SPEECH-DRIVEN EYEBROW MOTION SYNTHESIS WITH CONTEXTUAL MARKOVIAN MODELS

Yu Ding[†] *Mathieu Radenen* [‡] *Thierry Artières*[‡] *Catherine Pelachaud*[†]

[†] Université Pierre et Marie Curie (LIP6), Paris, France
[‡] CNRS-LTCI, Institut Mines-TELECOM TELECOM ParisTech, Paris, France

ABSTRACT

Nonverbal communicative behaviors during speech are important to model a virtual agent able to sustain a natural and lively conversation with humans. We investigate statistical frameworks for learning the correlation between speech prosody and eyebrow motion features. Such methods may be used to synthesize automatically accurate eyebrow movements from synchronized speech.

Index Terms— Hidden Markovian models, speech to motion synthesis, virtual agent

1. INTRODUCTION

Embodied conversational agents are autonomous entities with often a human-like appearance (see Figure 1) and endowed with communicative and expressive capabilities. As humans do they can communicate through various means such as speech, facial expressions, gesture, gaze, etc. Communicative behaviors are polysemic that is a same behavior may convey several meanings. For example, a head nod can convey agreement, mark an emphasis, or be a backchannel signal. Various studies have shown the tight relationship between speech and nonverbal behaviors production. For example, [1] found a strong correlation between the raise of F0 and of eyebrow movements. Nonverbal behaviors are important not only for the speaker as they are a mean of encoding her thoughts but also for the interlocutors that can perceive and decode these signals. Virtual humans ought to be capable of displaying such high quality behaviors. Our aim is to develop a model that drives the virtual agent's behaviors from the speech. It takes as input the spoken text the agent needs to say. It computes the facial expressions and other behaviors associated with the acoustic stream. While such an animation model does not rely on semantic information, the acoustic stream is by itself a reflector of communicative intentions. There are several applications that could benefit from such a computational model of communicative behaviors. The behaviors of avatars of human users in a 3D world could be driven by the users' speech. In video games non-player characters could also be animated similarly. It could aso be applied for computing the nonverbal behaviors of autonomous embodied conversational agents.

Nonverbal behaviors are not only higher level information such as emotional states and attitudes [2], but also with correlated with prosodic and acoustic features. [3] reported high correlation between head motion and the fundamental frequency (F0); on the other hand, [4] stated that "between 80 and 90 of the variance observed in face motion can be accounted for by the speech acoustics". Computational models such as those proposed by [5] and [6] rely on such links to learn the relationship between modalities. Most existing models of virtual agents' behaviors can be clustered into two main groups. In one group, models are based on theoretical models taken from domains such as psychology, emotion studies, linguistic [7]. On the other hand, statistical models have been applied to learn the correlation between speech and multimodal behaviors [8, 5, 9, 10, 11, 6, 12, 13, 14, 15, 16, 17]. These models make use of the tight relationship between acoustic and visual behaviors. While a few methods have already been proposed most of them lack variability in the produced animation.

We investigate here three statistical models to infer the facial signals from the speech signals. As a first start we focus on eyebrow motion. Our goal is to build a statistical system which is able to learn from training samples how to generate natural animation motion from speech features while authorizing realistic variability in the synthesized behaviors. We developed three statistical Markovian systems that all rely on *contextual models*, able to take into account contextual information (speech features here). In the remaining of this paper we first describe related works then we introduce our approaches and we finally report experimental results.

2. BACKGROUND AND RELATED WORKS

Basically we are interested in techniques that allow generating an output information stream (motion) from an input information stream (speech). We first recall major works on synthesizing smooth sequences from a HMM.

This work has been done within the context of the National French project ANR IMMEMO and of the European ITEA2 UsiXml project.

2.1. Using HMMs for synthesis

Synthesizing a realistic sequence of observations (called a trajectory hereafter) from a HMM is a key issue. Of course, synthesizing the most likely observation sequence given a particular state sequence yields a very unlikely piecewise constant trajectory. Integrating over all state sequences the corresponding piecewise constant trajectories gives a better result [14]. A key technique has been proposed by [18] to synthesize more realistic smooth trajectories from a standard HMM with Gaussian probability density functions. [18] proposed few variants of a generic method that we do not detail here but which is a building block in few methods for speech-to-motion synthesis that we will discuss in the following, including ours. We will distinguish between a general synthesis case (named *Integrated method* hereafter) that consists in synthesizing a trajectory from the HMM by integrating trajectories over all possible state sequence, and a more restricted synthesis that considers only one state sequence (possibly the most likely or whatever), which we will name *single method*.

2.2. Speech to motion synthesis

Few researchers have presented data-driven approaches to synthesize speech animation, including body and facial animation. [13, 12] and [10] generate automatically body motions from spoken speech. Given the tight relationship between acoustic phonemes and visual visemes, speech is also used to drive lip motion in [8, 15]. While these works mainly focus on speech content, other works are particularly interested in synthesizing nonverbal communicative behaviors during speech, such as head and eyebrow motion.

A key idea that was followed by a number of researchers has been to use Gaussian distribution on feature vectors including speech and motion features to capture the correlation between these two types of features. [11, 16] used Gaussian Mixture Model (GMM) while [19] and [5, 9, 6, 17] used HMMs. This latter approach is probably the most popular for synthesizing behaviors from speech (we will use this as a baseline in our experiments). It consists in designing a Gaussian *joint* HMM, named λ hereafter, working on concatenated observation vectors for the two streams (i.e. a frame at time t is $x_t = \left[x_t^1 x_t^2\right]'$ where x_t^i stands for the feature vector at time t for stream i). A key point is that one can build from the *joint* HMM a Gaussian HMM for every stream, named λ_1 and λ_2 by keeping only parameters related to the stream. Note that these models λ_i, have the same architecture and share transition probabilities. Based on this, once a joint HMM is trained, one can synthesize a trajectory for the second stream from the observation sequence of the first stream as follows. Using λ_1 one determines the most likely state sequence. Then using λ_2 one can determine a synthesized trajectory for the second stream using the *single* method of [18]. Alternatively, one may use λ_1 to compute the probability distribution over all state sequence given the stream 1 observation sequence.

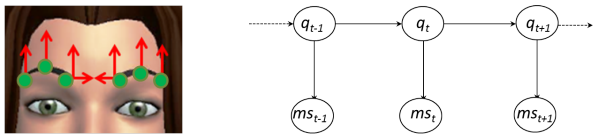

Fig. 1. Left: Illustration of the extracted facial animation parameters (arrows illustrate displacements). Right: Representation of a HMM used for speech to motion synthesis in [6] as a dynamic Bayesian network. Motion and speech features are coupled in observation frames and their interdependency is modeled through covariance matrices.

Then using λ_2 one can determine a synthesized trajectory for the second stream using the *integrated* method.

3. SPEECH TO MOTION SYNTHESIS USING CONTEXTUAL MARKOVIAN MODELS.

We present below three approaches that are based on contextual HMMs [20]. We introduce first contextual HMMs and show how they can be used to infer motion from speech. This reference method generalizes in particular the method in [6]. Then we present two new approaches that improve on this baseline. The three proposed modeling are illustrated in Figure 2 as dynamic Bayesian networks. In the following we consider a training dataset where every observation sequence is a sequence of frames x_t's that are composed of motion features m_t and of speech feature s_t.

3.1. Contextual HMMs (CHMMs)

Our first system is based on Contextual hidden Markov models (CHMMs). CHMMs have been initially proposed for recognizing gestures with the idea of using contextual information related to the physiology of the person realizing the gesture or the amplitude of the gesture [21], [20]. Assume that we are given a set of external (contextual) variables θ (vector of dimension c) for any observation sequence $\mathbf{x} = (x_1, ..., x_T)$ where x_t's are d-dimensional feature vectors. A CHMM is a HMM whose means and covariance matrices depend on θ. For instance the mean $\hat{\mu}_j$ (d-dimensional vector) of the Gaussian distribution in state j is defined as:

$$\hat{\mu}_j(\theta) = W_j^{\mu}\theta + \bar{\mu}_j \quad (1)$$

with W_j^{μ} a $d \times c$ matrix, and $\bar{\mu}_j$ is an offset vector. Training is performed via the Generalized EM algorithm. Note that θ may be dynamic and vary with time [20]. In such we get a sequence of θ_t together with the sequence of observation and Gaussian pdfs have time-varying means and time-varying covariance matrices. For instance the mean in state j equals:

$$\hat{\mu}_j(\theta_t) = W_j^{\mu}\theta_t + \bar{\mu}_j \quad (2)$$

To design a speech-to-motion system we learn one CHMM with speech features as (dynamic) contextual variables (i.e. pdfs are conditioned on speech features) and with both motion and speech features as observations as in [6]. Note that when used as contextual variables we use short term means of the speech frames computed on a sliding window of length 10 (we note these features $\bar{s}$).

Once such a model is trained one can determine a CHMM on speech λ_s only by ignoring pdf parameters on motion features. Also one can use the speech signal to determine a CHMM on motion whose parameters are modified by the speech stream, we note this model $\lambda_{m/s}$. Actually it is a CHMM with time varying parameters (e.g. the mean of a Gaussian changes with time). At the synthesis step, speech features are first processed with λ_s to find the most likely state sequence, then we use the *single method* of [18] (cf. section 2.1) to synthesize a trajectory along this state sequence with $\lambda_{m/s}$. While this approach is close to [6] however we use contextual HMMs instead of HMMs which allows capturing complex dependencies between speech and motion, yielding improved synthesis as we will demonstrate.

3.2. Fully Parameterized HMMs (FPHMMs)

We have developed a new extension of CHMM, named FPHMM, by parameterizing transition probabilities including the initial state distribution with external variables θ_t. In addition to means and covariance matrices already parameterized in CHMMs, transition probabilities also depend on external variables here. The state transition distribution $a_{i,j}$ from i^{th}= state to j^{th} state at time t is defined as:

$$a_{i,j}(\theta_t) = \frac{e^{logA_{ij} + W_{ij}^{tr}\theta_t}}{\sum_{j'} e^{logA_{ij'} + W_{ij'}^{tr}\theta_t}} \quad (3)$$

where $W_{i,j}^{tr}$ is a c-dimensional vector and A_{ij} may be viewed as an offset value. Hence transition probabilities change at every time step according to the contextual variables. Such a model is interesting in our case when one exploits speech features as contextual variables. It allows defining more directly the sequence of states as a function of the speech signal.

To design a speech-to-motion synthesis system we learn a FPHMM that takes speech features as external variables and motion features only as observation. Thereby, speech features influence directly state transition probabilities and emission probability distributions. This model is trained via likelihood maximization with a GEM algorithm. For synthesis, speech features (external variables) are used to determine a probability distribution over all the hidden states at each time step using only transition probabilities. Then we use the *integrated method* of [18] to generate the most likely animation with the motion HMM $\lambda_{m/s}$.

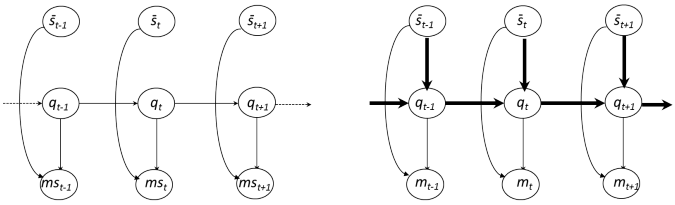

Fig. 2. Representation of a CHMM (left) and of a FPHMM (right) as DBNs. CHMM use short term speech features to modify the pdfs while FPHMM model more directly the motion as a function of the speech. State at time t is noted q_t and short term mean of speech feature vectors (when speech is used as contextual variable) is noted $\bar{s}_t$. Note that a PFHMM-CRF are similar to PFHMM (right) but where the dependencies indicated by thick lines are modeled with a CRF.

3.3. Combining FPHMMs and CRFs (FPHMMs-CRFs)

At last, we have investigated the combination of Fully Parameterized HMMs and of Conditional Random Fields (CRFs) [22], named FPHMMs-CRFs, where the CRF is used similarly as in [13]. The FPHMM has the same architecture as in previous case, it takes speech features as external variables and motion features as observation. The CRF has the same architecture as the FPHMM. It takes speech features as input and it outputs a state sequence or a probability distribution over state sequences, which will be used to synthesize the motion features.

For training, we first learn a FPHMM as described in section 3.2. Then for each training sequence $\mathbf{x}^i$, we determine the most likely state sequence $\mathbf{h}_{s^i}$ in the motion FPHMM $\lambda_{m/s}$. Then the CRF is trained using the set of $(\mathbf{s}^i, \mathbf{h}_{\mathbf{s}^i})$ as training dataset. For synthesis, a speech signal $\mathbf{s}$ is input to the CRF to get a probability distribution over hidden state sequences in $\lambda_{m/s}$. Speech features are also used to determine $\lambda_{m/s}$ from the Fully Parameterized HMM. Then, given the distribution on hidden states as output by the CRF, we synthesize with $\lambda_{m/s}$ a smooth trajectory using the *integrated* method of [18]. This approach not only overcomes the limitations from the assumptions of standard HMM by Fully Parameterized HMM but also takes the advantages of CRF as a discriminative model for inferring accurate probability distribution over all hidden state sequences.

4. EXPERIMENTS

4.1. Datasets

Experiments have been performed on the Biwi 3D Audiovisual Corpus of Affective Communication database (B3D / AC) [23]. 14 subjects were invited to speak 80 short English sentences. In total, this corpus includes 1109 sequences, each lasting 4.67s long on average. We used a part of this

corpus corresponding to 240 sentences from three subjects. We manually annotated the data with respect to five labels $\mathbf{L} = \{c_1, ..., c_5\}$ that consist in combination of Action Units[2] (including a *no move* label). A sequence of observation is then labeled as a sequence of labels (a specific combination of action units) together with their boundaries, just like a speech signal is annotated in phones. Every training sequence consists then in a triple $(\mathbf{s}, \mathbf{m}, \mathbf{y})$ of a sequence of speech feature vectors (of length T), a sequence of motion feature vectors (of length T) and a sequence of labels $\mathbf{y}$ (of length T, with $\forall t, y_t \in \mathbf{L}$). We preprocessed each sequence to get a speech stream and an eyebrow motion stream at the same rate of 25 frames (i.e. feature vectors) per second (fps). For the motion stream we gathered four features for each eyebrow corresponding to four facial animation points (FAPs) as defined by the MPEG-4 standard [25] (see Figure 1); these features move with respect to a neutral pose according to FAPs values. We computed average values for the 4 FAPs for each each brow. Concerning speech we used prosodic features (pitch and RMS energy) which we extracted with PRAAT [26]. We used augmented feature vectors both for motion and for speech streams by adding first and second order derivatives of static features (i.e. velocity and acceleration). Hence we get 6 dimensional frames for speech and 12 dimensional frames for motion. In contextual models, the speech feature $\bar{s}$ used as contextual variables are short term means of the speech frames computed on a sliding window of length 10 (found by trials and errors to give the best results).

4.2. Results

We performed experiments with our approaches and with the method in [6] that exploits HMMs. We considered as many models as there are eyebrow motion classes(5). We used an ergodic model for the *no motion* class and left-to-right models for the other classes. We trained the models with a dataset including speech and motion features for each sentence. We first trained independently class models (whatever the models used, HMM, CHMM, PFHMM and PFHMM-CRF) using corresponding segments of training sequences. Then we combined these submodels into a global model which is reestimated on whole sentences.

For the test we use the sequence of speech features only. We primarily evaluated our methods with respect to a reconstruction error, i.e. the mean squared error between the syntesized motion signal (from the speech signal) and the real motion signal (MSE criterion). To gain more insight on the behavior of the methods we also evaluated the methods with respect to their labeling quality, i.e. the recognition of the sequence of labels. We computed the recognition accuracy with respect to the Hamming distance (H criterion) and to the edit distance (E criterion) between recognized and manually annotated sequences of labels. Reported results are averaged results over 20 random splits of the dataset into 80% for training and 20% for testing, together with standard deviation.

[2]An Action Unit AU as defined by [24] is a minimal visible muscular contraction (e.g. raise eyebrow). Facial expressions are described as a combination of AUs and express emotional state (anger, fear, sadness, surprise...).

Model	#states	MSE	Acc (H)	Acc (E)
HMM [6]	3	0.67 (0.052)	37% (4.7)	45% (4.2)
	5	0.59 (0.042)	43% (4.7)	49% (4.4)
	7	0.56 (0.056)	53% (5.7)	51% (4.3)
CHMMs	3	0.51 (0.055)	55% (4.8)	49% (4.4)
	5	0.49 (0.064)	58% (5.7)	50% (4.9)
	7	0.47 (0.056)	59% (4.5)	50% (3.4)
FPHMM	3	0.55 (0.042)	60% (5.3)	57% (4.7)
	5	0.46 (0.051)	61% (5.1)	61% (3.8)
	7	0.45 (0.037)	63% (3.0)	62% (3.7)
FPHMM-CRF	3	0.47 (0.054)	58% (4.2)	60% (3.7)
	5	0.44 (0.061)	61% (4.0)	65% (3.8)
	7	0.39 (0.051)	66% (4.1)	64% (3.7)

Table 1. Performance of the models with respect to the synthesis quality (MSE) and to labelling accuracy where accuracy is computed by evaluating Hamming distance (H) and edit distance (E). Performances are averaged results gained on 20 experiments (standard deviations are given in brackets).

Model	#states	MSE	Acc (H)
HMM [6]	3	0.43 (0.055)	73% (4.7)
	5	0.39 (0.051)	75% (4.4)
	7	0.36 (0.063)	78% (4.7)
CHMMs	3	0.37 (0.057)	77% (5.0)
	5	0.31 (0.061)	81% (4.7)
	7	0.30 (0.061)	82% (5.0)
FPHMM	3	0.33 (0.043)	80% (4.1)
	5	0.28 (0.048)	83% (5.3)
	7	0.25 (0.052)	84% (4.9)
FPHMM-CRF	3	0.31 (0.044)	81% (5.8)
	5	0.26 (0.040)	84% (5.5)
	7	0.23 (0.038)	84% (5.4)

Table 2. Similar results as in Table 1 but where we assume the sequence of labels of each test observation sequence is known (but not the time boundaries).

Table 1 reports the performance, on the test set, of the four methods with respect to the three evaluation criteria and for a number of states per class model ranging from 3 to 7. As can be seen in Table 1 our three novel approaches (CHMM, PFHMM and PFHMM-CRF) performs better than conventional HMMs used by [6] and the performance with PFHMM-CRF is the best. Table 2 reports similar results in a slightly different setting. We computed the same performance criterion as in table 1 but in that case the sequence of labels was assumed known for every test sequence (but not the time boundaries between labels). Of course the H and MSE obtained here show significant improvements compared to table 1 but the gap is not so big. This means that even if the system does not always recognize labels, it does not affect too much the synthesized motion stream.

5. CONCLUSION

We have investigated few approaches for speech to motion synthesis. Our results show that contextual models are significantly better than a benchmark method in the field. Moreover our method combining a new extension of contextual HMMs and CRFs outperforms all other methods under investigation.

6. REFERENCES

[1] D.L.M. Bolinger, *Intonation and Its Uses: Melody in Grammar and Discourse*, University Press, 1989.

[2] P. Ekman, *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*, Owl Books, mar 2004.

[3] T. Kuratate, K. G. Kuratate, P. E. Rubin, E. V. Bateson, and H. Yehia, "Audio-visual synthesis of talking faces from speech production correlates," in *EUROSPEECH*, 1999.

[4] H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Linking facial animation, head motion and speech acoustics," *Journal of Phonetics*, vol. 30, no. 3, pp. 555 – 568, 2002.

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

[6] G. Hofer, H. Shimodaira, and J. Yamagishi, "Speech driven head motion synthesis based on a trajectory model," in *ACM SIGGRAPH 2007 posters*, 2007.

[7] E. Bevacqua, K. Prepin, R. Niewiadomski, E. de Sevin, and C. Pelachaud, "GRETA : Towards an Interactive Conversational Virtual Companion," in *Artificial Companions in Society: perspectives on the Present and Future*, pp. 1–17. 2010.

[8] M. Brand, "Voice puppetry," in *Proceedings of conference on Computer graphics and interactive techniques*, 1999, pp. 21–28.

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

[10] C. C. Chiu and S. Marsella, "How to train your avatar: A data driven approach to gesture generation," in *IVA*, 2011, pp. 127–140.

[11] M. Costa, T. Chen, and F. Lavagetto, "Visual prosody analysis for realistic motion synthesis of 3d head models," in *Proc. of ICAV3D*, 2001, pp. 343–346.

[12] S. Levine, C. Theobalt, and V. Koltun, "Real-time prosody-driven synthesis of body language," *ACM Trans. Graph.*, vol. 28, no. 5, 2009.

[13] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun, "Gesture controllers," *ACM Trans. Graph.*, vol. 29, no. 4, 2010.

[14] Y. Li and H. Y. Shum, "Learning dynamic audio-visual mapping with inputoutput hidden markov models," *IEEE Trans. on Multimedia*, pp. 542–549, 2006.

[15] J. Xue, *Acoustically-driven talking face animations using dynamic bayesian networks*, Ph.D. thesis, Los Angeles, CA, USA, 2008, AAI3351613.

[16] B. H. Le, X. Ma, and Z. Deng, "Live speech driven head-and-eye motion generators," *IEEE Trans. on Visualization and Computer Graphics*, vol. 18, pp. 1902–1914, 2012.

[17] S. Mariooryad and C. Busso, "Generating human-like behaviors using joint, speech-driven models for conversational agents," *IEEE Trans. on Audio, Speech & Language Processing*, vol. 20, no. 8, pp. 2329–2340, 2012.

[18] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *ICASSP*, 2000, pp. 1315–1318.

[19] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1330–1345, 2008.

[20] M. Radenen and T. Artières, "Contextual hidden markov models," in *ICASSP*, 2012, pp. 2113–2116.

[21] A. D. Wilson and A. F. Bobick, "Parametric hidden markov models for gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, pp. 884–900, 1999.

[22] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001, pp. 282–289.

[23] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool, "A 3-D Audio-Visual Corpus of Affective Communication," *Multimedia, IEEE Transactions on*, vol. 12, no. 6, pp. 591–598, Oct. 2010.

[24] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Consulting Psychologists Press, 1978.

[25] I.S. Pandzic and R. Forcheimer, *MPEG4 Facial Animation - The standard, implementations and applications*, John Wiley & Sons, 2002.

[26] P. Boersma and D. Weeninck, "Praat, a system for doing phonetics by computer.," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.