

# Speaker Localisation Using Audio-Visual Synchrony: An Empirical Study

H.J. Nock, G. Iyengar, and C. Neti

IBM TJ Watson Research Center, PO Box 218, Yorktown Heights, NY 10598. USA.

**Abstract.** This paper reviews definitions of audio-visual synchrony and examines their empirical behaviour on test sets up to 200 times larger than used by other authors. The results give new insights into the practical utility of existing synchrony definitions and justify application of audio-visual synchrony techniques to the problem of active speaker localisation in broadcast video. Performance is evaluated using a test set of twelve clips of alternating speakers from the multiple speaker CUAVE corpus. Accuracy of 76% is obtained for the task of identifying the active member of a speaker pair at different points in time, comparable to performance given by two purely video image-based schemes. Accuracy of 65% is obtained on the more challenging task of locating a point within a  $100 \times 100$  pixel square centered on the active speaker’s mouth without no prior face detection; the performance upper bound if perfect face detection were available is 69%. This result is significantly better than two purely video image-based schemes.

## 1 Introduction

Several recent papers discuss the idea of “audio-visual synchrony”, which considers the strength of relationship between audio signals and video image sequences. For example, a soundtrack containing drumbeats and an image sequence showing the person beating that drum would be strongly related or strongly “synchronous”. Similarly, if one (or both) faces in Figure 1 are saying the words heard in the speech soundtrack, then the audio and video signals are “synchronous”. Conversely, if the soundtrack for Figure 1 has a voiceover unrelated to images on screen then the audio and video signals are not “synchronous”.



**Fig. 1.** Keyframe from TREC Video Track 2002 Corpus

Many definitions of audio-visual synchrony are possible. Those proposed thus far fall somewhere between two extremes. At one extreme are *generic* definitions of synchrony using weak models: these assign good scores to any audio and video signals displaying consistency without regard to the type of audio and video signal under consideration. Examples include [9], which defines synchrony as the Gaussian-based

mutual information between audio energy and pixel intensities. At the other extreme are definitions of synchrony that are *specific* to particular types of signal; most existing work of this type considers evaluating consistency between facial movements and speech. Examples include [14], which suggests the speech signal could be used to synthesise a talking head and differences between synthesised and actual faces compared. The other measures proposed fall somewhere between these two extremes. Suggestions include [14], which uses Canonical Correlation Analysis on a set of training data to find the linear projection of audio and automatically-detected video face data onto a single axis that maximises the (linear) correlation between the projected variables; synchrony of new data is evaluated by calculating the (linear) correlation between the new audio and video in this space. The paper by [6] uses a pre-trained time-delay neural network classifier to predict whether audio and video features at a particular frame correspond to a person talking.

In practice, the appropriate definition of audio-visual synchrony may vary with the intended application. Our research is currently focussing on synchrony measures useful for deciding whether facial movements are related to speech signals, since a good measure for evaluating consistency of face-speech relationships would have many potential applications. An important part of this research is to consider whether these audio-visual synchrony measures scale to larger and more challenging test sets than those used by other authors. In addition to evaluating these measures using relatively large artificial test sets, this paper considers the use of synchrony measures for speaker localisation in eg. broadcast video. A robust solution to this problem would benefit multimedia retrieval applications in two ways<sup>1</sup>. Firstly, a more robust method for speaker localisation in broadcast video would allow wider application of IBM's audio-visual speech recognition systems (eg. [4]), which are more robust than audio-only speech recognition in noisy environments. This should improve transcription accuracy in traditionally challenging non-studio environments, improving speech-based-indexing performance in those video segments. Secondly, systems for automatic semantic concept annotation in video (eg. [1]) could exploit speaker localisation information for distinguishing monologues, dialogues and voiceovers. To illustrate, consider Figure 1 for which (a) absence of synchrony implies a voiceover shot, (b) high synchrony localised to the right face implies a monologue and (c) an alternating pattern of high synchrony implies a dialogue. Similar ideas were used in our TREC Video Track 2002 automatic monologue annotator [10].

The paper is structured as follows. Section 2 reviews selected measures of audio-visual synchrony and investigates their behaviour on artificial test sets. Motivated by these findings, Section 3 describes our synchrony-based approach to active speaker localisation in broadcast video and presents empirical results on the CUAVE corpus. The paper ends with conclusions and future work.

## 2 Preliminary Studies

This section reviews the generic and specific speech and face consistency measures defined in [11] and discusses their empirical behaviour on artificial test sets.

---

<sup>1</sup> Potential applications outside the digital library arena include face and voice-based biometrics, systems for marking speaker-level metadata such as speaker turns in video, systems for assessing dubbing quality and systems for animating lip movements in cartoon characters.

## 2.1 Definitions of Audio-Visual Synchrony

Assume we are given a test video clip with a speech soundtrack and a moving face in the video. Let  $a_t \in \mathcal{R}^n$  be a feature vector describing the acoustic signal at time  $t$  eg. a vector of mel-frequency cepstral coefficients. Let  $v_t \in \mathcal{R}^m$  be a feature vector describing the image at time  $t$  eg. a vector of discrete cosine transform coefficients of the face region in the frame at  $t$ . Let  $\mathcal{S}_1^T = ((a_1, v_1), \dots, (a_T, v_T))$  represent the joint sequence of audio and video feature vectors. Our goal is to define a measure of synchrony between sequences  $\mathcal{A}_1^T = a_1, \dots, a_T$  and  $\mathcal{V}_1^T = v_1, \dots, v_T$  derived from a test clip.

**Generic Measures.** Let us ignore the information that  $\mathcal{A}_1^T$  and  $\mathcal{V}_1^T$  correspond to audio containing speech and images containing a face. Consider each vector in  $\mathcal{S}_1^T$  to be an independent sample from some joint distribution  $p(A, V)$ , rather than explicitly modelling temporal dependence in the individual sequences. As suggested in [9], one measure of audio-visual synchrony is the mutual information  $\mathcal{I}(A; V)$  between random variables  $A$  and  $V$ <sup>2</sup>. Since the  $p(A)$ ,  $p(V)$ ,  $p(A, V)$  forms are unknown in practice, assumptions must be made. Our earlier work investigated two simple assumptions, both allowing straightforward evaluation of mutual information [11]: discrete distributions and, as in [9], continuous multivariate Gaussian distributions. We note here only that assumption of discrete distributions requires a preparation phase which constructs codebooks to quantize  $a_t$ ,  $v_t$  and  $(a_t, v_t)$  prior to discrete distribution estimation at test time; in contrast, assumption of multivariate Gaussian distributions allows parameter estimation at test time without prior preparation. We term these implementations *Discrete Mutual Information* (“*Discrete MI*”) and *Gaussian Mutual Information* (“*Gaussian MI*”).

**Face-and-Speech Specific Measures.** Require now that  $\mathcal{A}_1^T$  and  $\mathcal{V}_1^T$  correspond to speech audio and images containing faces. Assume we know the word sequence  $\mathcal{W}$  spoken in audio  $\mathcal{A}_1^T$ . We define likelihood  $p(\mathcal{S}_1^T | \mathcal{W})$  as a measure of synchrony. In practice  $\mathcal{W}$  may be unknown; audio-only speech recognition gives a reasonable approximation. Similarly, the form of  $p(\mathcal{S}_1^T | \mathcal{W})$  is unknown in practice; one implementation uses Hidden Markov Models (HMMs) trained on joint sequences of audio- and visual- data, such as HMMs for audio-visual speech recognition (eg. [4]). We term this implementation *Audio-Visual Likelihood* (“*AV-LL*”).

## 2.2 Empirical Behaviour of Audio-Visual Synchrony

This section discusses empirical findings about the definitions of audio-visual synchrony discussed above. Our experiments use artificial test sets constructed from the IBM ViaVoice<sup>TM</sup> audio-visual database [13], comprising full-face frontal video and audio of multiple speakers reading prompts from a large vocabulary in a continuous speech fashion. These experiments therefore assume prior existence of (good) face and speech detection and focus upon the success of different measures in assessing synchrony between facial movements and speech.

Audio features  $a_t$  are extracted as follows. 24 Mel-frequency cepstral coefficients (MFCCs) are extracted from the audio signal at a 100Hz rate. These features are

<sup>2</sup> Mutual Information (see [5]) is a measure of dependence between random variables or, phrased differently, the amount of information one random variable tells us about another one. To illustrate for the discrete case, let  $A$  and  $V$  be discrete random variables with joint distribution  $p(A, V)$  and marginal distributions  $p(A)$  and  $p(V)$ . Then  $\mathcal{I}(A; V) = \sum_{a \in A} \sum_{v \in V} p(a, v) \log \frac{p(a, v)}{p(a)p(v)}$ , which is also the Kullback-Leibler distance (ie. relative entropy) between the joint and product distributions.

used directly in the Discrete and Gaussian MI schemes; before presentation to the audio-visual HMMs, at every  $t$ , 9 consecutive MFCC vectors are concatenated, projected to a lower dimensional space using linear discriminant analysis (LDA) and rotated by a maximum likelihood linear transform (MLLT, [8]) to give a 60-dimensional vector.

Video features  $v_t$  are extracted as follows. Operating at a 60Hz sampling rate, a normalised mouth region-of-interest (ROI) is extracted from each frame and then compressed using a discrete cosine transform (DCT). The 24 highest energy DCT coefficients are retained and linearly interpolated to give a 100Hz frame rate, matching the audio processing. These features are used directly in Discrete and Gaussian MI schemes; before presentation to the audio-visual HMMs, at every  $t$ , 15 consecutive frames are concatenated and LDA and MLLT transforms applied to give a 41-dimensional vector.

**Experiment 1: Choose synchronised speaker from a small set.** Previous work used an artificial test set, intended to simulate video-conferencing or CSPAN panel discussion scenarios in which we want to identify the talking face among several in a shot. A total of 1016 “true” speech and face combinations of four seconds in length were extracted from the database; for each of these “true” cases, one, two or three “confuser” examples were formed by pairing the “true” audio with four seconds of video from randomly chosen speaker(s). Table 1 briefly reviews the results, which show that Gaussian MI significantly outperforms the other two definitions of synchrony for this task; reasons are discussed in [11]. We conclude that where good face and speech detection is available, Gaussian MI will solve at least two practical problems in video analysis. Firstly, it solves the speaker localisation problem that arises when we have already identified one or more faces on screen but need to identify the talking face. Secondly, it gives information for classifying video shots already known to contain talking face(s) into monologues or dialogues via the pattern of activity amongst those faces.

Number of Confusers	Discrete MI	Gaussian MI	AV-LL
1	70	91	72
2	53	85	53
3	47	82	45

**Table 1.** Synchronised Speaker Detection % Accuracy

**Experiment 2: Classifying speakers as synchronous or non-synchronous.** A second experiment examined whether Gaussian MI - the most successful synchrony measure in **Experiment 1** - is useful for absolute classification tasks such as distinguishing between voiceovers and monologues. An artificial test set of 1016 four second examples was constructed, in which 254 examples contain synchronised faces and speech and 762 do not. Despite the earlier success of this measure in correctly ranking synchronised speakers, the absolute classification performance was very poor. The result is explained by Figure 2, which shows the distribution of synchrony scores for the synchronous and non-synchronous test clips. The classes are highly overlapping based on this measure and no good decision boundary can

be defined. We conclude Gaussian MI alone is not adequate for video annotation applications such as distinguishing voiceover shots from monologue shots.

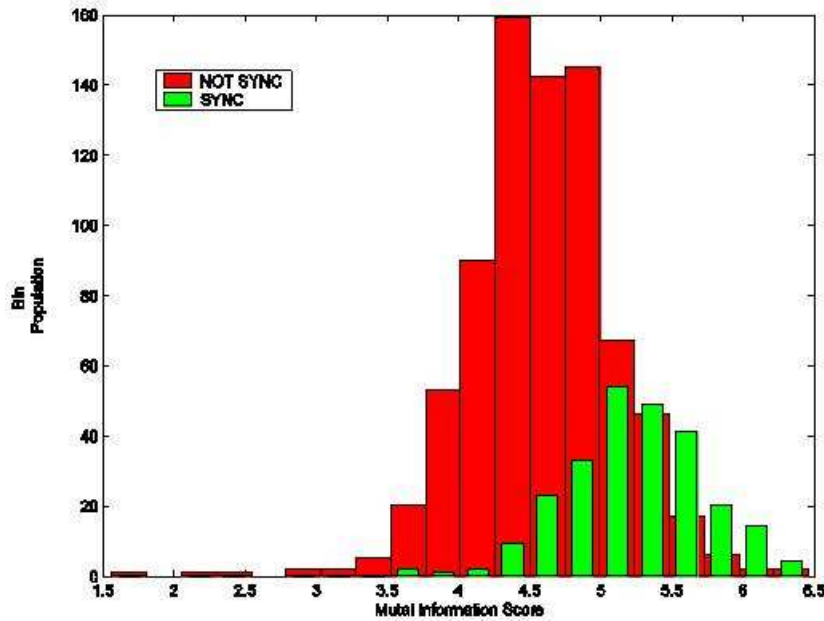


Fig. 2. Per-class histograms of Gaussian MI Scores

### 3 Using Synchrony for Speaker Localisation

This section studies Gaussian MI further by applying it to speaker localisation on a real test set.

#### 3.1 Corpus

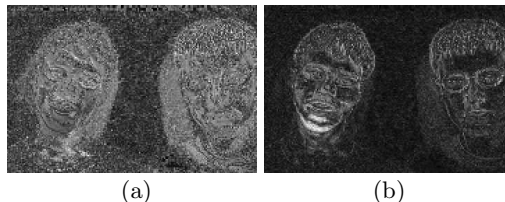
Experiments use the CUAVE corpus [12], a speaker-independent, multiple speaker corpus of connected and continuous digit strings designed to support research into audio-visual speech recognition in adverse conditions. The first ten clips from the **groups** partition form a validation set (where required) and the second twelve clips form a held-out test set. Each clip involves two speakers, arranged as in Figure 3; clips begin with the speakers taking turns in reading digit strings and end with both speakers simultaneously reading different digit sequences. We defer the challenge of determining whether one or two people are speaking at once until future work and consider only the parts of each clip in which a single person speaks at any one time. CUAVE has properties making it distinctly non-trivial for localisation work: for example, the left speaker in Figure 3 often mouths parts of the text which is being uttered by the right speaker. In addition, the single speaker portion of each clip is on average 20-25 seconds long. Thus this dataset is larger and more varied than used by other authors: the largest comparable test set of which we are aware comprises eight 2-2.5 second utterances of “How’s the weather in Taipei?” from different speakers [7].



**Fig. 3.** Example two-person clip from CUAVE

### 3.2 Pixel-wise Gaussian Mutual Information

In these experiments, we will (by nature of the dataset) always have perfect a-priori speech detection. However, in contrast to the artificial dataset experiments earlier, we will not always assume a-priori face detection. This tests the hypothesis that face detection may not be an essential prerequisite for successful speaker localisation. However, we will always present comparative results for the case where we have *perfect* a-priori face detection to see whether performance improves. Our basic approach will be to calculate Gaussian MI between individual pixels and the audio. Specifically, we replace  $v_t$  in the earlier discussion by  $v_{txy}$ , a value related to pixel  $x, y$  in the frame at time  $t$ , and assume Gaussian forms for  $p(A)$ ,  $p(V)$  and  $p(A, V)$  where  $V$  now generates a pixel value  $v_{txy}$ . Thus, we will now obtain a mutual information value  $\mathcal{I}(A; V)$  for each pixel. The full set of per-pixel mutual information values estimated for a test clip can be thought of as a *Mutual Information Image*: Figure 4 shows examples, where lighter pixels indicate higher mutual information values. We localise the speaker by searching the Mutual Information Image for compact regions having high mutual information with the audio<sup>3</sup>. This pixel-wise Gaussian MI implementation is similar to [9], but differs in choice of  $a_t$  and  $v_{txy}$  as is now described.



**Fig. 4.** Mutual Information Images: (a) Pixel Intensities (b) Pixel Intensity Changes

Preliminary experiments showed results improve when choosing  $a_t$  to be a mel-frequency cepstral coefficient vector rather than audio energy. Results also improve by choosing  $v_{txy}$  to be related to grey-scale pixel intensity *changes* rather than the grey-scale pixel intensity, as illustrated by Figure 4. Image (a) is a Mutual Information Image calculated when  $v_{txy}$  is the grey scale pixel intensity. High mutual information occurs around both heads rather than being isolated around the speaker’s mouth; this trend is seen in most CUAVE Mutual Information Images when  $v_{txy}$  is pixel intensity. The paper by [2] suggests this problem is reduced by defining  $v_{txy}$  to be related to pixel intensity *changes* and our experimental results

<sup>3</sup> We find computation of a full mutual information image of dimensionality matching the original image(s) is not essential in practice; a naive subsampling of the original image by factors of 10 or more prior to forming the Mutual Information Images does not significantly degrade results for either of the tasks discussed later in this section.

confirm this finding. Specifically, [2] defines  $v_{txy}$  to be the *pixel intensity change* value  $F^{IC}(x, y, t)$  defined as

$$F^{IC}(x, y, t) = \sum_{l,m=-1}^1 F(x+l, y+m, t+1) - F(x+l, y+m, t-1)$$

where  $F(x, y, t)$  is the original image pixel  $(x, y)$  value in the frame at time  $t$ . We refer to the image at a fixed time  $t$  whose  $(x, y)$  pixel values correspond to  $F^{IC}(x, y, t)$  values as the *Intensity Change Image*  $F_t^{IC}$ . Figure 4 (b) shows the Mutual Information Image analogous to (a) but calculated using  $v_{txy} = F^{IC}(x, y, t)$ ; mutual information is now highest around speaker’s mouth and jawline, as we would like.

### 3.3 Detecting the Active Speaker

Our first experiment considers whether Pixel-wise Gaussian MI can be used to detect the active speaker (ie. left or right) at one second intervals throughout the test clips. Chance performance is 50%. To solve this problem, a window of two seconds length is used and a Mutual Information Image calculated; the estimate of active speaker is obtained by considering total mutual information in the left of the image relative to the total in the right half. The higher of these values is assumed to indicate the active speaker. The window is then shifted by one second and the same procedure repeated. (Preliminary experiments investigated the effects of different window lengths and window shifts upon performance but no significant variation was noted.) Estimates are scored at one second intervals through each clip, a total of 252 test points. Table 2(a) shows results for two cases: where  $v_{txy}$  is the pixel intensity and where  $v_{txy}$  is the pixel intensity change. Performance for both schemes is significantly above chance; the higher active speaker detection for the latter case is not unexpected given the discussion in the previous subsection and experiments in the remainder of the paper use only pixel intensity changes. Further analysis shows one third of errors in the “pixel intensity change” case occur close to speaker turn points eg. when the left speaker stops speaking and the right speaker starts. This is not surprising: estimates at those points use pixel-wise mutual information estimated across a window spanning some data from an “active” left speaker and some from an “active” right speaker. One possible solution is to detect speaker turn points using an audio-based technique (eg. [3]) and adjust estimates in these regions. As baselines, we compare against two simple video-only techniques which make an estimate of the active speaker at time  $t$  based on the Intensity Change Image  $F_t^{IC}$ . In the first scheme (*Intensity Change Image Sums*), we simply compare the total pixel intensity changes on the left and right halves of image  $F_t^{IC}$ ; this gives performance 77%. In the second scheme (*Intensity Change Image X-Projection Peak*), we sum the intensity changes in each column of  $F_t^{IC}$  and use the column with the maximum sum to identify the active speaker; this gives a performance of 81%. We conclude that for the simple task of determining the active speaker, the video-only X-Projection Peak technique is adequate and there is little benefit from using the more computationally expensive (and non-causal) Pixel-wise Gaussian MI. However, there are many assumptions made in this experiment, including a known number of speakers in known regions on screen and a lack of background motion; it is not clear that video-only performance would be maintained when these conditions do not hold, whereas informal experiments

CLIP	Pixel Intensity $v_{txy}$	Intensity Change $v_{txy}$
G11	50	63
G12	32	64
G13	44	50
G14	82	91
G15	50	75
G16	52	85
G17	47	94
G18	55	64
G19	53	47
G20	78	93
G21	67	83
G22	64	95
ALL	57	76

CLIP	$T = 100$	$T = 200$
G11	44	69
G12	46	68
G13	19	25
G14	86	91
G15	65	70
G16	57	57
G17	94	94
G18	65	71
G19	41	41
G20	89	93
G21	75	79
G22	74	79
ALL	65	71

(a) Active Speaker Detection

(b) Active Speaker Mouth Localisation

**Table 2.** Results (% Accuracy)

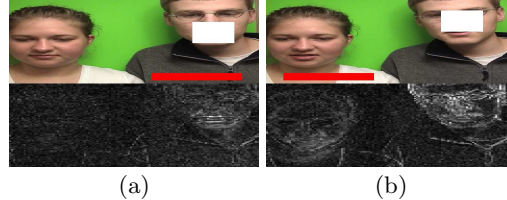
on other data sets indicate that Pixel-wise Gaussian MI maintains good performance in such situations. For more challenging tasks, such as active speaker mouth localisation in the next section, Pixel-wise Gaussian MI gives clear benefit.

None of these results use a-priori face detection. We repeat Pixel-wise Gaussian MI and Intensity Change Image Sums experiments, this time assuming perfect head detection. This time the techniques compare average (rather than total) mutual information or intensity change within the head regions. Results change by less than 1.5% for each technique. We conclude a-priori face detection is useful but not essential for speaker localisation on data such as CUAVE which does not have high background motion.

### 3.4 Detecting Active Speaker’s Mouth

A second set of experiments investigated a more challenging task: to locate the mouth region of the active speaker during the test clips. We begin by computing Mutual Information Images at each test point as above; then, we locate the active speaker’s mouth within each Mutual Information Image by searching for the  $M \times N$  region with the highest concentration of mutual information values within some fraction  $f$  of the maximum mutual information value in the full Mutual Information Image<sup>4</sup>. Parameters  $M$ ,  $N$ ,  $f$  are tuned using the validation set. Figure 5 shows two images from a demo illustrating results on held-out data: for each image, the top half shows the original video, the bottom is the Mutual Information Image, the wide but narrow (red) rectangle is placed under the true speaker and the small (white) square indicates the best  $M \times N$  region found by the algorithm. As we would hope, the optimal  $M, N$  previously found on the validation set correspond to a region about the size of a speaker’s mouth. To quantify results, an estimate of the mouth region centre is defined as “correct” if it falls within a  $T \times T$  pixel square centred on the “true” mouth centre, for  $T$  of 100 and 200. Figure 6 illustrates “correct” regions for each  $T$  using white squares in the upper images. Estimates are scored at one second intervals throughout each clip, a total of 252 test points. Table 2(b) shows results.

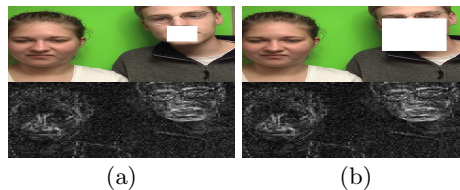
<sup>4</sup> Smoothing of mouth estimates between frames is not used in these experiments, since our test points are separated in time, but is an obvious direction for future work.



**Fig. 5.** Speaker Localisation Examples: (a) Successful (b) Unsuccessful

As baselines, we again compare against two video-only techniques which make an estimate of the active speaker’s mouth at time  $t$  based on the Intensity Change Image  $F_t^{IC}$ . In the first scheme (*High Intensity Change Region*), we locate the active speaker’s mouth at time  $t$  by searching the intensity change image  $F_t^{IC}$  for the  $M \times N$  region with the highest concentration of intensity change values within some fraction  $f$  of the maximum; this gives performance 50% at  $T = 100$  and 52% at  $T = 200$ . In the second scheme (*Intensity Change Image X- and Y-Projection Peaks*), we sum the intensity changes in each row and column of  $F_t^{IC}$  and use the row and column with maximum sums to locate the mouth; this gives a performance of 49% at  $T = 100$  and 51% at  $T = 200$ . For this task, we conclude that Pixel-wise Gaussian MI performs significantly better. This result is plausible: the X- and Y-Projection scheme has less information available than the High Intensity Change Region scheme and the High Intensity Change Region scheme has less information available than Pixel-wise Gaussian MI, which uses a longer temporal window and incorporates audio information.

None of these results use a-priori face detection. We repeat the Pixel-wise Gaussian MI and High Intensity Change Region experiments, now assuming existence of perfect head detection. We constrain the techniques to search only within the head region. At  $T = 100$ , High Intensity Change Region improves to 54% (a gain of 2%) and Pixel-wise Gaussian MI improves to 69% (a gain of 4%). We conclude that for best speaker localisation performance, good a-priori face and speech detection is essential. However, it is interesting that Pixel-wise Gaussian MI alone gives reasonable speaker localisation performance on data such as CUAVE which has no background motion.



**Fig. 6.** Speaker Localisation Correctness Regions: (a)  $T=100$  (b)  $T=200$

## 4 Conclusions

This paper presented an empirical study of audio-visual synchrony measures and their application to speaker localisation in video. The artificial dataset experiments support two conclusions. Firstly, Gaussian MI outperforms the other synchrony

measures proposed and potentially solves specific practical problems in video analysis, such as identifying the active speaker from a set of faces on screen or distinguishing monologues from dialogues when it is known one of the two occurs in the shot. Secondly, Gaussian MI is not suitable for making absolute decisions about degree of synchrony, such as distinguishing voiceovers from monologues. Experiments on CUAVE support two further conclusions. Pixel-wise Gaussian MI gives performance close to two video-only techniques on the task of active speaker localisation; later informal experiments suggest it scales better to other tasks. For active speaker mouth localisation, the additional visual context and audio information available to Pixel-wise Gaussian MI leads to performance significantly better than two video-only techniques. Future work will follow three directions: to identify new definitions of synchrony suitable for distinguishing monologues from dialogues, to develop methods for robustness to background video motion and to consider efficiency issues (such as whether a similar approach could be implemented in the MPEG compressed domain).

**Acknowledgements** We thank Gerasimos Potamianos and Miroslav Novak for technical assistance and the anonymous reviewers for their comments.

## References

1. W. Adams, G. Iyengar, C.-Y. Lin, M. R. Naphade, C. Neti, H. J. Nock, and J. R. Smith. Semantic Indexing of Multimedia Content Using Visual, Audio and Text Cues. *Eurasip Journal on Applied Signal Processing*, 2:170–185, 2003.
2. T. Butz and J.-P. Thiran. Feature Space Mutual Information In Speech-Video Sequences. In *Proc. ICME*, Lausanne, Switzerland, 2002.
3. S. Chen and P. Gopalakrishnan. Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion. In *Proc. DARPA Broadcast News Transcription & Understanding Workshop*, VA, USA, 1998.
4. J. Connell, N. Haas, E. Marcheret, C. Neti, G. Potamianos, and S. Velipasalar. A Real-Time Prototype for Small-Vocabulary Audio-Visual ASR. In *ICME (Submitted)*, 2003.
5. T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.
6. R. Cutler and L. Davis. Look Who’s Talking: Speaker Detection using Video and Audio Correlation. In *Proc. ICME*, NY, USA, 2000.
7. J. W. Fisher III and T. Darrell. Informative Subspaces for Audiovisual Processing: High-Level Function from Low-Level Fusion. In *Proc. ICASSP*, 2002.
8. R. Gopinath. Maximum Likelihood Modeling with Gaussian Distributions for Classification. In *Proc. ICASSP*, volume 2, pages 661–664, WA, USA, 1998.
9. J. Hershey and J. Movellan. Using Audio-Visual Synchrony to Locate Sounds. In *Proc. NIPS*, 1999.
10. G. Iyengar, H. Nock, and C. Neti. Audio-Visual Synchrony for Detection of Monologues in Video Archives. In *Proc. ICASSP*, Hong Kong, 2003.
11. H. Nock, G. Iyengar, and C. Neti. Assessing Face and Speech Consistency for Monologue Detection in Video. In *Proc. ACM Multimedia*, Juan-les-Pins, France, 2002.
12. E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy. Moving Talker, Speaker-Independent Feature Study and Baseline Results Using the CUAVE Multimodal Speech Corpus. *Eurasip Journal on Applied Signal Processing*, 11:1189–1201, 2002.
13. G. Potamianos, J. Luettin, and C. Neti. Hierarchical Discriminant Features for Audio-Visual LVCSR. In *Proc. ICASSP*, pages 165–168, 2001.
14. M. Slaney and M. Covell. FaceSync: a linear operator for measuring synchronization of video facial images and audio tracks. In *Proc. NIPS*, 2001.