REALISTIC MOUTH ANIMATION BASED ON AN ARTICULATORY DBN MODEL WITH CONSTRANDED ASYNCHRONY

Dongmei Jiang$^1$, Ilse Ravyse$^2$, Peizhen Liu$^1$, Hichem Sahli$^2$, Werner Verhelst$^2$

VUB-NPU Joint Research Group on Audio Visual Signal Processing (AVSP)
1. Northwestern Polytechnical University, Xi’an 710072, P.R. China
2. Vrije Universiteit Brussel, Department ETRO, Pleinlaan 2, 1050 Brussels, Belgium

jiangdm@nwpu.edu.cn, nwpuaslpliu@gmail.com
{icravyse, hsahli, wverhels}@etro.vub.ac.be

ABSTRACT

In this paper, we propose an approach to convert acoustic speech to video realistic mouth animation based on an articulatory dynamic Bayesian network model with constrained asynchrony (AF_AVDBN). Conditional probability distributions are defined to control the asynchronies between the articulators such as lips, tongue and glottis/velum. An EM-based conversion algorithm is also presented to learn the optimal visual features given an auditory input and the trained AF_AVDBN parameters. In the training of the AF_AVDBN models, downsampled YUV spatial frequency features of the interpolated mouth image sequences are extracted as visual features. For reproducing the mouth animation sequence, from the learned visual features, a spatial upsampling and a temporal downsampling are applied. Both qualitative and quantitative results show that the proposed method is capable of producing more natural and realistic mouth animations, and the accuracy is further improved compared to the state of the art multi-stream Hidden Markov Model (MSHMM) and articulatory DBN model without asynchrony constraint (AF_DBN).

Index Terms— AF_AVDBN, AF_DBN, asynchrony, conditional probability distribution, mouth animation

1. INTRODUCTION

Computer animated talking faces have become a popular research topic in human-computer interaction. According to the underlying face model, talking faces can be categorized into 3D-model-based animation and image-based rendering of models. Image-based approaches try to achieve photorealistic performance, where an important issue is synthesizing mouth images that tightly match the content and intensity of the input speech. Some researchers adopted a machine learning strategy by considering mouth synching as an audio-to-visual conversion problem. E.g. [1] exploits the Hidden Markov Model inversion (HMMI) technique in audio to visual conversion for an MPEG-4 facial animation system. The correlation between audio and visual speech is modeled by a MSHMM with diagonal covariance matrices. Given an audio input and the trained Gaussian parameters, visual parameters are learned based on the Maximum Likelihood Estimation (MLE) of some auxiliary function. Later, Lucas et al. [2] expanded the HMMI technique to a general case of full covariance matrices, and proposed a speech driven MPEG-4 compliant facial animation system. In [3], a similar audio to visual conversion theory is performed on an audio visual dynamic Bayesian network model with articulatory features (AF_DBN), in which all the articulatory features (AFs) are asynchronous without any constraint. Natural and realistic mouth animations are obtained. However, the assumption that all the AFs move with unlimited asynchrony along the sentence does not fit the physical mechanism of the articulator organs. This has been considered in [4] where Livescu et al. proposed an audio visual DBN model with articulatory features (AF_AVDBN) for speech recognition, in which the maximum asynchrony between the AFs can be controlled. They demonstrated the validity of the model by producing high recognition rates. However, in [4] Livescu et al. only discussed the structure of the AF_AVDBN model, they did not define clearly the conditional probability distributions (CPDs) of the nodes.

In this paper, based on the AF_AVDBN structure and the use of Discrete Cosine Transform (DCT) coefficients as visual feature of [4], we, i) define the CPD for each node, and present speech recognition experiments to verify the capabilities of AF_AVDBN vs. AF_DBN and MSHMM; ii) propose an acoustic speech to realistic mouth animation conversion, based on color DCT coefficients as visual features and the AF_AVDBN models trained with an audio visual digit speech database. Experimental results show that using the proposed CPD, the AF_AVDBN model not only provides higher speech recognition rates, but also learns more accurate visual features than MSHMM and the AF_DBN model.

The remainder of the paper is organized as follows: audio visual features extraction is addressed in Section 2. Section 3 describes the AF_AVDBN model and defines the CPDs of the nodes. Section 4 introduces the visual feature learning and the mouth animation sequence reproduction. While experiments and analysis are given in Section 5, we
2. AUDIO VISUAL SPEECH FEATURES

The used AF_AVDBN and AF_DBN models are being trained using the GMTK [5] toolkit, on an audio visual digit database. From the audio speech, 42 audio features, corresponding to 13 perceptual linear prediction (PLP) features and energy, plus their first and second order differential coefficients, are extracted with a frame length of 25 milliseconds and frame shift of 10 milliseconds. To obtain the visual speech features with the same frame rate as the audio features (100 frames/second), the front view face image sequences are processed in three steps.

i) Face tracking and lip contour extraction. The Constrained Bayesian Tangent Shape Model of [6] has been used for the detection and tracking of a shape model defined by 83 facial feature points, over a facial image sequence. For each image, a 64x64 mouth region of interest (ROI) is extracted, based on the tracked feature points of the lip contours. The extracted ROIs form a mouth image sequence of 25 frames/s (Fig. 3, rows 1 and 4).

ii) Temporal upsampling of the mouth image sequences. The frame rate of the above obtained mouth image sequences is increased 4 times. A linear bi-variate 2-D radial basis function interpolates [7] the image points that do not belong to the lip contour points from a nearby original image. The forward and backward interpolation results are summed according to the passed time duration:

\[
I^D_{tp} = (1 - \text{res}_t)[W[I(t_{(r)}), m[I(t_{(r)})]]] + \text{res}_t[W[I(t_{(r)})], 1 - m[I(t_{(r)})]]
\]

where res = t(t) - t(r) with \( I \) being the floor function. \( I_{(t)} \) and \( m_{(t)} \) are the mouth image and motion vectors of the lip contour points at frame \( I(t) \), respectively.

iii) Visual feature extraction. We describe the spatial frequencies of the mouth pattern in each image by the coefficients of its DCT [4]. This linear transform is capable of highly decorrelating the pattern frequencies. A similar DCT method as in the JPEG standard allows us to incorporate color. The RGB images are converted to Y (luminance), UV (chrominances) color space following the CCIR 601 transform. Although we downsample each 64x64 images in a 4:2:2 scheme to reduce the amount of features, we do not use the 8x8 blocks for the DCT as in JPEG. Here, we compute the DCT coefficients of the entire mouth image for which we downsample Y to 32x32, and further to 16x16, for the calculation of the frame-difference features, using the MPEG-4 downsample filter, and U and V to 16x16 using the 2-tap and 3-tap filters[8]. Next, we select the first \( n \) coefficients in a zigzag scan order. The number of coefficients, \( n \), is set as \( n_Y = 78, n_U = 10, \) and \( n_V = 10 \), to keep a psychovisual contrast as in the JPEG quantization table[9]. Finally, for each frame of the interpolated (step ii) mouth image sequence, the \( n_Y + n_U + n_V \) DCT coefficients are concatenated together with the 21 frame difference coefficients to form a 119 dimensional visual feature vector.

3. AF_AVDBN MODEL AND CPDS

Fig.1 shows the recognition model of AF_AVDBN where the chunk part is repeated every frame with the evolution of the audio visual features. L, T, G are articulatory features denoting the states of lips (lip location and lip opening), tongue tip and tongue body, as well of Glottis and velum. The other nodes are:

- LPosition(LP)/TPosition(TP)/GPosition(GP): position (index) of the AF in the sentence.
- LTransition(LT)/TTransition(TT)/GTransition(GT): transition condition of an AF; a value 1 indicates that the AF transits at the next frame.
- ChecksynLCLT/ChecksynLCLTG: check if L, T and G satisfy the constraint of asynchrony.
- Audio/Visual Obs (\( \phi \)): audio or visual features

In the training process, the AF transcriptions are obtained from mapping the phonemes-by inquiring a phoneme-AF table [4], and the maximum index \( N \) of one AF in a sentence can be obtained.

By setting \( CLLT = LB - TB \) as the asynchrony between \( L \) and \( T \), \( CLTG = GP - (LP + TP) \) as the asynchrony between \( G \) and the mean position of \( L, T \), we define the following CPDs to control the asynchronies.

\[
p(LP = i) = \begin{cases} 1 & \text{if } k = i = 0 \text{ and } m \leq -S \\ 1 & \text{if } k = i = 0 \text{ and } m \in [-S, S] \\ 1 & \text{if } k = i = 1 \text{ and } m \leq -S \\ 1 & \text{if } k = i = 1 \text{ and } m \in [-S, S] \\ 0 & \text{otherwise} \end{cases}
\]

\[
p(TP = i) = \begin{cases} 1 & \text{if } k = i = 0 \text{ and } m \leq -S \\ 1 & \text{if } k = i = 0 \text{ and } m \in [-S, S] \\ 1 & \text{if } k = i = 1 \text{ and } m \leq -S \\ 1 & \text{if } k = i = 1 \text{ and } m \in [-S, S] \\ 0 & \text{otherwise} \end{cases}
\]

where \( S \) is the allowed maximum asynchrony. The definition of \( p(GP = i) = \begin{cases} 1 & \text{if } m \leq -S \end{cases} \) is the same as in Equation (2). The above CPDs indicate that when an AF drops behind the other AF (or \( G \) drops behind the mean position of \( L, T \) for more than \( S \) indices, the state of the AF is forced to change; On the contrary, when it leads ahead the other AF for more than \( S \) indices, its state is (forced to) maintained; When the asynchrony \( m \) does not exceed the limitation \( S \), the AF does not reach its last index in the sentence and is allowed to transit, the AF state will change and its index increments by 1, otherwise the state and index remains.
The probability of emitting the audio visual observation features of each articulatory feature state is modeled as a Gaussian mixture model (GMM):

\[
p(O_t^n | L_t = q, T_t = m, G_G = n) = p(O_t^n | S_t = j) = p(O_t^n | T_t = j) \cdot p(O_t^n | S_t = j)
\]

\[
= \prod_{d(a,v)} \left[ \sum_{k=1}^{M_d} e_{jk}^d N(O_t^d, \mu_{jk}^d, \Sigma_{jk}^d) \right]
\]

(4)

where \( N() \) is the Normal distribution, with mean \( \mu_{jk} \), and covariance matrix \( \Sigma_{jk} \).

Fig.1. the AF_AVDBN recognition model

4. MOUTH ANIMATION BASED ON AF_AVDBN

4.1. Audio to visual conversion based on AF_AVDBN

Let \( \psi \), denoting the set of all hidden variables (LP, TP, GP, LT, TT, GT, L, T, G, CLT, CLTG) at frame \( t \). The probability of an audio visual speech \( (o^a, o^v) \) evolving along a hidden variable path \( \Psi = (\psi_1, \psi_2, ..., \psi_T) \) can be concisely defined as

\[
P(o^a, o^v, \psi_T | \Psi) = \prod_{t=1}^{T} p(o^a_{t-1} | L_t, T_t, G_t) p(o^v_{t-1} | L_t, T_t, G_t) p(\psi_t | \psi_{t-1})
\]

(5)

Given an input audio sequence \( o^a \) and the trained model set \( \lambda \), the Maximum Likelihood (ML) criterion is iteratively maximizing an auxiliary function \( \Omega(\lambda; o^a, o^v, \psi^v) \) defined as:

\[
\Omega(\lambda; o^a, o^v, \psi^v) = \sum_{\psi \in \Theta} \log P(o^a, o^v, \psi | \lambda)
\]

(6)

where \( o^a \) and \( o^v \) are the old and estimated visual parameter sequences respectively in one iteration. The optimal visual feature \( o^v \) can be obtained by setting the derivative of \( \Omega(\lambda; o^a, o^v, \psi^v) \) with respect to \( o^v \) equal to zero, i.e.

\[
\frac{\partial \Omega(\lambda; o^a, o^v, \psi^v)}{\partial o^v} = \sum_{\psi \in \Theta} P(o^a, o^v, \psi | \lambda) \frac{\partial \log P(o^v | L_t, T_t, G_t)}{\partial o^v} = 0
\]

(7)

where \( \psi_r \) and \( k \) are possible values of the hidden variables at frame \( t \).

From Equation (7), \( o^v_r \) is re-estimated as

\[
o^v_r = \frac{\sum_{\psi_r \in \Omega} P(o^a, o^v_r, \psi_r | \lambda) e_{\psi_r,k}^v \left( \Sigma_{\psi_r,k}^{-1} \right) o^v_r - \mu_{\psi_r,k}^v}{\sum_{\psi_r \in \Omega} P(o^a, o^v_r, \psi_r | \lambda) e_{\psi_r,k}^v \left( \Sigma_{\psi_r,k}^{-1} \right)}
\]

(8)

where \( P(o^a, o^v_r, \psi_r | \lambda) \) is the probability of the audio visual sequence \( (o^a, o^v_r) \) passing through \( \psi_r \). In our implementation, we firstly perform speech recognition on the AF_AVDBN model, with \( (o^a, o^v) \) as input, to get the N-best paths, then we estimate \( o^v \) replacing the outer sum over all possible states of the hidden variables at frame \( t \), by the sum over the states of frame \( t \) along the N best paths.

4.2. Mouth animation sequence reproduction

The inverse DCT on the learned visual features \( o^v \) provides us downsampled YUV images at audio frame rate. The synthesized mouth sequence is reproduced by a spatial upsampling and a temporal downsampling as follows. The luminance and chrominance channels are upsampled with the 2-tap filters [8] before conversion to RGB. To reconstruct mouth images at video frame rate, we blend the corresponding pixel values of a frame \( I_m \) at video frame rate and its surrounding frames \( I_{m+k} \) at audio frame rate.

This correspondence is estimated as a dense optical flow field \((u_k, v_k)\) which maps \( I_{m+k} \) to \( I_m \) [10]. The blending of the motion-compensated images is defined as:

\[
I^{blend}_{m}(x, y) = \sum_{k=-2}^{2} b_{\alpha, \beta} \left( \frac{k}{3} \right) \cdot I_{m+k}(x+u_k, y+v_k)
\]

(9)

and the blending function \( b_{\alpha, \beta}(\cdot) \) is chosen as the parametric rational \( \alpha \) continuous blending function as in [11], where \( \alpha = 0.2, \beta = 1 \) control the shape of the blending function.

5. EXPERIMENTS AND ANALYSIS

For assessing the proposed approach, an audio visual digit database is used containing 11 digits of oh and zero to nine. For each digit, 40 sentences of uttering the digit are used as training set, and the other 20 sentences as testing set. Table 1 shows the speech recognition rates on the AF_AVDBN models of setting the maximum asynchrony limitation \( S \) as 1 and 2 respectively, as well on the MSHMM and AF_DBN models. Note that, the MSHMM and the AF_AVDBN model with \( S = 1 \) (AF_AVDBN_1) provide the highest recognition rates, thus AF_AVDBN_1 is chosen for the later audio to visual conversion.
Table 1. Speech recognition rates (%)

<table>
<thead>
<tr>
<th></th>
<th>MSHMM</th>
<th>AF DBN</th>
<th>AF_AVDBN_1</th>
<th>AF_AVDBN_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate</td>
<td>97.27</td>
<td>92.73</td>
<td>97.27</td>
<td>95.45</td>
</tr>
</tbody>
</table>

6. CONCLUSIONS AND FUTURE WORK

This paper proposes an audio to visual conversion scheme based on the articulatory DBN model with constrained asynchrony (AF_AVDBN). Conditional probability distributions are defined to control the asynchrony between the articulatory features. Experimental results show that the proposed method is capable of producing more natural and realistic mouth animations, and the accuracy is further improved compared to the state of the art MSHMM and the AF_DBN model without asynchrony constraint. In the future work, we will extend the structure of the AF_AVDBN model to make mouth animations for continuous speech.

Fig.2. trajectories of visual features (amplitude vs. frame)

Fig.3. original mouth images (rows 1, 4), reproduced mouth images from AF_AVDBN_1 (rows 2, 5) and AF_DBN (rows 3, 6)

For each of the 220 testing speech sentences, three mouth animations have been reproduced, from the visual features learned, using AF_AVDBN_1, AF_DBN and MSHMM, respectively. The mean relative distance (MRD) between the real and the learned visual features have been calculated using Equation (10), where $N_k$ is the frame number of the $k$th mouth sequence, $v_{kj}^v$ and $\hat{v}_{kj}^v$ are the $j$th learned and actual visual feature parameters of frame $t$ respectively.

$$MRD = \frac{\sum_{t=1}^{220} \sum_{k=1}^{N_k} \left| v_{kj}^v - \hat{v}_{kj}^v \right|^2}{\sum_{t=1}^{220} \sum_{k=1}^{N_k} v_{kj}^v}$$

The MRD scores are 3.1923, 3.5431 and 3.9502 for AF_AVDBN_1, AF_DBN and MSHMM, respectively, which shows that much more accurate visual features can be learned from the AF_AVDBN_1 model. Fig.2 shows the time trajectories of two visual feature coefficients. Indeed, the learned visual features using AF_AVDBN_1 resemble the real parameters much better than those from AF_DBN.

Finally, a visual evaluation of the synthesized mouth images is given in Fig.3. The reproduced mouth shapes from AF_AVDBN_1 are very much alike the real ones, even with only little mouth pattern frequency coefficients. Moreover, the results from AF_AVDBN_1 are more accurate than those from AF_DBN (see frames 6-8 of ‘five’ and frames 4-7 of ‘three’).

7. ACKNOWLEDGMENT

This research has been conducted within the research fund from National Natural Science Foundation of China (60703104), and the joint VUB-NPU AVSP research group.

8. REFERENCES