AN ARTICULATORY APPROACH TO VIDEO-REALISTIC MOUTH ANIMATION

Lei Xie and Zhi-Qiang Liu

{xielei, zq.liu}@cityu.edu.hk
School of Creative Media
City University of Hong Kong
Kowloon, Hong Kong S.A.R., China

ABSTRACT

We propose an articulatory approach which is capable of converting speaker independent continuous speech into video-realistic mouth animation. We directly model the motions of articulators, such as lips, tongue, and teeth, using a Dynamic Bayesian Network (DBN)-structured articulatory model (AM). We also present an EM-based conversion algorithm to convert audio to animation parameters by maximizing the likelihood of these parameters given the input audio and the AMs. We further extend the AMs with introduction of speech context information, resulting in context dependent articulatory models (CD-AMs). Objective evaluations on the JEWEL testing set show that the animation parameters estimated by the proposed AMs and CD-AMs can follow the real parameters more accurately than that of phoneme-based models (PMs) and their context dependent counterparts (CD-PMs). Subjective evaluations on an AV subjective testing set, which collects various AV contents from the Internet, also demonstrate that the AMs and CD-AMs are able to generate more natural and realistic mouth animations and the CD-AMs achieve the best performance.

1. INTRODUCTION

Computer animated talking faces entertain humans in various multimedia applications, such as virtual agents, newsreaders, video games, and videophones. Different from text-driven ones which use synthesized voices with few prosody, speech-driven talking faces are capable of converting real speech to facial animations with high fidelities of both audio and video. The essential problem of speech-driven talking faces is *mouth-synching*: the synchronization of the animated mouth with speech. Various approaches have been proposed [1], which generally can be categorized to physics-based, sample-based, and statistical-model-based. Even a mouth animation system without the corresponding face still can help hearing impaired people to better communicate with machines through lipreading.

Recently, due to the success in modelling speech, many researchers have used hidden Markov models (HMMs) to solve the mouth-synching problem. Some of them converted speech into linguistic units (such as phonemes) using an HMM-based Viterbi recognizer, and mapped these units to pre-defined mouth animation parameters [2]. Others directly estimated the visual parameters from acoustic speech and bypassed the Viterbi search which was considered as lack of robustness to ambient noise. Choi *et al* [3] presented a Baum-Welch HMM inversion approach, which trained audio visual phoneme HMMs (we call this model PM) and animation parameters were directly generated by Baum-Welch iterations.

This research has been supported in part by a research grant CityU 1247/03E from Hong Kong UGC RGC.

Although these HMM-based approaches can provide reasonable mouth-synching performance, they still lack natural audio visual synchronization compared with real recordings. This is probably because these approaches adopt the conventional phoneme-based speech modelling schemes, which does not incorporate any knowledge of the source that articulates speech.

In this paper, we present an articulatory approach to realize video-realistic mouth animation. We directly model the movements of articulators, such as lips, tongue and teeth, using articulatory models (AMs). Speaker independent continuous speech is converted to mouth animation based on an EM-based conversion algorithm. As autosegmental phonology indicates [4], speech may be better described by asynchronous motion of articulators than by rigid line-up of phonemes. Articulatory modelling has many advantages such as being better able to predict coarticulation effects. Therefore, a mouth animation system may benefit from directly describing the action of articulators and achieve more realistic performance.

2. ARTICULATORY MODELS

2.1. Model Structure

Articulators are formally known as speech organs such as glottis, vocal folds, velum, lips and tongue. During the last decade, there has been much interest in modelling articulators for speech. Recently, due to the great expressive power of Dynamic Bayesian Networks (DBNs) in modelling speech phenomena, Bilmes *et al.* [5] proposed a prototype of DBN-based articulatory model for speech recognition. We extend this model to solving mouth-synching problem.

Fig. 1 shows the repeating structure of our articulatory model (AM), where two successive time frames are given. Similar to that in [5], a layer representing various articulators is inserted between linguistic states and observation variables. In each frame t, there is a set of articulators $\Psi_t = \{\psi_t^1, \psi_t^1, \cdots, \psi_t^D\}$, each of which depends on the current state variable q_t and its own value in the previous frame. The time dependency between successive articulator values is to model the natural continuity constrains.

Different from [5], our model removes the complicated syntax layers for word decoding, since we are not interested in the syntax the utterance conveys, but how the visual observations match the acoustic speech. Another difference is that our model includes two observation streams ($\mathbf{o}_t^a$ and $\mathbf{o}_t^v$) each of which describes one modality of speech–audio and video. Since the audio and visual observations are originated from the same articulator source, only a unique articulator layer is incorporated. As several articulators such as *velum* cannot be observed visually, visual observations are up-linked only to those *visible* articulators. Not only does this multi-stream structure mimic the true human articulatory system to a certain extent, but also encapsulates the *synchrony* between the audio

and video and the *asynchrony* between different articulators; and thus may lead to a better mouth-synching performance.

2.2. Model Parameters

The conditional probability distributions (CPDs) associated with variables for each frame describe the following probabilities:

- $P(q_t|q_{t-1})$: state transition probability,
- $P(\psi_t^i|\psi_{t-1}^i, q_t), i = 1, 2, \cdots, D$: articulator generation probability, and
- $P(\mathbf{o}_t^s|\Psi_t^s), s \in \{a, v\}$: audio (visual) observation emission probability.

Since q_t and ψ_t^i are discrete variables, the probabilities $P(q_t|q_{t-1})$ and $P(\psi_t^i|\psi_{t-1}^i, q_t)$ are described in a tabular way, known as conditional probability tables (CPTs). $P(\mathbf{o}_t^s|\Psi_t^s)$ is described as Gaussian mixtures:

$$P(\mathbf{o}_t^s|\Psi_t^s) = \sum_{k=1}^{K} \omega_{\Psi_t^s k} \mathcal{N}(\mathbf{o}_t^s; \mu_{\Psi_t^s k}, \mathbf{\Sigma}_{\Psi_t^s k}) \;\; s \in \{a, v\}, \qquad (1)$$

where $\mathcal{N}(\mathbf{o}_t^s; \mu_{\Psi_t^s k}, \mathbf{\Sigma}_{\Psi_t^s k})$ is a multivariate Gaussian with mean vector $\mu_{\Psi_t^s k}$ and covariance matrix $\mathbf{\Sigma}_{\Psi_t^s k}$, and $\omega_{\Psi_t^s k}$ denotes the mixture weight for the k^{th} Gaussian. Each allowed combination of articulator values is implemented via a Gaussian mixture.

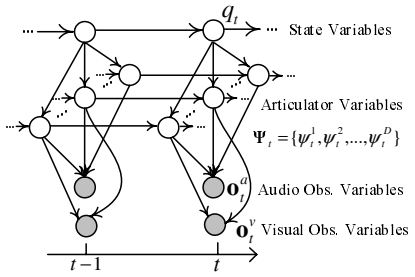

Fig. 1. Articulatory Model

2.3. Articulator Variables and Parameter Training

Speech production is a complicated process which involves various articulators. Prior to modelling, we must consider what articulatory information can be encapsulated. We use pseudo-articulatory information since they are widely used in the speech processing literature where statistical classifications are applied to acoustic speech, resulting in abstract classes representing articulator configurations.

According to [6], we define our articulator variables (with discrete values) as follows[1]: *voicing* (on, off), *velum* (open, closed), *manner* (closure, sonorant, fricative, burst), *lip rounding* (rounded, slightly rounded, mid, wide), *tongue show* (touching top teeth, near alveolar ridge, touching alveolar, others), *teeth show* (on,off).

In order to constrain the articulatory variables to their intended meanings, we use a two-step parameter training procedure. Firstly, we train the articulator generation probabilities $P(\psi_t|\psi_{t-1}^i, q_t)$ using probability factoring and supervised learning; secondly, we train the state transition probabilities $P(q_t|q_{t-1})$ and the observation emission probabilities $P(\mathbf{o}_t^s|\Psi_t^s)$ using the EM algorithm together with the pre-trained articulator generation probabilities.

[1]Lip rounding, tongue show and teeth show are visible articulator variables

To get a reasonable guess of $P(\psi_t^i|\psi_{t-1}^i, q_t)$, we factor it into the following form:

$$P(\psi_t^i|\psi_{t-1}^i, q_t) = \frac{P(\psi_t^i|q_t)P(\psi_t^i|\psi_{t-1}^i)}{N(\psi_t^i)}. \qquad (2)$$

Thus we build two separate CPTs for:

- $P(\psi_t^i|q_t)$: state-to-articulator mapping probability and
- $P(\psi_t^i|\psi_{t-1}^i)$: articulator transition probability.

The final CPT of $P(\psi_t^i|\psi_{t-1}^i, q_t)$ is constructed by multiplying the appropriate items in the above two CPTs and normalized by a constant $N(\psi_t^i)$.

Since we use 47 English phonemes as the linguistic units, we manually examine each phoneme state's articulatory characteristics to determine the best mapping to the articulator values and created a phoneme-state-to-articulators mapping table [7]. Based on this table, we use a supervised learning method [7] to train $P(\psi_t^i|q_t)$ and $P(\psi_t^i|\psi_{t-1}^i)$.

3. EM-BASED AUDIO TO VISUAL CONVERSION

Since our DBN-structured articulatory model is to output appropriate visual parameters given the only audio input, we need an audio to visual conversion algorithm based on an optimal criterion. We use the Maximum Likelihood (ML) criterion to find the optimal visual parameters $\hat{\mathbf{O}}^v$, maximizing the likelihood of visual parameters given the audio data $\mathbf{O}^a$ and the trained articulatory models λ. The EM algorithm is used in our ML approach.

According to the EM algorithm, the optimal visual parameter sequence $\hat{\mathbf{O}}^v$ can be found by iteratively maximizing the auxiliary function $\mathcal{Q}(\lambda, \lambda; \mathbf{O}^a, \mathbf{O}^v, \mathbf{O}^{v'})$, i.e.,

$$\hat{\mathbf{O}}^v = \arg \max_{\mathbf{O}^{v'} \in \mathcal{O}^v} \mathcal{Q}(\lambda, \lambda; \mathbf{O}^a, \mathbf{O}^v, \mathbf{O}^{v'}), \qquad (3)$$

where $\mathbf{O}^v$ and $\mathbf{O}^{v'}$ denote the old and new visual parameter sequences in the visual parameter space $\mathcal{O}^v$ respectively.

As described in EM, given the model with audio visual observation streams $\mathbf{O}^a$ and $\mathbf{O}^v$ in Fig. 1, the complete-data likelihood function is $P(\mathbf{O}^a, \mathbf{O}^v, q, \Psi|\lambda)$. Given Ψ and q and a trained model set λ, according to the Markov property of independent relationships between variables, the complete-data likelihood can be formed as

$$P(\mathbf{O}^a, \mathbf{O}^v, q, \Psi|\lambda) = \prod_{t=1}^{T} \left[P(q_t|q_{t-1}) \prod_{i=1}^{D} P(\psi_t^i|\psi_{t-1}^i, q_t) P(\mathbf{o}_t^a|\Psi_t^a) P(\mathbf{o}_t^v|\Psi_t^v) \right]. \qquad (4)$$

Then $\mathcal{Q}(\lambda, \lambda; \mathbf{O}^a, \mathbf{O}^v, \mathbf{O}^{v'})$ can be expressed as

$$\begin{aligned} &\mathcal{Q}(\lambda, \lambda; \mathbf{O}^a, \mathbf{O}^v, \mathbf{O}^{v'}) \\ &= \sum_q \sum_\Psi P(\mathbf{O}^a, \mathbf{O}^v, q, \Psi|\lambda) \log P(\mathbf{O}^a, \mathbf{O}^{v'}, q, \Psi|\lambda) \\ &= \sum_q \sum_\Psi P(\mathbf{O}^a, \mathbf{O}^v, q, \Psi|\lambda) \{ \sum_{t=1}^{T} \log P(q_t|q_{t-1}) \\ &\quad + \sum_{t=1}^{T} \sum_{i=1}^{D} \log P(\psi_t^i|\psi_{t-1}^i, q_t) \\ &\quad + \sum_{t=1}^{T} \log P(\mathbf{o}_t^a|\Psi_t^a) + \sum_{t=1}^{T} \log P(\mathbf{o}_t^{v'}|\Psi_t^v) \}. \end{aligned} \qquad (5)$$

By taking the derivative of $\mathcal{Q}(\lambda, \lambda; \mathbf{O}^a, \mathbf{O}^v, \mathbf{O}^{v'})$ with respect to

$\mathbf{o}_t^{v'}$ to zero, i.e.,

$$\frac{\partial \mathcal{Q}(\lambda,\lambda;\mathbf{O}^a,\mathbf{O}^v,\mathbf{O}^{v'})}{\partial \mathbf{o}_t^{v'}}$$
$$=\sum_q \sum_{\Psi} P(\mathbf{O}^a,\mathbf{O}^v,q,\Psi|\lambda)\frac{\partial}{\partial \mathbf{o}_t^{v'}}\left[\log P(\mathbf{o}_t^{v'}|\Psi_t^v)\right]$$
$$=\sum_{q_t}\sum_{\Psi_t}\sum_k P(\mathbf{O}^a,\mathbf{O}^v,q_t,\Psi_t|\lambda)\omega_{\Psi_t^v k}\Sigma^{-1}_{\Psi_t^v k}(\mathbf{o}_t^{v'}-\mu_{\Psi_t^v k})=0, \quad (6)$$

where q_t and Ψ_t ($\Psi_t^v \subset \Psi_t$) denote possible values of state variable and possible value sets of articulatory variables at time t respectively. We define

$$\gamma_t(q_t,\Psi_t) = P(\mathbf{O}^a,\mathbf{O}^v,q_t,\Psi_t|\lambda) \quad (7)$$

as the *occupation probability*. We can find the re-estimated inputs $\mathbf{o}_t^{v'}$ by iteratively computing Eq. (8) until convergence.

$$\mathbf{o}_t^{v'} = \frac{\sum_{q_t}\sum_{\Psi_t}\sum_k \gamma_t(q_t,\Psi_t)\omega_{\Psi_t^v k}\Sigma^{-1}_{\Psi_t^v k}\mu_{\Psi_t^v k}}{\sum_{q_t}\sum_{\Psi_t}\sum_k \gamma_t(q_t,\Psi_t)\omega_{\Psi_t^v k}\Sigma^{-1}_{\Psi_t^v k}}. \quad (8)$$

The occupation probabilities $\gamma_t(q_t,\Psi_t)$ can be computed using the *frontier algorithm* [8].

4. CONTEXT DEPENDENT ARTICULATORY MODELS

For continuous speech with a large vocabulary, context dependent speech units, known as *triphones* and *biphones*, are better used. More accurate relationships between articulators and speech observations can be modelled using the context information although no direct dependency exists between the linguistic states and the speech observations in our model(see Fig. 1). This is because context dependent modelling can categorize speech into more accurate speech atoms, and the training data for the observation emission probabilities will thus become more accurate.

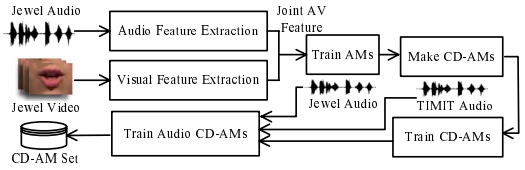

Fig. 2. Training procedure for CD-AMs

Fig. 2 shows the training procedure for the context dependent articulatory models (CD-AMs). We use the TIMIT speech database and the JEWEL audio-visual database [9] in the training since our object is to achieve speaker independent mouth-synching and abundant audio samples from various speakers are needed to get good distributions of acoustic signals. we use MFCCs as well as their velocity and acceleration derivatives as audio features, leading to a set of 39 parameters for each frame. The principal component analysis (PCA) is implemented to the R, G, and B channels of JEWEL mouth images separately, generating a set of 90 visual parameters (30 for each channel) for each image frame. We up-sample visual parameters from 25Hz to 100Hz to meet the audio feature rate.

Firstly we use synchronized audio-visual features extracted from the JEWEL training set to train AMs. This training process ensures the synchronization of the audio and visual signals. Secondly, we convert the phoneme-based transcriptions to equivalent word internal triphone and biphone transcriptions. For example, a phoneme sequence of "$\cdots$ *w ao sh sp w ao t axr* $\cdots$" is converted to "$\cdots$ *w+ao w-ao+sh ao-sh sp w+ao w-ao+t ao-t+axr t-axr* $\cdots$". Totally we collect 3932 context dependent units (triphones, biphones and phonemes). Correspondingly, the state-to-articulator mapping probability for the new CD-AM is

$$P^{CD}(\psi^i|q=\xi_j) = \sum_{j=1}^{J} a_j P(\psi^i|q=l_j) + bP(\psi^i|q=p_j) + \sum_{j=1}^{J} c_j P(\psi^i|q=r_j), \quad (9)$$

where ξ denotes a J-state CD-AM with the form $l-p+r$, and $\sum_j a_j + b + \sum_j c_j = 1$. This new mapping probability is designed to introduce the co-articulation effects from the neighboring sounds. The weights a, b, c are experientially chosen from a development set.

Finally we use the TIMIT training audio together with the JEWEL training audio (totally 5072 utterances) to train the audio observation emission probabilities of CD-AMs ($P(\mathbf{o}_t^a|\mathbf{\Psi}_t^a)$), while keep the visual observation emission probabilities ($P(\mathbf{o}_t^v|\mathbf{\Psi}_t^v)$) unchanged.

5. EXPERIMENTS

To evaluate the performance of the proposed articulatory approach, we have carried out objective and subjective experiments. We have compared our technique with Choi's phoneme-based model (PM) [3] and its context dependent counterpart (CD-PM), since they all directly estimate visual parameters from acoustic speech, and use similar audio to visual conversion mechanisms.

Although the proposed EM-based conversion algorithm in Section 3 is able to estimate visual parameters using all possible linguistic states (all q_t), the computing is time-consuming. We instead used a sub-optimal approach that an *N*-Best state list was computed for each utterance frame by a separate speech recognizer built by the HTK toolkit [10]. As a result, only the N most likely states were involved in the estimation iterations (Eq. (8)) and the computing time was significantly reduced with a comparable performance.

We trained a set of 47 three-state,left-to-right phoneme HMMs for the PM system and correspondingly a set of 3932 context dependent HMMs with the same structure for the CD-PM system. The *N*-Best state lists were also used in their HMMI-based conversion [3]. For the AM and CD-AM system, considering the physical constraints of the articulators and the database context, we trained 472 and 236 sets of Gaussian mixtures for the audio and visual respectively. For all the testing systems, a five-continuous-Gaussian-mixture was used for each HMM state or each combination of articulator values.

5.1. Objective Evaluations

We have carried out objective evaluations on the testing set of JEWEL database. Fig. 3 depicts the estimated visual parameter time trajectories (the most significant PCA component for the R channel), compared with the actual parameters extracted from the original video for a testing sentence. The curves show that all the testing systems were able to generate visual parameter sequences that follow the approximate shapes of the actual parameter curves. However, the PM system produced obvious estimation errors. For example, large errors lie in the neighborhoods of frame 52, 75 and 275. By taking account of the context information, the CD-PM system was observed reduced errors. The curve generated by the proposed AM system was able to follow the actual one more closely, and the best matching performance was achieved by the CD-AM system. This is mainly because our articulatory approach can describe better the coarticulation effect than the conventional phoneme-based approaches tested.

To make quantitative evaluations, we have calculated the *percentage normalized mean error* (PNME) on the whole testing set. The PNME is defined as

$$PNME = \frac{\sum_{k=1}^{72}\sum_{i=1}^{I_k}\sum_{j=1}^{90}\left|\hat{C}_{ijk}-C_{ijk}\right|}{\sum_{k=1}^{72} I_k \times 90} \times 100\%, \quad (10)$$

Table 1. PNMEs for the 4 testing systems

System	PM	CD-PM	AM	CD-AM
PNME	10.12	9.04	8.87	8.24

Table 2. Subjective evaluation results

System	Score (5=Excellent, 1=Bad)					Average
	5	4	3	2	1	(MOS)
PM	2	7	10	7	4	2.9
CD-PM	10	8	8	1	3	3.7
AM	12	8	7	3	0	4.0
CD-AM	16	5	8	1	0	4.2

where I_k denotes the frame number of sentence k (totally 72 testing sentences). C_{ijk} and $\hat{C}_{ijk}$ denote the j^{th} actual and estimated visual parameter after normalization for frame i in sentence k respectively. The calculated PNMEs for the 4 testing systems are shown in Table 1. As can be seen clearly, the proposed AM and CD-AM can effectively reduce the estimation errors in terms of PNME as compared with the conventional phoneme-based models (PM and CD-PM). Especially, the CD-AM system reduced PNME by as much as 19% as compared with the PM system. Analysis shows that many large estimation errors of the PM and CD-PM systems occur during the onset, offset of speech due to their lack of good prediction of coarticulation, leading to higher PNMEs. The results show that directly model articulatory motions can achieve visual parameters that match well with the actual ones extracted from the original videos.

5.2. Subjective Evaluations

To subjectively evaluate the performance of the proposed approach, we built an AV subjective testing set. We collected 30 AV snippets with lengths from 15 secs to 100 secs from the Internet concerning contents from different subjects about news reports, distinguished speeches, weather reports, news briefing, etc. We synthesized each mouth video from the corresponding data set audio using the 90 estimated PCA coefficients. The synthesized mouth videos were overlaid onto the original videos (see Fig. 4), and 10 relatively inexperienced viewers were involved to rank the performance of the mouth animation in terms of naturalness of the mouth matching the audio accompanied. We used a five-point assessment, where 5 means "excellent" and 1 means "bad".

Table 2 summarizes the average scores over the 10 viewers. It shows that the viewers give relatively high scores to the synthesized videos generated by the AM and CD-AM systems, and the CD-PM system can achieve comparable performance with the proposed AM system. But with the introduction of context modelling, the proposed CD-AM system performs the best with a mean opinion score (MOS) of 4.2. Averagely, 16 synthesized videos generated by the CD-AM system were given the highest score. Analysis shows that for the PM and CD-PM systems, lower scores are more likely given to the relative longer videos, and this indicates that unnatural artifacts are more easily detected by the viewers in a long time period. Some synthesized mouth animation videos can be found online at http://www.cityu.edu.hk/rcmt/mouth-synching/mouth-synching.htm.

6. CONCLUSIONS

We present a novel articulatory approach to video-realistic mouth animation. Motivated by the fact that speech-related facial animation is originated by articulation, we directly model the articulator motions using a DBN-structured articulatory model (AM). We also present an EM-based conversion algorithm to convert audio to visual parameters by maximizing the likelihood of the visual parameters given the audio data and the articulatory model. To realize mouth-synching for speaker independent continuous speech with a large vocabulary, we further refine the model by introducing speech context information, leading to context dependent articulatory model (CD-AM). Objective and subjective evaluations show that articulatory modelling is promising, and our models (AM and CD-AM) outperform the phoneme-based models tested (PM and CD-PM). The mouth animation generated by the CD-AMs has achieved more natural and realistic performance. For future work, we'd like to investigate how to stitch the mouth animation with the movement of the whole face seamlessly, and thus achieve a lifelike talking face.

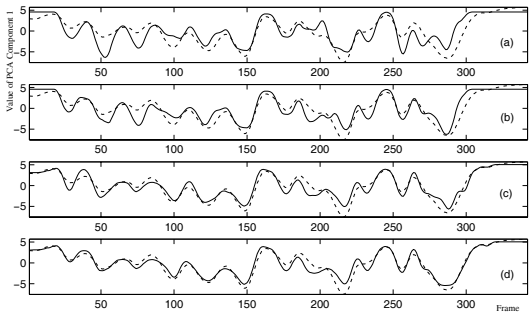

Fig. 3. Estimated visual parameter time trajectories (the 1st PCA component for the R channel) for a testing sentence. Dotted lines indicate the actual parameters, and solid lines indicate the estimated parameters by (a)PMs, (b)CD-PMs, (c)AMs, and (d)CD-AMs. All the curves are smoothed by the moving average method.

Fig. 4. Some snapshots of the synthesized mouth animation overlaid onto the videos from the AV subjective testing set.

7. REFERENCES

[1] J. Ostermann and A. Weissenfeld, "Talking faces–technologies and applications," in *Proc. of ICPR'04*, vol. 3, Aug. 2004, pp. 826–833.

[2] R. R. Rao, T. Chen, and R. M. Mersereau, "Audio-to-visual conversion for multimedia communication," *IEEE Trans. Ind. Electron.*, vol. 45, no. 1, pp. 15–22, Feb. 1998.

[3] K. Choi, Y. Luo, and J. Hwang, "Hidden markov model inversion for audio-to-visual conversion in an MPEG-4 facial animation system," *Journal of VLSI Signal Processing*, no. 29, pp. 51–61, 2001.

[4] J. Goldsmith, *Autosegmental and Metrical Phonology*, W. Hardcastle and N. Hewlett, Eds. Oxford: Basil Blackwell, 1990.

[5] J. A. Bilmes, G. Zweig, T. Richardson, K. Filali, K. Livescu, P. Xu, K. Jackson, Y. Brandman, E. Sandness, E. Holtz, J. Torres, and B. Byrne, "Discriminatively structured graphical models for speech recognition," *Technical Report of JHU 2001 Summer Workshop*, 2001.

[6] H. T. Edwards, *Applied Phonetics: The Sounds of American English*, 2nd ed. Singular, 1997.

[7] L. Xie, "Research on key issues of audio visual speech recognition," Ph.D. dissertation, Northwestern Polytechnical University, Sept. 2004.

[8] K. Murphy, "Dynamic bayesian networks: representation, inference and learning," Ph.D. dissertation, University of California, 2002.

[9] *http://www.cityu.edu.hk/rcmt/mouth-synching/jewel.htm.*

[10] S. Young, G. Evermann, D. Kershaw, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.2)*. Cambridge: Cambridge University Engineering Department, 2002. [Online]. Available: http://htk.eng.cam.ac.uk/