC. He was selected by the School of Engineering of Chile as the best Electrical Engineer graduated in Chile in 2003. At USC, he received a Provost Doctoral Fellowship from 2003 to 2005 and a Fellowship in Digital Scholarship from 2007 to 2008. He received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain). He is the co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. At UTD, he leads the Multimodal Signal Processing (MSP) laboratory [http://msp.utdallas.edu]. His research interests are in digital signal processing, speech and video processing, and multimodal interfaces. His current research includes the broad areas of affective computing, multimodal human-machine interfaces, modeling and synthesis of verbal and nonverbal behaviors, sensing human interaction, and machine learning methods for multimodal processing.