Multimodal Dialog Systems Research at Illinois

Stephen E. Levinson, Thomas S. Huang, Mark A. Hasegawa-Johnson,
Ken Chen, Stephen Chu, Ashutosh Garg, Zhinian Jing,
Dafeng 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 naive user in a noisy environment. This paper will review our recent work in seven fields related to multimodal semantic understanding of speech: audio-visual 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, viseme, 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.

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.

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
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 speed autocorrelation $\hat{R}(\tau)$ by optimally selecting and adding together the high-SNR sub-band autocorrelations. In our work [6], we have replaced Meddis and Hewitt’s optimum selection algorithm by an optimal scaling algorithm. Specifically, we estimate the sub-band SNR $v_k$ using a standard pitch prediction coefficient, i.e.

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

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

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

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 sub-vector using a quantitative optimality criterion. In our research, we select a 39-dimensional feature sub-vector from a list of 180 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.

### 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.
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 latter 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 literature. 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
<table>
<thead>
<tr>
<th>SNR</th>
<th>10dB</th>
<th>20dB</th>
<th>30dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4.03</td>
<td>45.81</td>
<td>99.10</td>
</tr>
<tr>
<td>V</td>
<td>42.95</td>
<td>42.95</td>
<td>42.95</td>
</tr>
<tr>
<td>A+V</td>
<td>10.58</td>
<td>72.79</td>
<td>99.74</td>
</tr>
<tr>
<td>CHMM</td>
<td>35.32</td>
<td>86.58</td>
<td>93.32</td>
</tr>
</tbody>
</table>

Table 2: Results 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 $p(d_q|q_t,k_t)$. 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 $(k_t, cslow, medium, fast)$ 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 $O((NT(N+T)))$, where $N$ is the number of states, $T$ is the number of frames in the input signal, and $(O(N^2T))$ is the complexity of an HMM without explicit duration. The forward algorithm computes

$$
\alpha_t^t(j) = \begin{cases} 
F(O_1, \ldots, O_t, j \text{ commences at } t+1) \\
\sum_j \alpha_t(i) a_{ij} 
\end{cases}
$$

$$
\alpha_t(i) = \begin{cases} 
P(O_1, \ldots, O_t, i \text{ ends at } t) \\
\sum_d \alpha_{t-d}^t(i) p(d|i)p(O_{t-d+1}, \ldots, O_{t})
\end{cases}
$$

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 [10, 11].

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


