ACOUSTICALLY-DRIVEN TALKING FACE SYNTHESIS USING DYNAMIC BAYESIAN NETWORKS

Jianxia Xue[1], Jonas Borgstrom[1], Jintao Jiang[2], Lynne E. Bernstein[2], Abeer Alwan[1]

[1]University of California, Los Angeles, CA 90095, USA, {jxue, jonas, alwan}@ee.ucla.edu
[2]House Ear Institute, Los Angeles, CA 90057, USA, {jjiang, lbernstein}@hei.org

ABSTRACT

Dynamic Bayesian Networks (DBNs) have been widely studied in multi-modal speech recognition applications. Here, we introduce DBNs into an acoustically-driven talking face synthesis system. Three prototypes of DBNs, namely independent, coupled, and product HMMs were studied. Results showed that the DBN methods were more effective in this study than a multilinear regression baseline. Coupled and product HMMs performed similarly better than independent HMMs in terms of motion trajectory accuracy. Audio and visual speech asynchronies were represented differently for coupled HMMs versus product HMMs.

1. INTRODUCTION

Highly-intelligible talking face synthesis systems could facilitate speech comprehension in noise and enhance human-machine interactions. Acoustically-driven talking face synthesis systems would be advantageous in low bandwidth applications such as wireless communications and internet video conferencing. In recent years, dynamic Bayesian networks (DBNs) have emerged as a powerful and flexible theoretical framework for multi-modal stochastic processes [1]. Different DBN configurations have been applied to audio-visual speech recognition ([2], [3], [4], [5]), and audio-visual speaker identification [6], etc. To the best of the authors' knowledge, this is the first study to compare different DBN configurations systematically for acoustically-driven talking face synthesis.

Among various configurations of DBNs, three were chosen for this study: independent HMMs (I-HMMs), coupled HMMs (C-HMMs), and product HMMs (P-HMMs). I-HMMs and P-HMMs represent the two extreme cases of state transition integration: complete independence and complete dependence, respectively. C-HMMs correlate the audio and visual speech models using conditionally independent audio-visual hidden state transitions. The three DBN configurations were implemented and evaluated in an acoustically-driven talking face synthesis context. Basic model selection parameters were studied under the synthesis framework with quantitative and qualitative evaluations of the synthesized talking face.

The paper is organized as follows. Section 2 describes the synthesis system. Section 3 describes the three DBN prototypes. The experimental setup and results are presented in Section 4. Discussion and conclusions are in Section 5.

2. TALKING FACE SYNTHESIS SYSTEM

The talking face synthesis system comprised four major components: feature extraction, acoustic-to-optical mapping, optical feature inversion, and deformation (see Fig. 1). The performance of the synthesizer was evaluated objectively using a set of measurements.

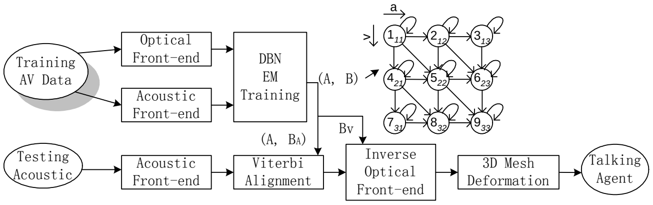


Figure 1. Acoustically-driven talking face synthesis system. **A** represents the transition probability parameters, and **B** represents the observation probability parameters.

2.1. Feature extraction

Two representations of speech acoustics were used: Linear Predictive Cepstral Coefficients (LPCCs, A_{LPCC}) for backend modeling and Line Spectral Pairs (LSPs, A_{LSP}) for optical feature transformation (see Eq. 1). A previous study [7] showed that LSPs resulted in better linear estimation of optical features than LPCCs. Our pilot studies confirmed that the combination of the two acoustic feature representations was better than either of the single representations in the synthesis framework.

Optical data (details in Sec. 4.1) were pre-processed in five steps: compensation for missing data, eye-brow motion, and head motion, noise removal, and head-size normalization.

Optical feature extraction comprised three steps. Let V_{Disp} be the normalized displacement features relative to a neutral facial gesture and obtained from the preprocessed optical data V. Let W_{LMS} be the matrix for a global transformation from A_{LSP} to V_{Disp}. Firstly, W_{LMS} was estimated via least-mean square (LMS) estimation [7]. Then the residual optical signal V_R was obtained as follows:

$$V_R = V_{Disp} - A_{LSP} W_{LMS} . \qquad (1)$$

Finally, principal component analysis (PCA) was applied to V_R for data dimension reduction. Reduced optical features V_{RPC} were used for back-end modeling, and the corresponding inverse principle component transformation W_{IPCA} were used for visual feature inversion.

2.2. Acoustic-to-optical mapping

The mappings from the acoustical feature vectors A_{LPCC} to the optical feature vectors V_{RPC} were modeled using the three DBN setups: I-HMMs, P-HMMs, and C-HMMs. For all configurations, context-independent phoneme models were trained through the typical HMM training procedures of Viterbi initialization, isolated model re-estimation, and embedded model re-estimation using the modified expectation maximization (EM) algorithm. The acoustical and optical models were trained separately on phoneme units for I-HMMs, but jointly for C-HMMs and P-HMMs. In synthesis, the trained acoustic-to-optical mapping models were applied to A_{LPCC} for Viterbi forced alignment, using incomplete feature inference of DBNs (details in Section 3).

2.3. Optical feature inversion

In this operation, residual optical feature vectors were converted back to optical position feature vectors. The Gaussian means of the optical features were placed in the middle of the corresponding inferred hidden visual states. Cubic polynomial interpolations were followed to generate continuous residual optical features $\hat{V}_{RPC}$. Then the normalized displacement features were estimated as follows:

$$\hat{V}_{Disp} = A_{LSP}W_{LMS} + \hat{V}_{RPC}W_{IPCA} \, . \tag{2}$$

Finally, the position trajectories were recovered by adding the neutral marker positions to the displacement trajectories.

2.4. Deformation

Radial basis functions [8] were modified for the position mapping from M key points (see Sec. 4.1 for details) to 3D face model vertices as follows:

$$\begin{aligned} p_j^k(t) &= p_j^k(0) + \sum_{m=1}^{M} w_m^k(t)\varphi_{jm}^k(t) \\ &= p_j^k(0) + \sum_{m=1}^{M} w_m^k(t)\exp\left(-\frac{(p_j^k(0)-v_m^k(0))^2}{2\sigma_m^2(t)}\right), \end{aligned} \tag{3}$$

where k refers to the k-th dimension of the 3-D position data, $p_j(t)$ is the position of vertex j, $v_m(t)$ is the position of key point m at time t, t of 0 corresponds to the neutral facial gesture, $\{w_m^k(t)\}$ are the key-point weights for the k-th dimension obtained from the linear equations:

$$\begin{pmatrix} \varphi_{11}^k(t) & \cdots & \varphi_{1M}^k(t) \\ \vdots & \ddots & \vdots \\ \varphi_{M1}^k(t) & \cdots & \varphi_{MM}^k(t) \end{pmatrix} \begin{pmatrix} w_1^k(t) \\ \vdots \\ w_M^k(t) \end{pmatrix} = \begin{pmatrix} v_1^k(t)-v_1^k(0) \\ \vdots \\ v_M^k(t)-v_M^k(0) \end{pmatrix}, \tag{4}$$

and $\sigma_m(t)$ is obtained by solving the equation:

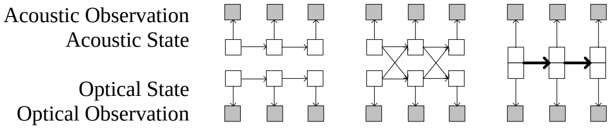
$$\exp\left(-\frac{\min\limits_{l=1\ldots M,\, l\neq m}\|v_l(t)-v_m(t)\|_2^2}{2\sigma_m^2(t)}\right) = \tau \, , \tag{5}$$

where τ is a threshold determined empirically.

2.5 Evaluation measurements

Synthesized optical data were compared to the original recordings using several measurements to objectively evaluate the synthesizers. The measurements were normalized Manhattan and Euclidean distance (N-M and N-E), Kullback-Leibler distance (K-L), and Pearson-r correlation (Corr). The first three measurements represent the deviation from the original data, while the last measurement corresponds to the goodness of fit to the original data.

3. DBNS FOR INCOMPLETE DATA INFERENCE



(a) I-HMM (b) C-HMM (c) P-HMM

Figure 2. Three configurations of DBNs [1].

3.1. Dynamic Bayesian networks

Dynamic Bayesian networks are directional graphical models. They are also a generalized form of the traditional hidden Markov models (HMMs) in the sense that DBNs allow multiple hidden state Markov chains. The physical concept of multi-modal speech processing can be more easily represented by the DBNs than in single-chain HMMs, as shown in Fig. 2.

The state transition probabilities are integrated differently across the three DBN configurations for I-HMMs in Eq. 6, C-HMMs in Eq. 7, and P-HMMs in Eq. 8 [1]:

$$a(\mathbf{i}|\mathbf{j}) = \prod_{s=1}^{2} a^s(i_s|j_s) = \prod_{s=1}^{2} P(q_t^s = i_s | q_{t-1}^s = j_s)\, , \tag{6}$$

$$a(\mathbf{i}|\mathbf{j}) = \prod_{s=1}^{2} a^s(i_s|\mathbf{j}) = \prod_{s=1}^{2} P(q_t^s = i_s | \mathbf{q}_{t-1} = \mathbf{j})\, , \tag{7}$$

and

$$a(\mathbf{i}|\mathbf{j}) = P(\mathbf{q}_t = \mathbf{i}|\mathbf{q}_{t-1} = \mathbf{j})\, , \tag{8}$$

where **i** and **j** are the current and previous hidden state vectors, respectively, and s is the index of a hidden Markov chain that corresponds to the acoustic or optical feature space.

Definitions for P-HMMs were found to be different in previous studies ([1], [2], [3]). In this study, for the purpose of comparing different levels of hidden state transition integration, the observation probabilities for the acoustic and optical feature spaces of each speech unit were all integrated independently in the three DBN configurations, as in [3]:

$$b_t(\mathbf{i}) = \prod_{s=1}^{2} b_t^s(i_s) = \prod_{s=1}^{2} P(\mathbf{O}_t^s \mid q_t^s = i_s)\, , \tag{9}$$

where $b_t^s(i_s)$ is the observation probability of state i_s in chain s, and O_t^s is the observation at time t in chain s.

For the three methods, given the maximum inter-chain state asynchrony (MICSA) and the number of hidden states for each model, the joint state transition probability matrices follow the same non-zero structures. However, in addition to the transition probability constraint, relationships among the elements are different for each prototype. In I-HMMs,

$$\mathbf{A} = \mathbf{A}^a \otimes \mathbf{A}^v \, , \tag{10}$$

where **A** is the DBN transition matrix, $\mathbf{A}^a$ and $\mathbf{A}^v$ are the HMM transition matrices trained independently from the acoustic and optical feature spaces of a speech unit, and "$\otimes$" represents the Kronecker product. In C-HMMs,

$$a([i^a, i^v] \mid \mathbf{j}) = (\sum_{k^a=1}^{N^a} a([k^a, i^v] \mid \mathbf{j})) \cdot (\sum_{l^v=1}^{N^v} a([i^a, l^v] \mid \mathbf{j})), \qquad (11)$$

where N^a and N^v are the number of hidden states for the two Markov chains. In the P-HMMs, no other relationships exist among the elements in **A**.

The EM algorithms [1] were adapted into a single chain HMM implementation with state observation binding for both C-HMMs and P-HMMs and joint-state transition constraint for C-HMMs. Both C-HMMs and P-HMMs should yield better training accuracy than I-HMMs, because of the inherent state dependencies and the integrated training procedures. Given sufficient training data, P-HMMs should yield the best training accuracy. However, P-HMMs require the most training data for reliable parameter estimation.

3.2. DBNs for incomplete feature inference

In the synthesis system, complete features (acoustic and optic) were used in training. However, incomplete features (acoustic only) were used in inference. In this study, the Viterbi algorithm was modified for the incomplete feature inference as follows:

$$\tilde{q}(t) = \arg\max_i \tilde{\phi}_i(t), \qquad (12)$$

where $\tilde{\phi}_i(t)$ is the partial forward probability of observation $\mathbf{O}_t$ at state i. Let $\phi_i(t)$ be the corresponding complete forward probability. $\tilde{\phi}_i(t)$ and $\phi_i(t)$ have the relationship:

$$\begin{aligned}\tilde{\phi}_{\mathbf{i}}(t) &= \sum_{\mathbf{j}} \tilde{\phi}_{\mathbf{j}}(t-1) a(\mathbf{i} \mid \mathbf{j}) p(\mathbf{O}_t^a \mid q_t^a = i_a) \\ &= \frac{\phi_{\mathbf{i}}(t)}{p(\mathbf{O}_t^v \mid q_t^v = i_v)} + \sum_{\mathbf{j}} \varepsilon_{\mathbf{j}}(t-1) a(\mathbf{i} \mid \mathbf{j}) p(\mathbf{O}_t^a \mid q_t^a = i_a),\end{aligned} \qquad (13)$$

where $\varepsilon_{\mathbf{j}}(t)$ represents the partial forward probability error $\tilde{\phi}_{\mathbf{j}}(t) - \phi_{\mathbf{j}}(t)$. The incomplete feature inference error $\tilde{q}(t) - q(t)$ cannot be represented by an analytical function with regards to transition matrix **A**. Physical interpretation of state transition integrations implies that C-HMMs have less cross-model dependency between the audio and visual models, potentially providing less inference error than P-HMMs when training accuracies of the two methods are identical. The competition between C-HMMs and P-HMMs in the present application is related to the tradeoff between training accuracy and incomplete feature inference error. Given limited training data, C-HMMs have the potential of performing as well as P-HMMs.

4. EXPERIMENTS

4.1. Database and experiment setups

The database comprised 320 sentences by a single talker using an audio-visual recording setup [9]. The optical data were obtained from simultaneously recorded 3D positions of 20 markers (see Fig. 3). The sampling rates of the optical and the clean acoustic data were 120 Hz and 44.1 KHz, respectively. Manual phoneme segmentations were obtained using the acoustic signal. The edited 3D polygon face meshes (originated from www.digimation.com) had 1915 vertices.

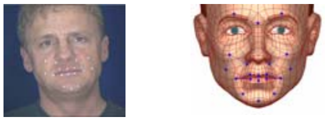


Figure 3. Recorded marker positions (left) and deformed face model and key-points (right).

The frame rate for acoustic feature extraction was 120 Hz. The dimensions of acoustic feature vectors A_{LPCC} and A_{LSP} were 15 and 17, respectively. The nose-ridge marker was removed after head-motion compensation. Optical feature vectors O_{Disp} had 57 elements. The dimension of O_{RPC} was reduced to 20, accounting for 99% of the variance. W_{LSP} was a 17x57 matrix estimated from 2 minutes of training data. W_{IPCA} was a 20x57 matrix obtained using the entire optical training data.

In each DBN configuration, 41 context-independent phoneme DBN models were trained. The same numbers of hidden states were used in the two Markov chains. The maximum intra-chain state asynchrony was one. Model selection parameters, $[N^a, N^v]$ and maximum inter-chain state asynchrony (MICSA) were studied.

4.2. Baseline

The baseline was obtained using a multilinear regression method [7]. The acoustic and optical feature vectors were A_{LSP} (17 elements) and O_{Disp} (57 elements), respectively. For each set of training and test data used in DBNs, the same training sentences were used to train acoustic-to-optical regression models for each phoneme using the manual segmentation, and then these regressors were applied to the corresponding phonemes in the test sentences.

4.3. Results

Due to limited data, a resampling procedure was applied to protect against bias in the acoustic-to-optical mapping. That is, one set was left out for testing and the remaining sets were used for training; a rotation was then performed to guarantee that each utterance was tested once. The results were averaged from the 320 utterances. Results from four measurements were consistent in the relationship among different methods as shown in Table 1. In the remainder of the section, the correlations are used for performance evaluation. Paired t-tests (df = 319) with Bonferroni correction for multiple comparisons ($p < 0.05$) were applied on the correlation vectors of all the methods or conditions. All the DBN methods performed significantly better than the baseline ($p < 0.05$). C-HMMs and P-HMMs performed similarly better than I-HMMs ($p < 0.05$). Context-independent modeling limited the overall performances.

C-HMMs generated the highest state path entropy (see Table 2). The state path distribution resulting from C-

HMMs showed a higher percentage of paths corresponding to facial motion events beginning before and ending after acoustic events than to paths where acoustic events preceded facial ones (36% of VCA vs. 5% of APV3).

Table 3 shows the correlation results of the three DBN approaches with different numbers of joint states, which are a function of $[N^a, N^v]$ and MICSA. The latter parameter had a significant effect on the results ($p < 0.05$). As the complexity of the model increased, results with C-HMMs approached those with P-HMMs. As the joint states reached 16, the results of C-HMMs and P-HMMs degraded due to insufficient training data. In some resampling trials, P-HMMs failed in training for the same reason. These observations confirmed the theoretical comparison between the two DBN configurations as in Sec. 3.2.

Figure 4 shows the motion trajectories of a synthesized sentence. Results during connected speech were better than during acoustic silence due to the temporal differences. Animation demos are available at www.icsl.ucla.edu/~jxue.

Table 1. Quantitative evaluations using $[N^a, N^v]$ of [3, 3] and MICSA of 2 using training-to-testing ratio of 1.875:1.

	N-M	N-E	K-L	Corr
Baseline	.324	.058	.274	.179
I-HMM	.280	.049	.231	.440
C-HMM	.254	.044	.201	.549
P-HMM	.251	.044	.193	.559

Table 2. Average state path entropies with $[N^a, N^v]$ of [3, 3], and MICSA of 1. The upper bound of the entropy is 3.459 bits. DP refers to the most frequent state paths. APV3 refers to acoustic events (state transitions) ahead of facial events in mode 3. VCA refers to facial events starting before and ending after acoustic events.

	I-HMM	C-HMM	P-HMM
Entropy (bits)	1.354	2.985	2.592
DP (appearance%)	APV3 (70%)	VCA (36%)	APV3 (29%)

Table 3. Pearson-r correlation as a function of the number of joint states using training-to-testing ratio of 7:1.

$[N^a, N^v]$	[3,3]		[4,4]		
MICSA	1	2	1	2	3
#JointState	7	9	10	14	16
I-HMM	.427	.448	.464	.419	.422
C-HMM	.524	.543	.534	.562	.561
P-HMM	.548	.558	.536	.569	.563

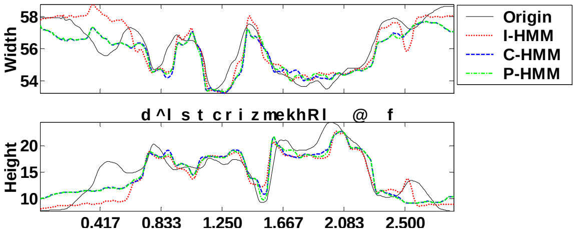


Figure 4. An example of the synthesized sentence "Dull stories make her laugh." The plots are the trajectories of lip spreading width (top) and lip opening height (bottom) in millimeters.

5. DISCUSSION AND CONCLUSIONS

The three tested DBN methods were superior to the multilinear regression method. C-HMMs and P-HMMs generated similarly better results than I-HMMs, suggesting the effectiveness of the state dependency structure in the first two methods. C-HMMs generated higher state transition path entropy and captured more state asynchrony between the audio and visual models than P-HMMs. Maximum inter-chain state asynchrony had a greater effect on synthesis accuracy than the numbers of hidden states in the two Markov chains. This study demonstrated the potential for DBNs in acoustically-driven talking face synthesis. Formal perception tests are in preparation for visual intelligibility evaluation of the synthesis system. In future work, combining DBN methods and visual feature re-estimation and optimization methods with context-dependent modeling will be pursued to improve system performance.

6. ACKNOWLEDGMENTS

This research was supported in part by NSF grants IIS 0312810 (to A. Alwan) and IIS 0312434 (to L.E. Bernstein).

7. REFERENCES

[1]. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP J. on Applied Signal Process.*, vol. 11, 1-15, 2002.

[2]. S. Nakamura, "Statistical multimodal integration for audio-visual speech processing," *IEEE Trans. Neural Networks*, vol. 13, no. 4, 854 – 866, 2002.

[3]. P. S. Aleksic and A. K. Katsaggelos, "Product HMMs for audio-visual continuous speech recognition using facial animation parameters," *Proc. ICME*, vol. 2, 481-484, 2003.

[4]. S. M. Chu and T. S. Huang, "An experimental study of coupled hidden Markov models," *Proc. ICASSP*, vol. 4, 4100-4103, 2002.

[5]. G. Potamianos, C. Neti, and S. Deligne, "Joint audio-visual speech processing for recognition and enhancement," *Proc. AVSP*, 95-104, St. Jorioz, France, 2003.

[6]. S. Lucey, T. Chen, S. Sridharan, and V. Chandran, "Integration strategies for audio-visual speech processing: Applied to text-dependent speaker recognition," *IEEE Trans. Multimedia*, vol. 7, no. 3, 495-506, 2005.

[7]. J. Jiang, "Relating optical speech to speech acoustics and visual speech perception," *Ph.D. Dissertation*, UCLA, 2003.

[8]. M.D. Buhmann, *Radial Basis Functions: Theory and Implementation.* Cambridge University Press, 2003.

[9]. L.E. Bernstein, E.T. Auer, B. Chaney, A. Alwan, and P.A. Keating, "Development of a facility for simultaneous recording of acoustic, optical 3-D motion and video, and physiological speech data," *J. Acoust. Soc. Am.,* 107, 2887, 2000.