AUDIO-VISUAL SPEECH RECOGNITION BY SPEECHREADING

Xiaozheng Zhang, Russell M. Mersereau, and Mark A. Clements

ABSTRACT

Speechreading increases intelligibility in human speech perception. This suggests that conventional acoustic-based speech processing can benefit from the addition of visual information. This paper exploits speechreading for joint audio-visual speech recognition. We first present a color-based feature extraction algorithm that is able to extract salient visual speech features reliably from a frontal view of the talker in a video sequence. Then, a new fusion strategy using a coupled hidden Markov model (CHMM) is proposed to incorporate visual modality into the acoustic subsystem. By maintaining temporal coupling across the two modalities at the feature level and allowing asynchrony in the state at the same time, a CHMM provides a better model for capturing temporal correlations between the two streams of information. The experimental results demonstrate that the combined audio-visual system outperforms the acoustic-only recognizer over a wide range of noise levels.

1. INTRODUCTION

Hearing-impaired people use speechreading as a primary source of information for speech perception. Even listeners with normal hearing can enhance their speech perception by seeing the speaker's face, particularly in noisy conditions. The benefit gained from the presence of the visual signal has been quantitatively estimated to be equivalent to an increase of 15dB in the SNR when noisy environments are encountered [1]. This is due in part to the complementary nature of the audio and visual aspect of speech.

The first attempt to use vision to aid speech recognition was done by Petajan in 1984 [2]. He demonstrated that visual speech yields information that is not always present in the acoustic signal and enables improved recognition accuracy over purely acoustic-based systems. Since then, there has been increasing interest in supplementing acoustic recognizers with the visual modality to overcome their limitations. While yielding excellent results in a controlled environment, the performance of acoustic-only systems degrades dramatically in the real world in noisy environments, such as in an automobile, or in a typical office with noise from ringing telephones, fans, and human conversations. Robust automatic speech recognition has long been an engineering goal for several decades. The use of the additional visual information has opened new possibilities.

Automatic speechreading is primarily directed at two research areas—the design of a visual front end where visual speech features are accurately and reliably extracted, and the development of an effective strategy to integrate the two separate information sources. In this paper, we examine both of these issues.

This paper is organized into three distinct parts. Section 2 describes the novel visual front end that we use to extract the visual features. Section 3 addresses the problem of audio-visual integration and introduces the coupled hidden Markov model (CHMM) for fusing the two speech modalities. Finally, some initial experiments on audio-visual speech recognition and performance evaluations are presented in Section 4 for both speaker-dependent and speaker-independent cases.

2. VISUAL ANALYSIS

2.1 Previous Work

Most visual speech information is contained in the lips. Thus, visual analysis in automatic speechreading usually focuses on lip feature extraction. Existing approaches to visual feature extraction generally fall under two main categories: image-based techniques and explicit feature extraction.

In the image-based approach, the whole image containing the mouth area is used as a feature either directly [3], after some preprocessing such as a principal component analysis [4] or vector quantization [5]. In a more recent study [6], the image was processed by a discrete cosine transformation followed by a linear discriminant analysis projection and maximum likelihood linear transform feature rotation. The advantage of an image-based approach is that no information is lost, however, it is left to the recognition engine to determine the relevant features in the image. A common criticism of this approach is that it tends to be very sensitive to changes in illumination, position, camera distance, rotation, and speaker.

The alternative to the image-based method aims at explicitly extracting relevant visual speech features. Here, model-based methods are commonly considered where a geometric model of the lip contour is applied. Typical examples are deformable templates, "snakes", and active shape models (ASM). Recently, an active appearance model (AAM) extending the ASM was proposed [7]. It adds a statistical model of gray-level appearance. However, most of these methods use intensity-based images. The difficulty with these approaches usually arises when the contrast is poor along lip contours, which occurs quite often under natural lighting. In particular, edges on the lower lip are difficult to distinguish because of shading and reflection. The algorithm is difficult to extend to various lighting conditions, different skin colors, and people with facial hair. In addition, it is difficult to detect the teeth and tongue using intensity information only, because the skin-lip and lip-teeth edges are highly confusable.
2.2 Our Approach

We propose a color-based approach for lip feature extraction. Color is an important identifying feature for the lips. Prominent colors can be used as a far more efficient search criterion for detecting and extracting certain objects, e.g., red for identifying the lips. In our previous work [8], we derived a modified version of the hue representation for lip images. Hue is easily justified because of its color constancy across genders and races and its high discriminative power for detecting the lips. Thus, the first step in our analysis is a transformation from RGB to modified HSV.

Figure 1 shows an overview of the visual front end for the feature extraction. It consists of three visual analysis stages: lip region localization, lip segmentation, and a final lip feature extraction.

![Figure 1: Visual Processing Steps.](image)

In [9], we describe in detail how we locate the speaker's mouth region reliably from a color video sequence by using hue, saturation, and motion information. Next, we combine both color and edge information to segment the lip from its surroundings by using a Markov random field framework (MRF). Under the MRF modeling assumption, image interpretation is formulated as a problem of maximizing the a posteriori probability of correct labeling given prior knowledge and actual observed data. Finally, the key points that define the lip position are detected and the relevant visual speech parameters are derived. Fig. 2 shows examples of the extracted feature key points.

![Figure 2: Measured feature points on the lips.](image)

Based on the extracted feature key points, the following geometric dimensions of the lips are derived: mouth width \( w_2 \), upper/lower lip width \( (h_1,h_3) \), lip opening height/width \( (h_2,w_1) \), and the distance between the horizontal lip line and the upper lip \( (h_4) \). An illustration of the geometry is shown in Fig. 3.

![Figure 3: Illustration of the extracted geometric features of the lips.](image)

In addition to the geometric dimensions of the lips, we also detect the visibility of the tongue and teeth [9]. The parameter for the tongue is the total number of lip-color pixels that lie within the inner lip contour, while the parameter for the teeth is the total number of white pixels that lie within the bounding box.

We applied the feature extraction algorithm on the Carnegie Mellon University database [10] with ten test subjects. The database includes head-shoulder full frontal face color video sequences of a person talking. The feature extraction algorithm works well for the data sets, which contain video sequences for several hours. In a few cases, a few pixels of inaccuracy have been observed.

3. AUDIO-VISUAL INTEGRATION

3.1 Previous Approaches

The second major issue in an automatic speechreading system is how to incorporate the visual component into an acoustic speech recognizer so that optimal performance can be achieved by using both modalities together.

For engineering applications, two AV integration models are commonly used in automatic speechreading systems: early integration (EI) and late integration (LI). In early integration, audio-visual fusion is performed in the feature space to form a composite feature vector. Recognition is based on the augmented feature vector. In late integration, each modality is first preclassified independently of the other. The resulting audio and visual recognition scores are then combined using a rule.

Late integration offers several advantages over early integration because its implementation is simple and it does not require synchronization of the acoustic and visual features. In late integration, each independent subsystem can be developed and trained separately. However, the use of separate models assumes conditional independence between the two feature sets and, therefore, it fails to model the correlations between the visual and acoustic channels. Early integration provides a more general model by integrating the two components before recognition. However, the classification is based on training a single HMM using the concatenated audio and visual feature vectors. It forces the same state sequences upon the audio and visual components, which does not correspond to the way that people talk. Often the lips start moving before voicing commences. Therefore an early integration model restricts the asynchrony between the two streams of information that naturally occurs in speech production.

3.2 Our Approach

We propose the use of a generalized model—the coupled hidden Markov model (CHMM) to model the audio-visual interaction for speech recognition. The coupled hidden Markov model was first introduced by Brand in 1996 and was successfully used for modeling Tai Chi gestures [11]. In a coupled HMM, as shown in Fig. 4, the traditional left-right HMM is expanded to a model containing two Markov chains, representing the audio and visual channels. The coupling between the two subprocesses is introduced by conditional probabilities between the hidden state variables \( Pr(S_t^{AV} | S_{t-1}^A, S_{t-1}^V) \) and...
Pr(S,v|o,s). On the one hand, this architecture relaxes the restriction of the early integration by allowing asynchrony between the two channels. On the other hand, unlike late integration, it incorporates temporal coupling terms across the two sub-systems.

To solve the inference problem in a coupled HMM, we employ the approximate approach proposed by Boyen and Koller [12]. The key ingredient of the BK algorithm is the propagation of an approximate probability distribution over the entire system using factored products defined over independent clusters. The accumulated error arising from the repeated approximation was proved to remain bounded indefinitely over time. The BK algorithm has been shown to be an efficient approach to solving inference problems in general dynamic Bayesian networks.

For learning parameters in the CHMM, forward and backward variables are first approximated. The BK algorithm represents the forward variable as a product of marginal random variable over two sub-processes. The approximated forward variable at time \( t \) is then propagated through the transitional model and conditioned on evidence at time \( t+1 \) using the junction tree algorithm [13]. To allow the algorithm to continue, the forward variable at \( t+1 \) is approximated using a random variable that admits a compact representation by computing marginals over each cluster. The same procedures can be applied to approximating the backward variable \( b \). These two variables are then used in an EM algorithm that learns the model in an iterative manner.

4. AUDIO-VISUAL SPEECH RECOGNITION

We performed our experiments on audio-visual speech recognition using the audiovisual database available from Carnegie Mellon University [10]. This database includes ten test subjects (three females, seven males); speaking 78 isolated words repeated ten times. In our experiment, we use the data set consisting of the 31 “number” words: one, two, three, four, five, six, seven, eight, nine, ten, eleven, twelve, thirteen, fourteen, fifteen, sixteen, seventeen, eighteen, nineteen, twenty, thirty, forty, fifty, sixty, seventy, eighty, ninety, hundred, thousand, million, billion.

In the visual sub-system, we used the six geometric features defined in Fig. 3 and the two parameters for the teeth and tongue. Delta features (first derivatives of the features over consecutive frames) were also included in the visual features, forming a 16-dimensional feature vector. They were computed by using a regression formula using a few frames before and after the current frame. The visual feature vectors were preprocessed by normalizing with respect to the average mouth width, \( w_r \), of each speaker to account for the difference in scale between different speakers and different recording settings for the same person.

In the acoustic subsystem, we used twelve mel-frequency cepstral coefficients (MFCCs) and their corresponding delta parameters as the features—a 24-dimensional feature vector. The MFCCs were derived from FFT-based log spectra with a frame period of 11 msec. and a window length of 25 msec.

We conducted tests for both speaker-dependent and speaker-independent tasks. For the speaker-dependent task, the test was evaluated by using a leave-one-out procedure. For the speaker-independent task, we used different speakers for training and testing.

In all cases, the HMMs have ten states, and we modeled the observation vectors using two Gaussian mixtures for the speaker-independent task. Because of the limited training data available, we used a single Gaussian mixture in the speaker-dependent case. For early integration (EI), the classification was based on training a traditional HMM on the concatenated audio-visual observation vectors. Since the video has a frame rate of 33 ms, to match the audio frame rate of 11 ms linear interpolation was applied to the visual features to fit the data values between the existing feature data points. For the late integration fusion (LI), the combined score takes the following form was computed using the formula

\[
\log P_{av} = \lambda \log P_a + (1 - \lambda) \log P_v,
\]

where \( P_a \) and \( P_v \) are the probability scores of the audio and visual components and the weighting factor \( \lambda \) was set to 0.7 in our experiments. For comparison, we also include results for a multistream HMM (MS), which is characterized by its output distribution:

\[
b_i(o|a) = \left\{ b_i(o|a) \right\}_{v}
\]

where the exponents \( \lambda \) and \( \gamma \) are the weighting factors for each stream. We set \( \lambda = 0.7 \) and \( \gamma = 0.3 \) in our experiments.

Model training and Viterbi decoding of the HMMs were implemented using the HTK Toolkit [14]. The BK algorithm for the coupled HMM was implemented using the Bayes Net Toolbox [15]. Prior to employing the BK algorithm, it is essential that the model parameters be well initialized. For this, we applied a traditional EM algorithm on the two separate HMMs, and uses the model parameters trained on the separate HMMs as the initial parameters in the CHMM.

In the following, we present our experimental results on audio-visual speech recognition over a range of noise levels using these four models. Artificial white Gaussian noise was added to simulate various noise conditions. The experiment was conducted under a mismatched condition—the recognizers were trained at 30dB SNR, and tested under varying noise levels. Tables 1 and 2 summarize the recognition performance using the four integ-
tion schemes for the speaker-dependent and speaker-independent tasks, respectively. As can be seen, all four integration models demonstrate improved recognition accuracy over audio only performance. While the recognition accuracy of the CHMM is very close to the best results from three other models in a speaker-dependent task, the CHMM consistently outperforms others in the speaker-independent task. This might indicate that the CHMM requires a larger training data for better model parameter estimates.

Table 1: Audio-visual speech recognition performance in the speaker-dependent mode. The numbers represent the percent correct recognition.

<table>
<thead>
<tr>
<th>SNR</th>
<th>video</th>
<th>audio</th>
<th>El</th>
<th>Li</th>
<th>MS</th>
<th>CHMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0dB</td>
<td>48.94</td>
<td>7.48</td>
<td>39.10</td>
<td>14.74</td>
<td>27.29</td>
<td>39.34</td>
</tr>
<tr>
<td>10dB</td>
<td>48.94</td>
<td>34.90</td>
<td>72.19</td>
<td>57.06</td>
<td>75.06</td>
<td>75.27</td>
</tr>
<tr>
<td>30dB</td>
<td>48.94</td>
<td>84.89</td>
<td>88.48</td>
<td>91.23</td>
<td>91.52</td>
<td>91.23</td>
</tr>
</tbody>
</table>

Table 2: Audio-visual speech recognition performance in the speaker-independent mode. The numbers represent percent correct recognition.

<table>
<thead>
<tr>
<th>SNR</th>
<th>video</th>
<th>audio</th>
<th>El</th>
<th>Li</th>
<th>MS</th>
<th>CHMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0dB</td>
<td>26.90</td>
<td>8.29</td>
<td>15.84</td>
<td>7.13</td>
<td>11.77</td>
<td>20.84</td>
</tr>
<tr>
<td>10dB</td>
<td>26.90</td>
<td>31.13</td>
<td>45.03</td>
<td>32.45</td>
<td>42.77</td>
<td>50.18</td>
</tr>
<tr>
<td>30dB</td>
<td>26.90</td>
<td>68.03</td>
<td>68.74</td>
<td>63.52</td>
<td>68.65</td>
<td>72.19</td>
</tr>
</tbody>
</table>

5. SUMMARY
In this paper, we demonstrated an automatic speechreading system for an audio-visual speech recognition. By combining the visual speech features extracted from our visual front end and a traditional acoustic front end, we performed the bimodal speech recognition using a coupled hidden Markov model. The combined system demonstrated significant performance improvement over an audio only subsystem. This gain is most distinct in low SNR, where traditional ASR performs poorly.

Acknowledgments
We would like to acknowledge the use of the audiovisual data base [10] from the Advanced Multimedia Processing Lab at Carnegie Mellon University.

REFERENCES