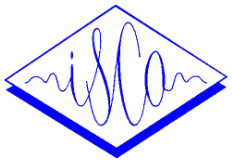

Auditory-Visual Speech Processing
(AVSP'99)
Santa Cruz, CA, USA
August 7-10, 1999

ON THE USE OF VISUAL INFORMATION FOR IMPROVING AUDIO-BASED SPEAKER RECOGNITION

*Andrew Senior, Chalapathy V. Neti and Benoît Maison**
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598

ABSTRACT

Audio-based speaker identification degrades severely when there is a mismatch between training and test conditions either due to channel or noise. In this paper, we explore various techniques to fuse video based speaker identification with audio-based speaker identification to improve the performance under mismatch conditions.

1. INTRODUCTION

Humans identify speakers based on a variety of attributes of the person which include acoustic cues, visual appearance cues and behavioural characteristics (such as characteristic gestures, lip movements). In the past, machine implementations of person identification have focussed on single techniques relating to audio cues alone (speaker recognition), visual cues alone (face identification, iris identification) or other biometrics. More recently, researchers are attempting to combine multiple modalities for person identification [3]. Speaker identification is an important technology for a variety of applications including security, and more recently as an index for search and retrieval of digitized multimedia content (for instance in the MPEG7 standard). Audio-based speaker recognition accuracy under acoustically degraded conditions (such as background noise) and channel mismatch (telephone) still needs further improvements. To make improvements in such degraded conditions is a hard problem. We have begun [6] to investigate the combination of audio-based processing with visual processing for speaker recognition to improve the accuracy in acoustically degraded conditions in the broadcast news domain. The use of two independent sources of information can bring significantly increased robustness to both speech and speaker recognition since signal degradations in the two channels are uncorrelated [2]. Furthermore, the use of visual information allows a much faster speaker identification than possible with acoustic information. In this paper, we present results of various methods to fuse person identification based on visual information with identification based on audio information for TV broadcast news video data (CNN and CSPAN) provided by the linguistic data consortium (LDC).

*In reverse alphabetical order

2. METHOD

The system carries out speaker identification independently on the acoustic and visual signal. The results for the two modes are then combined together to arrive at a final speaker identity, and a list of scores for all the registered speakers indicating their similarity to the test speaker.

2.1. Visual speaker identification

The visual mode of speaker identification is implemented as a face recognition system. Faces are found and tracked in the video sequences, and recognized by comparison with a database of candidate face templates. This section describes the detection, tracking and recognition processes.

2.1.1. Face detection

Faces can occur at a variety of scales, locations and orientations in the video frames. In this system, we make the assumption that faces are close to the vertical, and that there is no face smaller than 66 pixels high. However to test for a face at all the remaining locations and scales, the system searches for a fixed size template in an image pyramid. The image pyramid is constructed by repeatedly downsampling the original image to give progressively lower resolution representations of the original frame. Within each of these sub images, we consider all square regions of the same size as our face template (typically 11x11 pixels) as candidate face locations. A sequence of tests is used to test whether a region contains a face or not. These are summarized below and described in more detail in another paper [4].

First, the region must contain a high proportion of skin-tone pixels, and then the intensities of the candidate region are compared with a trained face

model. Pixels falling into a pre-defined cuboid of hue–chromaticity–intensity space are deemed to be skin tone, and the proportion of skin tone pixels must exceed a threshold for the candidate region to be considered further.

The face model is based on a training set of cropped, normalized, grey-scale face images. Statistics of these faces are gathered and a variety of classifiers are trained based on these statistics. A Fisher linear discriminant trained with a linear program is found to distinguish between faces and background images, and 'Distance from face space' (DFFS) [7] is used to score the quality of faces given high scores by the first method. A high combined score from both these face detectors indicates that the candidate region is indeed a face. Candidate face regions with small perturbations of scale, location and rotation relative to high-scoring face candidates are also tested and the maximum scoring candidate among the perturbations is chosen, giving refined estimates of these three parameters.

In subsequent frames, the face is tracked by using a velocity estimate to predict the new face location, and face models are used to search for the face in candidate regions near the predicted location and with similar scales and rotations. A low score is interpreted as a failure of tracking, and the algorithm begins again with an exhaustive search.

2.1.2. Face recognition

Having found the face, K facial features are located using the same techniques (linear discriminant and DFFS) used for face detection. Features are found using a hierarchical approach where large-scale features, such as eyes, nose and mouth are first found, then sub-features are found relative to these features. As many as 29 sub-features are used, including the hairline, chin, ears, and the corners of mouth, nose, eyes and eyebrows. Prior statistics are used to restrict the search area for each feature and sub-feature relative to the face and feature positions respectively. At each of the estimated sub-feature locations, a Gabor Jet representation [8] 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 (the sine wavelength and Gaussian standard deviation with fixed ratio) and orientation (of the sine wave). 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:

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

A simple average of these similarities, $S_i = 1/K \sum_1^K \mathcal{S}_{ik}$, gives an overall measure for the similarity of the test face to the face template in the database.

2.2. Audio-based speaker identification

The IBM Speaker identification system uses two techniques: a model-based approach and a frame-based approach [1]. In the experiments described here, we use the frame-based approach for speaker identification based on audio. Briefly, the frame-based approach can be described as follows:

Let M_i be the model corresponding to the i^{th} enrolled speaker. M_i is represented by a mixture Gaussian model defined by the parameter set $\{\mu_{i,j}, \Sigma_{i,j}, p_{i,j}\}_{j=1,..n_i}$, consisting of the mean vector, covariance matrix and mixture weights for each of the n_i components of speaker i's model. These models are created using training data consisting of a sequence of K frames of speech with d-dimensional cepstral feature vectors, $\{f_m\}_{m=1,..,K}$. 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, $\{f_n\}_{n=1,..,N}$. We use the following frame-based weighted likelihood distance measure, $d_{i,n}$ in making the decision:

$$d_{i,n} = -\log \left[\sum_{j=1}^{n_i} p_{i,j} p(f_n | \mu_{i,j}, \Sigma_{i,j}) \right] \tag{2}$$

The total distance, D_i of model M_i from the test data is then taken to be the sum of the distances over all the test frames.

$$D_i = \sum_{n=1}^{N} d_{i,n} \tag{3}$$

2.3. Fusion

In general, mode-fusion or the integration of different modes of information can be achieved by any of the following methods of data fusion [5].

- data fusion — this involves integration of different modalities in raw form e.g. video camera and microphone outputs.
- feature fusion — features are extracted from the raw data and subsequently combined, e.g. for speaker recognition cepstral features and facial Gabor jet features could be combined.

- decision fusion — this is the fusion at the most advanced stage of processing and involves combining the decisions of two different classifiers making independent decisions about the identity of the speaker-based on audio and visual features

In general, decision fusion provides a higher degree of robustness, but is accompanied by possible loss of information. An optimal fusion policy of using one of these fusion strategies or some combination of the three strategies needs to be investigated. For this paper, we have experimented with the technique of decision fusion and combine the scores based on visual information (face-identification) and audio information (based on audio speaker identification).

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 and the best-scoring candidate is chosen. Since the weights need only to be defined up to a scaling factor, we can define the combined score S_i^{av} as a function of the single parameter α:

$$S_i^{av} = \cos\alpha \; D_i + \sin\alpha \; S_i \tag{4}$$

The mixture angle α has to be selected according to the relative reliability of audio identification and face identification. One way to achieve this is to optimize α in order to maximize the audio-visual accuracy on some training data. Let us denote by $D_i(n)$ and $S_i(n)$ the audio ID and video ID score for the ith enrolled speaker $(i = 1 \ldots P)$ 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 one otherwise. The cost function $C(\alpha)$ to be minimized is the empirical error rate [9], that can be written as

$$C(\alpha) = \frac{1}{N}\sum_{n=1}^{N} T_{\hat{\imath}}(n) \quad \text{where} \quad \hat{\imath} = \arg\max_i S_i^{av}(n), \tag{5}$$

and where

$$S_i^{av}(n) = \cos\alpha \; D_i(n) + \sin\alpha \; S_i(n). \tag{6}$$

In order to prevent over-fitting, one can also resort to the smoothed error rate [10] defined as

$$C'(\alpha) = \frac{1}{N}\sum_{n=1}^{N}\sum_i T_i(n)\frac{\exp^{\eta S_i^{av}(n)}}{\sum_{j=1}^{P}\exp^{\eta S_j^{av}(n)}}, \tag{7}$$

When η is large, all the terms of the inner sum approach zero, except for $i = \hat{i}$, and $C'(\alpha)$ approaches the raw error count $C(\alpha)$. Otherwise, all the incorrect hypotheses (those for which $T_i(n) = 1$) have a contribution that is a decreasing function of the distance between their score and the maximum score. If the best hypothesis is incorrect, it has the largest contribution. Hence, by minimizing the latter cost function, one tends to maximize not only the recognition accuracy on the training data, but also the margin by which the best score wins. This function also presents the advantage of being differentiable, which can facilitate the optimization process when there is more than one parameter.

3. RESULTS

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

As pointed out earlier, the key challenge for audio-based speaker identification is to improve performance when there is a significant mismatch between testing and training conditions either due to background noise or channel mismatch. To investigate the benefit of combining video information under these conditions we artificially generated mismatch between training and test conditions. Noise mismatch was created by adding speech noise to the audio signal at a signal-to-noise ratio of about 10 dB.

Table 1 shows the recognition accuracy for different testing conditions and fusion techniques. The first two rows give the accuracy of audio-only ID and video-only ID. The next four rows show the results of several linear fusion experiments. Since training data is needed for the optimization of the fusion weights, the 154 clips have been split into two sets of 77, with occurrences of the same speaker evenly divided. The fusion weights have been trained on set 1, then tested on set 2, and conversely. The total number of tests is 154, like in the first two rows. Hard optimization refers to the raw error count of Eq. (5), while soft optimization refers to the smoothed cost function of Eq. (7). For noisy data, rows 3 and 4 refer to fusion weights optimized on clean data (of set 1, when testing on set 2, and conversely), i.e. fusion mismatch conditions, while rows 5 and 6 refer to fusion weights optimized on noisy data.

Linear joint audio-visual identification significantly improves the accuracy on noisy audio data, while it does slightly worse on clean data. A detailed analysis

	Acoustic Condition	Clean	Noise mismatch
1	Audio ID only	92.9%	77.9%
2	Video ID only	63.6%	63.6%
3	Linear fusion Hard opt.	90.9%	81.2%
4	Linear fusion Soft opt.	92.2%	82.5%
5	Matched fusion Hard opt.	n.a.	83.8%
6	Matched fusion Soft opt.	n.a.	84.4%

Table 1. Audio-visual speaker ID

of the results shows that the amount of training data is insufficient to properly train the fusion weights in the latter case.

Detailed examination of the audio and visual scores suggest that the simple fusion technique used based on a linear combination of the audio and visual scores is sufficient for the data set on which the experiments were carried out. We are investigating other techniques based on estimates of the confidence of the classifiers to determine the weights of the linear combination.

REFERENCES

[1] H. Beigi, S. H. Maes, U.V. Chaudari and J.S. Sorenson, IBM model-based and frame-by-frame speaker recognition. *Speaker Recognition and its Commercial and Forensic Applications*, Avignon, France, 1998.

[2] C. Benoit, T. Guiard-Marigny, B. Le Goff and A. Adjoudani, Which components of the face do humans and machines best speechread? In D. G. Stork and M. E. Hennecke (Eds.), *Speechreading by humans and machines* (pp. 315-328). New York: Springer, 1996.

[3] J. Bigun, B. Duc, F. Smeraldi, S. Fischer and A. Makarov Multi-modal person authentication. In H. Wechsler, J. Phillips, V. Bruce, F. Fogelman Soulié, T. Huang (Eds.), *Face recognition: From theory to applications. Berlin: Springer-Verlag.*, 1998.

[4] A. Senior, Face and Feature Finding for a Face recognition System. In *Proceedings of the Audio and Video-based Person Authentication'99*, March 1999.

[5] D. L. Hall, Mathematical Techniques in multisensor data fusion. Artech House, 1992.

[6] C. Neti and A. Senior, Audio-visual speaker recognition for the broadcast news domain. *Proceedings of the ARPA HUB4 workshop*, Washington D.C., March 1999.

[7] M. Turk and A. Pentland, Eigenfaces for Recognition. *Journal of Cognitive Neuro Science* Vol. 3 No. 1 pp. 71–86 1991.

[8] L. Wiskott and C. von der Malsburg, Recognizing Faces by Dynamic Link Matching. *Proceedings of the International Conference on Artificial Neural Networks* pp. 347–352 1995.

[9] V. N. Vapnik, The Nature of Statistical Learning Theory Springer, 1995.

[10] H. Ney, On the Probabilistic Interpretation of Neural Network Classifiers and Discriminative Training Criteria *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 17 No. 2 pp. 107-119 1995.