Multimodal Dialog Systems Research at Illinois

Stephen E. Levinson, Thomas S. Huang, Mark A. Hasegawa-Johnson,
Ken Chen, Stephen Chu, Ashutosh Garg, Zhinian Jing,
Danfeng Li, John Lin, Mohamed Omar, and Zhen Wen

June 5, 2002

Abstract

Multimodal dialog systems research at the University of Illinois seeks to develop algorithms and systems capable of robustly extracting and adaptively combining information about the speech and gestures of a nave user in a noisy environment. This paper will review our recent work in seven fields related to multimodal semantic understanding of speech: audiovisual speech recognition, multimodal user state recognition, gesture recognition, face tracking, binaural hearing, noise-robust and high-performance acoustic feature design, and recognition of prosody.

1 Introduction

The purpose of this paper is to summarize ongoing multimodal speech and dialog recognition research at the University of Illinois. A multimodal speech recognition system can be described in two distinct stages: (1) robust audiovisual feature extraction, and (2) speech and user state recognition using dynamic Bayesian networks. Features are extracted from audiovisual input in order to optimally represent phonetic, visemic, gestural, and prosodic information. Our specific ongoing research projects include binaural hearing (array processing on a mobile platform), biomimetic noise-robust acoustic feature extraction, maximum mutual information acoustic feature design, and face tracking. Customized Dynamic Bayesian networks have been designed for three different recognition tasks: audiovisual speech recognition using coupled HMMs, user state recognition using hierarchical HMMs, and recognition of speaking rate using hidden-mode explicit-duration acoustic HMMs.

Image and Speech Processing research at the University of Illinois is currently tested in two ongoing research prototype environments. The first research prototype environment is an experimental computing facility for teaching children about physics. The second research environment is an autonomous robot, Illy, who acquires language through the semantic association of audio, visual, and haptic sensory data. Prior to implementation on one or both of these platforms, most of our algorithms are tested using standard or locally acquired datasets.

2 Pre-Processing

2.1 Binaural Hearing

Our research on binaural hearing addresses the extraction of noise-robust audio from a two-microphone array mounted on a physically mobile platform (a language-learning autonomous robot). The source localization algorithm is based on a two channel Griffiths-Jim beamformer [3] and a new phase unwrapping algorithm for accurate estimation of time difference of arrival measures [8]. The new phase unwrapping algorithm is trained using many measurements of TDOAs in order to create an accurate spatial map of TDOA pattern as a function of arrival azimuth and elevation. These can then be used both to cancel interfering noise and to get a faithful representation of the desired speech signal. Preliminary results show that a speech signal can be accurately located in noisy laboratory room within a few milliseconds and with ten degree accuracy at a distance of 2-4 meters (acoustic far field).

In the current implementation, detection of a speech signal triggers physical rotation of the receiver platform (the robot's "head") so that it faces the primary talker. By physically aligning the "head" of the robot with the direction of primary source arrival, we are able to use extremely efficient off-axis cancellation algorithms for improved SNR [9].

2.2 Acoustic Features

Standard speech recognition features (including MFCC, PLP, and LPCC) result in isolated digit

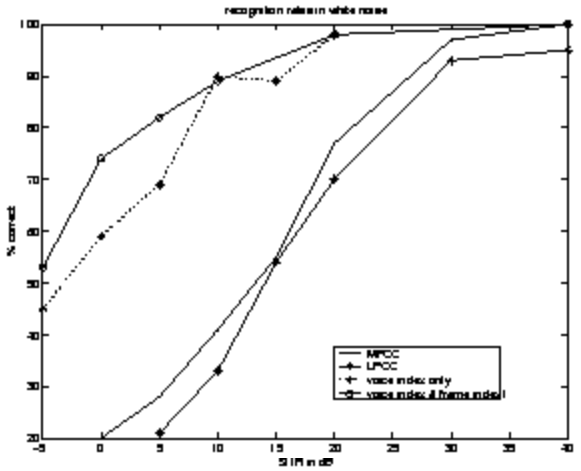

Figure 1: WER: isolated digit recognition in white noise with two standard feature sets, MFCC and LPCC, and two novel feature sets, LPCC with voice index and with frame index (from [6]).

recognition error rates of approximately 60% at 10dB SNR, and nearly 80% at 0dB SNR. In 1992, Meddis and Hewitt proposed a biomimetic method for recognition of voiced speech in high noise environments [10]. Meddis and Hewitt proposed filtering a noisy speech signal into many bands, computing the autocorrelation function $R_k(\tau)$ in each sub-band, and then estimating the speech autocorrelation $R(\tau)$ by optimally selecting and adding together the high-SNR sub-band autocorrelations. In our work [6], we have replaced Meddis and Hewitt's optimal selection algorithm by an optimal scaling algorithm. Specifically, we estimate the sub-band SNR v_k using a standard pitch prediction coefficient, i.e.

$$v_k = \frac{\text{Speech Energy in Band k}}{\text{Total Energy in Band k}} \approx \frac{R_k(T_0)}{R_k(0)} \qquad (1)$$

where T_0 is the globally optimum pitch period. The maximum likelihood estimate of the noise-free speech signal autocorrelation is then

$$\hat{R}(\tau) = \sum_k v_k R_k(\tau) \qquad (2)$$

In isolated digit recognition experiments, the use of equations 1 and 2 reduced word error rate by more than a factor of three in white noise at 10dB through -10dB, and by more than a factor of two in babble noise at the same SNRs (Figure 1).

The phonological features implemented at a speech landmark influence the acoustic spectrum at distances of 50-100ms [4, 19]. Complete representation of a 100ms spectrogram requires a 120-dimensional acoustic feature vector. It is not possible to accurately train observation PDFs of dimension 120 using existing data sets, but it is possible to select a subvector using a quantitative optimality criterion. In our research, we select a 39-dimensional feature subvector from a list of 160 candidate features in order to optimize the mutual information between features and phoneme labels [12]. Optimality is determined using a clean speech database (TIMIT) with no language model, but the resulting optimality generalizes. As shown in Table 1, the resulting MMIA (maximum mutual information acoustic) feature vector outperforms all standard feature vectors under at least three conditions: in quiet and at 10dB SNR, without a language model and with an optimized phoneme bigram. Larger improvements may be obtained by testing the 5-10 best feature vectors generated during the mutual information search. The best recognition accuracy, obtained using the feature set with second-best mutual information, was 62% with no language model in quiet conditions.

	No LM		Phone Bigram	
Features	35dB	10dB	35dB	10dB
LPCC	56	40	59	46
MFCC	58	42	63	48
FM	58	42	62	46
MMIA	59	43	63	49

Table 1: Phoneme recognition correctness in four conditions. Features selected using a maximum mutual information criterion (MMIA) provide superior performance in all four conditions.

2.3 Face Tracking

Research has shown that facial and vocal-tract motions are highly correlated during speech production [20]. Speech recognition using both audio/visual features is shown to be more robust in noisy environments [5]. Analysis of non-rigid human facial motion is a key component for acquiring visual features for audio/visual speech recognition.

In the past several years, research in our group has led to a robust 3D facial motion tracking system [16]. A 3D non-rigid facial motion model is manually constructed based on piecewise Bezier volume deformation model (PBVD). It is used to constrain the noisy low-level optical flow. The tracking is done in a multi-resolution manner such that higher speed could be achieved. It runs at 5 fps on an SGI Onyx2 machine. This tracking algorithm has been successfully used for audio-visual speech recognition and bimodal emotion recognition.

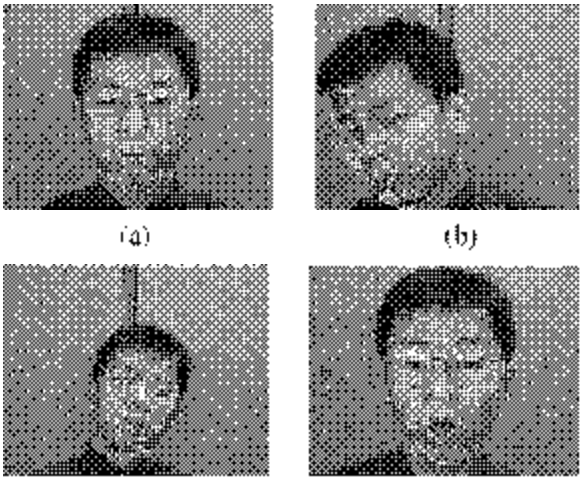

Figure 2: Demonstration of our face tracking system.

2.4 Gesture Recognition

Hand gestures are capable of delivering information not presented in speech [14]. Controlling gesture can be used to provide commands to the system. Navigation gestures provide information for manipulating virtual objects, and for selecting point objects or large regions on the screen. Conversational gestures provide subtle cues to sentence meaning in normal human interaction. Automated hand tracking and gesture recognition can help improve the performance of human-machine interface.

We have investigated both appearance-based gesture recognition (using neural network-based pattern recognition techniques) and model-based gesture recognition [18, 17]. In model-based recognition, the configuration of a hand model is first determined by providing a set of joint angle parameters. The 2D projection of this hand model, determined by the translation and orientation of the model relative to a viewing portal, is compared with the hand image from input video. Estimate of the correct input hand configuration is determined by the best matching projection. A complete description of the global hand position and all finger joint angles requires specification of 21 joint angles. Using both known anatomical constraints and PCA to reduce dimensionality, we can initially reduce the dimensionality of the gestural description from 21 to 7 independent dimensions while keeping 95% of the information. In this 7-dimensional space, it is possible to define 28 basis configurations, consisting of the configurations in which each finger is either fully folded or completely extended. A close examination of the motion trajectories between these basis states shows that natural hand articulations seem to lie largely in the linear manifold spanned by pairs of basis states. We believe that, based on these preliminary results, it will be possible to map all observed gestures into a low-dimensional gestural manifold, resulting in efficient and accurate gesture recognition.

3 Dynamic Bayesian Networks

3.1 Lip Reading

The focus of our research in lip reading is a novel approach to the fusion problem in audio-visual speech processing and recognition. Our fusion algorithm is built upon the framework of coupled hidden Markov models (CHMMs). CHMMs are probabilistic inference graphs that have hidden Markov models (HMMs) as sub-graphs. Chains in the corresponding inference graph are coupled through matrices of conditional probabilities modeling temporal dependencies between their hidden state variables. The coupling probabilities are both cross chain and cross time. The later is essential for capturing temporal influences between chains. In a bimodal speech recognition system, two-chain CHMMs are deployed, with one chain being associated with the acoustic observations, the other with the visual features. Under this framework, the fusion of the two modalities takes place during the classification stage. The particular topology of the CHMM ensures that the learning and classification are based on the audio and visual domains jointly, while allowing asynchronies between the two information channels.

In essence, CHMMs are directed graphical models of stochastic processes and are a special type of Dynamic Bayesian Networks (DBNs). The DBNs generalize the HMMs by representing the hidden states as state variables, and allow the states to have complex interdependencies. The DBN point of view facilitates the development of inference algorithms for the CHMMs. Specifically, two inference algorithms are proposed in this work. Both of the algorithms are exact methods. The first is an extension of the well-known forward-backward algorithm from the HMM literatures. The second is a strategy of converting CHMMs to mathematically equivalent HMMs, and carrying out learning in the transformed models.

The benefits of the proposed fusion scheme are confirmed by a series of preliminary experiments on audio-visual speech recognition. Visual features based on lip geometry are used in the experiments. Furthermore, comparing with an acoustic-only ASR system trained using only the audio channel of the same dataset, the bimodal system consistently demonstrates improved noise robustness across

SNR	10dB	20dB	30dB
A	4.03	43.61	99.10
V	42.95	42.95	42.95
A+V	10.58	72.79	99.74
CHMM	35.32	86.58	93.32

Table 2: Result of experiments in audiovisual speech recognition (measured in %word accuracy). A indicates the audio-only system; V indicates the visual-only system; A+V indicates a bimodal system using early integration; and CHMM indicates the CHMM-based system.

a wide range of SNR levels.

3.2 Prosody

Our approach to the recognition of prosody is the use of a "hidden mode variable" [13] to condition the explicit duration PDFs of a CVDHMM [7]. In our prototype algorithm, the state space consists of parallel phonetic state variables (q_t) and prosodic state variables (k_t). The dwell time of state q_t is a random variable d_q with PDF depending $p(d_q|q,k)$. At the end of the specified dwell time, the phonetic variable always changes state (no self-loops), but the prosodic state variable may or may not change state. Thus, for example, if ($k_t \in$slow, medium, fast) represents speaking rate, it may be reasonable to allow k_t to change state at any word boundary with a small probability.

In order to allow efficient experiments, we have modified HTK to make use of Ferguson's EM algorithm for explicit-duration HMMs [1, 2]. Ferguson's algorithm is an order of magnitude faster than most algorithms for the explicit-duration HMMs. The computational complexity of the algorithm is $\mathcal{O}(NT(N+T))$, where N is the number of states, T is the number of frames in the input signal, and $(\mathcal{O}(N^2T))$ is the complexity of an HMM without explicit duration. The forward algorithm computes

$$\begin{aligned}
\alpha_t^*(j) &= P(O_1,\ldots,O_t, j \text{ commences at } t+1) \\
&= \sum_j \alpha_t(i)a_{ij} \\
\alpha_t(i) &= P(O_1,\ldots,O_t, i \text{ ends at } t) \\
&= \sum_d \alpha_{t-d}^*(i)p(d|i)p(O_{t-d+1},\ldots,O_t|i)
\end{aligned}$$

3.3 User State Recognition

Integration of a large number of sources for the purpose of multimodal user-state recognition can be accomplished using a hierarchical dynamic Bayesian network (figure 3). In a hierarchical DBN, each modality (audio, lip reading, gesture, and prosody) is modeled using a modality-dependent HMM. Each modality-dependent HMM is searched in order to generate the N transcriptions that best match the observed data in the given modality. The likelihood of each transcription is then estimated using a constrained forward-backward algorithm, generating the probability of state residency during every frame. These probabilities are fed forward to the supervisor HMM, which integrates them to determine a single transcription of the sentence in order to maximize the a posteriori transcription probability. By imposing a prior on the probability distributions learned by the model for the purpose of increasing conditional entropy, we have demonstrated a 10% increase in user state classification performance [15, 11].

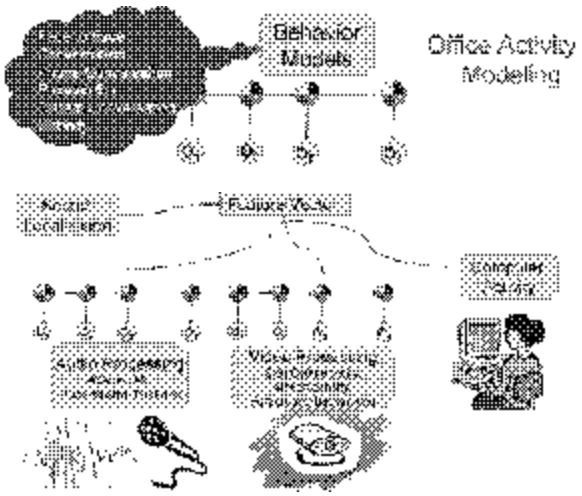

Figure 3: Architecture for detecting events in the office scenario

4 Conclusions

Our research is intended to elucidate both the theoretical and the practical requirements for effective multimodal speech understanding systems. The use of speech in multimodal systems will increase our theoretical understanding of the problems of sensor fusion and representations of multimodal signals. Increased theoretical understanding, in turn, will enable us to produce practical results that can be directly used in state-of-the-art speech recognition systems and as part of larger systems for advanced human-machine communication.

References

[1] Ken Chen. Em algorithm for prosody-dependent speech recognition. Final Project Report, CS 346, 2002.

[2] John D. Ferguson. Variable duratin models for speech. In J.D. Ferguson, editor, *Proceedings of the Symposium on the Application of Hidden Markov Models to Text and Speech*, pages 143–179. Princeton University Press, Princeton, NJ, 1980.

[3] L.J. Griffiths and C.W. Kim. An alternative approach to adaptive beamforming. *IEEE Trans. Antennas and Propagation*, AP-30(1):27–34, 1982.

[4] Mark Hasegawa-Johnson. Time-frequency distribution of partial phonetic information measured using mutual information. In *Proc. Int. Conf. Spoken Lang. Proc.*, volume IV, pages 133–136, Beijing, 2000.

[5] Marcus E. Hennecke, David G. Stork, and K. Venkatesh Prasad. Visionary speech: Looking ahead to practical speechreading systems. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines: Models, Systems, and Applications*, pages 331–350. Springer, New York, 1996.

[6] Zhinian Jing and Mark Hasegawa-Johnson. Auditory-modeling inspired methods of feature extraction for robust automatic speech recognition. In *Proc ICASSP*, 2002.

[7] Stephen E. Levinson. Continuously variable duration hidden Markov models for speech analysis. In *Proc. ICASSP*, pages 1241–1244, 1986.

[8] Danfeng Li and Stephen E. Levinson. Adaptive sound source localization by two microphones. In *Proc. ICASSP*, page 143, Salt Lake City, UT, 2001.

[9] Chen Liu, Bruce C. Wheeler, William D. O'Brien Jr., Robert C. Bilger, Charissa R. Lansing, and Albert S. Feng. Localization of multiple sound sources with two microphones. *J. Acoust. Soc. Am.*, 108(4):1888–1905, 2000.

[10] Ray Meddis and Michael J. Hewitt. Modeling the identification of concurrent vowels with different fundamental frequencies. *J. Acoust. Soc. Am.*, 91(1):233–245, 1992.

[11] Nuria Oliver and Ashutosh Garg. MIHMM: mutual information hidden markov models. In *Proceedings of the Nineteenth International Conference on Machine learning*, Sydney, 2002.

[12] M. Kamal Omar and Mark Hasegawa-Johnson. Maximum mutual information based acoustic features representation of phonological features for speech recognition. In *Proc ICASSP*, 2002.

[13] M. Ostendorf, B. Byrne, M. Fink, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfield. Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. In *CSLU Workshop 1996*, March 1997.

[14] V. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gesture for human computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:677–695, 1997.

[15] Vladimir Pavlovic, Ashutosh Garg, James M. Rehg, and Thomas S. Huang. Multimodal speaker detection using error feedback dynamic bayesian networks. In *IEEE Computer Vision and Pattern Recognition*, 2000.

[16] H. Tao and T. Huang. Explanation-based facial motion tracking using a piecewise bezier volume deformation model. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1999.

[17] Ying Wu and Thomas S. Huang. Self-supervised learning for visual tracking and recognition of human hand. In *Proc. AAAI National Conf. on Artificial Intelligence*, pages 243–248, 2000.

[18] Ying Wu and Thomas S. Huang. View-independent recognition of hand postures. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume II, pages 88–94, 2000.

[19] Howard Yang, Sarel van Vuuren, and Hynek Hermansky. Relevancy of time-frequency features for phonetic classification measured by mutual information. In *Proc. ICASSP*, Phoenix, AZ, 1999.

[20] Hani Yehia, Philip Rubin, and Eric Vatikiotis-Bateson. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26:23–43, 1998.