Joint processing of audio and visual information for multimedia indexing and human-computer interaction

C. Neti, B. Maison, A. Senior, G. Iyengar, P. Decuetos, S. Basu and A. Verma
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598

Abstract

Information fusion in the context of combining multiple streams of data e.g., audio streams and video streams corresponding to the same perceptual process is considered in a somewhat generalized setting. Specifically, we consider the problem of combining visual cues with audio signals for the purpose of improved automatic machine recognition of descriptors e.g., speech recognition/transcription, speaker change detection, speaker identification and speaker event detection. These happen to be important descriptors for multimedia content (video) for efficient search and retrieval. A general framework for considering all of these fusion problems in a unified setting is considered.

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 this paper, we discuss the general problem of fusing these multimodal streams of information to arrive at a coherent decision of human identity and activity. Use of visual information to improve audio-based technologies such as speech recognition, speaker recognition, speech event detection and speaker change detection is a specific example of this endeavor.

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].

- 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

An optimal fusion policy of using some of these fusion strategies remains the holy grail of research [5, 6, 10]. In this paper, we restrict our considerations to audio-visual information fusion [8, 12, 11, 9, 7].

2 Speechreading

The potential for joint processing of audio and visual information for speech recognition is well established on the basis of psychophysical experiments.

Here, in a simpler version of the general fusion problem the set of objects to be recognized can be taken to be the speech utterances. These have different realizations in the acoustic domain and in the visual domain. In the acoustic domain the basic (atomic) symbolic units

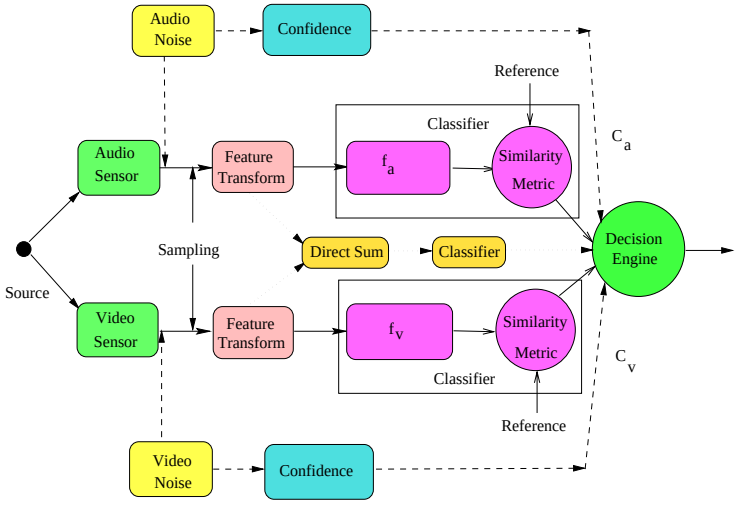

Figure 1: Audio-visual information fusion

associated with the utterances are the *phonemes* that are delineated in linguistics theory, whereas in the visual domain the elemental units are the so called *visemes* borrowed from the psychoacoustic literature. Visemes provide information that complements the phonetic stream from the point of view of confusability. For example, "mi" and "ni" which are confusable acoustically, especially in noise situations, are easy to distinguish visually: in "mi" lips close at onset, whereas in "ni" they do not. The unvoiced fricatives "f" and "s" which are difficult to distinguish acoustically belong to two different viseme groups.

Our focus and interest is in demonstrating meaningful improvements for realistic tasks such as broadcast news transcription for audio/video indexing, large vocabulary dictation and speech reading for the hearing/speech impaired.

To make precise mathematical definitions, we denote by $x_a \in \mathbf{R^m}$ the audio feature vectors and by $x_v \in \mathbf{R^n}$ the video feature vectors.

2.1 Early fusion or feature fusion

Here, the strategy is to combine the two streams of information at an early stage and possibly exploit a single classifier. To be specific, we consider vectors $x = x_a \oplus x_v \in \mathbf{R^{m+n}}$ in the larger space

$$\mathbf{R^{m+n}} = \mathbf{R^m} \oplus \mathbf{R^n}$$

where components of x come from the components of x_a and x_v respectively.

We then define a class of maps

$$f_i : \mathbf{R^{n+m}} \to \mathbf{R}$$

such that $f_i(x)$ becomes a score on the basis of which the symbolic units are detected. See Figure 1 for details (dotted line).

2.2 Late fusion or decision fusion

Here, since the symbolic units are different in the two domains, different classifiers f_a and f_v need to be exploited. Decision fusion then involves combining the results of these classifiers in an intelligent fashion with due regard to the *confidence* that can be attributed the results of the two classifiers. See Figure 1 for details.

The function of the classifiers is to assign numerical scores (e.g, class probabilities) via the class of maps:

$$f_{ai} : \mathbf{R^m} \rightarrow \mathbf{R}$$

$$f_{vi} : \mathbf{R^n} \rightarrow \mathbf{R}$$

and then to combine the outcomes of the classifiers via the fusion maps:

$$F_{C_a,C_v,i} : \mathbf{R} \times \mathbf{R} \rightarrow \mathbf{R}$$

where a fusion map F may depend on the confidence parameters C_a and C_v associated with the audio and video streams of information and is denoted by

$$F_{C_a,C_v,i}(f_a(x_a), f_v(x_v)) \tag{1}$$

Example: An example of this in the case of speech recognition is:

$$F_{C_a,C_v,i}(f_{ai}(x_a), f_{vi}(x_v)) = [f_{ai}(x_a)]^{C_a} \cdot [f_{vi}(x_v)]^{C_v} \tag{2}$$

where C_a and C_v depends on the confidence parameters C_a, C_v and it is conceivable that the constraint

$$C_a + C_v = 1 \tag{3}$$

is adopted for the purpose of normalization. This product separable F_{C_a,C_v} assumes that the two streams of information are independent, especially when $f_a(x_a)$ and $f_v(x_v)$ are interpreted as probabilities of occurrences of the symbolic units associated with the two streams. In practice such an independence assumption could be debated, especially since the two streams are realizations of the same perceptual process synchronously observed in time.

The importance of C_a and C_v in the fusion equation above can be highlighted by the following experiments on the effect of visual noise on the phonetic classification performance.

2.3 Effect of Visual Noise

The face tracking system occasionally fails to track the face in the video sequence. This can be either due to mismatch between training and test conditions of the candidate face is unlike any of the training examples, implying inability of the face model to generalize. In addition, the face tracking can also be poor, where the located face does not align accurately with the actual face in the video stream. In situations when the tracking completely fails, the visual data is represented by *visual silence*. However, in poor tracking, the visual processing results in geometry errors (e.g, nose tip classified as a lip) which gives rise to noise in the visual data. We note here that this noise is different from the signal noise(i.e, noise in video stream, per se).

We designed a supervised classifier to prune the visual noise due to poor tracking. This classifier is a Gaussian mixture model trained on a small subset of PCA projections (typically 20-25 dimensions). We classify the extracted PCA lip projections in a sequence and consider only those sequences that have a high percentage of good lips.

The performance of the lip classifier is presented in Table 1

We note here that in the context of this experiment, we are interested in an estimate of the visual noise. For this purpose, it is adequate to get a lip classification percentage that is close to the true percentage of lips in the data. It is not necessary to consider the false alarm and false reject numbers.

To understand the effect of visual noise we carried out phonetic classification experiments using 5000 sentences spoken by 45 speakers for training and 500 sentences for testing. The

Seq	True Lip%	Classification (%)	
		Lip	Non Lip
Spkr1	100	96.05	3.72
Spkr2	68.9	66.4	33.4
Spkr3	36.5	35.8	63.9

Table 1: Lip classifier results for Test datasets

results suggest that visual noise can have a significant impact on classification performance. For example, the visual phonetic classification performance improves from 11.68% to 22.98% by considering clips with more than 90% good lip images.

3 Speaker Recognition

Here we combine image or video based visual signatures with audio feature based speaker identification for improved person authentication.

3.1 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:

$$\mathcal{S}_{ik} = \frac{\sum_j a(j)a_i(j)}{\sqrt{\sum_j a(j)^2 \sum_j a_i(j)^2}} \tag{4}$$

An average of these similarities,

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

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

3.2 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} as in (5) 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.

$$f_{ai} = \sum_{n=1}^{N} d_{i,n} \tag{5}$$

3.3 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. Recalling (2), we can define the combined score $F^i \equiv F^i_{C_a,C_v}$ as a function of the single parameter α:

$$F^i = C_a f_{vi} + C_v f_{ai} \;\text{ with }\; C_a = \cos\alpha,\;\; C_v = \sin\alpha \tag{6}$$

The angle α has to be selected according to the relative reliability of audio and face identification (note that in (6) a scaling different from (3) is adopted). 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. As per Vapnik theory of empirical errors one can minimize the cost function $C(\alpha)$ given by

$$C(\alpha) = \frac{1}{N}\sum_{n=1}^{N} T_{\hat{\imath}}(n), \;\text{ where }\; \hat{\imath} = \arg\max_i F_i(n) \tag{7}$$

and $F^i(n)$ is as in (6) with $f_{ai} = f_{ai}(n)$ and $f_{vi} = f_{vi}(n)$. For a 77 speaker video broadcast database, with audio-only accuracy of 78% and with video-only accuracy of 64%, a fused accuracy of 84.4% was obtained [1].

4 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. 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. In such a scenario, gathering the joint audio-visual statistics and leveraging this to generate more accurate audio-segmentations (which in turn is desirable for accurate speech transcription and retrieval) seems to be of interest.

A likelihood criterion penalized by the model complexity, namely the BIC criterion has been used. Let $\mathcal{X} = \{x_{ai} : i = 1, \cdots, N\}$ be the audio feature vectors for which we are seeking a statistical model. Let $\mathcal{M}$ be the class of candidate models, $L(\mathcal{X}, \mathcal{M})$ be the likelihood function for the model $M \in \mathcal{M}$, and $\#(M)$ be the number of parameters in the model M. For an empirically chosen weight λ, the BIC procedure maximizes

$$BIC(M) = \log L(\mathcal{X}, M) - 0.5\lambda \times \#(M) \times \log N \tag{8}$$

with respect to M.

4.1 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":

$$f'_a(i) = R(i) - \lambda P \tag{9}$$

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

$$R(i) = N log|\Sigma| - N_1 log|\Sigma_1| - N_2 log|\Sigma_2|, \tag{10}$$

$P = 0.5(d + 0.5d(d+1)) \log N$ 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 $f'_a(i) > 0$.

4.2 Videor-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:

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

between the adjoining vectors x_{vi}^k and $x_{v(i-1)}^k$, 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

$$f'_v(i) = |\bar{g}_v(i) - g_v(i)| - \Theta > 0. \tag{11}$$

4.3 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 (11) 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 (9) 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} as in (1). 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 well known 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 show that at a recall rate of about 67% (percentage of actual speaker changes detected), the precision improves from 95% to 97%.

5 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 [2]. 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. An intelligent method of combining the two methods may be to compute the following two probability densities

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

as, say, mixtures of Gaussian pdfs. A simple fusion strategy (cf. (1)) is to use the linear combination:

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

We are, at present, building a practical system that aims to detect the user's intent to speak to a computer. Our method relies on the premise that when a user is using natural spoken language for information interaction (with information displayed on a desktop display), he faces the computer before he speaks. In such a scenario, the first step is to detect a frontal face as seen through a simple desktop video camera mounted on the monitor. We use a method based on more general techniques for face and facial feature detection on one image to detect frontality of facial pose and infer speech intent. We are currently exploring the second step: which uses a measure of visual speech energy based on mouth activity to combine with a measure of audio energy (based on the cepstral C0 coefficient) to determine speech events more robustly, especially in the presence of background acoustic noise. The whole system is designed to intuitively turn on the microphone for speech recognition without needing to click on a mouse, thus improving the human-like communication between the user and his computer.

6 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 attempted to outline a unified framework for fusion of audio and visual information by focusing on the problems of speech recognition, speaker recognition, speaker change detection and speech event detection.

References

[1] B. Maison, C. Neti and A. Senior, IEEE MMSP Workshop, 1999.

[2] P. Decuetos, C. Neti and A. Senior, IEEE Int. Conf. on Acoustics Speech and Signal Processing., 2000.

[3] S. Basu, C. Neti, N. Rajput, A. Senior, L. Subramaniam, A. Verma, IEEE MMSP Workshop, 1999.

[4] A. Verma, T. Faruquie, A. Senior, C. Neti and S. Basu, Automatic Speech Reco. & Understanding Workshop, 1999.

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

[6] E. Mandler and J. Schurman, Pattern Recognition & Artificial Intelligence, E. S. Gelsema and L. N. Kanal (ed.), Elsevier Science Publishers, 1988.

[7] Javier R. Movellan & Paul Mineiro, UC SanDiego, CogSci Tech. Rep. no. 97-01.

[8] Gerasimos Potamianos and Hans Peter Graff, Proc. ICASSP, pp.3733-3736, 1998.

[9] Patrick Verlinde and Gerard Chollet, Proc. of AVSP, 1999

[10] Josef Kittler, Mohamed Hatef, Robert Duin and Jiri Matas, IEEE Trans. on PAMI, vol.20, n0.3, March 1998.

[11] S. Ben-Yacoub, Y. Abdeljaoued and E. Mayoraz, IDIAP Research Report 99-03.

[12] P. Teissier, J. Robert-Ribes, J-L. Schwartz and A. Guérin-Dugué, IEEE Trans. SAP, vol.7, no. 6, pp. 629-642.