PERCEPTUAL INTERFACES FOR INFORMATION INTERACTION: JOINT PROCESSING OF AUDIO AND VISUAL INFORMATION FOR HUMAN-COMPUTER INTERACTION

C. Neti, G.Iyengar, G. Potamianos, A. Senior, B. Maison

IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, U.S.A.

ABSTRACT

We are exploiting the human perceptual principle of sensory integration (the joint use of audio and visual information) to improve the recognition of human activity (speech recognition, speech event detection and speaker change), intent (intent to speak) and human identity (speaker recognition), particularly in the presence of acoustic degradation due to noise and channel. In this paper, we present experimental results in a variety of contexts that demonstrate the benefit of joint audio-visual processing.

1. INTRODUCTION

Humans use a variety of modes of information (audio, visual, touch and smell) to recognize people and understand their activity (speech, emotion, etc). In our lab, we have begun investigating methods to combine audio and visual information to improve the robustness and naturalness of human computer interfaces. In particular, our short term focus is to exploit visual information to improve audio-based technologies such as speech recognition, speaker recognition, speech event detection and speaker change detection in the presence of acoustic mismatch due to noise and channel.

Our long term objective is to build perceptual interfaces for information interaction that use multiple sensory information sources (initially acoustic and visual) for active acquisition, recognition and interpretation of human activity and intent (Perceptual Computing).

The applications for this work include accurate audio transcription for efficient search and retrieval of multimedia content and improved human/computer interfaces that use multiple modes for robust recognition of human activity (speech, gesture, etc) in realistic environments like automobiles and public information kiosks, where background noise is a serious problem for recognition technologies based only on acoustics.

In Figure 1, we illustrate schematically the steps involved in audio-visual fusion. A single percept, such as the identity of a person or his speech activity, is captured by audio and visual sensors. The two streams are then sampled at different rates. These sampled signals are then represented in some feature spaces followed either by feature fusion (visual and audio features are concatenated to form a single feature set, by interpolating the slower stream to equalize the rates) and a decion is made using a single classifier or by combining independent decisions on the two streams (decision fusion) using multiple classifiers.

Note that in each case before feature representation, the region of interest (for the percept being considered) has to be extracted. For both speech and speaker recognition this entails detecting a face in the scene, followed by extracting the region or regions of interest. In the first section, we describe our methodology for face detection, followed by experimental results that demonstrate the benefit of joint audio-visual processing for speech recognition, speaker recognition, speaker-change detection and speech-event detection. In each section, we will describe the extraction of appropriate regions of interest and their representations.

2. FACE DETECTION AND FACIAL FEATURE EXTRACTION

We use the face detection and facial feature localization method described in [11]. Given a video frame, face detection is first performed by employing a combination of methods, some of which are also used for subsequent face feature finding. A face template size is first chosen (11×11 pixels, here), and an image pyramid over the permissible scales (given the frame size and the face template) is used to search the image space for the possible face candidates. Since the video signal is in color, skin-tone segmentation is first used to narrow this search to candidates that contain a significantly high proportion of skin-tone pixels. Every remaining face candidate is given a score based on both a two-class Fisher linear discriminant and its *distance from face space* (DFFS). All candidate regions exceeding a threshold score are considered as faces.

Once a face has been found, an ensemble of facial feature detectors are used to extract and verify the locations of 26 facial features, including the lip corners and centers. The search for these features occurs hierarchically. First, a few “high”-level features are located, and, subsequently, the 26 “low”-level features are located relative to the high level feature locations. The feature locations at both stages are determined using a score combination of prior statistics, linear discriminant and DFFS [11].

3. SPEECH RECOGNITION

It is well known that humans fuse information from both the audio and visual stimuli to recognize speech [1], as well as that the visual modality contains some complementary information to the audio modality [2].

Our focus and interest is in demonstrating meaningful improvements for realistic tasks such as broadcast news transcription for audio/video indexing, large vocabulary dictation [3, 4, 9, 10] and speechreading for the hearing/speech impaired.

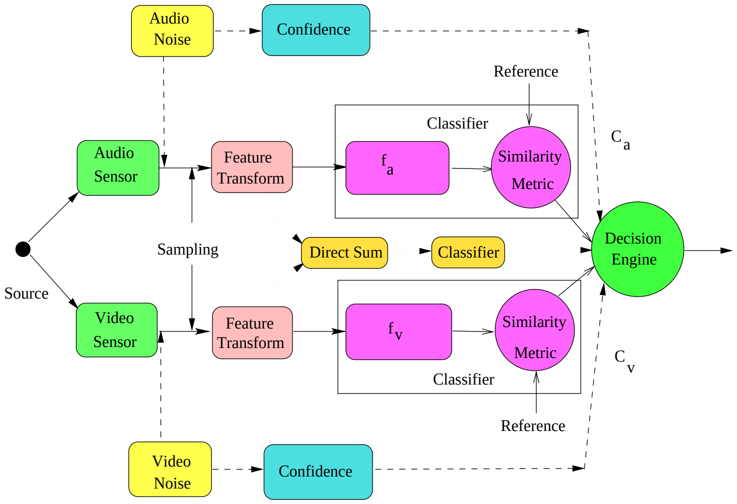

Figure 1: Audio-visual information fusion

We have been collecting a multi-subject, continuous, large vocabulary, audio-visual database, using ViaVoice$^{\text{TM}}$ training utterance scripts (VVAV data). Currently, it consists of 285 subjects and close to 50 hours of speech (about 30,000 utterances). The database contains full frontal face color video of the subjects with minor face-camera distance and lighting variations. The video is captured at a resolution of 704 $\times$ 480 pixels (interlaced), a frame rate of 60 Hz, and it is MPEG2 encoded to about 0.5 MBytes/sec. The audio is captured at a sampling rate of 16 KHz, and it is time-synchronous to the video stream. For the experiments in this paper, we selected a database with 20 utterances for 162 subjects and randomly split them into 16 training and 4 test utterances per subject, thus creating a *multi-subject* 2,592 utterance *training set* (5.5 hours) and a 648 utterance *test set* (1.4 hour).

We first process the video data to extract the mouth region-of-interest (ROI). The statistical face detection and feature localization templates [11] are trained using 10 video frames for 162 database subjects, each marked with the 26 facial feature locations. The face detection performance tested on all 15,350 database video sequences (containing approximately 3.5 million frames) is 99.5% correct, assuming that one face is present per video frame. Given the face detection and mouth feature localization results, ROI is estimated and visual features are computed.

We summarize the results (*word error rate* (WER)) for some large vocabulary, continuous speech recognition (LVCSR) preliminary experiments on this database. See [9] for details. We consider a HMM based LVCSR system with a vocabulary of 60000 words and a tri-gram language model. Visual speech is represented by a cascade transform representation [9] of the mouth region which comprises of a image-based transform followed by projection into a discriminant feature space. Standard cepstral features are used to represent the audio stream. We obtain an audio-only WER of 13.94%, a visual-only WER of 87.60%, and an audio-visual WER of 13.78%. Feature fusion is known to fail to improve ASR performance in small vocabulary tasks [8], therefore the above results in this LVCSR experiment are not surprising. Nevertheless, significant ASR improvement can be achieved when the audio is corrupted by noise. For example, and for audio corrupted by “babble noise” at the signal–to–noise ratio of 10 dB, the *matched* noisy audio WER improves from 41.57% to 31.30% by incorporating the visual information. We are currently implementing decision fusion instead of feature fusion as the audio-visual LVCSR fusion strategy, by means of the *multi-stream* HMM ([8]). In addition, we are investigating various stream “confidence” estimation techniques [10]. We believe that significant LVCSR WER reduction can be achieved by such techniques, even in the clean audio case.

4. SPEAKER RECOGNITION

Here we combine image or video based visual signatures (face identification, [11]) with audio feature based speaker identification for improved person authentication (see [5] for details).

4.1. Image based speaker identification

A set of K facial features are located. These include large scale features and small scale sub-features. Prior statistics are used to restrict the search area for each feature and sub-feature. At each of the estimated sub-feature locations, a Gabor Jet representation is generated. A Gabor jet is a set of 2-dimensional Gabor filters — each a sine wave modulated by a Gaussian. Each filter has scale and orientation. We use five scales and eight orientations, giving 40 complex coefficients ($a(j), j = 1, \ldots, 40$) at each feature location.

A simple distance metric is used to compute the distance between the feature vectors for trained faces and the test candidates. The distance between the i^{th} trained candidate and a test candidate for feature k is defined as:

$$\mathcal{S}_{ik} = \frac{\sum_j a^k(j) a_i^k(j)}{\sqrt{\sum_j a^k(j)^2 \sum_j a_i^k(j)^2}} \qquad (1)$$

An average of these similarities,

$$f_{vi} = 1/K \sum_{k=1}^{K} \mathcal{S}_{ik}$$

gives an overall measure for the similarity of the test face to the face template in the database.

4.2. Audio-based speaker identification

The frame-based approach for audio based speaker identification can be described as follows.

Let M_i, the model corresponding to the i^{th} enrolled speaker, be represented by a mixture Gaussian model defined by the parameter set $P_i(\mu_i, \Sigma_i, p_i)$, consisting of the mean vectors μ_i, covariance matrices Σ_i and mixture weight vectors p_i. The goal of speaker identification is to find the model, M_i, that best explains the test data represented by a sequence of N frames $\{f_n\}_{n=1,..,N}$. The total distance, f_{ai} of model M_i from the test data is then taken to be the sum of the "distances" $d_{i,n} = -\log P_i(f_n|\mu_i, \Sigma_i, p_i)$ of all the test frames measured as per likelihood criterion, i.e.,

$$f_{ai} = \sum_{n=1}^{N} d_{i,n} \tag{2}$$

4.3. Fusion

Given the audio-based speaker recognition and face recognition scores, *audio-visual speaker identification* is carried out as follows: the top N scores are generated-based on both audio and video-based identification schemes. The two lists are combined by a weighted sum. Subsequently the best-scoring candidate is chosen. We can define the combined score $F^i \equiv F^i_{C_a,C_v}$ as a function of the single parameter α:

$$F^i = C_a f_{vi} + C_v f_{ai} \;\; \text{with} \;\; C_a = \cos\alpha, \;\; C_v = \sin\alpha \tag{3}$$

The angle α has to be selected according to the relative reliability of audio and face identification. For this, one may optimize α to gain maximum accuracy on some training data. To elaborate on this, denote by $f_{ai}(n)$ and $f_{vi}(n)$ the respective scores for the ith enrolled speaker computed on the nth training clip. Let us define the variable $T_i(n)$ as zero when the nth clip belongs to the ith speaker and equal to unity otherwise. One can minimize the number of empirical errors $C(\alpha)$ given by

$$C(\alpha) = \frac{1}{N}\sum_{n=1}^{N} T_{\hat{\imath}}(n), \;\; \text{where} \;\; \hat{\imath} = \arg\max_i F^i(n) \tag{4}$$

and $F^i(n)$ is the joint score with $f_{ai} = f_{ai}(n)$ and $f_{vi} = f_{vi}(n)$.

All the experiments were carried out on CNN and CSPAN video data collected as part of the ARPA HUB4 broadcast news transcription task by the linguistic data consortium (LDC). We digitized 20-40 second clips of anchors and reporters with frontal shots of their faces from the video tapes into MPEG2 format. The training data contained 76 clips of 76 speakers while the test data consisted of 154 additional clips from the same 76 speakers

The results of combining audio and visual information for speaker recognition using linear fusion methods are shown in Table 1. See [5] for details.

5. SPEAKER CHANGE DETECTION

Speaker change detection is a valuable piece of information for speaker identification and as metadata for search and retrieval of multimedia content. We are currently exploring the use of visual speaker and scene change information to remove the limitiations of audio-based speaker change detection (See [6] for details). Our hypothesis is that the performance of audio only or video only techniques can be further improved by exploiting the joint statistics between the audio stream and its associated video. There is significant correlation between audio and video speaker changes in a newscast scenario, for example. Frequently, the video scene change follows shortly after an audio change.

	Acoustic Condition	Clean	Noisy	Telephone
1	Audio ID only	92.6%	77.7%	50.0%
2	Video ID only	81.1%	81.1%	81.1%
3	Linear fusion	93.2%	87.8%	87.2%

Table 1: Audio-visual speaker ID

5.1. Audio-based speaker change

The problem of detecting a transition point at time i is to choose between two models of the data: one where the data set is modeled by a single Gaussian process i.e., $x_{a1}\cdots x_{aN} \sim N(\mu,\Sigma)$, or by two distinct Gaussian processes $x_{a1}\cdots x_{ai} \sim N(\mu_1,\Sigma_1)$ and $x_{a(i+1)}\cdots x_{aN} \sim N(\mu_2,\Sigma_2)$. Here, the obvious notation μ for the mean vector and Σ for the covariance matrix has been used.

The BIC based model selection procedure considers the difference between the BIC values associated with the two models as a "classifier":

$$f'_a(i) = R(i) - \lambda P \tag{5}$$

where $R(i)$ is the maximum likelihood ratio statistics:

$$R(i) = N log|\Sigma| - N_1 log|\Sigma_1| - N_2 log|\Sigma_2|, \tag{6}$$

$P = 0.5(d+0.5d(d+1))\log N$ is the penalty, d is the dimension of the vectors x_{ai}'s, and $\lambda = 1$. We consider i to be a transition point if $f'_a(i) > 0$.

5.2. Video-based speaker change

While for video based scene change detection a statistical model based criterion such as the BIC criterion could also be used, we describe an alternate procedure. Consider the n dimensional color histogram generated by the video feature vectors $x_{vi} \in \mathbf{R^n}$ ($n = 64$ in our experiments), and consider a Kullbach-Liebler type divergence criterion:

$$g_v(i) = -\sum_{k=1}^{n} x^k_{vi} \log \frac{x^k_{v(i-1)}}{x^k_{vi}}$$

between the adjoining vectors x^k_{vi} and $x^k_{v(i-1)}$, where the superscript k denotes the kth component of vectors.

We then compute the average $\bar{g}_v(i)$ of $g_v(i)$ over a fixed number N of samples in the past of i and consider i to be a transition point if for a threshold Θ

$$f'_v(i) = |\bar{g}_v(i) - g_v(i)| - \Theta > 0. \tag{7}$$

5.3. Fusion

The fusion problem now is to intelligently combine two probabilities. One of these is the probability $f_v = \Pr(f'_v(i) > 0 | \{x_{vi}\}_{i=1}^N)$ that $f'_v(i)$ in (7) given N video feature vectors from the past is positive. The other is the probability $f_a = \Pr(f'_a(i) > 0 | \{x_{ai}\}_{i=1}^N)$ that $f'_a(i)$ in (5) computed based on audio data $\{x_{ai}\}_{i=1}^N$ is positive. The fusion strategy then is to devise an adequate fusion map F_{C_a,C_v}. In the particular case under consideration, a fusion strategy is to solve the optimization problem

$$F_{C_a,C_v}(i) = \arg\max_{i,\Delta} \{C_a f_a(i) + C_v f_v(i+\Delta)\}$$

where Δ is a parameter that accounts for the fact that the speaker change in audio signal precedes the speaker change in the video signal.

In 31 minutes of a television panel discussion that we analyzed, 67% of the audio speaker changes were immediately followed (within 3 seconds) by a corresponding video change. Our initial results on CSPAN video content suggest that fusion helps improve the precision for the same recall rate. For example, at a recall rate of 88.5%, audio-only precision is 83.3% while audio-visual precision is 95%.

6. SPEECH EVENT DETECTION

Speech recognition systems have opened the way towards an intuitive and natural human-computer interaction (HCI). However, current HCI systems using speech recognition require a human to explicitly indicate one's intent to speak by turning on a microphone using the keyboard or mouse. One of the key aspects of naturalness of speech communication involves the ability of humans to detect an intent to speak. For recent experiments on this we refer to [7]. Humans detect an intent to speak by a combination of visual and auditory cues. Visual cues include physical proximity, frontality of pose, lip movement, etc. Automatic detection of speech onset can be carried out using silence/speech detection or based on audio energy alone. One method of combining the two methods may be to compute the following two probability densities

$$f_a = Pr(\text{speech}|x_a), \quad \text{and} \quad f_v = Pr(\text{speech}|x_v)$$

as, say, mixtures of Gaussian pdfs. And, a simple fusion strategy is to use the linear combination:

$$F_{C_a,C_v} = C_a f_a + C_v f_v.$$

6.1. Experiments

Using a subset of the VVVAV database (described in the speech recognition section) containing 10hrs of training data comprising of 76 speakers, we were able to improve speech event detection (speech/silence classification) from 73% audio-only at 10 dB SNR (speech noise) to about 79.24% by combining with visual information.

In addition, we have built a practical system that uses visual cues such as proximity, head pose and visual mouth activity to intuitively manipulate the microphone state during speech interaction.

7. CONCLUSIONS

Fusion of multiple sources of information is a mechanism to robustly recognize human activity and intent in the context of human computer interaction. In this paper, we have presented results that suggest that meaningful improvements can be obtained by fusion of audio and visual information for speech recognition, speaker recognition, speaker change detection and speech event detection.

8. REFERENCES

1. H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746-748, 1976.
2. D.W. Massaro and D.G. Stork, "Speech recognition and sensory integration," *American Scientist*, vol. 86, pp. 236-244, 1998.
3. S. Basu, C. Neti, N. Rajput, A. Senior, L. Subramaniam, A. Verma, Audio-visual large vocabulary continuous speech recognition in the broadcast domain. IEEE MMSP Workshop, 1999.
4. A. Verma, T. Faruquie, A. Senior, C. Neti and S. Basu, Late-integration in audio-visual continuous speech recognition. Automatic Speech Reco. & Understanding Workshop, 1999.
5. B. Maison, C. Neti and A. Senior, Audio-visual speaker recognition for video broadcast news: some fusion techniques. IEEE MMSP Workshop, 1999.
6. G. Iyengar and C. Neti, Speaker Change detection using joint audio-visual statistics. RIAO 2000, Paris, France, April 2000.
7. P. Decuetos, C. Neti and A. Senior, Audio-visual Intent-to-speak detection for Human Computer Interaction. IEEE Int. Conf. on Acoustics Speech and Signal Processing., 2000.
8. G Potamianos and H. P. Graff, Discriminative training of HMM stream exponents for audio-visual speech recognition. Proc. ICASSP, pp.3733-3736, 1998.
9. G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu, "A cascade image transform for speaker independent automatic speechreading," to appear: *Proc. Int. Conf. Multimedia Expo.*, New York, 2000.
10. G. Potamianos and C. Neti, Stream Confidence estimation for audio-visual speech recognition. Proc. ICSLP, Beijing, 2000. *to appear*.
11. Senior, A.W., Recognizing faces in broadcast video. IEEE International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 105–110, 1999.