ELSEVIER

Information Sciences 123 (2000) 127–148

INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL

www.elsevier.com/locate/ins

AVIS: a connectionist-based framework for integrated auditory and visual information processing

Nikola Kasabov [a,*], Eric Postma [b], Jaap van den Herik [b]

[a] *Department of Computer and Information Science, University of Otago, P.O. Box 56, Dunedin, New Zealand*

[b] *Department of Computer Science, Universiteit Maastricht, St Jacobstraat 6, Maastricht, The Netherlands*

Received 5 November 1997; accepted 15 April 1999

Abstract

This paper presents a general framework that facilitates the exploration of a single information-processing system in which auditory and visual information are integrated. The framework allows for learning, adaptation, knowledge discovery, and decision making. An application of the framework is a person-identification task in which face and voice recognition are combined in one system. Experiments are performed using visual and auditory *dynamic* features which are synchronously extracted from visual and auditory information flows. The experimental results support the hypothesis that the recognition rate is considerably enhanced by combining visual and auditory dynamic features.

Keywords: Multimodal information processing; Integrated systems; Fuzzy neural networks

* Corresponding author. Fax: +64-3-4798311.

E-mail addresses: nkasabov@otago.ac.nz (N. Kasabov), postma@cs.unimaas.nl (E. Postma), herik@cs.unimaas.nl (J. van den Herik).

0020-0255/00/$ - see front matter
PII: S0020-0255(99)00114-0

1. Multimodal information processing

Information from different modalities can support the performance of a computer system originally designed for a task with a unimodal nature. So, a system for speech recognition may benefit from an additional visual information stream. For instance, visual information from the lips and the eyes of a speaker improves the spoken-language recognition rate of a speech-recognition system substantially [3,7,15,16,21,23]. The improvement per se is already obvious from the use of two sources of information (i.e., sound and images). However, an integration of the two information streams into a multimodal information system may be even more effective.

Research on multimodal speech-recognition systems started a few years ago and has shown promising results. A notable example is the successful recognition of words pronounced in a noisy environment, i.e., the 'cocktail party problem' (also known as the 'source separation problem') [2,8,24]. The additional visual information can be used for solving important problems in the area of spoken-language recognition, such as the segmentation of words from continuous speech and the adaptation to new speakers or to new accents.

Conversely, auditory information and textual input (possibly synchronised with the auditory signal) can be used to enhance the recognition of images. For instance, a challenging task in this area is the identification of moving objects from their blurred images and their sounds [4,20,22,25]. Obviously, the auditory information does not have to be speech or sounds within the audible spectrum of human perceivers. It could also be a *signal* characterised by its frequency, time, and intensity (e.g., the echolocation of dolphins). Two questions must be answered. First, how much auditory or textual input information is required in order to support or improve an image-recognition process significantly? Second, how should several flows of information become synchronised? Since we believe that a proper contribution of the distinct information streams leads to better results we aim at the integration of multimodal information.

Integrating auditory and visual information in one system requires consideration of the following four questions:

1. Auditory and visual information processing are both multilevel and hierarchical (ranging from an elementary feature level up to a conceptual level). So, *at which level* and to *what degree* should the two information processes be integrated?
2. How should *time* be represented in an integrated audio-visual information processing system? This problem relates to the synchronisation of two flows of information. There are several possible scales of integration, e.g., milliseconds, seconds, minutes, years, etc.
3. How should *adaptive learning* be accomplished in an integrated audio-visual information-processing system? Within one modality, the system should

adapt each of its modules dependent on the information processing in the other modalities.
4. How should *new* knowledge (e.g., new rules) be acquired about the auditory and the visual inputs from the real world?

This paper describes a general framework for integrating auditory and visual information to answer these questions. The application of the framework is illustrated on a person identification task involving audio-visual inputs.

The outline of the paper is as follows. In Section 2, the connectionist framework for integrated audio-visual information processing (AVIS) is presented. Section 3 describes an experimental system for studying audio-visual person identification (PIAVI). In Section 4 two case studies are performed to assess the benefit of combining auditory and visual information processing in a person-identification task. Section 5 discusses the results of the case studies. Finally, Section 6 answers the four questions posed above and concludes by stating that AVIS forms a suitable framework for studying multimodal information processing.

2. AVIS: a connectionist framework for integrated auditory and visual information processing systems

Below we describe our connectionist framework AVIS, which combines the principles from two preceding unimodal models. One model originates from multilingual adaptive speech processing [10] and the other from image processing using dynamic features [18,19]. The global architecture of AVIS is illustrated in Fig. 1, and consists of three subsystems:
1. an auditory subsystem;
2. a visual subsystem;
3. a higher-level conceptual subsystem.

Each of them is specified below, followed by a description of the modes of operation (See Fig. 2).

2.1. The auditory subsystem

The auditory subsystem consists of five modules. Below we give the main characterisations.

(a) *The auditory pre-processing module* transforms the auditory signal into frequency features, such as mel-scale coefficients. It accounts for time at a low level of synchronisation (i.e., milliseconds). Frequency, time and intensity features are spatially (tonotopically) represented as a sequence of vectors (i.e., a matrix). The functioning of the pre-processing module may be compared to the functioning of the cochlea.

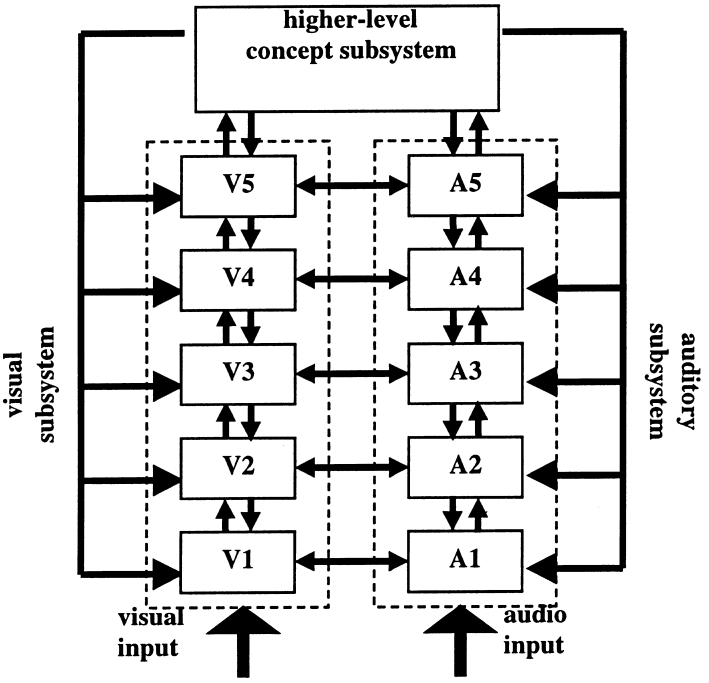

Fig. 1. A block diagram of the framework for auditory visual information processing systems (AVIS).

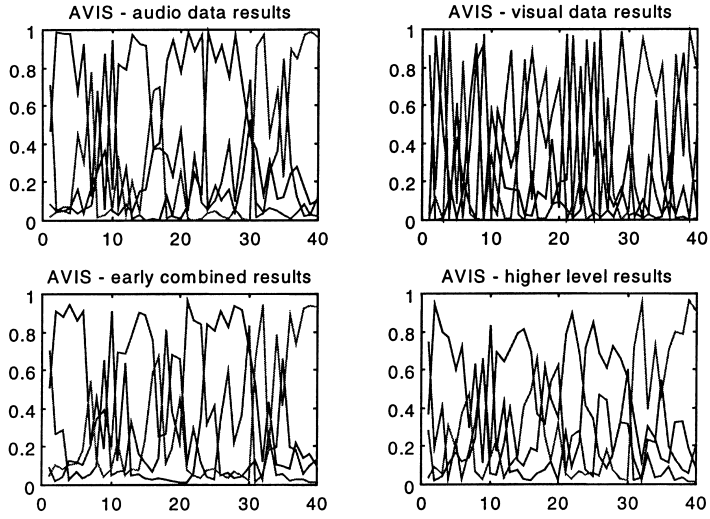

Fig. 2. The results obtained in case study 1. The test frames are shown on the x-axis (first 10 for person one, etc.). The output activation values are shown on the y-axis.

(b) *The elementary-sound recognition module* is a basic building block of the subsystem. It is extendable so that new classes of sounds can be added during operation. A phoneme is adequately represented by a population activity pattern, i.e., an activity pattern distributed over a cluster of neurons. The position of the cluster centre can change through learning.

(c) *The dynamic-sound recognition module* accounts for the dynamical changes in the auditory information. The auditory cortex of the human brain functions analogously.

(d) *The word-detection module* attempts to identify the words. It uses a dictionary of pre-stored words. In the human brain the auditory detection of words is part of the cortical language areas [13].

(e) *The language-structure detection module* accounts for the order in which words are recognised. It uses linguistic knowledge, language knowledge, and domain knowledge as well as feedback from the higher-level conceptual subsystem.

2.2. The visual subsystem

The visual subsystem also consists of five modules. A characterisation follows below:

(a) *The visual pre-processing module* mimics the functioning of the retina, the retinal network, and the lateral geniculate nucleus (LGN).

(b) *The elementary-feature recognition module* is responsible for the recognition of features such as the curves of lips or the local colour. The peripheral visual areas of the human brain perform a similar task.

(c) *The dynamic-feature recognition module* detects dynamical changes of features in the visual input stream. In the human brain, the processing of visual motion is performed in area V5/MT.

(d) *The object-recognition module* recognises elementary shapes and their parts. This task is performed by the infero-temporal (IT) areas of the human brain.

(e) *The object-configuration recognition module* recognises configurations of objects such as faces. This task is performed by the IT and parietal areas of the human brain.

2.3. The higher-level conceptual subsystem

The higher-level conceptual subsystem takes its inputs from all modules of the lower-level subsystems and activates the clusters of neurons representing concepts (e.g., familiar persons) or meanings. The clusters of neurons are connected to the action part of the system (corresponding to the motor areas of the brain). In a person-identification task, the conceptual subsystem takes

information from all the modules in the auditory and visual subsystems and makes a decision on the identity of the person observed.

2.4. Modes of operation

Our framework AVIS allows the auditory subsystem as well as the visual subsystem to operate as a separate subsystem. Their distinct outputs will then be combined in the higher-level subsystem. In addition, each subsystem in isolation is able to accommodate both unimodal and bimodal input streams. Altogether, AVIS can operate in six main modes of operation.

- The *unimodal auditory mode*: the auditory subsystem processes auditory input only (e.g., spoken-language recognition from speech).
- The *cross-modal auditory mode*: the auditory subsystem processes visual input only (e.g., speech recognition from lip movements).
- The *bimodal auditory mode*: the auditory subsystem processes both visual and auditory inputs (e.g., spoken-language recognition from speech and lip movements).
- The *unimodal visual mode*: the visual subsystem processes visual input only (e.g., face recognition).
- The *cross-modal visual mode*: the visual subsystem processes auditory input only (e.g., an image-recognition system trained on audio-visual inputs recalls images from their associated sounds).
- The *bimodal visual mode*: the visual subsystem processes both visual and auditory inputs (e.g., recognising a speaker by his speech and face).

Furthermore, each of the six modes can be combined with conceptual processing in the conceptual subsystem. There are various strategies for combining multimodal sources of information. We propose the principle of *statistically based specialisation* for taking decisions based on different sources of information (i.e., modalities). In general, the auditory and visual subsystems deal with different parts of a task. For instance, take a person-identification task, then the auditory subsystem is responsible for recognising a person's voice and the visual subsystem for recognising a person's face. Each of the subsystems makes its own contribution to the overall task. The conceptual subsystem weights the contributions of the two subsystems according to their (average) recognition rates. The weights have values that are proportional to the probability of each subsystem to produce a correct classification. For example, if the recognition probability of the visual subsystem is 0.7, and the recognition probability of the auditory subsystem is 0.5, then the weights of the two inputs to the conceptual subsystem are 0.7/1.2 and 0.5/1.2 for the visual and auditory subsystems, respectively. Hence, the conceptual subsystem assigns more weighted 'trust' to the visual subsystem. The principle of statistically based specialisation can be readily implemented in a connectionist way.

3. PIAVI: an experimental system for person identification based on integrated auditory and visual information processing

The AVIS framework can be applied to tasks involving audio-visual data. As an instantiation of the AVIS framework, we have developed the person identification based on auditory and visual information (PIAVI) system which identifies moving persons from dynamic audio-visual information.

3.1. PIAVI's subsystems

The global structure of PIAVI resembles the structure of AVIS. However, in PIAVI the auditory and visual subsystems are treated as single modules rather than as sequences of modules. Each of the subsystems is responsible for a modality-specific subtask of the person-identification task. The visual subsystem processes visual data associated with speech, i.e., lip reading, or facial expressions. The inputs to this subsystem are raw visual signals. These signals are pre-processed, e.g., by normalising or edge-enhancing the input image. Further processing subserves the visual identification of the person's face. The auditory subsystem of PIAVI comprises the processing stages required for recognising a person by speech. The inputs of the subsystem are raw audio signals. These signals are pre-processed, i.e., transformed into frequency features, such as mel-scale coefficients, and further processed to generate an output suitable for identification by speech. The conceptual subsystem takes inputs from the two subsystems and activates concepts. For instance, the activation of the concept 'Bill Clinton' may result from his voice being processed by the auditory subsystem and his face being processed by the visual subsystem.

3.2. PIAVI's modes of operation

PIAVI has four modes of operation, briefly described below:

(a) *The unimodal visual mode* takes visual information as input (e.g., a face), and classifies it. The classification result is passed to the conceptual subsystem for identification.

(b) *The unimodal auditory mode* deals with the task of voice recognition. The classification result is passed to the conceptual subsystem for identification.

(c) *The bimodal* (*or early-integration*) *mode* combines the bimodal and cross-modal modes of AVIS by merging auditory and visual information into a single (multimodal) subsystem for person identification.

(d) *The combined mode* synthesises the results of all three modes (a), (b) and (c). The three classification results are fed into the conceptual subsystem for person identification.

4. Two case studies

In two case studies we examine the bimodal processing of audio-visual information in PIAVI. The first case study consists of a preliminary investigation using a small dataset with the aim of assessing the beneficial effects of integrating auditory and visual information streams at an early locus of processing. The second case study employs a larger dataset to evaluate the relative efficiencies of unimodal and bimodal processing in solving the person-identification task.

4.1. Case study 1 [12]

The first case study aims at evaluating the added value of combining auditory and visual signals in a person-identification task. An additional goal is to assess the complexity of the task of identifying persons from *dynamic* auditory and visual input.

4.1.1. The dataset

Given the goals of the study, the dataset has to fulfil two requirements. First, it should contain multiple persons. Second, the persons contained in the dataset should be audible and visible simultaneously. To meet these two requirements, we downloaded a digital video containing small fragments of four American talk-show hosts, from CNN's web-site. The movie contains visual frames accompanied by an audio track. Each frame lasts approximately 125 ms. During most of the frames, the hosts are both visible and audible. The dataset is created as follows. Twenty suitable frames, i.e., frames containing both visual and auditory information, are selected for each of the four persons (hosts). The visual and auditory features are extracted from these 2.5-s fragments (20 frames).

4.1.2. Feature extraction

Person recognition relies on an integration of auditory and visual data. Although static images may suffice for person recognition [6], in our study we rely on dynamic visual information for two reasons in particular. First, dynamic features avoid recognition on the basis of unreliable properties, such as the accidental colour of the skin or the overall level of lighting. Second, the added value of integrating auditory and visual information at an early level lies in their joint temporal variation.

Our emphasis on dynamical aspects implies that the integration of auditory and visual information requires an extended period of time. The duration required for integration varies depending on the type of audio-visual event. For early integration, a duration of about 100 ms may suffice when short-duration visual events (e.g., the appearance of a light) are to be

coupled to short-duration auditory events (e.g., a sound). However, when dynamical visual events such as face and lip movements are to be coupled to speech, a duration of at least half a second is required. To accommodate early integration, we defined aggregate features encompassing the full duration (i.e., 125 ms) of the video segments for both modalities.

4.1.3. Visual features

The images (i.e., frames of video data) contained in each segment need to be transformed into a representation of the spatio-temporal dynamics of a person's head. It is well known that spatio-temporal features are important for person identification tasks (see, e.g., [14]). Behavioural studies show that facial expressions, probably person-specific, flicker rapidly across the face within a few hundred milliseconds [5]. Since the head moves in several directions during a segment, a means of compensating for these variations in a three-dimensional pose is required. Moreover, the head should be segmented from the background to remove background noise. To fulfil these requirements, we used a straightforward spatial-selection method. A single initial template was defined for each person in the dataset. The size of the template was set at $M \times N$ pixels, with $M = 15$ and $N = 7$. The templates intended to cover the entire head. The content of each template was cross-correlated with the content of the next video frame [19]. The best-matching $M \times N$ part of the next frame served as a starting point for the extraction of visual features and was defined as the new template.

A commonly-used technique for extracting features from images is based on principal-component analysis. For instance, in their lip-reading studies, Luettin et al. [14] employed principal-component analysis on the visual lip-shape data. However, a comparative study of (dynamic) features for speech-reading [8] showed that a 'delta' representation, based on the differences in grey values between successive frames, works better than a representation based on principal-component analysis. For this reason we used the delta representation to generate our visual features.

The visual features were obtained as follows. The absolute values of the changes of subsequent frames yield the elements of a delta image $\bar{\Delta}$, defined as

$$\bar{\Delta}(x,y) = |I(t+1,x,y) - I(t,x,y)| \tag{1}$$

with $I(t,x,y)$, the grey value of the pixel at co-ordinate (x,y) of frame t (t represents the frame number).

4.1.4. Auditory features

The audio signal is transformed into the frequency domain using standard FFT (256 points; sampling rate 11 kHz; one channel, one byte accuracy) in combination with a Hamming window, yielding a sequence of vectors

containing 26 mel-scale coefficients. Each vector represents an audio segment of 11.6 ms, with 50% overlap between the segments. The mel-scale vectors averaged over a duration of approximately 125 ms are represented as 'audio frames' (i.e., vectors containing averaged mel-scale coefficients). By subtracting subsequent frames, a delta representation is obtained that is similar to the visual representation. The auditory features are vectors containing three delta representations obtained at three different time lags.

4.1.5. Modelling the subsystems

The subsystems of PIAVI are modelled using fuzzy neural networks (FuNNs) [9]. Each of the input and output nodes of a FuNN has a semantic meaning. FuNNs are designed to facilitate the use of both data and fuzzy rules in a connectionist framework. They allow for easy adaptation, modification, rule insertion, and rule extraction. The unimodal visual mode of operation is modelled as a FuNN with 105 ($N \times M$) input nodes, 315 input membership functions (3 per input node, i.e., representing the fuzzy representations 'small', 'medium', and 'high'), 5 hidden nodes, 4 output nodes (for the four hosts), and 8 output membership functions (2 per output, i.e., representing the fuzzy classifications 'unlikely', and 'likely'). The unimodal auditory mode of operation is modelled as a FuNN with 78 (3×26) input nodes, 234 input membership functions, 5 hidden nodes, 4 output nodes, and 8 output membership functions. The bimodal, early-integration mode of operation is modelled by a FuNN with the same dimensions except for the input. There are 183 input nodes (105 for visual features plus 78 for auditory features) and 549 membership functions. Finally, in the combined mode of operation the two unimodal modes and the bimodal mode are combined. The higher-level concept subsystem is modelled according to the principle of statistically based specialisation. The criterion for classification is as follows. The output node with the largest activation defines the class assigned to the input pattern.

4.1.6. Experimental procedure

The FuNNs assembling PIAVI are trained by a modified backpropagation algorithm with a learning rate of 0.01 and a momentum of 0.8. Their training periods are 5000, 8000, and 6000 epochs for the unimodal auditory mode, the unimodal visual mode, and the early-integration modes of operation, respectively. The combined mode of operation is simulated by determining the appropriate weights of the three trained FuNNs. To assess the generalisation performance, the dataset is split into a training set and a test set, each containing 10 examples. The FuNNs are trained on the training set. The generalisation performance is defined as the classification performance on the test set.

4.1.7. Results

The results of the simulations are shown in Table 1. The performances achieved in the experiments are not very high. This may be due to (at least) three reasons: the limited number of examples contained in the training set, the limited temporal extent of the aggregated features, and the low quality of the dataset. Nevertheless, the overall recognition rate in the early-integration mode of operation is 12.5% higher than the recognition rate of the unimodal auditory mode, and 17.5% higher than the recognition rate in the unimodal visual mode of operation. The overall recognition rate achieved in the combined mode is 22% higher than the recognition rate in the unimodal visual mode, 17% higher than the recognition rate in the unimodal auditory mode, and 4% higher than the recognition rate in the early-integration mode of operation.

The experimental results do not allow us to reject the main hypothesis of this research, viz. that the AVIS framework and its realisation PIAVI achieve a better performance when auditory and visual information is integrated and processed together.

4.2. Case study 2

The second case study attempts to improve and extend the results obtained in the first case study by employing a larger (more realistic) dataset and by defining aggregate features representing longer temporal intervals.

4.2.1. The dataset

Existing large audio-visual datasets contain segments of persons in a highly-constrained setting. For instance, the TULIPS1 database [17] was created by carefully positioning persons in front of a camera and instructing them to count from one to four. Since we felt uncomfortable with such a unrealistic setting, we created our own dataset. We recorded CNN broadcasts of eight fully-visible and audibly-speaking presenters of sport and news programs (see

Table 1
The generalisation performances of case study 1

	Person 1	Person 2	Person 3	Person 4	% Recognition rate
The auditory subsystem only	7	5	3	9	60
The visual subsystem only	4	6	5	8	57.5
Early-integrated features module	7	8	3	8	67.5
Higher-level conceptual subsystem	7	8	4	9	70

Fig. 3). All recordings were captured into digital format (using a standard PC equipped with an Hauppage WinTV capture card). The digital video files so obtained were edited with a standard video editor. This yielded video segments of 1-s length at $F = 15$ frames per second. Each segment contains $F + 1$ frames. The visual and auditory features were extracted from these segments.

4.2.2. Visual features

As in the first case study, the $F + 1$ images (i.e., the frames of video data) contained in each segment are transformed into a representation of the spatio-temporal dynamics of a person's head. The extraction of visual features in this case study differed in three respects from the extraction in the first study. First, in segmenting the face from the background, a fixed template was used for each person, instead of redefining the template with each new frame. The size of the template is defined as $M \times N$ pixels, with $M = 40$ and $N = 20$. Fig. 4 shows the face template used for the person displayed in Fig. 3. Second, the temporal

Fig. 3. An example of a frame used in the dataset.

Fig. 4. The face template used for video's containing the person shown in Fig. 3.

extent of the aggregate features is extended over a 1-s period to accommodate temporal variations over a longer interval. The aggregate features were obtained as follows. The absolute values of the changes for a 1-s period (i.e., $F+1$ frames) are summed pixel-wise, yielding an average-delta image $\bar{\Delta}$, the elements of which are defined as

$$\bar{\Delta}(x,y) = \frac{1}{F}\sum_{t=1}^{F} |I(t+1,x,y) - I(t,x,y)| \tag{2}$$

with $I(t,x,y)$, the colour value of the pixel at co-ordinate (x,y) of frame t (t represents the frame number). Third, a compressed representation of the delta image is used instead of a representation based on all pixel values. The final aggregate visual features are contained in a vector $\boldsymbol{v}$, the elements of which are the summed row values and the summed column values of the average-delta image. Formally, the elements $v(i)$ of $\boldsymbol{v}$ are defined as

$$v(i) = \frac{1}{N}\sum_{j=1}^{N} \bar{\Delta}(j,i) \quad \text{for } 1 \leqslant i \leqslant M \tag{3a}$$

and

$$v(i) = \frac{1}{M}\sum_{j=1}^{M} \bar{\Delta}(i-M+1,j) \quad \text{for } (M+1) \leqslant i \leqslant (M+N). \tag{3b}$$

4.2.3. Auditory features

The auditory features are extracted according to the procedure described in the first case study, except for the aggregated features that are obtained by averaging over a 1-s interval.

4.2.4. Modelling the subsystems

The unimodal visual mode of operation is modelled as a FuNN with 60 $(=N+M)$ input nodes, 180 input membership functions, 10 hidden nodes, 8 output nodes, and 16 output membership functions. The bimodal mode of operation is modelled using a FuNN with the same dimensions except for the input. There are 86 input nodes and 258 membership functions. The higher-level concept subsystem is not modelled explicitly. It is (partly) contained in the output layers of the FuNNs. The criterion for classification is that the output node with the largest activation defines the class assigned to the input pattern.

4.2.5. Experimental procedure

The experimental procedure used for simulating the two modes of operation was as follows. The unimodal mode of operation was studied by presenting the visual examples to the appropriately-dimensioned FuNN

network (Experiments 1 and 2). In the bimodal mode of operation a FuNN with an extended input was used to accommodate the audio-visual input pattern (Experiment 3). In all experiments the learning rate and the momentum parameters were set to 0.01 and 0.8, respectively.

4.2.6. Experiment 1: unimodal processing

To assess the performance of PIAVI in the unimodal mode of operation the visual subsystem was trained until the optimal generalisation performance was reached (early stopping). The training set contained 385 examples (corresponding to a total of 385 s of video playing time). The test set contained 100 examples. The distribution of the examples in the training and test sets over the eight classes (persons) contained in the dataset is shown in Table 2.

4.2.7. Results

A perfect generalisation performance (100% correct classification on the test set) was obtained after 5000 epochs of training (4380 s on a Pentium 166 MHz computer). Upon completion, the RMS error was 0.27 on the training set, and 0.29 on the test set. PIAVI's performance on the unimodal task was unexpectedly good, even too good for potential improvement in the bimodal mode of operation. Given this result, it is very likely that the aggregate visual features contain dynamical information that is diagnostic for the identity of the person. Moreover, this information is hardly (if at all) affected by changes in the orientations of the heads. Possibly, the metrical information contained in the visual examples acts as a reliable diagnostic feature. Both the eyes and mouth are highly dynamic during speech, yielding large values at two vertical locations as represented in the first M elements of the visual vector (cf. Eq. (2)). The distance between these locations may have been a reliable diagnostic feature for person identification for the FuNN. These considerations are corroborated by the graphs in Fig. 5 displaying the $v(i)$ as a function of i for 10 examples of a single class. The peaks at $i \approx 15$ and $i \approx 30$ correspond to the (dynamics of the) eyes and mouth during a 1-s interval.

Table 2
Distribution of examples over the datasets for case study 2

Person	# Examples in the training set	# Examples in the test set
1	50	10
2	50	10
3	70	10
4	30	10
5	40	10
6	50	10
7	25	10
8	70	30

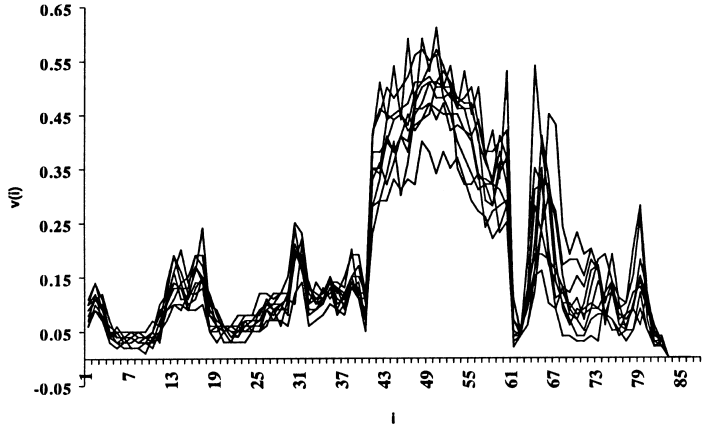

Fig. 5. Graphs showing $v(i)$ as a function of i for 10 examples of one class.

4.2.8. Experiment 2: unimodal processing with a small training set

The setting is as in Experiment 1, but here a smaller training set is used (5 examples per class) and the FuNN is trained for a short time, i.e., 500 epochs. The test set contains 25 examples per class, except for the class corresponding to speaker 1 which contains 5 examples only.

4.2.9. Results

An overall generalisation performance of 75% correct classification was obtained after 500 epochs (90 s simulation time). The RMS errors were 0.34 and 0.81 for the training and test sets, respectively.

4.2.10. Experiment 3: bimodal processing

To assess the performance of PIAVI in the bimodal mode of operation, 40 early-integrated audio-visual examples (five per class) are presented to the FuNN repeatedly, until the performance on the test set is optimal.

4.2.11. Results

An overall generalisation performance of 91% correct classification on the test set was obtained after 1000 epochs (150 s of simulation time). The final RMS errors on the training and test sets were 0.24 and 0.68, respectively. The individual generalisation performances of the bimodal mode of operation are shown in Table 3.

4.2.12. Summary and discussion of results

In the second case study, three experiments were carried out (see the summary of the results in Table 3). In Experiment 1, a perfect generalisation

Table 3
The generalisation performances for the three experiments of case study 2

Experiment	Refer. visual data (s)	Training time (epochs) (s)	FuNN structure	RMSE train/test	Test classification accuracy (%)	Test classification accuracy for each of the classes 1, 2, 3, 4, 5, 6, 7 and 8 (%)
1. Large visual dataset – Long training	385	4380 (5000)	60-180-20-16-8	0.27/0.29	100	100 for each class
2. Small visual dataset – Short training	40 (8×5)	90 (500)	60-180-10-16-8	0.34/0.81	75	100, 28, 92, 72, 60, 48, 100, 96
3. Integrated small visual & auditory dataset – Short training	40 (8×5)	150 (1000)	86-258-10-16-8	0.24/0.68	91	100, 88, 96, 80, 64, 96, 100, 100

performance is obtained when a large training set is used (i.e., visual data corresponding to 385 s) and a FuNN is trained for a long time (4380 s). This type of training is not suitable if fast, on-line training on short, noisy visual reference data is required. Reducing the size of the training set (i.e., corresponding to 40 s) and the amount of training time (i.e., to 90 s) in Experiment 2, led to a decrease in the generalisation performance (75%). Combination of the visual and auditory data in Experiment 3, yielded a major improvement in generalisation performance (e.g., 91% obtained within 150 s of simulation time).

The contribution of the non-linear FuNNs to the results becomes evident by considering the generalisation performances obtained with standard linear statistical methods such as canonical discriminant functions. Fig. 6 displays a plot of all auditory examples mapped on the first two discriminant functions. Fig. 7 shows the same plot for all visual examples. The mapping of the audio-visual examples are displayed in Fig. 8. A subset of these examples is plotted in Fig. 9 as a two-dimensional configuration obtained with a multidimensional scaling procedure. The low-dimensional configuration represents the high-dimensional configuration in audio-visual feature space by preserving the inter-point distances as much as possible. The generalisation performances obtained with the discriminant-functions model (measured using the leaving-one-out

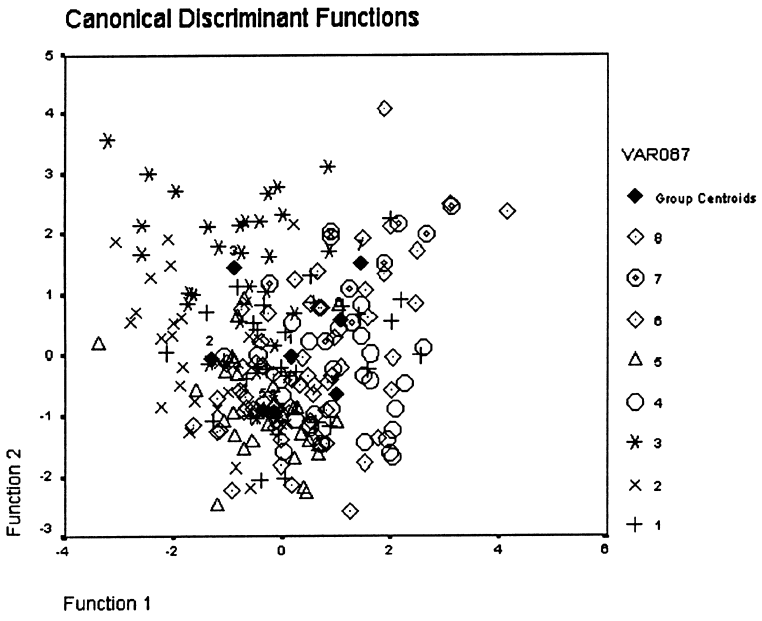

Fig. 6. Mapping of the auditory dataset on the first two discriminant functions.

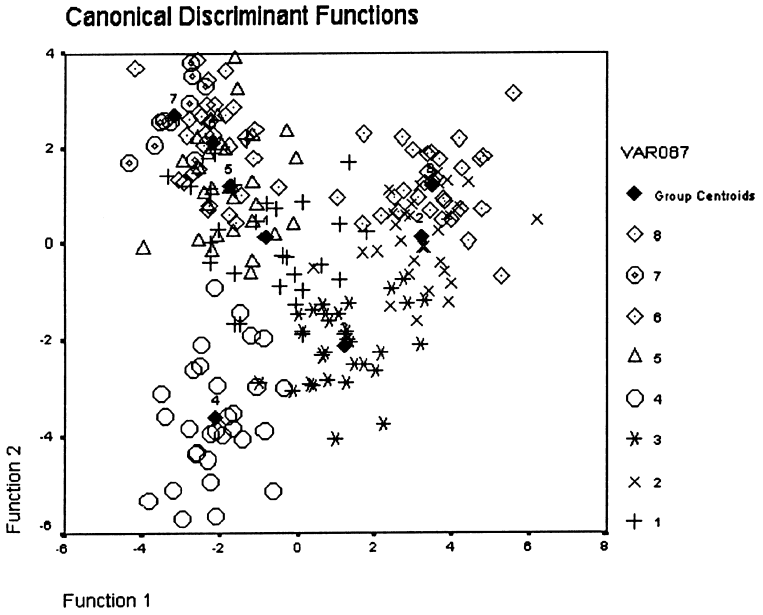

Fig. 7. Mapping of the visual dataset on the first two discriminant functions.

procedure) are 37.3%, 64.1%, and 66.4%, for the auditory, visual, and audio-visual examples, respectively. Evidently, the FuNNs contribute significantly to the generalisation performances obtained in this case study.

5. General discussion

The results of the two case studies prove the added value of integrating auditory and visual information for person identification. In the first case study, which used a small training set, combining the auditory and visual information enhanced the generalisation performance. In the second case study, the first experiment showed that, with a large number of training examples, unimodal processing on the basis of dynamical visual features leads to a perfect performance on a large dataset. This is an interesting phenomenon. Behavioural studies suggest that humans are not very good at identifying persons from their facial dynamics [14]. Nevertheless, the unimodal PIAVI system managed to deal with this task perfectly well. The second and third experiments showed that adding dynamic auditory input to the visual input enhances the identification performance considerably. In these experiments a smaller training set was used. From a practical viewpoint, the use of smaller training sets,

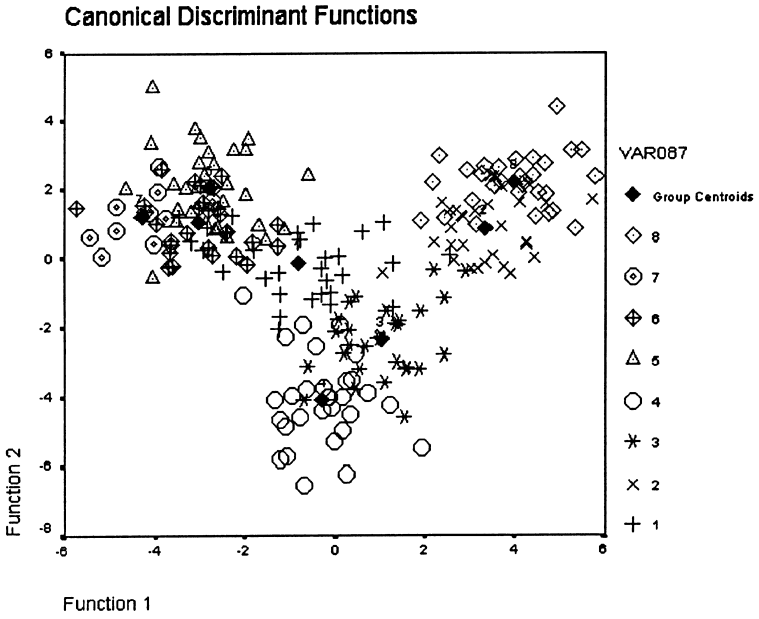

Fig. 8. Mapping of the integrated visual-auditory dataset on the first two discriminant functions.

facilitates the speed at which the PIAVI system in its bimodal mode of operation learns to classify persons from video data. On a standard desktop computer, the training times were 4386, 90, and 150 s, for the Experiments 1, 2, and 3, respectively. The latter two training times make on-line training on video data feasible. In the bimodal case, on-line training yields a satisfactory level of performance (>90%). Therefore, our results show the feasibility of applying the PIAVI system to on-line person identification tasks.

6. Conclusions and directions for further research

We have introduced the AVIS framework for studying the integrated processing of auditory and visual information. The framework facilitates the study of:

- different types of interaction between modules from hierarchically-organised subsystems for auditory and visual information processing;
- early and late integration of the auditory and the visual information flows;
- dynamic auditory and visual features;
- pure connectionist implementations at different levels of information processing, and

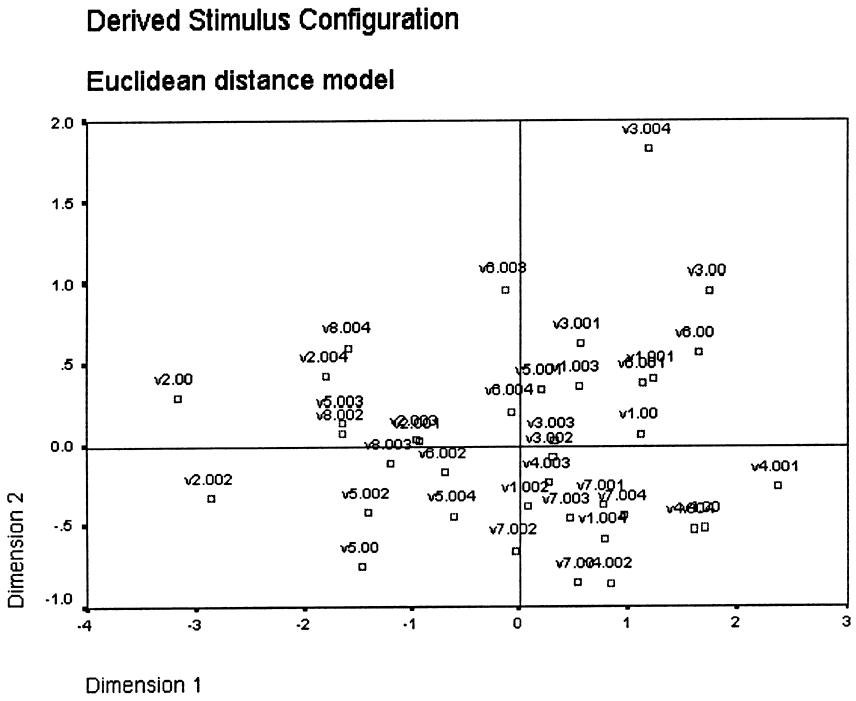

Fig. 9. Multidimensional scaling map of the Euclidean distance in the audio-visual space into two dimensions (vp.00n means the vector of the *n*th example of person *p*; only 5 examples per person have been mapped).

- fuzzy neural networks that allow for learning, adaptation, and rule extraction.

The integrated processing of auditory and visual information may yield:

1. an improved performance on classification tasks involving information from both modalities (cf. case study 1), and
2. reduced recognition latencies on these tasks (see case study two).

The AVIS framework accommodates many applications for solving difficult AI problems. Examples of such problems are: adaptive speech recognition in a noisy environment, face tracking and face recognition, person identification, tracking dynamic (moving) objects, recognising the mood or emotional state of subjects based on their facial expression and their speech, and solving the blind-source separation problem. Through solving these problems, the development of intelligent multimodal information systems is facilitated. Given the results obtained with PIAVI we conclude that AVIS forms a suitable framework for studying multimodal information processing.

For the integration of auditory and visual information we formulated four questions to be answered by the AVIS framework. The first question was:

at which level and to what degree should the auditory and visual information processes be integrated? The AVIS framework accommodates for integration at multiple levels and at various degrees. It seems that early integration works fine, but a further fine-tuning is required to obtain a better insight into the possibilities. The second question was: *how should time be represented in an integrated audio-visual information processing system*? In our two case studies we examined bimodal processing using an aggregate vector representation. with *time* included. This representation was especially effective when using longer time intervals. For shorter time intervals, other ways of representing time should be investigated. The third question was: *how should adaptive learning be realised in an integrated audio-visual information processing system*? In the AVIS framework we solved this question satisfactorily by introducing FuNNs. Other techniques may be possible, but a deeper investigation of FuNNs is tenable and should generate good results. The fourth question was: *how should new knowledge be acquired about the auditory and visual inputs of the real world*? Translating the hidden representations of the FuNNs into rules answers this question initially. Future research should elaborate on questions three and four, and focus on the required new learning techniques for on-line, adaptive learning.

The ECOS [11] technique can be applied to the realisation of the AVIS framework and will be the subject of future experiments. As AVIS is inspired by facts from psychology, more biologically proficient specimens of AVIS are envisaged [1].

References

[1] S. Amari, N. Kasabov (Eds.), Brain-like Computing and Intelligent Information Systems, Springer, Singapore, 1998.
[2] A.J. Bell, T.J. Sejnowski, An information–maximization approach to blind separation and blind deconvolution, Neural Computation 7 (1995) 1129–1159.
[3] C. Bregler, Y. Konig, Eigenlips for robust speech recognition, in: Proceedings of the IEEE Acoustic, Speech and Signal Processing Conference, Adelaide, 1997.
[4] V. Bruce, T. Valentine, When a nod's as good as a wink: The role of dynamic information in facial recognition, in: M.M. Gruneberg, P.E. Morris, R.N. Sykes (Eds.), Practical Aspects of Memory: Current Research and Issues, vol. 1, Wiley, Chichester, UK, 1988.
[5] V. Bruce, P.R. Green, M.A Georgeson, Visual Perception. Physiology, Psychology, and Ecology, third ed., Psychology Press, East Sussex, 1996.
[6] R. Brunelli, D. Falavigna, Person identification using multiple cues, IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (1995) 955–966.
[7] G.A. Calvert, E.T. Bullmore, M.J. Brammer, R. Campbell, S.C.R. Williams, P.K. McGuire, P.W.R. Woodruff, S.D. Iverson, A.S. David, Activation of auditory cortex during silent lipreading, Science 276 (1997) 593–596.
[8] M.S. Gray, J.R. Movellan, T.J. Sejnowski, Dynamic features for visual speechreading: A systematic comparison, in: M.C. Mozer, M.I. Jordan, T. Petsche (Eds.), Advances in Neural

Information Processing Systems, vol. 9, Morgan-Kaufmann, San Fransisco, CA, 1997, pp. 751–757.
[9] N. Kasabov, Foundations of Neural Networks, Fuzzy Systems and Knowledge Engineering, MIT Press, Cambridge, MA, 1996.
[10] N. Kasabov, A Framework for intelligent conscious machines utilising fuzzy neural networks and spatial–temporal maps and a case study of multilingual speech recognition, in: S. Amari, N. Kasabov (Eds.), Brain-like Computing and Intelligent Information Systems, Springer, Singapore, 1998, pp. 105–128.
[11] N. Kasabov, Evolving connectionist and fuzzy connectionist systems – theory and application for adaptive, on-line intelligent systems, in: N. Kasabov, R. Kozma (Eds), Neuro-fuzzy Techniques for Intelligent Information Systems, Springer, Berlin, 1999, pp. 11–146.
[12] N. Kasabov, E.O. Postma, H.J. van den Herik, AVIS: a connectionist-based framework for integrated audio and visual information processing, in: T. Yamakawa, G.Matsumoto (Eds.), Methodologies for the Conception, Design and Application of Soft Computing – Proceedings of the Iizuka '98, vol. 1, World Scientific, Japan, 1998, pp. 422–425.
[13] K. Kim, N. Relkin, K.-M. Lee, J. Hirsch, Distinct cortical areas associated with native and second languages, Nature 388 (1997) 171–174.
[14] J. Luettin, N.A. Thacker, S.W. Beet, Active shape models for visual speech feature extraction, in: D. G. Storck, M. E. Hennecke (Eds.), Speechreading by Humans and Machines, Springer, Berlin, 1996, pp. 383–390.
[15] D. Massaro, Perceiving Talking Faces, MIT Press, Cambridge, MA, 1997.
[16] D. Massaro, M. Cohen, Integration of visual and auditory information in speech perception, Journal of Experimental Psychology Human Perception and Performance 9 (1983) 753–771.
[17] J.R. Movellan, Visual speech recognition with stochastic networks, in: G. Tesauro, D.S. Touretzky, T. Leen (Eds.), Advances in Neural Information Processing Systems, vol. 7, MIT Press, Cambridge, MA, 1995, pp. 851–858.
[18] E.O. Postma, H.J. van den Herik, P.T.W. Hudson, Image recognition by brains and machines, in: S. Amari, N. Kasabov (Eds.), Brain-like Computing and Intelligent Information Systems, Springer, Singapore, 1998, pp. 25–47.
[19] E.O. Postma, H.J. van den Herik, P.T.W. Hudson, SCAN: a scalable model of covert attention, Neural Networks 10 (1997) 993–1015.
[20] V.S. Ramachandran, S.M. Anstis, Perception of apparent motion, Scientific American 254 (1986) 102–109.
[21] G.S. Russo, C.J. Bruce, Auditory receptive fields of neurons in frontal cortex of rhesus monkey shift with direction of gaze, Social Neuroscience Abstracts 15 (1989) 1204.
[22] D.L. Sparks, J.M. Groh, The superior colliculus: a window for viewing issues in integrative neuroscience, in: M.S. Gazzaniga (Ed.), The Cognitive Neurosciences, MIT Press, Cambridge, MA, 1995, pp. 565–584.
[23] D. Stork, M. Hennecke (Eds.), Speech Reading by Humans and Machines, Springer, Berlin, 1996.
[24] A. Waibel, M.Vo, P. Duchnovski, S. Manke, Multimodal interfaces, Artifical Intelligence Review, 1995.
[25] Z.Q. Wang, J. BenArie, Conveying visual information with spatial auditory patterns, IEEE Transactions on Speech and Auditory Processing 4 (1996) 446–455.