DIPARTIMENTO DI INFORMATICA
E SISTEMISTICA ANTONIO RUBERTI

SAPIENZA
UNIVERSITÀ DI ROMA

Multimodal speaker recognition in a conversation scenario

Maria Letizia Marchegiani
Fiora Pirri
Matia Pizzoli

Technical Report n. 7, 2009

Multimodal speaker recognition in a conversation scenario

Maria Letizia Marchegiani, Fiora Pirri, Matia Pizzoli
{marchegiani,pirri,pizzoli}@dis.uniroma1.it

June 16, 2009

Abstract

As a step toward the design of a robot that can take part to a conversation we propose a robotic system that, taking advantage of multiple perceptual capabilities, actively follows a conversation among several human subjects. The essential idea of our proposal is that the robot system can dynamically change the focus of its attention according to visual or audio stimuli to track the actual speaker throughout the conversation and infer her identity.

Keywords: multi-modal perception, human robot interaction.

1 Introduction

We propose a framework for real-time, multi-modal speaker recognition combining acoustic and visual features to identify and track people taking part to a conversation. The robot is provided with acoustic and visual perception: a colour camera and a pair of microphones oriented by a pan tilt unit. The robot follows a conversation among a number of people turning its head and focusing its attention on the current speaker, according to visual and audio stimuli. It tracks the participants exploiting a learning phase to settle both visual-audio descriptors and integration parameters. People are identified against a set of individuals whose audio and visual features are suitably structured in a knowledge base, as prior knowledge. The proposed scenario is quite general and people can change position, leave or join the conversation, thus we cannot exploit the current location of the speaker to infer her identity.

(a) (b)

Figure 1: On the left: the robot with a pair of microphones and a camera oriented by a pan-tilt unit following a conversation in the Lab. On the right the concept of the robotic head following the conversation is shown: pan angles are set according to the direction θ of the estimated voice source, relative to the zero pan position (the solid arrow). Besides the speaker (bold in the figure), other subjects are detected in the FOV, spanned by 2α.

Multi-people, multi-modal detection and tracking scenarios have been modelled in the context of *smart rooms*[Pentland, 1995], e-learning, meeting and teleconferencing support, but also in robot-person interface (see [Waibel et al., 2003] for a complete review of technologies for intelligent rooms). Recently, multi-modal features have been used in domestic environments in order to annotate people activities for event retrieval [Desilva et al., 2006], perform face and speech recognition, people tracking, gesture classification and event analysis (e.g. [Reiter et al., 2005]). A conversation scenario has been addressed in [Bennewitz et al., 2005]; here a robot interacts with several people performing changes in focus and showing different emotions.

The mentioned works do no directly address the problem of multi-modal identity estimation. To fill the gap we introduce a completely new model of a conversation scenario. In particular, four audio and visual descriptors of features are defined for both real time tracking and identification. Visual and audio identification is obtained by combining the outcome of these descriptors analysis with a generalised partial linear model (GPLM)[Müller, 2001,

Severini and Staniswalis, 1994]. Finally, the process undergoes a dynamic updating.

In Section 2 we introduce the data acquisition modalities with the representation of their different descriptors, both audio and video, and the general behaviour of the entire system. In Section 3 and 4 the acoustic and visual scene modelling is defined in details. In Section 5 the GPLM and the complete identification process are described. In Section 6 we only hint the updating problem and the dynamic clustering steps of the system, for lack of space. Finally the various experiments done and the results obtained are illustrated in Section 7.

2 Data acquisition

The knowledge base is a complex data structure that includes both the voice and visual features of R subjects, male and female, with $R = 30$. Each speaker's voice is modelled as a Gaussian mixture density (GMM). The models are trained with the first 18 Mel frequency cepstral coefficients (MFCC) of a particular set of part of speech, made up of very short word utterances of the English phonemes (a single word contains a single phoneme, such as: put, pet, do, etc.). These particular utterances allow to collect only a small set of vocal samples per speaker (two or three examples for phoneme), rather than a whole conversation. Furthermore experiments prove better performance on short words, in particular when the system works in real-time and the active speaker has to be recognised by a short observation sequence. The j-th phoneme pronounced by the i-th speaker is described by a mono-dimensional vector of the audio signal, and its relative MFCC by a 18-dimensional matrix S^i_j for each utterance. Given the number of phonemes N_f (in this case 44) and the number R of voice sampled, $\mathbf{S}^i = [S^i_1 S^i_2 ... S^i_j \dots S^i_{N_f}]$, with $i = 1, \dots, R$ and $j = 1, \dots, N_f$, indicates the complete features matrix of the speaker i.
Let χ be the number of Gaussian density in the mixture, let c_i be the weights components of the i-th model, with $\sum_{l=1}^{\chi} c_{il} = 1$, and let Σ_{il} and μ_{il}, $l = 1, \dots, \chi$ be the covariances and the means of each component of the i-th model, each voice model is completely specified by the parameters, i.e.

$$\lambda_i = (c_{i1}, \dots, c_{i\chi},\ \mu_{i1}, \dots, \mu_{i\chi},\ \Sigma_{i1}, \dots, \Sigma_{i\chi}) \tag{1}$$

The face appearance features of the R=30 people are coded in 2 coefficient matrices. Columns of both matrices encode the observations, namely the people features. In the first matrix, rows encode the Karhunen-Loève coefficient vectors of the *eigenfaces* [Turk and Pentland, 1991]. In the second matrix, rows encode the values of the non-parametric 2D face colour density, taken at each bin.

The acquisition process performs, on a continuous loop, the following tasks:

1. tracking people in the field of view over the frame stream, shifting the focus to include the current speaker into the field of view according to the angle θ determined by voice analysis;

2. extracting appearance based features, given the number of people in the current field of view and an hypothesis on the current speaker, returned by the voice analysis;

3. collecting the visual and voice descriptors to feed the multi-people identification process.

By "field of view"(FOV) we mean the width and the height of the scene that can be observed with the camera lens. In the following, given our audio set-up allowing only for horizontal speaker localisation, we refer the term FOV to the interval

$$FOV = [\theta - \alpha, \theta + \alpha], \qquad \text{with} \quad \alpha = \tan^{-1}(w/2f) \tag{2}$$

here f is the focal length of the camera, w is the image width and θ is the current pan angle of the estimated voice source (see Figure 1).

Variable	Meaning
R	Number of people whose voices and faces we have sampled. $R = 30$
N_f	Number of phonemes. $N_f = 44$.
S_j^i with $j = 1 \dots N_f$ and $i = 1 \dots R$	MFCC matrix of phoneme j pronounced by the speaker i, used to train the Gaussian Mixture model λ_i of the voice of each speaker
$\mathbf{S}^i = \left[S_1^i S_2^i \dots S_j^i \dots S_{N_f}^i\right]$ with $i = 1 \dots R$	MFCC matrix of all the utterances pronounced by the speaker i, used to train the model λ_i
λ_i with $i = 1...R$	GMM of the voice of speaker i
$\{ c_{il}, \vec{\mu_i}, \Sigma_i \}$	Weights, means and covariances of the model λ_i
θ	Angle between robot and its interlocutor
w	Width of the image
f	Focal length of the camera
$FOV = \gamma \in [\theta - \alpha, \theta + \alpha]$	Field of view where $\alpha = \arctan\left(\frac{w}{2f}\right)$

3 Acoustic scene modelling

In this section we present our approach to locate the active speaker and estimate the likelihood of the speaker features recorded during the conversation, with respect to the models created, as described in Section 2. The result is an ordered sequence of voice likelihoods that is suitably combined with the visual features, described in the next section, to produce the dataset further used to identify the speaker in the scene. We adopt the real-time algorithm proposed by [Murray et al., 2004], based on the time delay of arrival and the cross-correlation measure to compute, every 200 ms, the angle θ between the sound source and the robot on the horizon plane.

Each speaker's voice i is modelled as a GMM λ_i. Since each voice feature set i corresponds to a mixture, we also indicate the speakers with their corresponding voice model λ_i. The GMM are often selected for this kind of tasks, being able to describe a large class of sample distributions for acoustic classes relative to phonetic events, such as vowels, nasals or fricatives [Reynolds and Rose, 1995]. In order to obtain the models, we extract MFCC up to the 18^{th} order from the particular set of utterances described in the previous section. These features are robust against noise and use melfrequency scaling, which closely approximate the human auditory system. We use 18 coefficients as a trade off between complexity and robustness after 25 experiments. The parameters initialisation of the EM algorithm and the number of Gaussian components are provided by the the mean shift clustering technique ([Comaniciu and Meer, 2002]); we

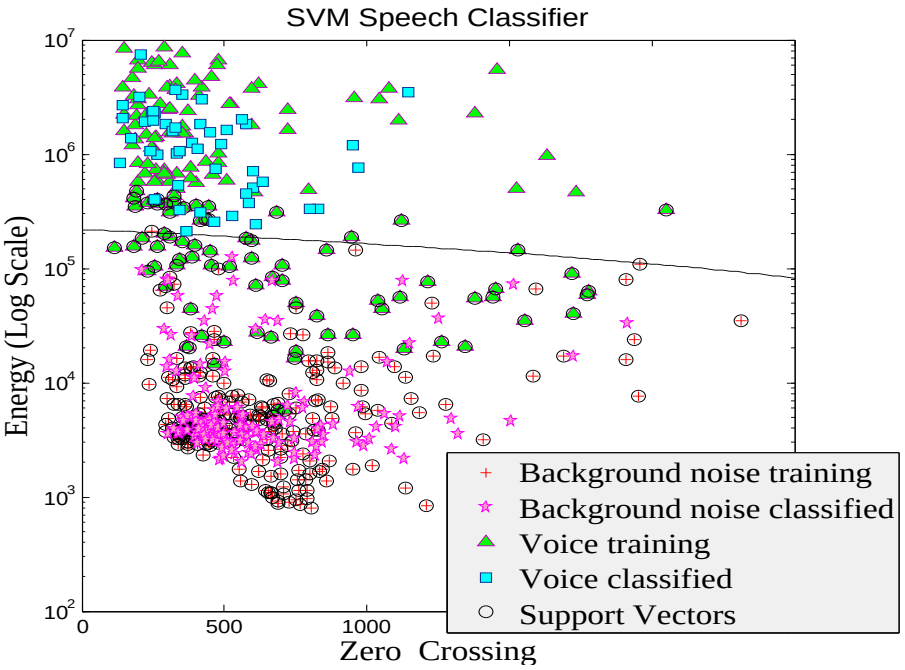


Figure 2: SVM classification of voice against background noise: training and testing.

get a varying number of components χ, with $7 \leq \chi \leq 15$. For each utterance x_t acquired during a conversation and the associated MFCC matrix $\mathcal{A}_t$, composed of N_A 18-dimensional coefficients, we obtain, through the complete features matrix and the GMM, a probability $p(x_t^j|\mathbf{S}^i)$ for all coefficient components x_t^j, $j = 1, \ldots, N_A$. Thus, the expectation is

$$E(\lambda_i|\mathcal{A}_t, \mathbf{S}^i) = \sum_{k=1}^{N_{At}} p(\lambda_i|x_t^k, \mathbf{S}^i)\; p(x_t^k|\mathbf{S}^i) \tag{3}$$

The identification process, on the other hand, also involves clustering of speakers voices labels (see Section 6). To prevent the robot from turning its head to follow noises or random sounds in the environment, we trained a linear classifier based on support vector machine (SVM), able to distinguish between speech and no-speech frames. We consider the short term energy and the zero crossing rate of the signal received ([Rabiner and Sambur, 1975], [Atal and Rabiner, 1976], [Childers et al., 1989]) as discriminant features for SVM classification. The short term energy is the sum of the squares of the amplitude of the fast Fourier transform of the samples in a frame and the zero crossing rate is the number of times the signal changes its sign within the same frame. The set that we use to train the classifier is composed of 10 different speech frames for each speaker in the knowledge base and the same amount of frames including silence or background noise (see Figure 2).

Acquisition and processing Acquisition and processing of voice features is performed in real time: both the voice detection and the localisation procedure work with frames of length $\Delta = 200ms$, the identification process with frames of length $5\Delta = 1s$. Specifically, given the signal $u(t)$, acquired by the microphones at time t, and the frame $u_\Delta(t)$, containing the values of $u(t)$ in the time interval $(t - \Delta,\ t)$, the SVM classification implements a filter that provides a signal $\hat{u}_\Delta(t)$, defined as follows:

$$\hat{u}_\Delta(t) = \begin{cases} u_\Delta(t-\Delta) & \text{if } u_\Delta(t) \text{ does not include a human voice} \\ u_\Delta(t) & \text{otherwise} \end{cases} \tag{4}$$

This filtered frame is used for speaker localisation and identification. The angle between the robot and its interlocutor is computed for each part of signal $\hat{u}_\Delta(t)$. On the other hand the segment of conversation which we link to an identity, among the sampled voices, is represented by the utterance x_t corresponding to 5 consecutive frames $\hat{u}_\Delta(t)$. On this premises, the acoustic scene modelling provides the list of the M most likely speakers. The first part $\hat{S}$ of the list includes the labels associated with the models λ_i, maximising $E(\lambda_i|\mathcal{A}_t, \mathbf{S}^i)$, given the utterance at time t and its MFCC matrix $\mathcal{A}_t$. The other values, modulo expectation, concern people indicated by the visual analysis, if not already in $\hat{S}$.

Variable	Meaning
χ	Number of components of the models λ
x_t	Audio signal captured by the microphones every $1s$
$\mathcal{A}_t$	MFCC matrix of the utterance in a conversation, the current relative speaker of which we have to estimate.
N_A	Number of coefficients in $\mathcal{A}_t$
$x_t^j,\ j = 1 \dots N_A$	j-th vector of the matrix $\mathcal{A}_t$
$u(t)$	Audio signal captured by the microphones at time t
Δ	Sampling interval of acoustic signal. $\Delta = 200$ms
$u_\Delta(t)$	Frame containing the values of $u(t)$ in the time interval $(t - \Delta, t)$
$\hat{u}_\Delta(t)$	Filtered signal by SVM classifier
M	Number of most probable speakers given by acoustic analysis
$\hat{S}$	List of the labels associated with the models λ_i maximising $E(\lambda_i\|\mathcal{A}_t, \mathbf{S}^i)$

4 Visual face descriptors

Visual scene analysis starts with a multi scale face detector. A cascade of classifiers is used to progressively discard areas that are not likely to include a face, by combining successively more complex and computationally expensive classifiers, to mimic the attention that selects areas deserving further processing. Once a face is detected, the face area is divided in regions of interest on which different detectors are scanned to locate the eyes and the mouth. If these detections succeed over a number of different frames the computational process enters the tracking state in which the eye and mouth detectors are scanned across a predicted face region that is computed from the previous frame by a face tracker,

based on mean shift. The core of visual feature extraction is the integration of the detection and tracking facilities, by a finite state machine with detection and tracking states. Pre-processing involves equalisation, aligning and segmentation of face images. For the alignment, we rely on the position of eyes and mouth: assuming all faces are frontally detected, the images are rotated and scaled to compensate the angle γ formed by the eyes. Being d the computed distance between eyes, $\sigma = \bar{d}/d$ is the scale factor needed to obtain the desired distance $\bar{d}$, (x_c, y_c) is the centroid of the eyes and mouth, $a = \sigma \cos\gamma$ and $b = \sigma \sin\gamma$. The transformation H that maps the original to the transformed image is expressed by the 2×3 matrix

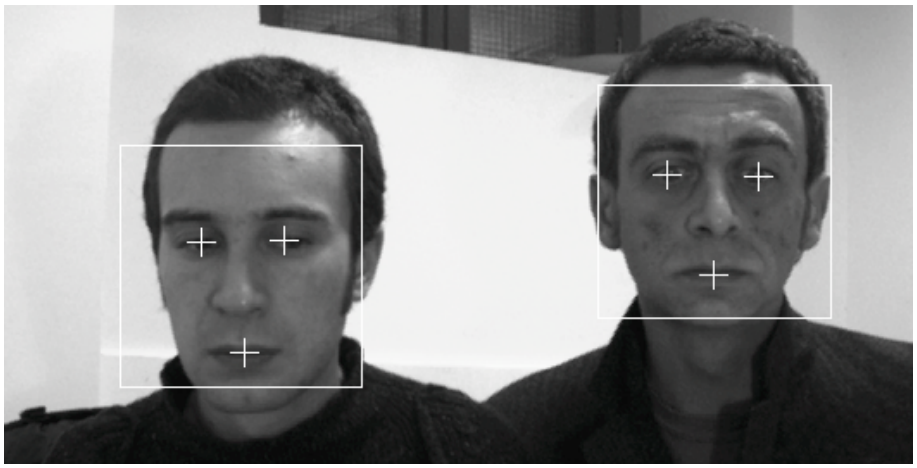
$$H = \begin{pmatrix} a & b & (1-a)x_c - by_c \\ -b & a & bx_c + (1-a)y_c \end{pmatrix} \tag{5}$$

A fixed size region of interest is centred on (x_c, y_c) to extract, from the transformed image, an aligned face image. In the following we introduce a set of descriptors providing a compact representation of a person's appearance and are suitable in this identification problem. Namely, three kind of visual descriptors are defined for each detected face in the scene: a probability that the subject is currently speaking based on mouth movements; a compressed representation of the intensity image; a non-parametric colour distribution. All descriptors refers to a specific region Q_i, $i = 1, \ldots, K_{FOV}$, with K_{FOV} the number of people visible in the camera FOV (see Table 1).

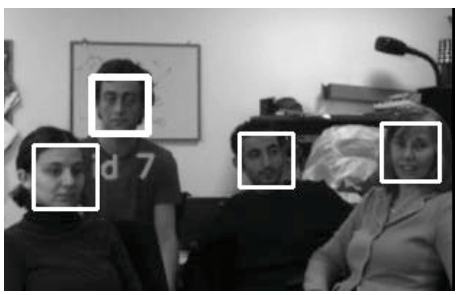
Visual speech descriptor. This descriptor sizes the significant mouth movements, in so contributing to the cross-relation between audio and visual features for the recognition of the speaker in the scene. Indeed, the problem is to evaluate the amount of pixel changes needed to tell that the mouth is articulating a phrase. To face this problem we define a binary mask M_B by thresholding differences of frames from subsequent time steps. Each pixel is treated as an i.i.d. observation of a binary random variable x representing the detected change. Assuming a Binomial distribution for the binary variables within the mask, we estimate the parameter μ using a Beta prior distribution characterised by the hyperparameters α and β. While the μ parameter accounts for the influence of the number of pixels that have changed over the all pixel set, the two binomial proportions α and β enforce or weaken the amount of changes according to the effective number of samples that seem to vary. The best values for our implementation are $\alpha > 0.7$ and $\beta < 0.2$. Let N_B be the size of the observations, with ρ the number of pixels set to 1 by the mask (those changing). The likelihood that the observations come from a windows in which the mouth has significantly moved, as for articulating speech utterances (of course also for smiling or yawning) is thus defined as

$$\sum_{x \in M_b} p(x|\mu_B, N_B - \rho + \beta, \rho + \alpha)\mu \tag{6}$$

Note here that μ is re-estimated from each detected M_B, and thus μ underlines the model which most likely is induced by mouth activity. In any case the M_Bs are chosen, among those having best expectation, also according to the chosen voice models.



(a)



(b)

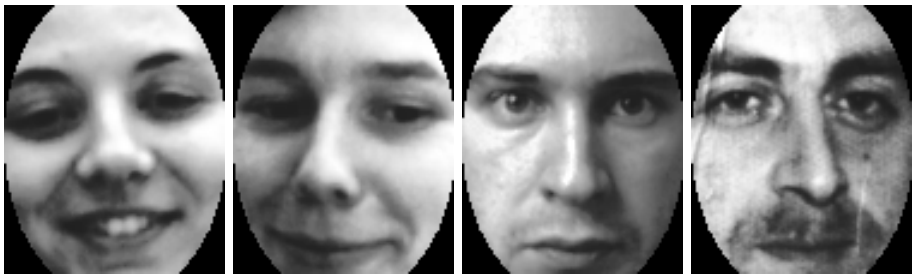
(c)

Figure 3: The figures illustrate the detection and tracking of facial features: faces detected in the field of view are tracked; eye positions are estimated and used for face normalisation, while mouth ROI is needed by the speaker detection process. Brighter pixels in the backprojected image indicate higher probability values; ellipses are centred and oriented according to the extracted eye and mouth positions.

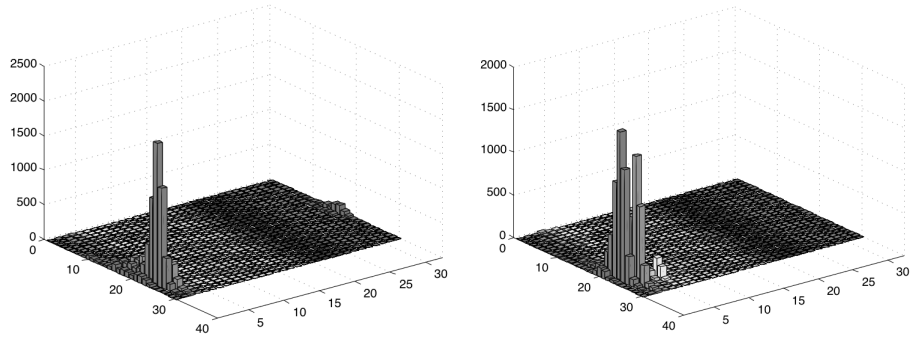
Face appearance feature descriptor. Karhunen-Loève(KL) coefficients provide efficient compression and are suitable to encode frontal faces that have been previously aligned. Compression is achieved by projecting D-dimensional face data into a D'-dimensional subspace spanned by the D' principal directions. Being $\mathbf{c}_i$ the D'-dimensional KL coefficient column vector representing the visual features of the i-th subject, we measure the similarity between i and j in face intensity images by computing the coefficient Mahalanobis distance $d_{\mathcal{M}}(\mathbf{c}_i, \mathbf{c}_j) = (\mathbf{c}_i^\top \mathbf{\Lambda}^{-1} \mathbf{c}_j)^{1/2}$, where $\mathbf{\Lambda}$ is the diagonal matrix, with the eigenvalues corresponding to each KL coefficient.

Face colour feature descriptor. The similarity in colour space is based on the Bhatthacharyya distance between non parametric densities of colour features. More precisely, given two histograms specified by the vectors of bin value $\mathbf{h}_j$ of the face colour features, $j = 1, \ldots, R$, the Bhatthacharyya distance is defined as $d_{\mathcal{B}}(\mathbf{h}_i, \mathbf{h}_j) = (1 - (\tilde{\mathbf{h}}_i^\top \mathbf{\Omega}^{-1} \tilde{\mathbf{h}}_j))^{1/2}$, here $\tilde{\mathbf{h}}$ is the vector obtained by computing the square root for every element of $\tilde{\mathbf{h}}$ and $\mathbf{\Omega}$ is the diagonal matrix mentioning the normalisation coefficients, such that $0 \leq d_{\mathcal{B}} \leq 1$. These colour descriptors, although are not robust against changes in face illumination conditions, compensate degradation of shape cues caused by poor resolution or changes in head orientation.

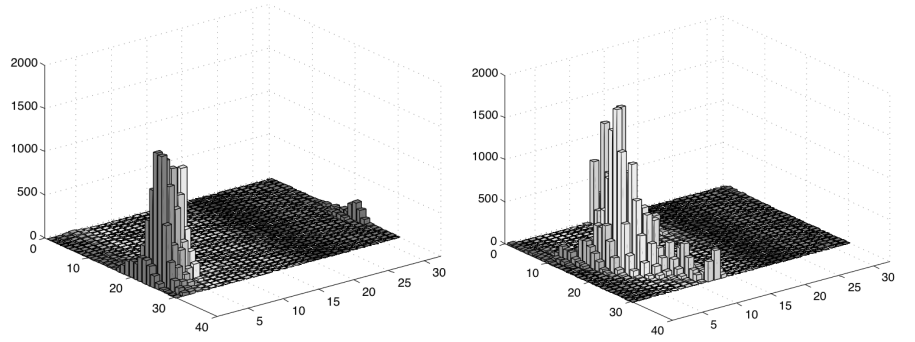
Variable	**Meaning**
γ	Angle between the eyes
d	Distance between the eyes
d	Desired distance between the eyes
σ	Scale factor to obtain the desired distance d
(x_c, y_c)	Coordinates of the centroid of the detected faces.
$a = \sigma \cos\gamma$	Transformation equation
$b = \sigma \sin\gamma$	Transformation equation
H	Transformation matrix
K_{FOV}	Number of people visible in the camera FOV
$Q_i,\ i = 1 \ldots, K_{FOV}$	Regions currently in the camera FOV
M_B	Binary mask for visual speech detection
x	Random variable representing the changes of the mouth.
μ	Binomial distribution parameter
α, β	Beta prior Hyperparameters
N_B	Size of the observations for visual speech detection
ρ	Number of pixels set to 1 by the mask
$\mathbf{c}_i$	KL coefficient column vector
$d_{\mathcal{M}}(\mathbf{c}_i, \mathbf{c}_j)$	Coefficient Mahalanobis distance
$\mathbf{\Lambda}$	Diagonal matrix with the eigenvalues D corresponding to KL coefficients
$\mathbf{h}$	Vector of bin values of the face colour features
$\tilde{\mathbf{h}}$	Vector obtained by computing the square root of every element of $\mathbf{h}$
$\mathbf{\Omega}$	Diagonal matrix mentioning the normalisation coefficients
$d_{\mathcal{B}}(\mathbf{h}_i, \mathbf{h}_j)$	Bhatthacharyya distance between $\mathbf{h}_i$ and $\mathbf{h}_j$



(a) face 1 (b) face 2 (c) face 3 (d) face 4



(e) face 1 (f) face 2



(g) face 3 (h) face 4

Figure 4: The preprocessing step involves equalisation, rotation and scaling. In figures (a)-(d): faces that have been processed by the system in order to extract the appearance and colour features. In figures (e)-(h): the correspondent 30×32 colour histograms for the H-S channels in the HSV colour space. The system uses this representation as a colour descriptor.

Q_i= *region label*	X_1	X_2	X_3	X_4	Y	*Peop. label*
Q_1	0.6909	0.0013	0.4419	0.505	1	A
Q_2	0.3090	0.0922	0.4652	0.505	0	A
Q_1	0.6909	0.1529	0.9516	0.334	0	B
Q_2	0.3090	0.1237	0.3638	0.334	0	B
Q_1	0.6909	0.0014	0.3954	0.161	0	C
Q_2	0.3090	0.1897	0.5641	0.161	0	C

Table 1: The descriptors table corresponding to a trial with two regions (Q_1, Q_2) in the camera FOV and three people labels A, B and C with best descriptors classification. From left to right Q_i is the region label, X_1 is the lips movements descriptor, X_2 is the normalised Mahalanobis distance between the Karhunen-Loève coefficients for the current observed regions, labelled Q_i, and the analogous coefficients stored for each identities in the Knowledge base. X_3 is the normalised Bhatthacharyya distance between the non parametric functions in colour space, sampled in the region Q_i and in the images recorded in the knowledge base. Finally X_4 is the voice descriptor and Y will take value 1 in correspondence of the estimated speaker. Here we assume that there are only two regions in the current camera FOV, labelled by Q_1 and Q_2 and that the people A, B, C chosen are those in the union of the voice set and the distance feature set with best classification. Data are repeated for each potential identity. The task is to identify which row is the correct one. The row will tell who is the current speaker and which is the region in the current camera FOV that corresponds to the speaker. This implies that the real speaker is identified by both the voice and the face. In this case the correct row is the first, thus the correct region label is Q_1 and the speaker is A.

5 Discovering people identities

We recall the reader that the problem we have to solve is online identification of the speaker. We have discussed in the previous sections the different descriptors of voice, lips movements, and face features. We have shown that the voice and the lips movements are defined by a probability distribution (Gaussian Mixture for voice and Beta-Binomial distribution for lips movements) while the other features are normalised distance measures with respect to data coded in the knowledge base.

In this section we discuss how to infer from these heterogeneous data the current speaker identity, presuming that the speaker changes in time. Now, data are collected during a time lapse $t : t + 1sec$ and descriptors are computed for each of these intervals. We define each time lapse a trial, hence from each trial a data structure is formed. This data structure is peculiar because the descriptors have been generated from different sample spaces.

1. The voice descriptor computes a probability with respect to MFCC codes, hence it returns the likelihood that A or B or C, etc. are speaking. It is clear that if the MFCC of two people with very similar voices are stored in the database, say A and Z, even if Z is not in the room, any time A will speak there will be a good chance that the voice estimator will return a high likelihood also for Z.

$X_i, i = 1, \ldots, 4$	Features descriptors: X_1: lips movements, X_2: Mahalanobis distance X_3: Bathacharyya distance, X_4: MFCC voice descriptor
Y	$Y = 1$ indicates the row of the recognised speaker
$A, B, C, ..$	Labels of people features, as recorded in the knowledge base
$\beta = (\beta_1, \beta_2)$	Parameters of the GPLM for X_2 and X_3 descriptors grouped as U
g	Function to be estimated for X_1 and X_4 grouped as T
$f(z) = \frac{\exp z}{1 + \exp z}$	Logistic distribution
FP	False Positive
TP	True Positive
FN	False Negative
TN	True Negative
FPR	False Positive Rate
m	number of regions identified in the camera FOV
$\hat{Y}$	The estimated Y by regression
$\hat{g}$	Original approximation of g
$\hat{\beta}$	initial value of β
$K_{\mathbf{h}}(X_j - X_i)$	Epanechnikov kernel
K	Number of people in the scene
$\mathscr{M}$	Smoother matrix defined using kernel approximation
h	Bandwidth for the kernel density estimation
h_w, h_k	Relative to Epanechnikov kernel
τ	Threshold used to establish convergence
$\mathbf{L}_t$	List of the labels of identified speakers
$\hat{S}$	List of the most probable speaker, provided by audio analisys
δK	Number of new people in the list $\hat{S}$
$\hat{S}_{mp}$	List of most probable labels in $\hat{S}$ in which there is, at most, one new label

2. The lips descriptor will tell who in the scene is moving the lips, so this feature needs to be combined with a voice and a face to be instantiated with a speaker. Indeed, people can articulate the mouth also, for example, for laughing and yawning.

3. The two normalised distances, on the other hand, tell who is plausibly in the camera FOV in the current trial, but cannot tell who is speaking.

A consistent data structure for these varied dataset is illustrated in Table 1. According to the structure of the descriptors, trials will nominally include all the people enumerated in the knowledge base (see Section 2). The chosen labels are those in the union of the sets of descriptors with best classification. Note that if B is in the set, because it has a good classification for the distance features, with respect to a region labelled Q_2, this does not imply that it keeps a good classification with respect to voice or to another region Q_j, $j \neq 2$. That is why we take the union of the sets, instead of the intersection, that might be the empty set. Note also that distances are normalised with respect to the whole dataset and not with respect to the chosen ones (see Section 4).

In order to find the correct row, given a trial, we use regression, in particular we use a semi parametric regression model. The model will estimate the parameters $\beta = (\beta_1, \beta_2)$ and a function g which, when applied to a trial, will return the row that most plausibly indicate the current speaker. To estimate these parameters, however, we had to resort to a training phase. Using the voices and the images in the knowledge base, and suitably adding errors for simulating environment influences, we have defined a set of 925 simulated trials 427 of which have been used for training and the remaining for testing. We can now introduce the model.

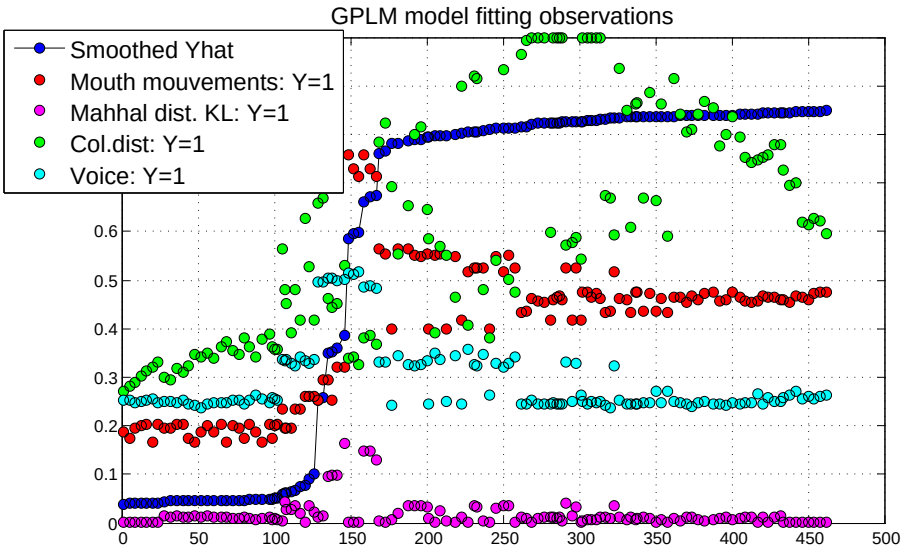


Figure 5: The behaviour of each descriptor, as indicated in the label, taken at the value $\hat{Y}$, indicated $YHAT$, chosen for $Y = 1$.

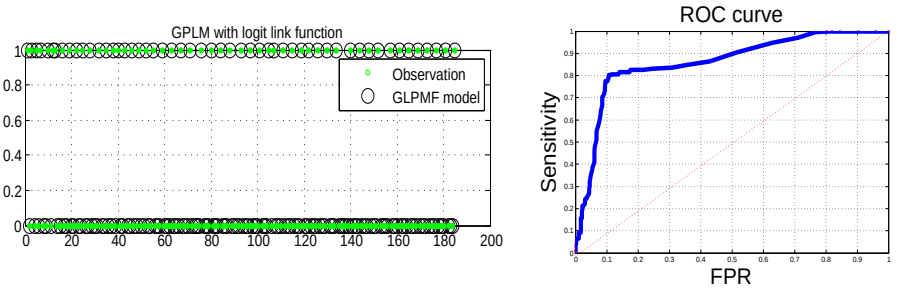


Figure 6: The table on the left illustrates 187 of the 498 matches obtained during testing with the GPLM. On the right, the ROC curve. The false positive rate and true positive rates are defined as $FPR = FP/(FP + TN)$, $TPR = TP/(TP+FN)$ (here FP are false positive, TP true positive, FN false negative and TN true negative). Sensitivity (TPR) is plotted in function of the false positive rate for different cut-off points. Each point on the ROC plot represents a sensitivity/specificity pair corresponding to a particular decision threshold for $\hat{Y}$: if the decision threshold is chosen to be 0.5 then $FPR < 0.1$, while if it is 0.9 then $FPR = 0.5$.

Initialisation: $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4$ are the descriptors for the 427(498) trials of the training(test) set $\mathbf{Y}$ the regressed vector $\mathbf{U} = (\mathbf{X}_2, \mathbf{X}_3), \mathbf{T} = (\mathbf{X}_1, \mathbf{X}_4)$, $Y_i = 1$ on all the correct Q_i, for each trial. $\hat{\beta} \leftarrow 0$, $\hat{g} \leftarrow f^{-1}((\mathbf{Y} + 0.5)/2)$ $\mu_{\hat{\beta}} \leftarrow \exp(\eta)/(1 - \exp(\eta))$ with $\eta = \mathbf{U}^\top \hat{\beta} + \hat{g}(\mathbf{T})$
Loglikelihood and derivatives for μ: $\mathcal{L}(y, \mu)$: $y \log(\mu) + (1 - y) log(1 - \mu)$ $\mathcal{L}'(y, \mu)$: $((y - \mu)/(\mu(1 - \mu)))\mu'$ $\mathcal{L}''(y, \mu)$: $(y - \mu)(\ \mu''/\mu(1 - \mu) - (1 - 2\mu)\mu'^2/(\mu(1 - \mu))^2) - \mu'^2/(\mu(1 - \mu)))$
Repeat estimate $\hat{\beta}, \hat{g}$, using a smoothing matrix $\hat{\mathscr{M}}$, see eq. (8) **until** $\mu_{\hat{\beta}}^{new} - \mu_{\hat{\beta}} < \epsilon$ here K is the number of trials, $\mathbf{1_K}$ is a vector of ones, $\otimes$ is the Kronecker product: $\mathbf{W} = diag(\mathcal{L}''(\mathbf{Y}, \mu_{\hat{\beta}}))$ $\mathbf{Z} = \mathbf{U}^\top \hat{\beta} + \hat{g} - \mathbf{W}^{-1} \mathcal{L}'(\mathbf{Y}, \mu_{\hat{\beta}})$ $\mathscr{M} = \mathscr{M}_1 ./ \mathscr{M}_2$, with $\mathscr{M}_1 = (\mathcal{L}''(\mathbf{Y}, \mu_{\hat{\beta}}) \otimes \mathbf{1_K})^\top . \hat{\mathscr{M}}$ and $\mathscr{M}_2 = \sum \mathscr{M}_1 \otimes \mathbf{1_K}$ $\hat{\mathbf{U}} = (\mathbf{I}_{K \times K} - \mathscr{M})\mathbf{U}$ $\hat{\mathbf{Z}} = (\mathbf{I}_{K \times K} - \mathscr{M})\mathbf{Z}$; $\hat{\beta} = (\hat{\mathbf{U}}^\top \mathbf{W} \hat{\mathbf{U}})^{-1} \hat{\mathbf{U}}^\top \mathbf{W} \hat{\mathbf{Z}}$ $\hat{g} = \mathscr{M}(\mathbf{Z} - \mathbf{U}^\top \hat{\beta})$ $\mu_{\hat{\beta}}^{new} \leftarrow \exp(\eta)/(1 - \exp(\eta))$ with $\eta = \mathbf{U}^\top \hat{\beta} + \hat{g}(\mathbf{T})$

Table 2: Estimation of $\hat{\beta}$ and $\hat{g}$ for the regression model $f\{\mathbf{U}^\top \beta + g(\mathbf{T})\}$, using the training set of $K = 427$ simulated trials and a test set of $K = 498$ trials.

Given the descriptors $X_1, \ldots, X_4$ a semi parametric regression model for inferring the row corresponding to the speaker is defined as:

$$E(Y|\mathbf{UT}) = P(Y = 1 \mid \mathbf{U}, \mathbf{T}) + \epsilon = f(X_2\beta_1 + X_3\beta_2 + g(X_1, X_4)). \tag{7}$$

Here f is the standard logistic distribution $f(z) = (exp(z)/(1 + exp(z)))$, $\mathbf{U} = (X_2, X_3)^\top$, $\mathbf{T} = (X_1, X_4)^\top$, and β and g are the parameters and function to be estimated. Note that we have grouped on one side the normalised distances $\mathbf{U} = (X_2, X_3)$, for which we want to estimate the parameters β_1 and β_2, and on the other side we have grouped the two probabilities $\mathbf{T} = (X_1, X_4)$. Differently from other regression models, the general non parametric regression model (7) is optimal for capturing the combination of linear and non-linear characters of the descriptors. Figure 5 illustrates the different behaviours of the features descriptors considering 427 trials. Here YHAT denotes the $\hat{Y}$ estimated by regression, that has been set to 1, with a decision threshold of 0.67. We estimate g and β according to the algorithm proposed by [Müller, 2001, Severini and Staniswalis, 1994], here for the logit case. The iterative steps of the algorithm are reported in Table 2.

The goal of an empirical analysis of the data collected is to use the finite set of observations obtained for training, that is, $(\mathbf{X}_{ji}, Y_j), j = 1, \ldots, 427, i = 1, \ldots, 4$ to estimate β, g. These values, together with the canonical logit f are used to finally predict a speaker identity. Estimation amounts to the following steps:

1. Analysis of predictors performance to prove their impact on Y. Estimation of the β and of the unknown function g using the training set of the 427

trials, from the 925 obtained by simulation (using the data collected in the knowledge base). Validation of g and β with the remaining 498 trials, for all the plausible $Q_i, i = 1, \ldots, m$ in the camera FOV (in our case $m = 2, 3, 4, 5$).

2. Prediction, in real time, of the speaker identity given the current observations and the knowledge of the current trial dimension (that is, $m^2 \times 5$, with m the number of identified regions in the camera FOV, Q_i, $i = 1, \ldots, m$), considering the whole dataset.

3. Convergence of the identification process after a burning period. Expectation of the features of each identified speaker can be used to track the conversation dynamically and refine the probability of the identity of each speaker using a dynamical model, not described here.

We consider each descriptor $X_1, \ldots, X_4$ as an explanation or predictor of the speaker identity. By looking at the performance of each explanation (see Figure 5) and also because we have two probabilities and two distances, we have chosen to group the two probabilities, that is, lips movements (X_1) and MFCC (X_4) with the non-parametric g. The iterative algorithm, with training data, starts with an initial approximation of $\hat{g} = f^{-1}((Y + 0.5)/2)$, with Y set to 1 on the correct regions labelled Q_i, and with initial values of $\hat{\beta}$ set to 0.

Now, to estimate μ a smoother matrix $\mathscr{M}$ is defined using kernel approximation. We have used the Epanechnikov kernel, defined with respect to the elements of $\mathbf{T}$

$$K_{\mathbf{h}}(X_j - X_i) = \prod_{w=1,2} (1/h_w)(3/4)(1 - ((X_{jw} - X_{iw})/h_k)^2) \cdot (\|(X_{jw} - X_{iw})/h\| \leq 0.75) \tag{8}$$

Here $h_w = 0.4$ and $w = 1, 2$ because $\mathbf{T}$ is $k \times 2$, with $K = 427$ in the training phase and $K = 498$ in the testing phase. Note that the kernel is evaluated with respect to the matrix $\mathbf{T}$, mentioning all the trials both in the training and testing phases. Then the smooth matrix $\mathscr{M}$, according to [Müller, 2001], can be formed by the following elements κ_{ij} of $\mathscr{M}$:

$$\frac{(\mathcal{L}''(Y, \mu_j))K_{\mathbf{H}}(X_j - X_i)}{(1/n)\sum_{j=1}^{n}(\mathcal{L}''(Y, \mu_j))K_{\mathbf{H}}(X_j - X_i)} \tag{9}$$

Convergence is achieved when difference of likelihood and the estimates of β is below a certain threshold τ. We used $\tau = 0.1E\text{-}004$ and for our set of trials 48 iterations were needed to converge on the data train set with $K = 427$. On data test the error is 0.4%. The error rate is, indeed, very low, as shown in the ROC curve displayed in Figure 6, reporting the behaviour of the estimator on data test.

6 Updating

One main problem for the online application of the system is the knowledge base dimension. If the knowledge base is large, then online acquisition for the

voice descriptors and the visual descriptors, concerning the two distances, is a quite hard task. Indeed, it requires a huge set of comparisons, since nothing is known about the people in the scene. So the question is: is there a time t at which the system knows who is in the scene and can rely on that for online identification?

Experiments show that a time t at which all people have spoken is difficult to predict, and if no constraint is put on the scene, some people can leave and new people can join. Thus there is not a fixed set that can be devised after a specified time.

To solve this problem and induce a partial knowledge of the people in the scene, we assume that changes are smooth: not all current people suddenly disappear nor are substituted altogether with new ones. So in a time lapse at most one person joins the conversation and one leaves, and partial updates can be inferred for voice and face similarities acquired up to time T. More specifically, for the same effective speaker, the list $\hat{S}$ of the most probable relative labels, estimated via the acoustic analysis, tends to involve the same ones. After a specified time T (burning period), clusters of different cardinality, for each different list $\hat{S}$, are generated, with the associated frequency of occurrence. Thus, if at time $t > T$ there are δK new people, with $\delta K \geq 2$, in the list, only the most probable labels $\hat{S}_{mp}$ are maintained, while the others are replaced with the labels mentioned in the most likely cluster, of the same cardinality. This includes $\hat{S}_{mp}$, according to the likelihood computed after the burning period. These clustering on voices is integrated with an analogous clustering on visual distances and are thus used for setting a dynamic model in which known states and unknown new states are devised. The dynamic of this process, unfortunately, cannot be further described here.

7 Experiments and Conclusion

The described framework has been tested in real conversation scenarios, involving several people standing around the robot and producing audio and visual stimuli (see Figure 1). Experiments with up to 5 people, after a burning period of 2-3min, and with one people change, have a mean error rate of a person every 10 experiments. The experiments set-up is a P3-DX robotic platform, by ActivMedia, extended by an alloy chassis supporting a laptop, which actually runs the software and presents a control interface, and the head, made of a pan tilt unit orienting the auditive and visual sensors: a pair of omnidirectional microphones, by Behringer, and a an AVT Marlin colour camera. Both the sensor types provide fast acquisition through the firewire bus. Computation is performed by a Pentium M based laptop computer, running the multi-modal data acquisition, segmentation, preprocessing and feature extraction C++ software, and a MATLAB engine for the regression and identification scripts. Audio signal is acquired and sampled at 44100 Hz, 16 bits/sample, using a Fireface 400 mixer interface, by RME Intelligent Audio Solution.

Training was performed off-line using acquired sequences from a knowledge base including 30 subjects (see Section 2). 925 randomly generated trials with cardinality 2, 3, 4 and 5 are built from the stored sequences, with the constraint that every generated set must contain the real speaker.

The heaviest computation is required by visual descriptors, mainly for the

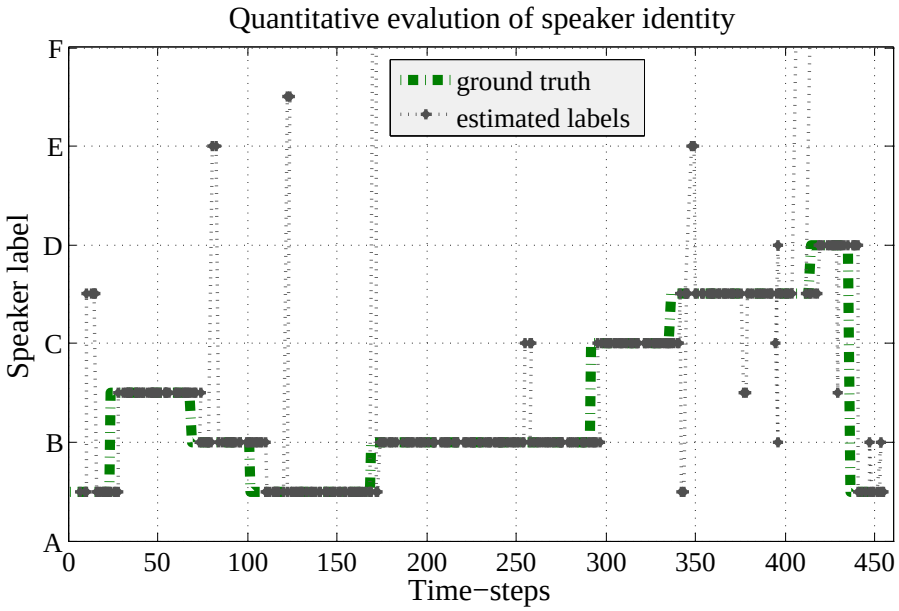


Figure 7: Quantitative evaluation of a speaker identity estimation performance over 450 time-steps. Figure shows the system tracking the current speaker identity and recovering from failure.

detection and tracking of the facial features. The system relies on a multiscale detector based on [Viola and Jones, 2004] and a mean shift tracker introduced in [Bradski, 1998], that uses the backprojection of a 30×32 colour histogram encoding the H and S levels in the HSV colour space (see Figure 3). The same histogram representation is also used to estimate the 2D colour pdf of a specified Q region, as described in Section 4. The back-projected image is conveniently used as a segmentation mask (see Figure 4) and we chose to rely on a mean shift tracking procedure that can make use of such a representation of faces. On the other hand, the choice for the detector was mainly motivated by the performance level achieved by the face/eye/mouth cascades during the experiments. The overall performance of the vision process depends on the number of people in the scene. The described set-up allowed the acquisition / feature extraction loop to run at ≈ 12 Hz in the case of only two people in the camera field of view, which decreases to ≈ 5 Hz if the people in the FOV are 5.

Online experiments prove that comparison of descriptors to the whole KB causes a sensible loss of performance, and brought the need to maintain a low number of known individuals in the KB. Evaluation was performed by setting up different conversation scenarios (Figure 1) and observing the robot shifting its attention towards the real speaker's direction and verifying his/her identity through the robot's GUI. We ignore audio/video data collected during the robot changes of position, to produce a sort of saccade. A quantitative performance analysis has been carried out on audio/video sequences gathered by the robot in such scenarios (see Figure 7). A total of about 1 hour of conversation, with 2, 3, 4, 5 people in the camera FOV has been collected and manually labelled: the

real speaker's identity has been assigned to each time-step to form the ground truth; the time step used amounts to 1 second. This experiment involved 10 people among the 30 stored in the knowledge base. It is worth noting that, since the sequence is collected directly by the robot, the number of errors in speaker identification is affected by errors in the speaker localisation process too. Along the entire sequence, the total error number is mediated by the number of time-steps and the resultant percentage of successful speaker identifications was $\approx 85\%$.

References

[Atal and Rabiner, 1976] Atal, B. and Rabiner, L. (1976). A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Signal Processing*, 24(3):201 – 212. 7

[Bennewitz et al., 2005] Bennewitz, M., Faber, F., Joho, D., Schreiber, M., and Behnke, S. (2005). Multimodal conversation between a humanoid robot and multiple persons. In *Proceedings of the Workshop on Modular Construction of Humanlike Intelligence at the Twentieth National Conferences on Artificial Intelligence (AAAI)*. 4

[Bradski, 1998] Bradski, G. R. (1998). Real time face and object tracking as a component of a perceptual user interface. In *4th IEEE WACV*. 19

[Childers et al., 1989] Childers, D., Hand, M., and Larar, J. (1989). Silent and voiced/unvoied/ mixed excitation(four-way),classification of speech. *IEEE Transactions on Acoustics, Speech, and. Signal Processing*, 37(11):1771–1774. 7

[Comaniciu and Meer, 2002] Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. In *IEEE TPAMI*. 6

[Desilva et al., 2006] Desilva, G. C., Yamasaki, T., and Aizawa, K. (2006). Interactive experience retrieval for a ubiquitous home. In *ACM CARPE*. 4

[Müller, 2001] Müller, M. (2001). Estimation and testing in generalized partial linear modelsa comparative study. *Statistics and Computing*, 11:299–309. 4, 16, 17

[Murray et al., 2004] Murray, J. C., Erwin, H., and Wermter, S. (2004). Robotics sound-source localization and tracking using interaural time difference and cross-correlation. In *AI Workshop on NeuroBotics*. 6

[Pentland, 1995] Pentland, A. P. (1995). Machine understanding of human action. In *M.I.T. Media Laboratory*. 4

[Rabiner and Sambur, 1975] Rabiner, L. and Sambur, M. (1975). An algorithm for determining the endpoints of isolated utterances. *The Bell System Technical Journal*, 54(2):297–315. 7

[Reiter et al., 2005] Reiter, S., Schreiber, S., and Rigoll, G. (2005). Multimodal meeting analysis by segmentation and classification of meeting events based on a higher level semantic approach. In *IEEE ICASSP*, pages 294–299. 4

[Reynolds and Rose, 1995] Reynolds, D. and Rose, R. (1995). Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE TSAP*, 3(1). 6

[Severini and Staniswalis, 1994] Severini, T. and Staniswalis, J. (1994). Quasi-likelihood estimation in semiparametric models. *J. Amer. Stat. Assoc.*, 89:501–511. 4, 16

[Turk and Pentland, 1991] Turk, M. A. and Pentland, A. P. (1991). Face recognition using eigenfaces. In *Proc. of IEEE CVPR*, pages 586–591. 5

[Viola and Jones, 2004] Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *IJCV*, 57(2):137–154. 19

[Waibel et al., 2003] Waibel, A., Schultz, T., Bett, M., Malkin, R., Rogina, I., Stiefelhagen, R., and Yang, J. (2003). Smart: the smart meeting room task at isl. In *Proc. of ICASSP*, pages 752–755. 4