AUDIOVISUAL SPEECH RECOGNITION WITH ARTICULATOR POSITIONS AS HIDDEN VARIABLES

*Mark Hasegawa-Johnson, Karen Livescu, Partha Lal and Kate Saenko**

University of Illinois at Urbana-Champaign, MIT, University of Edinburgh, and MIT
{jhasegaw@uiuc.edu,klivescu@csail.mit.edu,p.lal@sms.ed.ac.uk,saenko@csail.mit.edu}

ABSTRACT

Speech recognition, by both humans and machines, benefits from visual observation of the face, especially at low signal-to-noise ratios (SNRs). It has often been noticed, however, that the audible and visible correlates of a phoneme may be asynchronous; perhaps for this reason, automatic speech recognition structures that allow asynchrony between the audible phoneme and the visible viseme outperform recognizers that allow no such asynchrony. This paper proposes, and tests using experimental speech recognition systems, a new explanation for audio-visual asynchrony. Specifically, we propose that audio-visual asynchrony may be the result of asynchrony between the gestures implemented by different articulators, such that the most visibly salient articulator (e.g., the lips) and the most audibly salient articulator (e.g., the glottis) may, at any given time, be dominated by gestures associated with different phonemes. The proposed model of audio-visual asynchrony is tested by implementing an "articulatory-feature model" audiovisual speech recognizer: a system with multiple hidden state variables, each representing the gestures of one articulator. The proposed system performs as well as a standard audiovisual recognizer on a digit recognition task; the best results are achieved by combining the outputs of the two systems.

Keywords: Automatic Speech Recognition (ASR), Audiovisual Speech, Articulatory Phonology, Dynamic Bayesian Network (DBN)

1. INTRODUCTION

A large number of studies have demonstrated that speech recognition, by both humans and machines, benefits from visual observation of the face, especially at low signal-to-noise ratios (SNRs). For example, by integrating information from audio and video observations, it is possible to reduce the word error rate of automatic speech recognition (ASR) at low SNR [11]. The simplest method for reducing word error rate is to concatenate the audio and video observations, and to use the resulting vector as the observation in a hidden Markov model (HMM). The audible and visible correlates of a phoneme, however, may by asynchronous (e.g., Fig. 1). Motivated by the observed audio-visual asynchrony, several authors have proposed speech recognizers that use separate HMMs for the audio and video observations, with some type of connection between the transition probabilities of the two HMMs [3, 6].

Figure 1: An example of audiovisual asynchrony. The still image shown here (one frame from a video sequence) provides evidence that the talker has prepared her tongue tip and lips, respectively, for the first and second phonemes of the word "three." The audio signal recorded synchronous with this image contains only silence.

The theory of articulatory phonology [1] provides a new way of thinking about asynchrony. In the theory of articulatory phonology, the lexical entry for each word is composed of loosely ordered, indivisible mental constructs called "gestures." In normal speech production, all of the gestures in a word are always produced; there is no such thing as gesture deletion, substitution, or insertion. The wide range of pronunciation variability observed in real-world speech is modeled in three ways: (1) the strength or duration of a gesture may be reduced, (2) gestures may be produced in non-canonical order, or (3) several different gestures may act simultaneously on the same "tract variable" [2]. Any of these three causes

*This material is based upon work supported by the National Science Foundation under grant 01-21285 (F. Jelinek, PI). Any opinions, findings, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

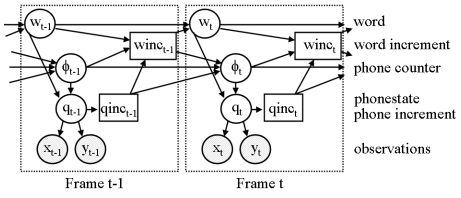
may result in a gesture that is not fully implemented by the articulators. In an articulatory phonology model, Fig. 1 does not exhibit asynchrony between the audio and video channels, but rather, among the articulatory positions of the lips, tongue, and glottis/lungs.

Articulatory phonology has inspired several recent automatic speech recognition experiments [8, 9, 5]. The goal of this paper is to test an audiovisual speech recognizer based on the "articulatory-feature based recognizer" of Livescu and Glass [5]. A system based on Livescu's model has been previously demonstrated for recognition of speech from video [10], but we believe that the experiments reported in this article are the first to demonstrate audiovisual fusion using such a system.

2. MODELS

Three classes of automatic speech recognition models are tested in this paper: hidden Markov models (HMMs; Fig. 2), coupled hidden Markov models (CHMMs; Fig. 3), and articulatory-feature models (AFMs; Fig. 4). All of these systems are implemented in the graphical modeling toolkit (GMTK) using the notation of a dynamic Bayesian network (DBN). The goal of this section is to describe DBN notation, and to define the models used in this paper.

Figure 2: DBN representation of a phone-based HMM speech recognizer



A DBN is a graph in which nodes represent random variables, and edges represent stochastic dependence of one random variable upon another. Fig. 2, for example, shows two frames from the DBN representation of a standard phone-based HMM speech recognizer. Fig. 2 specifies the following assumption about the random variables shown, where $p(\Lambda_t|\Lambda_{t-1})$ is a short-hand notation for the probability of all variables at time t, given all variables at time $t-1$:

$$\begin{aligned} p(\Lambda_t|\Lambda_{t-1}) &= p(w_t|w_{t-1}, \text{winc}_{t-1}) \\ &\times p(\phi_t|\phi_{t-1}, \text{qinc}_{t-1}, \text{winc}_{t-1}) \\ &\times p(q_t|\phi_t, w_t) \\ &\times p(\text{qinc}_t|q_t)p(\text{winc}_t|\text{qinc}_t, \phi_t, w_t) \\ &\times p(x_t|q_t)p(y_t|q_t) \end{aligned} \quad (1)$$

This paper follows several notational conventions exemplified in Fig. 2, and described in the next five paragraphs.

Shaded circles in a DBN represent observable vectors. In Fig. 2, the observable vectors are x_t and y_t. Subscript represents frame number; all of the experiments reported in this paper use 10ms frames. In all of the experiments in this paper, x_t is an acoustic observation vector composed of perceptual LPC coefficients, energy, and their first and second temporal derivatives. The video observation vector y_t includes the first 35 coefficients from a discrete cosine transform of the grayscale pixel values in a rectangle including the lips, upsampled from 30 to 100 frames per second, and their first temporal derivatives. The probability density functions $p(x_t|q_t)$ and $p(y_t|q_t)$ are mixture Gaussian PDFs, trained using expectation maximization.

Unshaded circles and rectangles represent hidden variables. There are three types of hidden variables: categorical variables, counters, and indicators.

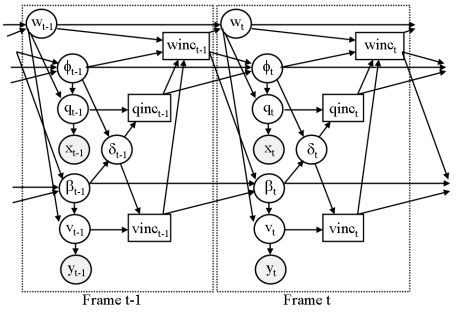
Categorical variables are represented using Roman letters. In Fig. 2, there are two categorical variables: w_t and q_t. w_t is the word label at time t, $w_t \in \mathcal{W}$, where $\mathcal{W}$ is the vocabulary. In all experiments reported in this paper, the vocabulary contains eleven "words:" the ten English digits, and silence. $q_t \in \mathcal{Q}$ is the phonestate label at time t. "Phonestates" are subsegments of a phone. In the experiments in this paper, every phone has three subsegments (beginning, middle, and end). The cardinality of set $\mathcal{Q}$ is therefore three times the number of phones: in our experiments, $|\mathcal{Q}| = 3 \times 42 = 126$.

Counters, represented using Greek letters, describe the ordinal position of a phone within a word. In Fig. 2, for example, ϕ_t is the position of phonestate q_t in word w_t. The probability distribution $p(\phi_{t+1}|\phi_t, \text{qinc}_t, \text{winc}_t)$ has an extremely limited form. ϕ_{t+1} is always equal to ϕ_t unless there is a phonestate transition or a word transition. If there is a word transition, ϕ_{t+1} is reset to $\phi_{t+1} = 1$. If there is a phonestate transition but no word transition, then $\phi_{t+1} = \phi_t + 1$.

Indicators are binary random variables: an indicator variable set to 1 indicates that some event occurs in frame t, otherwise the indicator variable is 0. An indicator variable has a two-letter name, and its node in the DBN is drawn as a rectangle. The indicator variables in Fig. 3 are winc_t ("word increment") and qinc_t ("phonestate increment"). Indicator variables are generally used to gather information from other variables, make a binary decision, and then distribute that information to variables in the next frame. For example, a phonestate increment occurs

with some probability $p(\text{qinc}_t = 1|q_t)$ that depends on the current phonestate label q_t. If $\text{qinc}_t = 1$, and if ϕ_t is equal to the number of phonestates in the dictionary entry for word w_t, then a word transition also occurs; otherwise, $\text{winc}_t = 0$.

Figure 3: DBN representation of a coupled HMM (CHMM) audiovisual speech recognizer

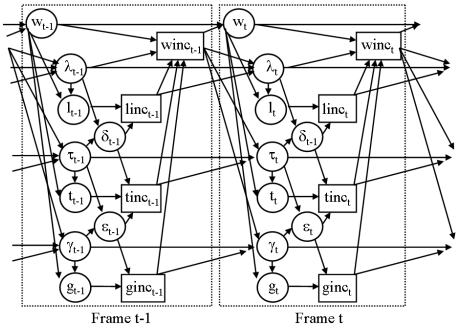


Many experiments have demonstrated that the word error rate of an audiovisual speech recognizer is reduced if the audible phonestate and the visible visemestate are allowed to be asynchronous [6, 3]. For example, a "coupled hidden Markov model" (CHMM) is a pair of hidden Markov models: one linked to the acoustic observation vector x_t, and one linked to the visual observation vector y_t. State variables in the two HMMs depend on one another; thus the two HMMs are allowed to be asynchronous, but not by very much.

Fig. 3 shows our implementation of a CHMM; to our knowledge, this is the first time a CHMM has been implemented using phone-like subword units. In our implementation, the acoustic observation vector x_t depends on the phonestate label q_t, which depends, in turn, on the phonestate counter ϕ_t; all of these variables are defined exactly as they were in Fig. 2. The visual observation vector y_t depends on the visemestate label v_t, which depends, in turn, on the visemestate counter β_t. The phonestate and the visemestate are allowed to be asynchronous, but not by very much. The degree of asynchrony between the phonestate and the visemestate is measured by the asynchrony variable, whose value is always $\delta_t = \phi_t - \beta_t$. If δ_t is greater than or equal to some preset limit δ_{max}, then the probability of a phonestate transition is $p(\text{qinc}_t = 1|\delta_t \geq \delta_{max}) = 0$. If δ_t is less than or equal to $-\delta_{max}$, then the probability of a visemestate transition is $p(\text{vinc}_t = 1|\delta_t \leq -\delta_{max}) = 0$. Transition probabilities under other circumstances are trained using expectation maximization.

The third speech recognition algorithm tested in our work is an articulatory-feature model (AFM),

Figure 4: DBN representation of an articulatory feature model (AFM) automatic speech recognizer



based on the model of Livescu and Glass [5]. Fig. 4 shows our implementation of an AFM; it differs from our CHMM in three ways.

First, there are three phone-like hidden variables, instead of two. In Fig. 4, l_t, the lipstate, specifies the current lip configuration, and depends on the lipstate counter λ_t. t_t, the tonguestate, specifies the current tongue configuration, and depends on the tonguestate counter τ_t. g_t, the glotstate, specifies the current state of the glottis, velum, and lungs, and depends on the glotstate counter γ_t. Asynchrony between the counters is measured using three asynchrony variables, of which two are shown in the figure: $\delta_t = \lambda_t - \tau_t$, and $\epsilon_t = \tau_t - \gamma_t$.

Second, the AFM differs from the CHMM because x_t and y_t depend on all three hidden state variables. Fig. 4 does not show x_t and y_t because space is limited, but the forms of their probability density functions are as follows. x_t depends on all three articulators: $p(x_t|l_t, t_t, g_t)$ is modeled using a mixture Gaussian PDF. y_t depends on the states of the lips and tongue, and $p(y_t|l_t, t_t)$ is modeled using a mixture Gaussian; the glottis/velum state variable g_t is assumed to have no visual correlates.

Third, and most important, the cardinality of the hidden state variables is considerably reduced. Cardinality of the hidden variables affects speech recognition accuracy and computational complexity. In general, system precision improves with greater cardinality, because the recognizer is able to represent a greater number of acoustic distinctions. Generalization error (error caused by differences between the training and test corpus) also increases with the cardinality of the hidden variables. The optimum balance between precision and generalization error is achieved when the hidden variables represent all, and only, the acoustic distinctions that can be generalized from training data to test data. In practice, achieving a reasonable balance between pre-

cision and generalization requires experimentation. In our CHMM, for example, the variables q_t and v_t are each drawn from set $\mathcal{Q}$, the set of all phonestates known to the speech recognizer, whose cardinality is $3 \times 42 = 126$. The variable $l_t \in \mathcal{L}$, on the other hand, represents only the set of English phoneme distinctions that are implemented using labial gestures; likewise $t_t \in \mathcal{T}$ represents phoneme distinctions implemented using the tongue, while $g_t \in \mathcal{G}$ represents distinctions implemented by the glottis, lungs, and soft palate. The sets $\mathcal{L}$, $\mathcal{T}$, and $\mathcal{G}$ are specified in Table 1. These sets were designed based on considerations of articulatory specificity (gestures must be local to the named articulator) and phone distinctiveness (each vector $[l_t, t_t, g_t]$ may correspond to no more than one value of the HMM phonestate label q_t), and were refined using a small number of preliminary experiments. In addition to the set of articulator-specific gestures, each of the three articulators was allowed to take the value "Silent," because during silence, the setting of the articulator is unspecified by phonetic requirements.

Table 1: State variables of the articulatory feature model are drawn from the sets shown: $l_t \in \mathcal{L}$, $t_t \in \mathcal{T}$, $g_t \in \mathcal{G}$. The operation "$\times \{1, 2, 3\}$" performs temporal segmentation, dividing each gesture into initial, medial, and final subgestures, thus the cardinalities of the three sets are $|\mathcal{L}| = 18$, $|\mathcal{T}| = 63$, and $|\mathcal{G}| = 12$.

Set	Gestures
$\mathcal{L}$	{Closed, Critical, Narrow, Protruded-Wide, Labial-Wide, Silent} $\times \{1, 2, 3\}$
$\mathcal{T}$	{Dental-Critical, Alveolar-Closed, Alveolar-Critical, Alveolar-Lateral, Alveolar-Narrow, Retroflex-Narrow, Palatal-Critical, Palatal-Narrow, Palatal-Narrow-Tense, Palatal-Mid-Narrow, Palatal-Mid, Palatal-Mid-Tense, Palatal-Wide, Velar-Closed, Velar-Critical, Velar-Narrow, Velar-Mid, Uvular-Narrow, Uvular-Mid, Pharynx-Narrow, Silent} $\times \{1, 2, 3\}$
$\mathcal{G}$	{Aspirated, Voiced-Oral, Nasal, Silent} $\times \{1, 2, 3\}$

3. EXPERIMENTAL METHODS

Experiments reported in this paper used the CUAVE database [7]. CUAVE is a database of isolated digits, telephone numbers, and read sentences recorded in high resolution video, with good lighting, in a quiet recording studio; acoustic noise is electronically added. Experiments in this paper used discrete speech: ten-digit sequences spoken with silence after each word. The corpus was divided arbitrarily into a training subset (60% of the talkers), a validation subset (20% of the talkers), and an evaluation subset (20% of the talkers); each subset was evenly divided between male and female talkers. Error rates using the evaluation subset were not computed for lack of time; results reported here are from the validation subset.

All recognition models used a uniform grammar, constrained to produce exactly ten words per utterance. All recognizers were trained using an SNR of ∞ (no added noise), and were tested at six different SNRs: ∞, 12dB, 10dB, 6dB, 4dB, and -4dB. Recognizer training consisted of three stages. First, the observation PDFs each recognizer was trained using noise-free training data, with a Gaussian representing each of the PDFs $p(x_t|l_t, t_t, g_t)$ and $p(y_t|l_t, t_t)$. Second, the number of Gaussians per PDF was doubled, the recognizer was re-trained using noise-free data, and the recognizer was tested using noise-free validation data. Minimum validation error was achieved with 32 Gaussians per PDF for the audio-only HMM, 16 Gaussians per PDF for the video and audiovisual HMMs, 2 Gaussians per PDF for the articulatory feature model, and 4-16 Gaussians per PDF for the CHMMs, as specified in Sec. 4. Third, the recognizer was tested using validation data representing each of the six SNRs (∞, 12, 10, 6, 4, and -4dB), and for each SNR, a video stream weight, ρ, was selected to minimize word error rate. ρ is used to determine the extent to which the observation probability depends on video vs. audio observations: the model used here is $p(x_t, y_t|l_t, t_t, g_t) = p(y_t|l_t, t_t)^{\rho} p(x_t|l_t, t_t, g_t)^{1-\rho}$.

Experimental results (reported in Sec. 4.) demonstrated that the articulatory feature model (AFM) performs comparably to the CHMM. Analysis revealed, however, that the two systems make different specific errors. When speech recognition systems have similar word error rate but different specific errors, it is common to allow the systems to correct each others' mistakes, using a system combination strategy called ROVER [4], in which the word transcriptions generated by the systems are aligned with one another, and majority voting determines the final ROVER output. If the specific error patterns of the systems are sufficiently complementary, the combined system will have lower word error rate than any component system. We used the NIST ROVER

implementation [4] to test the hypothesis that AFM and CHMM are complementary in this way.

4. RESULTS

Figures 5 through 8 describe the word error rates achieved, on validation data, under the experimental conditions described in Sec. 3.

Figure 5: Word error rates of video-only, audio-only, and audiovisual HMM recognizers in six different SNRs.

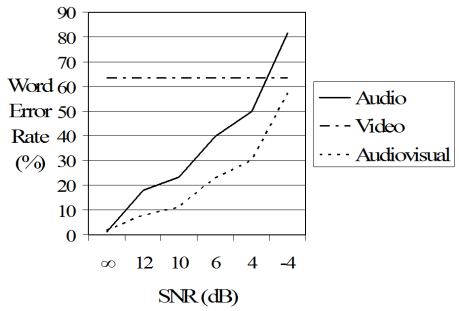


Fig. 5 plots the word error rate of HMM systems using audio-only, video-only, or audiovisual observations, as a function of signal to noise ratio. Word error rate of the video-only recognizer is about 60%, independent of acoustic SNR. Word error rate of the audiovisual recognizer is lower than that of the audio-only recognizer under every condition with acoustic noise. The difference is statistically significant (MAPSSWE test, $p < 0.05$).

Figure 6: Word error rates of CHMM recognizers with maximum allowed asynchrony ($\delta_{max} = \max|\phi_t - \beta_t|$) of 0, 1, 2, or an unlimited number of phonestates.

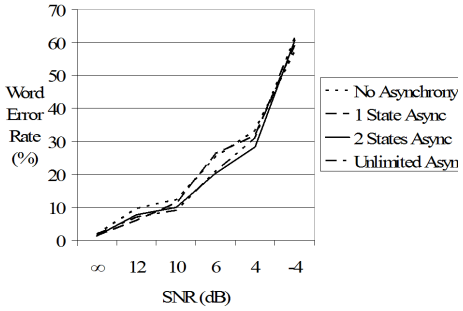


Fig. 6 plots word error rate as a function of signal to noise ratio for systems with maximum allowed asynchrony ($\delta_{max} = \max|\phi_t - \beta_t|$) of 0, 1, 2, or an unlimited number of phonestates (using 16, 16, 8, and 4 Gaussians per PDF, respectively). The system with no allowed asynchrony ($\delta_{max} = 0$) is an HMM, and its error rates are also plotted in Fig. 5. The system with unlimited asynchrony has no representation of dependence between the two modalities: both qinc_t and vinc_t are independent of δ_t. The system with $\delta_{max} = 2$ produces the lowest average word error rate (averaged across six SNRs). The difference between $\delta_{max} = 2$ and other conditions is statistically significant (MAPSSWE test, $p < 0.05$).

Figure 7: Word error rate of CHMM and articulatory feature model speech recognizers as a function of signal to noise ratio.

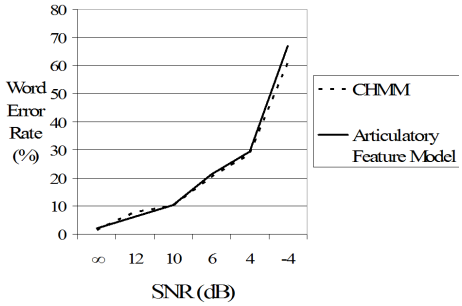
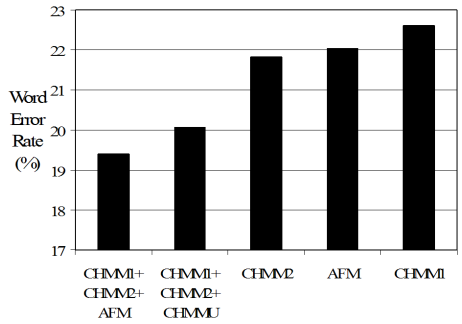


Fig. 7 plots word error rate versus signal to noise ratio of an articulatory feature model (AFM) speech recognizer with a maximum inter-articulator asynchrony of $\delta_{max} = \epsilon_{max} = 2$. For comparison, word error rate of the best CHMM from Fig. 6 ($\delta_{max} = 2$) is re-plotted on the same axes. The CHMM has a lower word error rate for some signal to noise ratios, and tends to have a lower word error rate on average, but the difference is not statistically significant.

Figure 8: Word error rates of three original systems (CHMM systems with $\delta_{max} = 1$ and $\delta_{max} = 2$, and AFM with $\delta_{max} = 2$) and two ROVER system combinations, averaged across all test-set SNRs.



Despite the similarity in their word error rates, the articulatory feature model and the CHMM are not identical. Fig. 8 shows word error rate of three recognizers, and of two different ROVER system combinations, averaged across all signal to noise ratios. The leftmost bar shows the word error rate achieved by a system combination using two CHMM systems ($\delta_{max} = 1$ and $\delta_{max} = 2$) and

one articulatory-feature model (AFM). The second bar shows the result of system combination in which the AFM has been replaced by another CHMM (the CHMM with unlimited δ_{max}). The word error rates achieved using system combination are lower than those achieved without system combination (MAPSSWE test, $p < 0.05$). The system combination that includes an articulatory feature model tends to have a lower word error rate, on average, than the system that includes only CHMMs, but the difference is not statistically significant.

5. CONCLUSIONS

This paper proposes a new paradigm for articulatory-feature based audiovisual speech recognition. Specifically, we propose that the apparent asynchrony between acoustic and visual modalities (noted in many previous AVSR papers) may be effectively modeled as asynchrony among the articulatory gestures implemented by the lips, tongue, and glottis/velum. Experimental tests using the CUAVE corpus provide tentative support for four conclusions. First, the combination of audio and visual features can reduce word error rate at low SNR. Second, word error rate of an audiovisual speech recognizer may be further reduced using a coupled hidden Markov model, and by allowing up to 2 states of asynchrony (about 2/3 of a phoneme) between the audio and video HMMs. Third, the benefits of the CHMM are also achieved by an articulatory feature model, in which asynchrony between the lips, tongue, and glottis/velum replaces the asynchrony between audio and video modalities. Fourth, the best results are achieved by combining the outputs of multiple systems, and there is a tendency for the combination of CHMM and AFM systems to outperform a system combination using only CHMMs.

The articulatory feature model described in this paper includes target specifications for every articulator, during every phoneme. One goal of future research will be the development of a model in which the lexical entries better approximate the minimally specified lexical entries used in theoretical studies of articulatory phonology [1]. The lexicon, in most theoretical studies, specifies only a small number of mandatory gestures, and each gesture may be associated with multiple phonemes (e.g., an Aspiration gesture may cover the entire onset of a syllable, as in the word "speech").

Our current model is also unsatisfactory because it contains no representation of prosody. The acoustic and articulatory correlates of a gesture are influenced by many prosodic context variables, including syllable structure, lexical stress, phrasal prominence, and prosodic phrase boundaries. The first two levels are necessary in order to distinguish the phones used in most HMM-based ASR: e.g., most systems differentiate nasal consonants in the nucleus vs. onset of a syllable, and most systems distinguish reduced vs. unreduced vowels. It might be possible to develop a comprehensive model, covering all of these levels, using the π-gesture theory of Byrd and Saltzman [2].

6. REFERENCES

[1] Browman, C. P., Goldstein, L. 1992. Articulatory phonology: An overview. *Phonetica* 49, 155–180.

[2] Byrd, D., Saltzman, E. 2003. The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. *J. Phonetics* 31, 149–180.

[3] Chu, S., Huang, T. S. 2000. Bimodal speech recognition using coupled hidden Markov models. *Proc. Interspeech* Beijing.

[4] Fiscus, J. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). *IEEE Workshop on ASRU* Santa Barbara.

[5] Livescu, K., Glass, J. 2004. Feature-based pronunciation modeling for speech recognition. *Proc. HLT/NAACL* Boston.

[6] Neti, C., Luettin, G. P. J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., Zhou, J. 2000. Audio-visual speech recognition: Final report. Technical Report WS00 Johns Hopkins University Center for Language and Speech Processing.

[7] Patterson, E., Gurbuz, S., Tufecki, Z., Gowdy, J. 2002. CUAVE: A new audio-visual database for multimodal human-computer interface research. *Proc. ICASSP* Orlando.

[8] Richardson, M., Bilmes, J., Diorio, C. 2000. Hidden-articulator Markov models: performance improvements and robustness to noise. *Proc. Interspeech* Beijing.

[9] Richmond, K., King, S., Taylor, P. 2003. Modeling the uncertainty in recovering articulation from acoustics. *Computer Speech and Language* 17(2-3), 153–172.

[10] Saenko, K., K, L., Siracusa, M., Wilson, K., Glass, J., Darrell, T. 2005. Visual speech recognition with loosely synchronized feature streams. *Proc. ICCV* Beijing.

[11] Silsbee, P. L., Bovik, A. C. 1996. Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Trans. Speech and Audio Processing* 4, 337–351.