Photo-Realistic Mouth Animation Based on an Asynchronous Articulatory DBN Model for Continuous Speech

He Zhang [1, 2], Dongmei Jiang [1, 2], Peng Wu [1, 2], Hichem Sahli [3]
VUB-NPU Joint Research Group on Audio Visual Signal Processing (AVSP)
[1] Northwestern Polytechnic University, Xi'an 710072
[2] Shaanxi Provincial Key Laboratory on Speech, Image and Information Processing
[3] Vrije Universiteit Brussel (VUB) - AVSP, Department ETRO, Pleinlaan 2, 1050 Brussels, Belgium
E-mail: zhanghe8642036@163.com, jiangdm@nwpu.edu.cn, min_forever@163.com, hichem.sahli@etro.vub.ac.be

***Abstract*— This paper proposes a continuous speech driven photo realistic visual speech synthesis approach based on an articulatory dynamic Bayesian network model (AF_AVDBN) with constrained asynchrony. In the training of the AF_AVDBN model, the perceptual linear prediction (PLP) features and YUV features are extracted as acoustic and visual features respectively. Given an input speech and the trained AF_AVDBN parameters, an EM-based algorithm is deduced to learn the optimal YUV features, which are then used, together with the compensated high frequency components, to synthesize the mouth animation corresponding to the input speech. In the experiments, mouth animations are synthesized for 80 connected digit speech sentences. Both qualitative and quantitative evaluation results show that the proposed method is capable of synthesizing more natural, clear and accurate mouth animations than those from the state asynchronous DBN model (S_A_DBN).**

I. INTRODUCTION

Photo realistic talking face animation has become a popular research topic in human-computer interaction, in which an important issue is synthesizing mouth animations matching the pronunciation of the input speech. Machine learning strategy has been adopted to solve mouth synching as an audio-to-visual conversion problem. E.g. [1] exploited the Hidden Markov Model inversion (HMMI) technique for an MPEG-4 facial animation system. Given an audio input and the trained multi-stream HMM (MSHMM) parameters, visual parameters are learned based on the Maximum Likelihood Estimation (MLE) of an auxiliary function. Later, Terissi 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 approach was performed using an audio visual articulatory DBN model with unlimited asynchronous (AF_DBN). Natural and realistic mouth animations have been obtained. However, the assumption that all the articulatory features (AFs) move with unlimited asynchrony along the sentence does not fit the physical mechanism of the articulator organs. This has been considered in [4] who proposed an audio visual articulatory DBN model with constrained asynchrony (AF_AVDBN) for speech recognition. High recognition rates have been obtained. However, the authors did not define clearly the conditional probability distributions (CPDs) of the nodes.

In our previous work[5], we proposed articulatory DBN models with constrained asynchrony for isolated words, and constructed mouth animations with YUV features learned from these models. Experimental results show that through the articulatory DBN models of isolated words, the accuracy of the constructed mouth shapes is improved compared to the state of the art MSHMM and AF_DBN based methods. Nevertheless, the clearness of the synthesized mouth images is not promising.

In this paper, based on the AF_AVDBN model, 1) we expand the articulatory DBN model based audio to visual conversion approach of isolated words to continuous speech; 2) in synthesizing the mouth images from the learned YUV features, we propose a high frequency components compensation method to improve the clearness of the synthesized mouth images. In the experiments, mouth animations have been synthesized for 80 connected digit speech sentences. Both qualitative and quantitative evaluation results show that clear and natural mouth animations can be obtained from AF_AVDBN with high frequency compensation, and the accuracy of dynamic mouth movements using AF_AVDBN are much higher than those from the state asynchronous DBN model (S_A_DBN), which has been proved to get higher performance than the sate synchronous DBN model (SS_DBN, the DBN implementation of MSHMM)[6].

The remainder of this paper is organized as follows: in section II, we summarize the audio and visual features. CPDs of the nodes in the AF_AVDBN model are defined in section III. Section IV introduces the visual feature learning algorithm and the high frequency components compensation method. Experimental results are analyzed in section V, and section VI discusses the conclusions and future work.

II. AUDIO VISUAL SPEECH FEATURES

From the audio speech, 42-dimension audio features, corresponding to 13 perceptual linear prediction (PLP) features and energy, plus their first and second order differential coefficients, are extracted with frame length of 50ms and frame shift of 40ms.

To get the visual features, firstly the Constrained Bayesian Tangent Shape Model (CSM)[7] is 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 12 tracked feature points of the outer lip contours. We describe the spatial frequencies of the mouth pattern in each image by the coefficients of its DCT[8]: the RGB images are firstly converted to Y (luminance), UV (chrominances) color space, then the DCT coefficients of the entire mouth image are computed for which Y is down-sampled to 32x32, U and V to 16x16. We select the first *n* DCT coefficients in a zigzag scan order, where *n* is set as *nY*=78, *nU*=10, and *nV*=10, respectively, to keep a psychovisual contrast as in the JPEG quantization table[8]. Finally a 98-dimension YUV visual feature vector is obtained for each mouth image with the frame rate of 25 frames/s.

III. AF_AVDBN AND DEFINITION OF CPDS

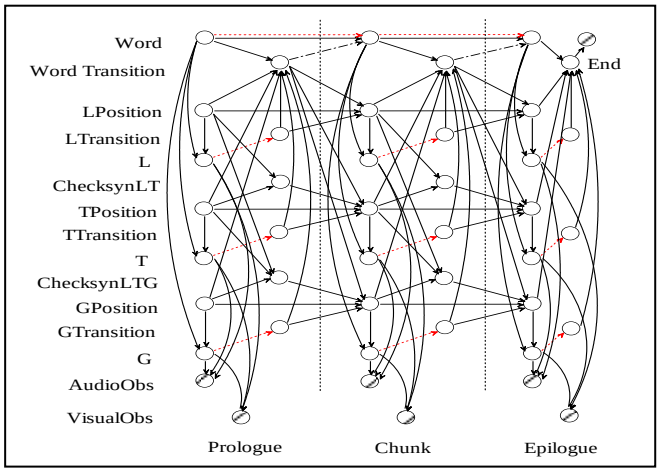

Fig. 1 The structure of AF_AVDBN

Fig.1 shows the structure of the AF_AVDBN model, where the *chunck* part is repeated every frame with the evolution of the audio visual features. The articulatory features (AFs) *L*, *T* and *G* denote the states of lips (lip location and lip opening), tongue tip and tongue body, as well of glottis and velum. The definitions of the nodes are:

- Word (W): word instance.
- Word Transition (WT): decides if the word transits in the next frame.
- LPosition(LP)/TPosition(TP)/GPosition(GP): position of the AF in the current word.
- LTransition(LT)/TTransition(TT)/GTransition(GT): decides if the AF transits. If LT, TT, or GT is 1, then the corresponding AF transits in the next frame.
- L/T/G: the AF instance.
- ChecksynLT(CLT)/ChecksynLTG(CLTG): check if L, T and G satisfy the asynchrony constraint.
- Audio/Visual Obs (o^a / o^v): 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 the AF in a word can be obtained.

By setting $CLT_t = LP_t - TP_t$ as the asynchrony between L and T, $CLTG_t = (LP_t + TP_t)/2 - GP_t$ as the asynchrony between G and the mean position of (L, T), the CPDs of the nodes related to the asynchronies are defined as follows.

$$p(LP_t = i \mid WT_{t-1} = j, LP_{t-1} = k, LT_{t-1} = l, CLT_{t-1} = m) = \begin{cases} 1 & i = k \quad and \quad j = 0 \quad and \quad l = 0 \\ 1 & i = k \quad and \quad j = 0 \quad and \quad l = 1 \quad and \quad k = N \\ 1 & i = k+1 \quad and \quad j = 0 \quad and \quad k < N \quad and \quad l = 1 \quad and \quad m \in [-S, S] \\ 1 & i = k \quad and \quad j = 0 \quad and \quad l = 1 \quad and \quad m \notin [-S, S] \\ 1 & i = 0 \quad and \quad j = 1 \\ 0 & otherwise \end{cases} \tag{1}$$

$$p(TP_t = i \mid WT_{t-1} = j, TP_{t-1} = k, TT_{t-1} = l, CLT_{t-1} = m) = \begin{cases} 1 & i = k+1 \quad and \quad j = 0 \quad and \quad l = 0 \quad and \quad k < N \quad and \quad m > S \\ 1 & i = k \quad and \quad j = 0 \quad and \quad l = 0 \quad and \quad m \in [-S, S] \\ 1 & i = k \quad and \quad j = 0 \quad and \quad k = N \quad and \quad l = 1 \\ 1 & i = k+1 \quad and \quad j = 0 \quad and \quad k < N \quad and \quad l = 1 \quad and \quad m \in [-S, S] \\ 1 & i = k \quad and \quad j = 0 \quad and \quad l = 1 \quad and \quad m < -S \\ 1 & i = 0 \quad and \quad j = 1 \\ 0 & otherwise \end{cases} \tag{2}$$

where S is the allowed maximum asynchrony. The definition of $p(GP_t = i \mid WT_{t-1} = j, GP_{t-1} = k, GT_{t-1} = l, CLTG_{t-1} = m)$ is similar as in (2). The above CPDs indicate that when T changes slower than L for more than S indices, the state of T is forced to transit. On the contrary, when T changes faster than L for more than S indices, its state is (forced to be) maintained; On the other hand, when the asynchrony m does not exceed the limitation S ($m \in [-S, S]$), if T (or L) does not reach its last index in the word and is allowed to transit, its state will change and the index increments by 1, otherwise the state and index remains.

The probability of emitting the audio visual observation features by the combined articulatory feature state j, is modeled as a multiply of Gaussian mixture models (GMMs).

$$p(O_t^{av} \mid L_t = q, T_t = m, G_t = n) = p(O_t^{av} \mid S_t = j) = p(O_t^a \mid S_t = j) \cdot p(O_t^v \mid S_t = j)$$
$$= \prod_{d \in \{a,v\}} \left[\sum_{k=1}^{M} c_{jk}^d N(O_t^d, \mu_{jk}^d, \Theta_{jk}^d) \right]^{\omega_d} \tag{3}$$

For each feature stream d, parameters c_{jk}^d, μ_{jk}^d and Θ_{jk}^d are the weight, mean and covariance matrix of the Gaussian mixture k of the state j, respectively. M is the number of Gaussian mixtures, fixed to 2 in our experiments. ω_d is the weight adjusting the influence of the stream d, with the constraint $\omega_a + \omega_v = 2$. In the training process of the AF_AVDBN model, ω_a and ω_v are both set to 1.

In the training process, for each stream d, one GMM parameter set λ^d of all possible combined articulatory

feature states, are estimated using the Expectation Maximization (EM) algorithm.

IV. Mouth Animation Based on AF_AVDBN

A. Visual Feature Learning Algorithm

let ψ_t be the set of all hidden variables $(W, WT, LP_t, TP_t, GP_t, LT_t, TT_t, GT_t, L_t, T_t, G_t, CLT_t, CLTG_t)$ 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:

$$P(O^a, O^v, \Psi|\lambda) = \prod_{t=1}^{T} p(o_t^a \mid L_t, T_t, G_t)\, p(o_t^v \mid L_t, T_t, G_t)\, p(\psi_t|\psi_{t-1}) \quad (4)$$

Given an input audio sequence O^a and the trained model set $\lambda = (\lambda^a, \lambda^v)$, the Maximum Likelihood (ML) criterion is used to find the optimal visual feature sequence by iteratively maximizing an auxiliary function $\Omega(\lambda; O^a, O^v, O^{v'})$ defined as:

$$\Omega(\lambda; O^a, O^v, O^{v'}) = \sum_{\Psi \in \Phi} P(O^a, O^v, \Psi|\lambda) \cdot \log\left[P(O^a, O^{v'}, \Psi|\lambda)\right] \quad (5)$$

where $o^{v'}$ is the newly estimated visual feature sequence, and o^v is the obtained visual feature sequence in the last iteration, respectively. The optimal visual feature $o_t^{v'}$ can be obtained by setting the derivative of $\Omega(\lambda; O^a, O^v, O^{v'})$ with respect to $o_t^{v'}$ equal to zero, i.e.

$$\frac{\partial \Omega(\lambda; O^a, O^v, O^{v'})}{\partial o_t^{v'}} = \sum_{\Psi \in \Phi} P(O^a, O^v, \Psi|\lambda) \frac{\partial \log\left[p(o_t^{v'} \mid L_t, T_t, G_t)\right]}{\partial o_t^{v'}} \quad (6)$$

$$= \sum_{\psi_t} \sum_{k} P(O^a, O^v, \psi_t|\lambda) \cdot c_{\psi_t k}^v \left(\Theta_{\psi_t k}^v\right)^{-1} \left(o_t^{v'} - \mu_{\psi_t k}^v\right) = 0$$

where *k* denotes the *k*th Gaussian mixture.

From (6), $o_t^{v'}$ is estimated as

$$o_t^{v'} = \frac{\sum_{\psi_t} \sum_{k} P(O^a, O^v, \psi_t|\lambda) \cdot c_{\psi_t k}^v \left(\Theta_{\psi_t k}^v\right)^{-1} \mu_{\psi_t k}^v}{\sum_{\psi_t} \sum_{k} P(O^a, O^v, \psi_t|\lambda) \cdot c_{\psi_t k}^v \left(\Theta_{\psi_t k}^v\right)^{-1}} \quad (7)$$

where $P(O^a, O^v, \psi_t|\lambda)$ is the probability of the audio visual sequence (o^a, o^v) passing through ψ_t.

In each iteration of estimating the visual features, we firstly perform speech recognition using GMTK[9] on the AF_AVDBN model, with the audio feature sequence o^a and the estimated visual feature sequence of the last iteration o^v as input. N-Best paths are deduced from the output file with the item *–verobs 80*. Then in (7), we estimate $o_t^{v'}$ by replacing the sum over all possible states of the hidden variables ψ_t, by the sum over their states in the N-Best paths.

B. High Frequency Compensation and Mouth Image Synthesis

In synthesizing the mouth animations, a method is designed to compensate for the lost high-frequency information in extracting the YUV features: for the training audio visual speech sequences, we firstly do speech recognition on the AF_AVDBN model, and get their best articulatory feature state labels, i.e the states of L, T and G, on each frame. Therefore for each combined articulatory feature state *S*, a data set can be constructed by collecting its mouth images $\{I_{S1}, ..., I_{Si}, ..., I_{SN}\}$ together with their YUV features $\left\{o_{S1}^v, ..., o_{Si}^v, ..., o_{SN}^v\right\}$.

On the other side, in the last iteration of estimating the visual features using (7), a best articulatory feature state path can be obtained, i.e. for the learned visual feature $o_t^{v'}$, the combined AF state S_t is known, therefore the distances between $o_t^{v'}$ and the YUV features $\left\{o_{S_t 1}^v, ..., o_{S_t i}^v, ..., o_{S_t N}^v\right\}$ of S_t can be calculated. Suppose $o_{S_t i}^v$ gets the minimum distance, then the corresponding mouth image $I_{S_t i}$ is selected as the compensation source, and its DCT coefficients are computed in the down-sampled YUV space as described in section II. In synthesizing the mouth image for frame *t*, the first *n* (*nY*=78, *nU*=10, and *nV*=10) coefficients in zigzag order of the DCT matrix is replaced with corresponding data in $o_t^{v'}$, then inverse DCT is made and up-sampled to get the 64*64 image in YUV space. Finally the YUV mouth image is converted to a 64*64 RGB mouth image.

V. Experiments and Analysis

In our experiments, an audio video database of connected digits has been used, with scripts from the Aurora5.0 speech database where each speech sentence contains 2 to 7 digits (oh and zero to nine). 100 sentences are randomly selected as the training set, and the other 80 sentences as the testing set.

Table I
Speech Recognition Rates (%)

AF_AVDBN (1)	AF_AVDBN (3)	MSHMM	S_A_DBN (1)	S_A_DBN (2)
92.65	94.29	90.24	92.28	92.28

To verify the performance of the AF_AVDBN model, we firstly do audio visual speech recognition experiments. The obtained recognition rates are given in Table I, where the numbers between brackets are the maximum allowed asynchrony. One can notice that the state asynchronous S_A_DBN with asynchrony constraints produces higher performance than the MSHMM model. Moreover, AF_AVDBN(3) gets the highest recognition rate, which is 2.01% and 4.05% higher than S_A_DBN and MSHMM respectively.

Fig. 2 Original and synthesized mouth image sequences. (row 1: original, row 2: from AF_AVDBN but without high-frequency compensation, row 3: from AF_AVDBN with high-frequency compensation, row 4: from S_A_DBN with high-frequency compensation)

Fig.2 depicts an example of a synthesized mouth image sequence. One can notice that with high frequency compensation, the mouth images are much clear so that the teeth are well shown. More over, the synthesized mouth shapes from AF_AVDBN(3) are very close to the original images, and the mouth movements are more precise than those from S_A_DBN(1) (see frames 1-3 and frames 8-11).

To measure objectively the accuracy of the learned visual features, the mean relative distance (MRD) between the real and the learned visual features have been calculated.

$$\mathrm{MRD} = \frac{\Sigma_{k=1}^{80}\Sigma_{t=1}^{N_k}\Sigma_{j=1}^{98}\left|\left(\hat{o}_{ktj}^{v} - o_{ktj}^{v}\right)/o_{ktj}^{v}\right|}{\Sigma_{k=1}^{80}N_k \times 98} \tag{8}$$

where N_k is the frame number of the kth mouth sequence. $\hat{o}_{ktj}^{v}$ and o_{ktj}^{v} are the jth learned and real visual feature parameters of frame t. The obtained MRD scores are 3.4655 for the visual features learned from AF_AVDBN(3), and 3.8188 for SA_DBN(1), respectively, showing that more accurate visual features can be learned from the AF_AVDBN(3) model.

TABLE II
SUBJECTIVE EVALUATION RESULTS

Model	Items	1	2	3	4	5	MOS
AF_AV DBN(3)	accuracy	0	15	110	214	111	3.94
	clearness	0	11	215	207	17	3.51
	naturalness	0	6	185	212	47	3.64
S_A_ DBN(1)	accuracy	2	41	125	207	75	3.69
	clearness	6	111	220	101	12	3.00
	naturalness	5	70	193	153	29	3.29

Subjective tests are also performed to evaluate the quality of the synthesized mouth animations. 15 students have been asked to perform an acceptability test on the arbitrarily chosen 30 of the 80 synthesized image sequences. The indicators are: accuracy of the synthesized mouth movements, clearness, and overall naturalness of the mouth animations. The Mean Opinion Score (MOS) is used as a measure on a five point scale: 1 (bad), 2 (poor), 3 (fair), 4 (good) and 5 (excellent). Table II gives the MOS of the synthesized mouth animations from AF_AVDBN(3) and S_A_DBN(1), respectively. One can notice that the overall MOS values of the mouth animations from AF_AVDBN(3) are much higher than those from the S_A_DBN(1) model, accurate mouth movements can be obtained matching the input speech.

VI. CONCLUSIONS AND FUTURE WORK

This paper proposes a continuous speech driven photo realistic mouth animation synthesis approach based on AF_AVDBN, and a high frequency components compensation method is designed to synthesize clear mouth images from the learned YUV features. Objective and subjective evaluation results show that the proposed method is capable of synthesizing more natural, clear and accurate mouth animations than those from the state asynchronous DBN model S_A_DBN.

In our future work, we would try to adopt active appearance model (AAM) features to obtain mouth animations with higher quality, and expand the proposed method to large vocabulary continuous speech.

ACKNOWLEDGMENT

This work is supported within the framework of the LIAMA-CAVSA project, the EU FP7 project ALIZ-E (grant 248116), Shaanxi Provincial Key International Cooperation Project(2011KW-04), and the NPU Foundation for Fundamental Research (NPU-FFR-JC200943).

REFERENCES

[1] K.Choi et al, "Hidden Markov Model Inversion for Audio-to-visual Conversion in an MPEG-4 Facial Animation System". *Journal of VLSI Signal Processing*, vol.29, pp.51-61, 2001.

[2] L. Terissi, J. Gomez, "Audio-to-Visual Conversion Via HMM Inversion for Speech-Driven Facial Animation", Lecture Notes on Artificial Intelligence, LNAI 5249, pp.33-42, 2008.

[3] L. Xie, Z.-Q. Liu, "Realistic Mouth-Synching for Speech Driven Talking Face Using Articulatory Modelling". IEEE Transactions on Multimedia, 9(3), pp.500-510, 2007.

[4] K. Livescu, et al, "Articulatory Feature-based Methods for Acoustic and Audio-Visual Speech Recognition: 2006 JHU Summer Workshop Final Report". Center for Language and Speech Processing, Johns Hopkins University, 2006.

[5] D. Jiang, I. Ravyse, P. Liu, H. Sahli, W. Verhelst, "Realistic Mouth Animation Based on an Articulatory DBN Model With Constrained Asynchrony", Proc. 35th IEEE Int. Conf. Audio, Speech and Signal Processing (ICASSP), pp.2478-2481, 2010.

[6] D. Jiang, P. Wu, F. Wang, H. Sahli. "Audio Visual Speech Recognition Based on Multi-Stream DBN Models with Articulatory Features", Proc. ISCSLP, pp.190-193, 2010.

[7] Y. Hou, H. Sahli, I. Ravyse, Y. Zhang, R. Zhao, "Robust Shape Based Head Tracking". Proc of the Advanced Concepts for Intelligent Vision Systems (ACIVS), pp.340–351, 2007.

[8] The Int. Telecommunication Union, "Information Technologie – Digital Compression and Coding of Continuous Tone Still Images", CCITT Recommendation T.81, Annex K, 1993.

[9] J. Bilmes, G. Zweig. "The Graphical Models Toolkit: An Open Source Software System for Speech and Time Series Processing", Proc. ICASSP, vol. 4, pp.3916-3919, 2002.