MULTIMODAL SPEAKER LOCALIZATION IN A PROBABILISTIC FRAMEWORK

Mihai Gurban and Jean-Philippe Thiran

Signal Processing Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
email: {mihai.gurban, jp.thiran}@epfl.ch
web: itswww.epfl.ch

ABSTRACT

A multimodal probabilistic framework is proposed for the problem of finding the active speaker in a video sequence. We localize the current speaker's mouth in the image by using the video and the audio channels together. We propose a novel visual feature that is well-suited for the analysis of the movement of the mouth. After estimating the joint probability density of the audio and visual features, we can find the most probable location of the current speaker's mouth in a sequence of images. The proposed method is tested on the CUAVE audio-visual database, yielding improved results, compared to other approaches from the literature.

1. INTRODUCTION

Speech production is multimodal in nature and, when communicating through speech, humans can augment the audio with visual information. The McGurk effect [1] is a good example: seeing a mouth that utters something different from what is heard can change the perception of the sound itself.

This multimodality can be exploited in different ways. For example, lip reading can be used to improve the quality of speech recognition, leading to audio-visual speech recognition [2]. Our purpose in this paper is to determine who is speaking, based on the audio-visual sequences of groups of speakers. To this end, we introduce a novel idea on how to extract visual features that are better suited to represent speech and, at the same time, more noise-tolerant. In the end, we will be able to draw some conclusions about the nature of the correlation between audio and video in the case of speech.

Knowing who is speaking is important for example in the case of a smart conference room. An automatic system with several cameras could switch views or change the focus depending on the speaker.

Several approaches to audio-visual speaker localization have been presented in the literature. Hershey and Movellan [3] use an estimate of the mutual information between the average acoustic energy and the pixel value, whose joint probability density function (pdf) they assume to be gaussian. Slaney and Covell [4] use Canonical Correlation Analysis to find a linear mapping which maximizes the audio-visual correlation on training data. They apply the same mapping on test data and measure the audio-visual correlation in the transformed space, obtaining a quantitative measure of audio-visual synchrony. This approach implicitly makes the same assumption that the joint pdf of audio and visual information is gaussian.

Audio-visual synchrony is also analyzed by Nock et al [5, 6]. The mutual information between the audio and the video is computed using two methods, one based on histograms to estimate the pdf, the other based on multivariate gaussians. For the second measure, they assume that the audio-visual data is gaussian locally, on short temporal windows.

Fisher et al. [7] use a nonparametric statistical approach to learn maximally informative joint subspaces for multimodal signals. Their method uses no prior model and no training data. In [8], the method is further developed, showing how the audio-visual association problem, formulated as a hypothesis test, can be related to mutual information-based methods.

Butz and Thiran [9, 10] propose an information theoretic framework for the analysis of multimodal signals. They extract an optimized audio feature as the maximum entropy linear combination of power spectrum coefficients. They show that the image region where the intensity change has the highest mutual information with the audio feature is the speaker's mouth. Besson et al. [11] use the same framework to detect the active speaker among several candidates. The measure that they maximize is the efficiency coefficient, i.e. the ratio between the audio-visual mutual information and the joint entropy. They use optical flow components as visual features, extracting them from candidate regions identified using a face tracker.

The disadvantage of methods that attempt to maximize an information theoretic measure *at test time* is that they need to use some time-consuming optimization procedure, such as gradient descent or a genetic algorithm. This means that, although these methods do not require a training procedure, the amount of computation that is needed during testing is important, making a real-time implementation unfeasible.

By contrast, our multimodal approach does use a training procedure. The joint pdf of the audio energy and a visual feature based on optical flow is learned. This ensures that the number of operations performed while testing is reduced, and thus a real-time implementation would be possible.

Another advantage of our approach is that, in contrast to methods that consider the audio and video of speech to have a gaussian joint pdf, we can model any kind of probability density. The gaussian mixture model that we use is an universal approximator of densities, even when using only diagonal covariance matrices, provided that enough gaussians are considered.

Moreover, in our case, no face tracker needs to be used, as testing is done on the entire image, not only the face or mouth region. An extracted mouth region is required, but only in the training step, when the joint pdf is estimated.

Finally, although the optical flow has been used before for speaker localization, our visual feature, which is the difference between vertical components of the optical flow, is novel. We argue that this feature is better at representing the

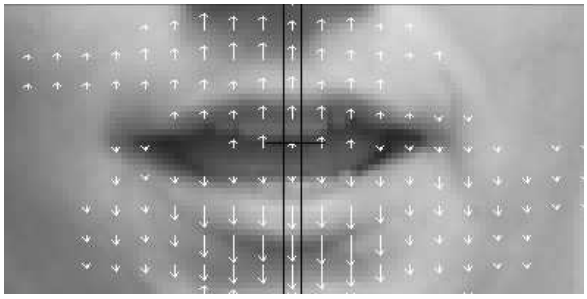

Figure 1: A frame from the training sequences, with the corresponding optical flow.

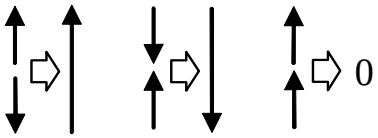

Figure 2: An illustration of the difference of motion vectors.

movement of the mouth, and, at the same time, more tolerant to the motion of the head, compared to simple optical flow, pixels or pixel differences (deltas).

2. ESTIMATING THE JOINT PROBABILITY DENSITY OF THE AUDIO AND VISUAL FEATURES

2.1 Feature extraction

As we want to model the dependency between the audio and the video signals in the case of speech, we need to extract temporally synchronized features from both streams. The audio feature that we use is the logarithm of the energy (log-energy) of the audio signal. From the video, in the training phase, we only use the rectangular region of the mouth. We extract visual features as follows. We compute the optical flow from the luminance component of the images. A single vertical column of points is selected at the center of the mouth region, and only the vertical components of the motion field are retained, as shown in figure 1. Our visual feature is the difference between the average optical flow on the top and bottom halves of this column.

What we observe is that the optical flow difference is closely related to the movement of the mouth. When the mouth is opening, the result is a large positive number, while when it is closing, the result is negative. However, when both vectors point in the same direction, they cancel each other out, as shown in figure 2. The advantage of this approach is that small movements of the head are neutralized. When the head is moving, the upper and lower components of the head motion cancel out, yielding only the mouth movement.

Our visual feature should also be tolerant to some amount of horizontal movement, as optical flow values are very similar on horizontal lines around the center of the mouth. This can be seen in figure 1. This tolerance to both vertical and horizontal displacement means that the extraction of the mouth region, required for training, does not need to be very accurate.

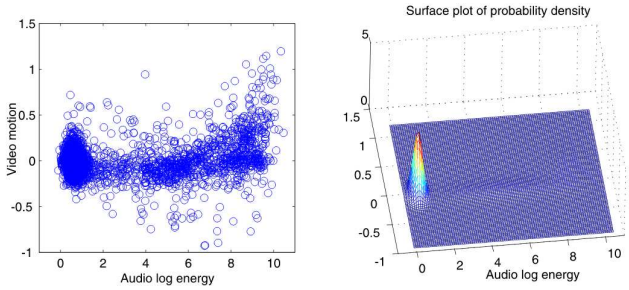

Figure 3: The distribution of audio-visual samples and their estimated pdf.

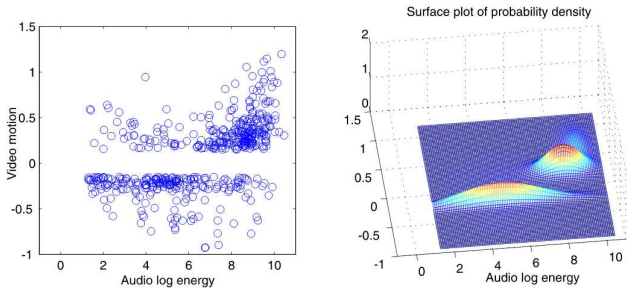

Figure 4: The distribution of audio-visual samples and their estimated pdf, after removing the parts where there is either silence, or very little motion.

2.2 The probability distribution

In order to estimate the joint pdf of features extracted from the training sequences, we need an appropriate model. If $F_v^{train}(t)$ is the visual feature for the training frame t, and $F_a^{train}(t)$ the corresponding audio feature, we want to estimate the probability density function $p(F_a^{train}, F_v^{train})$. Assuming that $p(F_a, F_v)$ is gaussian is too restrictive. Instead, we use a gaussian mixture model (GMM), trained with an Expectation-Maximization (EM) procedure [12]. As mentioned before, the GMM is a universal approximator, i.e. it can be used to represent any type of pdf, provided that enough components are included. Our trained model consists of four gaussians with diagonal covariance matrices, which proved to be a good representation for our data without overfitting it.

The distribution of the audio-visual samples taken from the training sequence, as shown in figure 3, has a high concentration of points around zero audio energy. This is caused by pauses between words. As can be seen, the estimated pdf has a high peak in the same area, while the distribution of the remaining points is poorly modelled.

When searching the correspondence between the sound and the movement of the mouth, the silent samples (low audio energy) do not convey any useful information. Therefore, we removed these samples through thresholding.

In general, image points with low relative movement (low value of the video feature) are characteristic for a static background, even when associated with a high audio energy. However, such points are also present in the training set consisting of mouth regions only. They appear either as a result of errors in the optical flow, or during the pronunciation of long vowels, when the mouth does not move much. As these

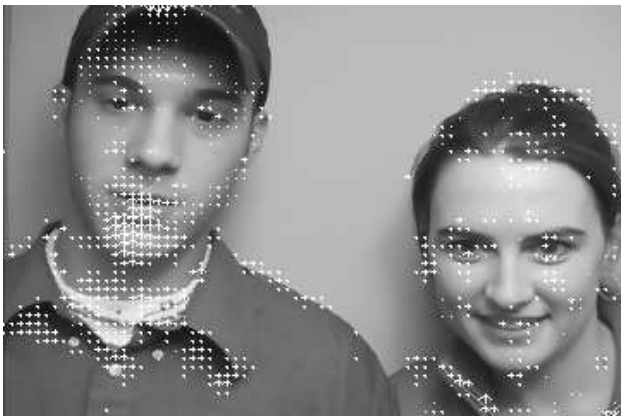

Figure 5: A frame from the test sequence, with the corresponding optical flow

samples can not help determine the location of the speaker, we removed them as well.

Figure 4 shows the distribution of the remaining samples. Their pdf has an interesting property, that is, high audio energy is more often associated to positive values of the visual feature, while lower audio energy is associated to negative values. Since our visual feature is the difference of vertical optical flow vectors, a positive value in the training samples represents the action of opening the mouth, while a negative one represents closing it, as can be seen from figure 2. This confirms the intuition that opening the mouth should lead to louder sounds than closing it.

We can infer from the discrimination, based on the audio, between positive and negative values of the visual feature, that the audio-visual approach can offer more information than video only. This clearly shows the advantages of multimodal analysis. Our method does more than just detecting regions of high relative motion. By associating this motion with a corresponding audio value, our algorithm can find the combination that most likely represents a speaking mouth.

However, the speed of the mouth's movement, as measured with the optical flow, can vary depending on the distance from the speaker to the camera. We normalized the values of the visual feature by scaling them with a factor proportional to the distance between the speaker's eyes. This scaling factor was computed once for each speaker, as the distance to the camera remains constant in our sequences.

3. FINDING THE ACTIVE SPEAKER

Our method of speaker localization is based on a maximum likelihood approach. We find the region of a test image where samples have the highest likelihood to have originated from our learned pdf. Our tests show that this region corresponds quite accurately to the active speaker's mouth.

The testing sequences consist of two speakers side by side, taking turns at speaking. They pronounce series of connected digits. Since we do not model the words themselves, it is not a requirement for testing to have the same vocabulary as the training set, but generally the same set of phonemes.

Our testing procedure is as follows. We compute the optical flow from the luminance of the frames. One such frame with the corresponding optical flow is shown in figure 5. Only vectors larger than 10% of the maximum motion vector

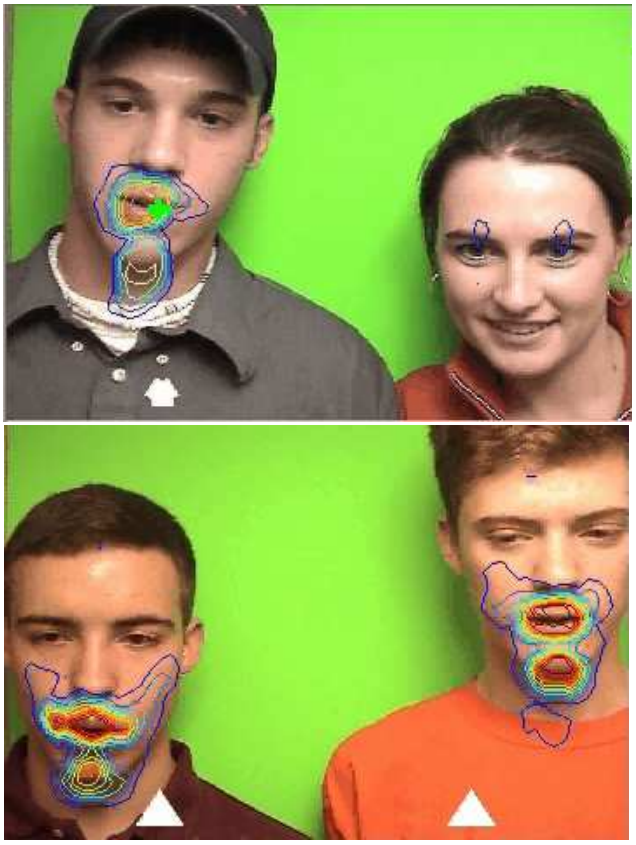

Figure 6: Isocontours of likelihood maps, superimposed on frames from the corresponding temporal windows.

in the image are represented. We used the LTI-Lib computer vision library (`http://ltilib.sourceforge.net`) to compute the optical flow and display it.

From the optical flow, only the vertical components are retained. We compute the value of the visual feature in all points on a grid (with a 10-pixel spacing), using the same method as in training. After selecting columns having the same height as the mouth regions from training, we compute the difference of average vertical optical flow between their top and bottom halves. The reason for using a grid is that the value of the visual feature does not differ much between neighboring points, and we considered the 10-pixel accuracy as sufficient for speaker localization.

For each video frame, the corresponding audio energy, together with the visual feature values on the points of the grid, are used to compute log-likelihoods from the learned joint pdf. If $F_a^{test}(t)$ is the audio feature for the test frame t, and $F_v^{test}(t,x,y)$ is the visual feature value at coordinates (x,y) in the same frame, then the obtained log-likelihood is:

$$l(t,x,y) = \log\left[p\left(F_a^{test}(t), F_v^{test}(t,x,y)\right)\right]$$

where p is the pdf obtained from training.

We sum the log-likelihoods resulting from several consecutive frames at each image coordinate on the grid. We use temporal windows of length W (2 seconds), with a $2W/3$ overlap, as shown in figure 7. The result of the summation is a 2D map, representing the likelihood that the active speaker's mouth is located at a certain coordinate in the image, during the time interval W. The algorithm outputs the

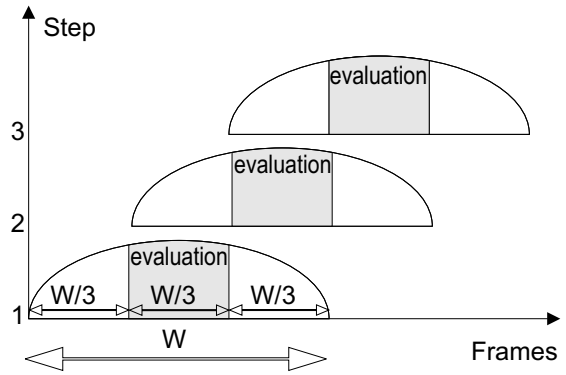

Figure 7: Overlapping temporal windows and the corresponding evaluation intervals. In our case, the length of the window is W = 2 seconds.

location of the detected active speaker as the (x,y) coordinates of the likelihood maximum:

$$L(x,y) = \sum_{t \in W} l(t,x,y)$$

$$(x_{speaker}, y_{speaker}) = \arg\max_{x,y} [L(x,y)]$$

Figure 6 shows the isocontours of such likelihood maps $L(x,y)$, superimposed on frames from the corresponding temporal windows. For the first image, the maximum likelihood point is emphasized by a cross, and, in this case, lies on the speaker's mouth, as expected. In the second image, both the left and the right person are simultaneously speaking, and, as can be seen, the two biggest local maxima of the likelihood function are on the speakers' mouths.

4. RESULTS

For our experiments, we use sequences from the CUAVE audio-visual database [13]. The video sequences are filmed at 30 fps, while the audio is sampled at 44kHz. The video is interlaced, leading to some comb-like artifacts, visible in the areas where there is significant motion. However, we can turn interlacing into an advantage. By unfolding the two separate fields and treating them as individual frames, we obtain video which has a doubled frame rate, 60fps, and half the vertical resolution. As we are interested in the movement of the mouth, having a higher frame rate more than compensates for the loss of vertical resolution, while the interlacing artifacts are eliminated. We compensate for the one-line shift between the fields through interpolation. All video sequences are filtered to remove noise and downsampled to half their resolution to speed up the processing.

Although the sampling rate of the audio is higher than the video frame rate, we need synchronized features. To this end, we compute the audio energy on short temporal windows, so as to obtain one audio feature value for every video frame.

The training sequence that we use belongs to the "individuals" part in the CUAVE database [14]. The speaker utters the English digits from "zero" to "nine" separately, for five times, with pauses between the repetitions. Testing sequences are from the "groups" section of the database. They consist of two speakers taking turns, and finally speaking simultaneously for a short time at the end. We ignored this final part of each sequence in testing.

Seq. no.	Including silence detection (%)	Only speaker localization (%)		Nock et al. [6] (%)
1	90	97		-
2	84	89		-
3	80	86		-
4	82	97		-
5	88	88		-
6	93	93		-
7	82	89		-
8	77	85		-
9	89	92		-
10	76	84		-
11	96	96		63
12	83	90		64
13	89	93		50
14	89	100		91
15	90	97		75
16	97	97		85
17	93	93		94
18	79	83		64
19	88	88		47
20	88	95		93
21	91	94		83
22	100	100		95
Avg.	87.4	92.1		75.3

Table 1: Speaker localization accuracy on the "groups" sequences of the CUAVE database, both with and without silence detection. Results from Nock et al. [6] are also included.

For quantitative results, we use the frame-level ground truth established by Besson et al. [15] for the "groups" sequences. They assign to each frame one of these three labels: *silence*, *left speaker* or *right speaker*. This is the ground truth used to obtain one set of results, for which we also detect silence in the audio.

A second type of ground truth is derived with the purpose of showing the performance of the speaker localization algorithm itself, without the silence detection. To obtain this different ground truth, we split every silent period marked in the old one, labelling each half with the nearest speaker label. With this second ground truth, we obtain a second set of results.

Although our method can quite accurately detect the position of the mouth, we only distinguish between the left and right speaker in our quantitative test. We base our choice on the horizontal position of the likelihood maximum. If it lies in the left half of the image, then we consider that the left speaker is active, and vice-versa.

We compare the detected speaker with the frame-level ground truth. This evaluation is done on the central part of the temporal window, as shown in figure 7. At the same time, the audio energy in the evaluation window is compared to the silence threshold used in training. If the majority of samples in the window are silent, we label it as silent. Otherwise, the label given is the detected active speaker. The detected label is compared to the one that forms the majority in the ground truth, within the temporal window. The first set of results presented in table 1 is obtained with this method. For the second set of results, the silence detection step is skipped,

and the second type of ground truth is used.

The 22 "groups" sequences from the CUAVE database with the superimposed likelihood isocontours can be downloaded from the author's webpage, at `http://itswww.epfl.ch/~gurban/eusipco06/`

The results obtained by Nock et al. [6] for multimodal speaker localization, using a gaussian mutual information measure on the same sequences, are also included in table 1. They make no attempt to detect silence, so their results should be compared to our second set, based on the two-label ground truth. The average performance they obtain is 76%. This result is lower than the 81% reported in the same paper using a visual-only method, so the multimodality did not improve performance in their case.

In contrast, our multimodal method did increase performance. Taking only the last 11 sequences into account, we obtain an average accuracy of 93.7%. This is better than the 81% reported for visual-only speaker localization, confirming that we are able to profit from the extra information present in the audio.

5. DISCUSSION AND FUTURE WORK

Our method leads to improved results compared to other approaches from the literature. Moreover, we are able to exploit the multimodality of speech, obtaining a better performance than that reported for a visual-only method. The training procedure makes the number of operations required for testing small, leading to a fast implementation.

Our novel visual feature is well-suited to represent the movement of the mouth, and is tolerant to some degree to both horizontal and vertical movement of the head. As testing is done on the entire image, there is no need for a face tracker.

The good results that we obtained on the CUAVE database could be justified by the small amount of background movement present in the sequences. On sequences with more movement in the background our method may perform worse, depending on the movement's amplitude and direction. To influence the results, there should be a motion difference oriented vertically and correlated in sign and amplitude with the audio energy, according to the joint pdf. However other methods of speaker localization, such as the ones using pixel differences, would be influenced by any type of movement, be it horizontal or vertical, and of any amplitude.

As future work, further improvement of both the visual and the audio features would make detection more reliable. A different method to extract the motion from the video could be used, as for example block matching. This might reduce the number of errors that appear because of poor estimation of the optical flow. For the audio, using features more related to speech, such as mel-cepstrum coefficients, would make our method invariant to differences in the loudness of the speakers' voices.

Acknowledgements

This work is supported by the Swiss National Science Foundation through the IM2 NCCR. The authors would like to thank Patricia Besson and Gianluca Monaci for fruitful discussions.

REFERENCES

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[2] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: an overview," in *Issues in audio-visual speech processing* (G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, eds.), MIT Press, 2004.

[3] J. Hershey and J. R. Movellan, "Audio vision: Using audio-visual synchrony to locate sounds," *Neural Information Processing Systems*, pp. 813–819, 1999.

[4] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," *Neural Information Processing Systems*, pp. 814–820, 2000.

[5] H. J. Nock, G. Iyengar, and C. Neti, "Assessing face and speech consistency for monologue detection in video," *Proc. ACM Multimedia*, 2002.

[6] H. J. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: An empirical study," *Proceedings of the International Conference on Image and Video Retrieval*, 2003.

[7] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," *Advances in Neural Information Processing Systems*, 2000.

[8] J. W. Fisher III and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Transactions on Multimedia*, vol. 6, no. 3, pp. 406–413, 2004.

[9] T. Butz and J. P. Thiran, "Feature space mutual information in speech-video sequences," *Proceedings of the IEEE International Conference on Multimedia and Expo*, vol. 2, pp. 361–364, 2002.

[10] T. Butz and J. P. Thiran, "From error probability to information theoretic (multi-modal) signal processing," *Signal Processing*, no. 85, pp. 875–902, 2005.

[11] P. Besson, M. Kunt, T. Butz, and J. P. Thiran, "A multimodal approach to extract optimized audio features for speaker detection," *Proceedings of European Signal Processing Conference (EUSIPCO)*, 2005.

[12] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[13] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Moving-talker, speaker-independent feature study and baseline results using the CUAVE multimodal speech corpus," *EURASIP Journal on Applied Signal Processing*, vol. 2002(11), pp. 1189–1201, 2002.

[14] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Moving-talker speaker-independent feature study and baseline results using the cuave multimodal speech corpus," *Journal of Applied Signal Processing*, no. 11, pp. 1189–1201, 2002.

[15] P. Besson, G. Monaci, P. Vandergheynst, and M. Kunt, "Experimental framework for speaker detection on the CUAVE database, EPFL-ITS Tech. Rep. 2006-003," tech. rep., EPFL, Lausanne, Switzerland, 2006.