FP6-506811

AMI

Augmented Multiparty Interaction

Integrated Project
Information Society Technologies

D4.2 Report on Implementation and Evaluation Results of Audio, Video, and Multimodal Algorithms

Due date: 31/12/2005
Project start date: 1/1/2004

Submission date: 31/12/2005
Duration: 36 months
Revision: 1

Lead contractor: TUM

Project co-funded by the European Commission in the 6th Framework Programme (2002-2006)		
	Dissemination Level	
PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

D4.2 Report on Implementation and Evaluation Results of Audio, Video, and Multimodal Algorithms

Editor: Marc Al-Hames, TUM

Abstract: WP4 is concerned with the automatic recognition from audio, video, and combined audio-video streams. Research topics include robust speech recognition for multiparty meetings, gesture and action recognition, emotion recognition, source localisation and object tracking, keyword spotting, and person identification. Deliverable D4.2 is a report on the implementation and evaluation of the different audio, video, and multimodal algorithms on common AMI datasets and with common evaluation schemes.

Contents

1 Introduction

WP4 is concerned with the automatic recognition from audio, video, and combined audio-video streams, with an emphasis on developing models and algorithms to combine modalities. Algorithms have been ported or implemented in the AMI domain and evaluated on the common AMI dataset. The models that have been applied include HMMs, Bayesian networks, neural networks, multistream approaches, and multisource decoding.

1.1 Involved partners

Table 1 shows the involved partners and resources in WP4 for the period from month 12 to 30.

Part.	UEDIN	DFKI	ICSI	TNO	BUT	TUM
PMnth.	36	4	45	14.6	72	65
Part.	IDIAP	USFD	UT			FC
PMnth.	53	12	25			4

Table 1: Involved partners and person-months in WP4

1.2 Splitting of work

Instead of dividing the tasks into speech, visual, and audio-visual groups it was decided to split the tasks into problem-based groups. Solutions are not distinguished by their approach (for example visual or audio identification of persons). Therefore different approaches can be evaluated and compared on a common data set with a given standard (for example how many persons have been identified correctly during the meeting?). We identified seven main questions and therefore split WP4 into seven sub-groups:

- Baseline speech recognition system (Sec. 2)
- Event spotting (Sec. 3)
- Localization and Tracking (Sec. 4)
- Gestures and actions (Sec. 5)
- Emotion recognition (Sec. 6)
- Person identification, segmentation, and clustering (Sec. 7 and 8)
- Focus of attention (Sec. 9)

1.3 Aim in the second year and outline of this deliverable

The expected result of WP4 is a set of multimodal recognizers for robust speech recognition, gesture and action recognition, emotion recognition, source localization, object tracking, and person identification.

In the first year we developed and ported a wide range of algorithms to the AMI domain. In the second year of the project, we decided about common interfaces and common evaluations for the different algorithms. This allows us to compare different approaches and algorithms on common AMI data and compare the achieved results to other state-of-the-art approaches. This deliverable D4.2 reports about the progress that has been made in the second year and the evaluation results. The methods and results are described in detail in Sec. 2 - 9, where each section describes the progress that has been made in one of the sub-groups (cf. Sec. 1.2).

2 Automatic speech recognition

Many people spend a considerable time in their working life in meetings, however the efficiency of meetings is often low and hence approaches for streamlining the process and for retaining and crystallising the right information have been developed. So far computers are rarely used to aid this process.

Meetings are an audio visual experience by nature, information is presented for example in the form of presentation slides, drawings on boards, and of course by verbal communication. The latter forms the backbone of most meetings. The automatic transcription of speech in meetings is of crucial importance for meeting analysis, content analysis, summarisation, and analysis of dialogue structure. Widespread Work on automatic recognition of speech in meetings started with yearly performance evaluations by the U.S. National Institute of Standards and Technology (NIST) [131]. Work on meeting transcription was initially facilitated by the collection of the ICSI meeting corpus [75] which was followed by trail NIST meeting transcription evaluations in Spring 2002. Further meeting resources were made available by NIST [46], Interactive System Labs (ISL) [21] and the Linguistic Data Consortium RT04s Meeting evaluations [131]. During last year the AMI project has collected audio data from many meetings which from the basis of experiments presented here.

2.1 Objectives

In general the objective for work in ASR is derived from the general objective to use machine based techniques to aid people in and outside of meetings to gain efficient access to content information. The objectives for work in automatic speech recognition are:

- to develop state-of-the-art speech recognition technology for meeting transcription;
- to enable research into meeting relevant topics into ASR;
- to provide a common working base for researchers;
- and to enable downstream processing by providing automatically annotated and transcribed data.

All of these objectives require a common and standardised evaluation scheme and unified testing procedures. For automatic speech recognition in general evaluation by word error rate measurement is standard. In order to ensure the objective of development of state of the art technology two requirements are derived:

- Generic universally tested core technology
  ASR system have essential generic core components that should perform well on related tasks.
- Particpation in international competitions.
  The proven way to ensure that technology under development is state-of-the-art is to enter international competitions in the field operated by groups outside of the AMI project.

Work for these objectives requires to work not only on AMI meeting data, but on data obtained from related research areas, projects, etc. Hence for project internal purposes, i.e. to ensure research compatibility and a common working base, we require data sets on evaluation procedures specific for AMI data.

- AMI data training and test data setup
  For AMI specific evaluation the definition of training and test corpora as well as system standards is necessary.

2.2 Evaluation method

Evaluation of ASR systems requires the specification of the training and test data to be used, and the specification of test conditions. In this section we outline the data used in the experiments described in the following sections. We further give details on how performance of individual algorithms and complete transcriptions systems is evaluated.

2.2.1 Data

Conversational Speech As the number of speech resources for meetings is still relatively small, similar to work presented in [146], a recognition system for conversational telephone speech (CTS) forms the starting point for our work on meetings. This approach was preferred to bootstrapping from Broadcast News (BN) systems (as for example in [136]) as the meeting style is expected to be colloquial rather than presentational.

Transcription of conversational telephone speech is very closely related to transcription of speech in meetings recorded with close-talking microphones. Hence we use the CTS data to evaluate performance of algorithms that target aspects of this more generic type of speech (rather than meeting specific aspects). We evaluate performance on the NIST Hub5E evaluations sets for the years 1998, 2001, and 2002.

General Meeting Data The ICSI Meeting corpus [75] is the largest meeting resource available consisting of 70 technical meetings at ICSI with a total of 73 hours of speech. The number of participants is variable and data is recorded from head-mounted and a total of four table-top microphones. A 3.5 hour subset of this corpus covering 30 minute extracts of 7 meetings was set aside for testing (icsidev). Further meeting corpora were collected by NIST [46] and ISL [21], with 13 and 10 hours respectively.Both NIST and ISL meetings have free content (e.g. people playing games or discussing sales issues) and number of participants. We also make use of the RT04s NIST evaluation set (*rt04seval*) which also includes meetings recorded by the LDC. As will be explained further on, we have participated in the 2005 NIST RT05s evaluations. Hence, further results will be presented on the RT05s NIST evaluation set (*rt05seval*).

AMI Meeting Data The AMI corpus collection and transcription effort is detailed in the WP2 description. Here we shall focus on the parts that are relevant for ASR work only. Naturally this is the audio data and associated word level transcripts. The practise of system development dictates that at certain points in time all data available at that moment is used for a complete development cycle. Due to the complexity of ASR systems these cycles take a long time. Hence not all available AMI data has been used yet. The complete set will be used for the next cycle.

The AMI corpus in its final form consists of meetings from 3 different meeting rooms, at The Centre for Speech Technology Research, Edinburgh (UEDIN), The IDIAP Research Institute, Switzerland (IDIAP) and TNO Human Factors, The Netherlands (TNO). Overall more than 100 hours of speech are to be transcribed. The meeting language is English. Each meeting normally has four participants and the corpus will be split into a scenario portion and individual meetings. The scenario portion involves the same participants over multiple meetings on one specific task. So far only data from scenario meetings was used. For training purposes approximately 15 hours of data (*amitrain05*) were used for meeting transcription experiments (unless noted otherwise). A development set (*amidev*) consisting of 8 meetings from 2 locations is used for testing.

Lecture Room Data Our objectives is to develop technology that is robust to the transfer to related areas. The CHIL project (Computers in the Human Interaction Loop, an EC IP project) works on transcription of seminars and lectures. In the context of the NIST RT05s evaluation we have also worked on this type of data. For the purpose of development of systems for transcription of lecture room speech a development set (*rt05slectdev*) was provided by CHIL. However this was provided very late and due to time constraints could only be used for language model (LM) optimisation. We further report results on the RT05s evaluation sets from the lecture room data (*rt05slecteval*). No training data is available.

The Multi-Channel Wall Street Journal Audio Visual Corpus The Multi-Channel Wall Street Journal Audio Visual Corpus (MC-WSJ-AV) is a corpus offering an intermediate task between simple digit recognition and large vocabulary conversational speech recognition. This corpus consists of read Wall Street Journal sentences taken from the test set of the WSJCAM0 database. The corpus is recorded in the instrumented meeting rooms constructed for the recording of the AMI Meetings Corpus at three sites: The Centre for Speech Technology Research, Edinburgh (UEDIN), The IDIAP Research Institute, Switzerland (IDIAP) and TNO Human Factors, The Netherlands (TNO). The sentences are read by a range of speakers (45 in total) with varying accents including a number of non-native English speakers. During recordings, all speakers wear lapel and headset microphones, and audio from two eight element microphone arrays is also captured. The rooms also provide synchronised video recordings including close-up views of the speakers' faces, as well as wide-angle views of the entire room.

The data consists of recordings made under three conditions:

1. Single Speaker Stationary: For this condition the speaker is asked to read sentences from six positions within the meeting room – four seated around the table, one standing at the whiteboard and one standing at the presentation screen.

2. Single Speaker Moving: For this condition the speaker is asked to move between the six positions while reading the sentences. The speaker begins reading at position 1 and moves to position 2 while reading the first sentence. They then move back to position 1 while reading the next, and continue alternating between these two positions with each sentence. This is repeated for each pair of adjacent positions.

3. Overlapping Speakers (Stationary): Here, two speakers are asked to simultaneously read their sentences from different positions within the room. The speakers remain in the same positions for the entirety of these recordings and separate recordings are made from each of the 15 pairs of positions.

The corpus is suitable for wide variety of research tasks. Here it is mostly used used for the development of microphone array ASR front-end processing systems.

2.2.2 Standardised Evaluations

As mentioned above, to ensure that the technologies under development are state of the art we decided tp participate in international evaluations of ASR systems for meeting transcription [109]. These evaluations are conducted by the U.S. National Institute for Standard and Technology (NIST) since 2002 on a yearly basis (with the exception of 2003). World-leading research groups in automatic speech recognition enter this competition which aims to provide a comparison between different approaches by provision of standardised common training and test sets, an evaluation schedule, and by organisation of a workshop to ensure information exchange. We have successfully competed in the NIST RT05s STT evaluations [109], yielding very competitive results on both conference meeting and lecture room transcription [59].

The NIST RT05 evaluations required the transcription of audio from meetings collected from headset microphones (Independent Headset Microphone, IHM) and from table-top microphones (Multiple Distant Microphones, MDM) with the latter forming the primary test condition. Data originates from the corpora collected by ICSI, NIST, ISL, CMU, AMI, and Virginia Polytechnic and State University. For training all publicly available corpora are allowed, for testing excerpts of two 10-minute excerpts from meetings from each of the aforementioned sources are given. For evaluation on lecture room speech only data collected by CHIL was available. Again 10 minute extracts of lectures are to be transcribed.

NIST has an ongoing commitment to meeting transcription and will also hold evaluations in 2006. We have decided that the RT evaluations allow a perfect and fair assessment of AMI technology in a wider framework and hence we will continue to participate in these evaluations for the duration of the project. Participation also brings access to the evaluation data (this is normally not publicly available). All major algorithmic developments are tested on the *rt05seval* and *rt04seval* test sets.

2.2.3 AMI-specific Evaluations

AMI specific evaluations are performed on AMI data alone (.e.g *amidev*). As all micriphone conditions are available for the complete corpus no special targeted sub-sets are defined. In the course of our next development cycle we will implement a larger development test set (planning of this set had an input on data collection) that will cover all aspects of the corpus in terms of meeting room coverage, scenario and speaker coverage.

Microphone array development in large parts makes use of the MC-WSJ-AV corpus, however final evaluation again will be obtained on the aforementioned development sets.

2.3 Results

In this section we present and discuss results on data outlined in the previous section and give details on the algorithms used to obtain these results. We also present the final results obtained in the NIST RT05s evaluations, including post evaluation analysis.

2.3.1 Results with the AMI CTS system

Acoustic modelling Font-ends make use of 12 MF-PLP [169, 66] coefficients and the 0th cepstral coefficient c_0. These are derived from a reduced bandwidth of 125-3800Hz. First and second order derivatives are added to form a 39 dimensional feature vector. Cepstral mean and variance normalisation is performed on complete conversation sides and hence are implicitly speaker specific. Acoustic models are phonetic decision tree state clustered triphone models with standard left-to-right 3-state topology. They were obtained using standard HTK [151] maximum likelihood training procedures (see for example [62]). The system uses approximately 7000 states where each state is represented as a mixture of 16 Gaussians. Speaker adaptive training is performed in the form of vocal tract length normalisation (VTLN) both in training and test. Warp factors are estimated using a parabolic search procedure, a piecewise linear warping function and a maximum likelihood criterion[62]. Speaker adaptation is perfermed using maximum likelihood linear regression (MLLR) [91] of the means and variances[45].

Feature transformation is applied in the form of smoothed heteroscedastic linear discriminant analysis (SHLDA) [22]. SHLDA is used to reduce a 52 dimensional formed by the standard feature vector plus third derivatives to 39 dimensions. HLDA estimation procedure[88] requires the estimation of full covariance matrices per Gaussion. SHLDA uses smoothing of the covariance estimates by interpolating with standard LDA type with-in class covariances.

$$\mathbf{\Sigma}_{sm} = \alpha\mathbf{\Sigma} + (1-\alpha)\mathbf{\Sigma}_{WC} \tag{1}$$

$\mathbf{\Sigma}_{sm}$ is the smoothed estimate of the covariance matrix and $\mathbf{\Sigma}_{WC}$ is the LDA type within-class matrix estimate based on an occupancy weighted average. Values for α of $0.8-0.9$ were found to yield satisfactory results.

Dictionaries The UNISYN pronunciation lexicon [43] forms the basis of dictionary development with pronunciations mapped to the General American accent. Normalisation of lexicon entries to resolve differences between American and British derived spelling conventions was performed yielding a 115k word base dictionary. Pronunciations for a further 11500 words were generated manually to ensure coverage

Corpus name	#words (MW)
Swbd/CHE	3.5
Fisher	10.5
Web (Switchboard)	163
Web (Fisher)	484
Web (Fisher topics)	156
BBC - THISL	33
HUB4-LM96	152
SDR99-Newswire	39
Enron email	152
ICSI meeting	1
Web (meetings)	128

Table 2: Size of various text corpora in million words (MW).

Hub5e eval sets	Bigram	Trigram	4-gram
Swbd	104.53	85.97	84.12
Swbd + HUB4	95.00	72.55	69.04
Swbd + HUB4 + Web	90.89	66.75	61.59

Table 3: Perplexities on the joint NIST Hub5E 1998/2001/2002 evaluation test sets (CTS).

of training data. For consistency and a simplified manual pronunciation generation process hypotheses generation procedures have been developed. Pronunciations for partial words are automatically derived from the baseform dictionary. Hypotheses for standard words were generated using CART based letter-to-sound rules. The CART based letter-to-sound prediction module was trained on the UNISYN dictionary using tools provided with the Festival speech synthesis software [16] using left and right context of five letters and left context of two phones. This gave 98% phone accuracy and 89% word accuracy on the base dictionary., for manually generated pronunciations the error rates were 89% and 51% respectively. Although the word accuracy is quite low on new words (many of which were proper names, partial words etc.), the phone accuracy remains relatively high.

Language modelling and Vocabulary Selection of vocabulary for recognition is based on a collection of in-domain words. However, in the case of insufficient data it is beneficial to augment this list with the most frequent words from other sources, for example Broadcast News (BN) corpora. This ”padding” technique was used for all dictionaries in this paper unless stated otherwise. The target dictionary size was 50000 words and the source of words was BBC news data, the Broadcast News 1996 Hub4 corpus (HUB4-LM96), and Enron data[85] (see table 3).

Language model training data for conversational speech is sparse. Hence models are constructed from other sources and interpolated (as in e.g. [62]). This is true for both CTS and meeting data. Hence we have processed a large number of different corpora to form the basis of our language models. The most important corpora are listed in Table 16. A full discussion of all source material would go beyond the scope of this paper. The most important non-standard data was found to be the the Web collected resources [18] and ICSI meetings. In total more than 1300 MW of text are used.

Each corpus was normalised using identical processes. Apart from standard cleanup we tried to ensure normalised spelling and uniform hyphenations across all corpora. For the training and testing of language models the SRI LM toolkit [149] was used to train models with Kneser-Ney discounting

eval01	VTLN	MLLR	non-HLDA	SHLDA
pass1			37.2	35.0
pass2	×		33.8	32.1
pass3	×	×	32.1	30.6

Table 4: %WER results on the NIST Hub5E 2001 evalution set.

	ICSI	NIST	ISL	AMI
Avg. Dur (sec)	2.42	3.98	3.21	3.95
#words	823951	157858	119184	154249
#unique wds	11439	6653	5622	4801

Table 5: Statistics for meeting corpora.

and Backoff. Table 3 shows perplexity results on the NIST Hub5e evaluation sets. Note the substantial reduction in perplexity by the additional web resources.

Decoding and overall system performance Decoding operates in three passes. The Cambridge University speech decoder HDecode is used for recognition with trigram language models. Table 4 shows results for each pass. The first pass yields a first level transcription which is used or VTLN warp factor estimation. In the second pass improved output is generated using VTLN trained models. The final output is obtained after MLLR adaptation using transforms for speech and silence. The table also gives a comparison of results with and without SHLDA. Trigram language models as described above were used in the experiments. A significant reduction in word error rate (WER) from both VTLN and SHLDA is observed.

2.3.2 Meeting Data

Language in Meetings Even though of general conversational nature, meeting data differs substantially from CTS. First of all the acoustic recoding condition is usually more complex as the speaker has no feedback on the recording quality. Speech signals of close-talking microphones are distorted by heavy breathing, head-turning and cross-talk. Table 5 shows raw statistics on several meeting corpora. Average utterance durations are larger than on CTS, however with great variation. We can also observe that corpus size is not a good predictor for the number of unique words in the corpus and hence complexity.

Vocabulary We shall loosely define a domain as a set of sub-corpora that, when used in a combined non-discriminative fashion, yield better performing models than the parts. This definition is not strict and will show a tendency to combine small corpora. However for the purpose of model training the question of how to use data is most important. Table 6 shows on the left hand side Out Of Vocabulary (OOV) rates using vocabulary derived from each meeting corpus. The OOV rates do not correlate perfectly with vocabulary sizes (Table 5).

On the right hand side the wordlists are padded as described in section 2.3.1 (this includes removal of obvious typographic errors). It is evident that overall the effect of vocabulary mismatch is greatly reduced uniformly for all cases. This suggest that only a very small amount of meeting specific vocabulary is necessary. Hence padding was used in all further experiments.

Content Apart from the raw word difference it is important understand the effect of the wide range of topics covered in the various meetings. A set of experiments was conducted to compare meeting resource

	Vocabulary sources							
	No padding				Padding to 50k			
	ICSI	NIST	ISL	AMI	ICSI	NIST	ISL	AMI
ICSI	0.00	4.95	7.11	6.83	0.01	0.47	0.58	0.57
NIST	4.50	0.00	6.50	6.88	0.43	0.09	0.59	0.66
ISL	5.12	5.92	0.00	6.68	0.41	0.37	0.03	0.57
AMI	4.47	4.39	5.41	0.00	0.53	0.53	0.58	0.30
COMB	1.60	4.35	6.15	5.98	0.16	0.42	0.53	0.55

Table 6: %OOV rates of meeting resource specific vocabularies. Columns denote the word list source, rows the test domain.

Test Corpus	ICSI	NIST	ISL	AMI	COMB
ICSI	68.17	74.57	73.76	77.14	67.97
NIST	105.91	100.87	102.01	105.95	101.25
iSL	104.68	99.45	98.45	106.39	102.86
AMI	115.56	114.26	114.41	88.91	94.08
LDC	97.78	90.66	88.87	92.44	93.84
COMB	107.46	105.93	105.73	90.62	92.74

Table 7: Cross meeting room perplexities on subsets of rt04seval and rt05samidev. COMB denotes training or testing using all meeting data.

optimised language models on the basis of the meeting resource specific (MRS) padded vocabularies. Language models are obtained by optimisation of interpolation weights for the components outlined in Table 16.

Table 7 shows perplexities on all corpora. In all cases that the best perplexities are achieved on the originating corpus, however with little margin. Note also that the MRS LMs significantly outperform the generic LMs only in the case of ISL and AMI. In general the perplexity of ICSI test data is very low. This appears to be a property of this data set.

2.3.3 Meeting transcription

Common for all meeting rooms is that audio is recorded either by close-talking microphones or via single or multiple distant microphones. The latter may be arranged in a fixed array configuration. Due to interaction between speakers the system must be capable of speech detection and and speaker grouping as well as recognition. In the following we first outline techniques for audio segmentation and microphone array processing, followed by a description of model training procedures and recognition results.

Automatic segmentation Speech activity detection (SAD) for close talking microphones poses a significant challenge. The high levels of cross-talk and non-speech noise (such as breath or contact noise) prohibit the use of threshold based techniques, the standard in more ‘friendly’ recording conditions. The system used here is a straight-forward statistical based approach with additional components to control cross-talk between channels. Statistical approaches to SAD typically use HMM or GMM based classifiers with special feature vectors such as channel cross-correlation and kurtosis (e.g. [119, 170]). A 14 dimensional PLP [66] feature vector is used to train a Multi-Layer-Perceptron (MLP) classifier with a 101 frame input layer, a 20 unit hidden layer and an output layer of two classes. Parameters are trained on 10 meetings from each meeting resource totalling around 20 hrs of data. Further 5 meetings from each corpus are used to determine early stopping of the parameter learning. The utterance segmentation uses Viterbi decoding and scaled likelihoods derived from the MLP and a minimum segment duration of 0.5 seconds.

Cross talk suppression is performed at the signal level using adaptive-LMS echo cancellation [105]. Additons to the basic system are: the use of multiple reference channels in cancellation; automatic channel delay estimation and offsetting of reference signals to account for this delay; automatic cross-talk level estimation; and ignoring of channels which produce low levels of cross-talk. Updates are further made on a per sample basis to account for non-stationary ‘echo’ path. On the classifier level additional features were introduced to aid the detection of cross-talk:

$$\begin{aligned}
\mathrm{RMS}_{norm}\left(x_{t-L}^{t+L}(i)\right) &= \log\left(\mathrm{RMS}\left(x_{t-L}^{t+L}(i)\right)\right) - \log\left(\sum_{j=1}^{N}\mathrm{RMS}\left(x_{t-L}^{t+L}(j)\right)\right), && (2)\\
Kur\left(x_{t-L}^{t+L}\right) &= \frac{E\left\{\left(x_{t-L}^{t+L} - E\left\{x_{t-L}^{t+L}\right\}\right)^4\right\}}{E\left\{\left(x_{t-L}^{t+L} - E\left\{x_{t-L}^{t+L}\right\}\right)^2\right\}^2}, && (3)\\
Cep\left(x_{t-L}^{t+L}\right) &= \max_{t=P_l-P_h}\left(\mathcal{F}\left(\log\left(\left|\ \mathcal{F}\left(x_{t-L}^{t+L}\right)\right|\ \right)\right)\right). && (4)
\end{aligned}$$

where x_{t-L}^{t+L} is the signal x windowed over $2 \cdot L$ samples and P_l and P_h are the minimum and maximum pitch period over which peak picking is carried out (corresponding to 50-300Hz). Eq. 2 describes across-meeting normalised RMS energy, Eq. 3 signal and spectrum kurtosis, and Eq. 4 as a voicing strength measure based on the maximum amplitude in the speech cepstrum in the range of frequencies 50-300Hz.

Microphone array processing Audio from multiple distant microphones (MDMs) can be used in variety of ways. The AMI baseline system uses an enhancement based approach. Recordings from a number of microphones placed in the meeting rooms are combined to arrive at a single, enhanced output file that is then used as input for recognition. The system is required to cope with a number of unknown variables: varying numbers of microphones; unknown microphone placement; unknown numbers of talkers; time variant skew between input channels introduced by the recording system; and different room geometry and acoustic conditions.

The MDM processing operates in a total of four stages. First gain calibration is performed by normalising the maximum amplitude level of each of the input files. Then a noise estimation and removal procedure is run. This in itself is a two pass process. On the first pass the noise spectrum $\Phi_{nn}(f)$ of each input channel is estimated as the noise power spectrum of the M lowest energy frames in the file ($M = 20$ was used for the current experiments). On the second pass a Wiener filter with transfer function $\frac{\Phi_{xx}(f)-\Phi_{nn}(f)}{\Phi_{xx}(f)}$ (where $\phi_{xx}(f)$ is the input signal spectrum) is applied to each channel to remove stationary noise. The noise coherence matrix Q, estimated over the M lowest energy frames, is also output at this time. In the third stage delay vectors between each channel pair are calculated for every frame in the input sample. The delay between two channels is the time difference between the arrival of the dominant sound source and is calculated by finding the peak in the Generalised Cross Correlation [86] between input frames across two channels. The delay vector is given as the delays for all pairs with respect to a single reference channel - there are therefore N delays in each vector, with the delay for the reference channel equal to 0. Further a vector of relative scaling factors is calculated, corresponding to the ratio of of frame energies between each channel and the reference channel. The start and end times in seconds, along with the delay and scaling factors are output for each frame. Finally The delay and scaling vectors are then used to calculate beamforming filters for each frame using the standard superdirective technique [36, 37]. The superdirective formulation requires knowledge of the noise coherence matrix. However this

Data	Bandwidth	Adaptation	#Iter	%WER
CTS	NB	-	-	33.3
ICSI	NB	-	-	27.1
ICSI	WB	-	-	25.3
ICSI	NB	MAP	1	26.5
ICSI	NB	MAP	8	25.8
ICSI	WB	MLLR + MAP	8	24.6
ALL	WB	MLLR + MAP	8	25.8

Table 8: %WER results on *icsidev*for several different training strategies and a trigram LM optimised for the ICSI corpus.

is not available as the microphone positions are not known. Either a unity coherence matrix may be used (leading to delay-sum filters) or the Q matrix estimate in the second stage may be used. Each frame is then beamformed using the appropriate filters and the output subsequently used for recognition.

Model building As outlined above the the fact that meeting resources are still comparatively small, bootstrapping from CTS models was used. However, as CTS data is only available at a bandwidth of 4kHz this poses additional questions on the initialisation and training procedure.

Bandwidth and Adaptation Table 8 shows recognition performance on the icsidev test set using various model training strategies. The baseline CTS systems yield a still reasonable error rate. Training on 8kHz-limited (NB) ICSI training data yields a WER of 27.1%. Using the full bandwidth (WB) reduces the WER by 1.8%. The standard approach for adaptation to large amounts of data is MAP [49]. As CTS is NB only, adaptation to WB ICSI data was performed using MAP adaptation in an iterative fashion. However the performance of the adapted NB system was still poorer than that of the system trained on WB data.

The results in table 8 show that MAP adaptation from CTS models while using wideband data is desirable. In our implementation the adaptation model set is used for two purposes: for computation of state level posteriors and to serve as a prior. Even if the former is performed well, NB models cannot be used to serve as prior directly. In order to overcome this problem the means of the CTS models were modified using block-diagonal MLLR transforms. One transform for speech and one for silence was estimated on the complete ICSI corpus using models trained on ICSI NB data. After an initial step with MLLR-adapted CTS models iterative MAP adaptation is resumed as before. The use of more detailed modelling of the transition from NB to WB by the use of more transforms was not found to yield a significant performance improvement. After 8 iterations a further 0.9% reduction in WER is obtained.

Meeting resource specific language modelling The language and vocabulary in meetings differs substantially. We have found evidence that his is also true for the acoustics [61]. However the advantage of having more data outweighs the differences. Hence we use acoustic models trained on the all meeting resources. Table 9 shows WER results using acoustic models trained on the complete meeting data and specific language models. An initial observation makes clear that on average the best strategy is to combine all the resources (similar to the acoustics). Further the variation of scores is modest whereby AMI data is distinct from all other resources. A moderate beneficial effect can be observed from using meeting room specific language models.

Independent Headset Microphone (IHM) Processing The sections above gave an outline of the components required for a baseline system on meeting transcription. The task of combining the

	TOT	ISL	ICSI	NIST	LDC
MRS ISL	40.2	44.7	25.8	34.1	53.8
MRS ICSI	40.2	45.2	25.1	34.7	53.5
MRS NIST	40.2	44.6	26.2	34.1	53.6
MRS AMI	41.0	45.1	26.9	35.8	54.2
COMBINED	40.0	44.5	25.6	34.4	53.4

Table 9: %WER on the *rt04eval* sets . TOT gives WERs overall, while MRS denotes the use of language models focusing on specific meeting rooms

System	CTS	VTLN	EC	TOT	F	M	ISL	ICSI	LDC	NIST
BASE	×			40.0	39.4	40.4	44.5	25.6	53.4	34.4
VTLN1	×	×		36.9	36.4	37.2	42.0	22.4	50.3	30.5
VTLN2		×		37.6	36.0	38.4	42.7	23.3	51.3	30.1
VTLN1 - SHLDA	×	×		36.0	35.1	36.5	41.0	21.8	50.5	27.4
EC1	×		×	40.3	39.5	40.7	44.7	25.9	54.8	33.1
VTLN-EC1	×	×	×	37.0	36.1	37.5	41.2	22.9	50.8	30.9
SEG1	×			50.8	51.1	50.6	50.4	38.2	73.3	37.4

Table 10: %WER on the rt04eval set using a combined tigram language model. CTS denotes CTS-adapted, EC echo cancellation.The table shows gender specific results (F/M) and results per meeting room . In the first section the reference segmentation of the data is used.

components in a sensible complex. For optimal performance many of the techniques cannot just simply be ”plugged” together.

Table 19 shows WER results using various model building techniques. Models are trained on a total of 96 hours of meeting speech. The baseline model yields 40% overall. By far the best performance is achieved on the ICSI portion of the data and performance is roughly gender balanced. Similar to CTS the use of VTLN yields a substantial improvement. Comparing the systems VTLN1 and VTLN2, the gain from CTS-adaptation remains even in conjunction with VTLN. The next part of the table shows the use of echo-cancelled (EC) data (as used for segmentation). Virtually no effect on recognition performance can be observed. The last section shows results with automatic segmentation (all other results are based on reference segmentation). The SEG1 system only makes use of the basic configuration, i.e. using an MLP only trained on PLP features.

MDM processing Almost all meeting corpora used a different approach to record speech with remote microphones. In the ICSI corpus microphones are not in fixed array configuration, the ISL corpus only uses one distant microphone, AMI uses a circular microphone array. Table 11 shows performance results with models trained on specific corpora. Overall the size and type of data used appears to have little impact on performance. Only the use of AMI training data appears to aid recognition on the AMI test set.

The enhancement based approach described in section 2.3.3 has the disadvantage that it cannot cope with overlapped speech. Since straight-forward removal of overlapping segments however would be far to restrictive. Instead word timings from forced alignment were used to identify overlaps. Speech segments were split, either at point of at least 100ms silence (ms10), of silence occurrence(ms0), or at arbitrary word boundaries (wb). These approaches reduce the original training set size of 96 hours to 56, 63 or 66 hours respectively. Table 12 shows associated WER results. Only a minor preference of an increase

	rt05seval					rt05samidev-n		
Combination	TOT	ISL	ICSI	LDC	NIST	TOT	UEDIN	IDIAP
ICSI,NIST	50.4	56.2	24.1	61.1	36.9	59.1	60.2	58.4
ICSI,NIST,ISL	50.6	56.2	22.9	61.8	37.2	59.1	60.0	57.6
ICSI,NIST,ISL,AMI	50.3	54.5	27.4	61.3	36.2	57.3	59.0	54.5

Table 11: %WER on *rt04seval* and *rt05seval*-AMI when training on various meeting resource combinations.

.

	rt05seval					rt05samidev-n		
System	TOT	CMU	ICSI	LDC	NIST	TOT	UEDIN	IDIAP
ms0	51.0	55.4	26.4	63.4	34.9	57.4	58.9	55.0
ms10	51.0	54.3	25.9	63.6	37.0	56.4	58.0	54.0
wb	50.7	56.5	24.3	61.9	36.4	56.3	58.2	53.4
VTLN - test - wb	50.4	55.9	26.3	61.7	34.7	-	-	-
VTLN - wb	47.2	51.4	20.6	60.2	31.3	-	-	-
wb icsiseg	55.2	59.5	32.2	66.7	40.5	-	-	-

Table 12: %WER on rt04seval and rt05samidev-n with different amounts of traiing data. ms0, ms10,and wb describe data preparation (see text).

in training set size is evident. However training set size has an impact on the effect of channel based normalisation schemes. Table 12 shows the performance after VTLN in both training and test, yielding improvements comparable to IHM.

Finally table 12 shows results for use of automatic segments as generated by the ICSI segmenter[146] which results in 5% absolute reduction in WER, mostly driven by an increase in the deletion rate. note that the greatest degradation was on the ICSI corpus.

2.3.4 NIST evaluation systems

This sections describes how the system components outlined above are combined into a single system and shows how the complete system performs on evaluation data.

System Architecture The system architecture overview presented in this section is generic to both the IHM and MDM systems. A more detailed description of system components is provided in the following section. The IHM and MDM systems differ only in the processing of the input audio and the use of input source specific acoustic models in the various processing stages.

The system operates in a total of 6 passes. Figure 1 shows a schematic representation of the processes. In the first pass (P1) the input data is segmented and transformed into a stream of 39 dimensional MF-PLP feature vectors[169]. Speech segments have a start and an end time as well as a channel/speaker label. A first recognition pass is conducted with acoustic models trained using maximum likelihood estimation (MLE) and a trigram LM (see Section 2.3.4). The resegmented output of this pass is used only for estimation of the vocal tract length normalisation (VTLN) warp factors on a per input channel basis. In the second pass (P2) the VTLN warp factors are determined and the audio data is recoded with these warp factors. Then a second decoding pass with acoustic models trained on VTLN data is performed. The P2 acoustic modelling includes a smoothed heteroscedastic linear discriminant analysis (SHLDA) input transform[22] and acoustic models are trained (in the IHM case) using the minimum

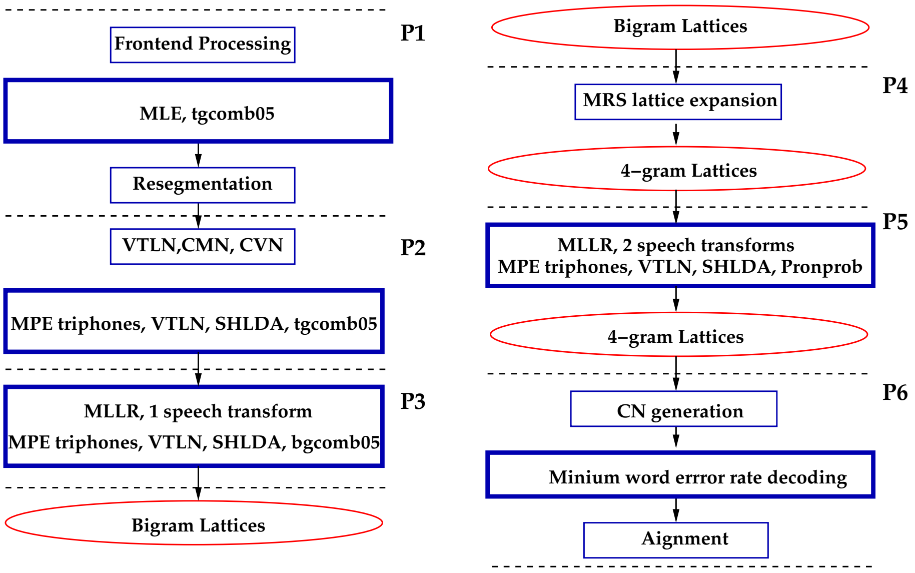


Figure 1: Processing stages of the 2005 AMI meeting transcription system.

phone error(MPE) criterion[122]. The output of P2 is used to adapt the acoustic model means and variances using maximum likelihood linear regression [45]. Two transforms, one for speech and one for silence are estimated. A third decoding pass (P3) uses MLLR adapted P2 models to generate bigram lattices. As all subsequent stages only process lattices to constrain the search space the use of a bigram in P3 avoids too harsh constraints.

In pass P4, the bigram lattices are first expanded using a trigram language model, followed by a second expansion using 4-gram LMs. For conference room data this expansion uses language models optimised for each meeting resource (MRS). The 4-gram lattices generated in P4 are used for rescoring in the following pass P5. Here models are adapted using up to two speech transforms using a regression class tree. Lattice rescoring further makes use of pronunciation probabilities estimated on the training data [63]. The output of this pass is a set of lattices which form the input to the final pass, P6. Here confusion networks [96] are formed and the most probable word from each confusion set is selected. The final output is then aligned using the P5 acoustic models.

IHM Front-end Processing Table 13 shows frame error rate results on the rt05seval before and after segmentation. Note that the relationship between false alarm and false reject rates differs substantially between meeting resources. The performance overall on the test data shows relatively high false reject rates. Smoothing the segment boundary estimates by padding allows to reduce the false reject rates significantly.

Acoustic Models Acoustic models are phonetic decision tree state clustered triphone models with standard left-to-right 3-state topology. Models are trained up to 16 mixture components using MLE with standard HTK [151] procedures and contain approximately 4000 states.

VTLN was applied both in training and testing, both on IHM and MDM. For training an iterative procedure was used alternating the estimation of warping factors and model parameter updates. For IHM initial warp factor estimates were obtained from CTS-adapted models. Experimental evidence shows improved WER performance with warp factor estimation at a reduced bandwidth of 3800Hz.

	AMI	ISL	ICSI	NIST	VT	TOT	TOT(REL)
RAW							
FA	1.29	1.52	0.71	1.49	3.70	1.64	2.00
FR	4.49	3.03	3.36	2.81	1.12	2.94	16.23
speech	24.40	28.84	13.79	15.56	14.83	18.12	
SMOOTHED							
FA	1.90	2.55	1.21	2.05	4.34	2.22	2.71
FR	3.80	2.01	2.71	2.18	0.83	2.30	12.69
speech	24.40	28.84	13.79	15.56	14.83	18.12	

Table 13: Segmentation performance (in %) on *rt05seval*. FA denotes false acceptance, FR false reject, and speech the percentage of speech in the reference. TOT gives the overall performance whereas TOT(REL) are relative to the associated class.

	TOT	Sub	Del	Ins	AMI	ISL	ICSI	NIST	VT
CTS adapted	39.1	20.0	13.4	5.7	39.9	35.1	36.0	46.9	37.6
CTS adapted, VTLN	36.9	18.5	13.0	5.5	37.0	33.1	34.4	45.2	34.8
VTLN	37.2	18.8	13.2	5.2	36.4	33.0	36.1	45.5	35.0
HLDA	35.7	17.8	13.4	4.6	36.0	31.0	33.9	43.3	34.6
SHLDA	35.6	17.7	13.3	4.5	35.6	30.3	34.5	42.8	34.7
SHLDA-MPE	32.9	15.8	13.3	3.8	32.8	27.8	32.3	39.8	31.9

Table 14: %WER on *rt05seval* IHM rescoring 4-gram lattices with pronunciation probabilities and various models. By default models are trained on meeting data only.

Initial experiments using IHM models for warp factor estimation on MDM data yielded a performance degradation. Hence IHM VTLN models were adapted to the MDM VTLN data where a single training iteration was found to yield good results that could not be improved further.

Feature space transformation was applied in the form of smoothed heteroscedastic linear discriminant analysis (SHLDA) [22]. The transform was used to reduce a 52 dimensional feature vector (standard plus third derivatives) to 39 dimensions. HLDA estimation procedure[88] requires the estimation of full covariance matrices per Gaussian. SHLDA in addition uses smoothing of the covariance estimates by interpolating with standard LDA type within-class covariance. The adaptation of CTS models when using SHLDA is non-trivial due to the reduced bandwidth of CTS data. To avoid further issues with discriminative training no CTS data was used in conjunction with SHLDA.

All further models were trained using the minimum phone error criterion [122]. The implementation of MPE used here is similar to that described in [122]. For this purpose numerator and denominator lattices were generated using the SHLDA models and a bigram LM interpolated with a unigram model that includes training set specific words. The phone times as obtained in recognition are used to improve speed in training. Only means and variances are modified and parameter update makes use of I-smoothing. Performance was found to stabilise after 10 training iterations[1].

Table 14 shows lattice rescoring results on rt05seval IHM for models of increasing complexity. Note the 0.3% performance degradation from the use of unadapted models which is compensated by 1.6% improvement from SHLDA. Another 2.8% absolute are gained by the use of MPE training. It can be observed that model improvement has little impact on the deletion rate. Table 15 shows equivalent

[1] Both SHLDA and MPE are developed as part of the STK HMM toolkit: `http://www.fit.vutbr.cz/speech/sw/stk.html`.

	TOT	Sub	Del	Ins	AMI	ISL	ICSI	NIST	VT
CTS adapted	32.2	21.3	6.8	4.1	31.7	31.6	25.3	38.2	34.7
CTS adapted, VTLN	30.2	19.9	6.4	3.9	29.2	30.4	23.5	36.9	31.2
VTLN	30.3	20.0	6.7	3.7	28.0	30.7	24.4	36.9	31.6
HLDA	28.7	18.7	7.0	2.9	27.0	27.7	23.3	34.8	31.1
SHLDA	28.7	18.8	7.0	2.9	26.7	27.5	23.7	34.4	31.7
SHLDA-MPE	25.8	16.7	6.9	2.3	24.1	24.4	20.8	32.4	27.6

Table 15: %WER on *rt05seval* IHM with manual segmentation (see Table 14).

Corpus	#words (MW)
Swbd/CHE	3.5
Fisher	10.5
Web (Swbd)	163
Web (fisher)	484
Web (fisher topics)	156
BBC - THISL	33
HUB4-LM96	152
SDR99-Newswire	39
Enron email	152
ICSI/ISL/NIST/AMI	1.5
Web (ICSI)	128
Web (AMI)	100
Web (CHIL)	70

Table 16: Size of various text corpora in million words (MW).

experiments conducted on manually segmented data. It is clear that the overall behaviour is similar, despite substantially lower error rates.

Vocabulary, Language Models and Dictionaries The recognition vocabulary is set to cover the 50000 most frequent words using a procedure outlined above. The same vocabulary was used both for lecture and conference room style meetings. Pronunciation probabilities are estimated from alignment of the training data[63].

As in previous work, LMs trained on a large number of corpora were used to derive meeting room specific and generic language models by optimisation of interpolation weights. The most important corpora are listed in Table 16. A full discussion of all source material would go beyond the scope of this paper. It is important to note that a collection of data from the web using tools and methods as provided by [18] was performed using both AMI and CHIL data as the basis. In both cases the proposed approach was altered to focus on previously unobserved contexts. This approach has in particular lead to a dramatic reduction in perplexity for lecture room data by more than 30%.

Table 17 shows perplexities for language models tuned to specific meeting resources as well as in combination. It is evident the meeting room specific models outperform the combined models. Hence the lattice expansion to 4-gram lattices (see Section 2.3.4) was performed using meeting resource specific models. This gave an additional 0.5% WER reduction on the *rt04seval* set.

	Language models					
Data source	ICSI	NIST	ISL	AMI	LDC	fgcomb05
ICSI	82.734	86.1662	87.3345	97.1024	109.86	84.1826
NIST	101.442	103.668	102.054	105.683	109.212	98.8722
ISL	110.124	110.99	106.66	119.327	114.483	108.588
AMI	92.9651	108.865	108.723	77.2817	101.714	84.1282
LDC	92.3824	92.761	87.6343	99.0105	84.2745	90.5354
AllDev	86.9236	93.2191	93.6604	92.0517	106.716	85.381

Table 17: Perplexities for 4-gram LMs on *rt04dev* and *rt05samidev*

CN decoding	Word time correction	IHM	MDM
		32.1	44.2
	×	31.2	42.2
×		31.5	44.0
×	×	30.6	42.0

Table 18: %WERs on rt05seval showing the effect of CN decoding. Word times are corrected by alignment.

Minimum Word Error Decoding Minimum word error rate decoding[96] is a widely used technique to counter the fact that the standard speech recognition objective function is to minimise sentence instead of word error rate which is the measurement metric. Table 18 compares the performance both on IHM and MDM. In both case the gain from this technique was found to be moderate. The table also shows the effect of correcting the word times by alignment. Standard decoding adds between-word silence to the end of a word, thus artificially lengthening words. Secondly, confusion network decoding uses heuristic rules to define word times. Hence again re-alignment is needed to correct the times.

Overall System Performance Table 19 shows WER results for the 2005 AMI meeting transcription system on a per pass basis. The result for P3 is higher than that for P2 due to the use of a bigram language model. The major reduction in WER at P6 can be explained by the use of alignment (see above). The high deletion rate is a main contributor to the error rate. Overall the WER reduction up to P6 is 10.5% absolute, however most of the gain is already obtained in P2.

The associated results on rt05seval MDM are shown in Table 20. Note that a similar improvement is obtained to that observed on IHM data, again with relatively high deletion rates. Particularly poor performance on VT data has a considerable impact on performance (only 2 distant microphones!).

Manual Segmentation In previous sections we have shown that automatic segmentation is still a main source of error.

Table 21 and 22 compare results with reference and automatic segmentation. Both on MDM and IHM the automatic segmentation naturally increases deletion rates, however the effect is far stronger on IHM where the overall difference between automatic and manual segmentation is 6.4%. The gain from confusion network decoding is further decreased with automatic segmentation. The absolute gain from P1 to P6 is similar in absolute terms, with or without manual segmentation.

2.3.5 Lecture Room Meetings

Lecture room meetings as included in the RT05s evaluations originate only from one recording site. Presentation sessions are mixed with question/answer meetings where more than one speaker talks. In

	TOT	Sub	Del	Ins	Fem	Male	AMI	ISL	ICSI	NIST	VT
P1	41.1	21.1	14.7	5.3	41.1	37.2	42.3	36.3	37.1	49.1	41.1
P2	33.1	15.9	13.4	3.9	33.1	28.2	33.4	27.2	32.8	39.5	32.8
P3	34.4	16.9	13.7	3.9	34.4	28.7	34.8	27.7	33.5	41.8	34.6
P4.tg	32.2	15.3	13.1	3.8	32.2	27.3	32.3	26.1	32.1	39.3	31.4
P4.fg	32.3	15.5	12.9	3.9	32.3	27.7	32.6	26.4	31.9	39.5	31.2
P5	32.1	15.3	12.8	4.0	32.1	27.4	32.7	26.3	31.8	39.1	30.5
P6	30.6	14.7	12.5	3.4	30.6	25.9	30.9	24.6	30.7	37.9	28.9

Table 19: %WER on rt05seval IHM.

	TOT	Sub	Del	Ins	Fem	Male	AMI	ISL	ICSI	NIST	VT
P1	53.6	32.1	17.3	4.1	53.6	56.4	46.5	50.2	48.2	53.6	63.0
P2	50.8	31.3	14.8	4.7	50.8	51.4	44.7	46.7	43.6	51.6	60.4
P3	50.4	31.1	14.6	4.7	50.4	53.0	44.7	47.0	45.2	48.9	59.7
P4.tg	48.4	30.0	13.6	4.8	48.4	49.4	43.9	44.8	42.5	46.9	57.2
P4.fg	47.9	29.5	13.7	4.7	47.9	49.3	42.4	45.0	41.8	47.4	56.6
P5	44.2	26.0	14.0	4.1	44.2	42.6	38.6	38.9	39.2	43.8	53.2
P6	42.0	25.5	13.0	3.5	42.0	42.0	35.1	37.1	38.4	41.5	51.1

Table 20: %WER on rt05seval MDM.

	IHM				MDM			
	ref		auto		ref		auto	
	TOT	Del	TOT	Del	TOT	Del	TOT	Del
P1	40.2	10.2	43.6	14.0	48.6	12.9	52.2	19.4
P2	33.6	11.2	37.5	14.8	44.8	13.9	49.0	17.2
P3	34.5	11.0	38.5	14.9	48.0	14.6	50.4	17.9
P4	30.1	9.4	34.5	13.5	44.4	13.6	46.7	15.8
P5	29.5	9.0	33.7	13.0	42.7	14.5	44.7	16.1
P6	29.0	9.2	33.2	11.6	41.7	14.4	43.1	15.6

Table 21: %WER on *rt04seval*

	IHM				MDM			
	refseg		autoseg		refseg		autoseg	
	TOT	Del	TOT	Del	TOT	Del	TOT	Del
P1	34.9	7.1	41.1	14.7	50.6	11.8	53.6	17.3
P2	26.0	7.1	33.1	13.4	46.4	11.4	50.8	14.8
P3	27.4	7.4	34.4	13.7	47.8	12.5	50.4	14.6
P4	24.5	6.4	32.3	12.9	45.1	11.5	47.9	13.7
P5	24.5	6.3	32.1	12.8	42.0	12.2	44.2	14.0
P6	24.2	6.4	30.6	12.5	40.7	12.3	42.0	13.0

Table 22: %WER summary for *rt05seval*

	IHM				MDM			
	TOT	Sub	Del	Ins	TOT	Sub	Del	Ins
P1	44.4	26.4	5.0	12.9	65.0	47.6	9.9	7.5
P2	33.0	19.1	5.2	8.7	60.0	43.4	10.0	6.7
P3	33.7	19.7	5.3	8.6	59.9	43.0	11.0	5.9
P4	31.4	18.2	4.8	8.3	58.8	42.2	10.1	6.5
P5	31.1	18.2	4.6	8.3	54.8	38.7	11.2	5.0
P6	30.4	17.7	4.6	8.0	53.5	37.2	11.6	4.7

Table 23: %WER on rt05slecteval.

this work no development work was performed due to lack of time. The system for conference room meetings was used as described except for language models optimised on the associated development data with additionally collected web-data. For MDM transcription only the four microphones on the table were used. Table 23 shows WERs both on IHM and MDM recordings. It is interesting to note that the WERs are in the same range as on lecture room data, however the overall gain of the passes is larger. Deletion rates are considerably lower on IHM compared to the results on conference room data.

2.3.6 Collecting Web-data for Language Modelling

The use of text retrieved from the internet for building domain specific language models has proven highly effective. Such data is collected by querying search engines for n-grams representative of the target domain. The choice of n-grams affects the quality of the language model built from the web-data directly. Considerations should include how common they occur in the target domain (T) and also how well represented they are in existing background corpora (B). A web-data collection (C) specific to a domain should represent T while minimising unnecessary overlap with B. We have developed a framework that gives rise to improved methods of selecting queries taking these points into consideration [166].

Experiments were performed on the AMI meeting corpus. The background text B consisted of 15M words from Switchboard, Fisher and ICSI meetings corpora. The T set consisted of 118 thousand words taken from a subset of the ES*a and ES*b recordings of the AMI corpus. The evaluation set E consisted of 90 thousand words from the corresponding ES*c recordings. Web-data collections (C) were obtained using the tools from the University of Washington [19] with some additional text normalisation to further improve the quality of the data.

The framework indicated several improved methods of choosing queries. Thus far, two have been investigated. One method is to rank all of the 4-grams in T using the log likelihood ratio of the two language models created from T and B and use the highest scoring 4-grams as queries. The other method of choosing queries is to find the most frequent 4-grams in T that do not occur in B. These two approaches were compared to a baseline which is to query for data using the most frequently occurring 4-grams in T irrespective of B.

Table 24 lists perplexity results in three sections. The first section gives the baseline perplexity on the evaluation set E using models constructed from B and T. The second section shows baseline results of a count based web-data collection obtained by searching for the 448 most frequent 4-grams of T irrespective of B (the least frequent having a count of 4), using one 4-gram per search. 5M words were collected, which, after normalisation, resulted in 3.9M words in C_{freq}. In the third section, 3.6M normalised words of web-data $C_{\not\subset B}$ were collected using the 432 most frequent 4-grams of T that were not found in B (the least frequent 4-gram query had a count of 2). The improvement is evident and shows that simply choosing the most frequent 4-grams is not the best approach.

Figure 3 is a histogram of the number of 4-grams that have a particular occurrence in T and a log likelihood ratio in a particular range. It shows how collecting C_{freq} is not optimal. There the least

Language model	PPL on E	Interpolation weights
B model	140.4	
T model	146.8	
B and T interp.	95.6	B=0.46; T=0.54
C_{freq} model	234.4	
B, C_{freq} interp.	119.1	B=0.70; C=0.30
B, T, C_{freq} interp.	91.3	B=0.35; T=0.51; C=0.14
$C_{\notin B}$ model	237.9	
B, $C_{\notin B}$ interp.	114.9	B=0.68; C=0.32
B, T, $C_{\notin B}$ interp.	90.4	B=0.35; T=0.50; C=0.15

Table 24: Perplexities on E of baseline language models and models derived from count based approaches.

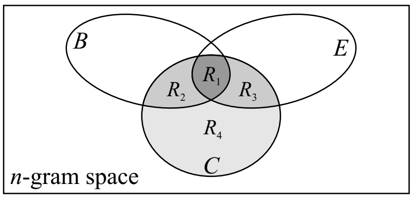


Figure 2: Partitioning the n-gram space

frequent 4-gram queried had a count of 4. It can be seen from figure 2b that some of those queried 4-grams have negative log-likelihood ratios. This violates one of the conditions for improvement in the framework and, therefore, may affect the resulting web-data language model negatively.

Table 25 shows the results of using the log likelihood ratio (**??**) for choosing search queries. All n-grams in T were grouped by their log likelihood ratio according to the ranges shown in column 1 of the table and separate collections where made for each range using single 4-gram queries. For comparison purposes, the sizes of the (normalised) collections were limited to a maximum of about 4M words. 4-grams with ratios greater than 12 were also queried but they returned very little data. The distribution of the queries across each log likelihood range can be inferred from figure 3.

Table 25, column 3 shows the change in perplexity after each web-data model is interpolated with the background model. It indicates that, for AMI data, there is a "sweet spot" in the log likelihood ratio ranges between 2 and 6, each of which have lower perplexities than the C_{freq} collection. The framework indicated that queries based on 4-grams with higher likelihood ratios should yield better models than lower ratio queries. However, this is true only up to a point as higher likelihood ratio n-grams generally occur less frequently so the gain achieved by boosting them is lessened. This is clearly shown in figure 3. The result may also be partly related to the relatively small number of words returned by searches for high likelihood ratio queries: it is not uncommon for high ratio queries to return no results.

The mass distribution columns of table 25 show the proportions of the web-data intersecting with E and B. The regions R_1 to R_4 are defined with the aid of figure 2. The results of $(R_1 + R_3)$ show that only a small proportion of the collected web-data actually intersects with E: approximately 1% of the 4-gram mass and between 7% and 9% of the 3-gram mass. This small percentage overlap is unsurprising as E is only small but it may also indicate that some additional filtering of the web-data may be necessary. Interestingly, there is a substantial difference between the percentage overlap of 3-grams and of 4-grams in R_1 and R_2. It suggests that querying for 4-grams actually returns many more topic relevant 3-grams.

LLR range	Number of words collected	ppl on E after interp. with B	Overall interp. weight	% 4-gram mass distribution of collected data				% 3-gram mass distribution of collected data			
				R_1	R_2	R_3	R_4	R_1	R_2	R_3	R_4
B	–	140.4	0.610								
1 – 2	3,763,221	122.0	0.103	1.37	13.26	0.06	85.31	8.85	30.59	0.12	60.44
2 – 3	3,988,426	117.7	0.081	1.28	12.26	0.08	86.38	8.47	29.23	0.19	62.11
3 – 4	4,004,998	118.3	0.049	1.14	11.62	0.08	87.16	7.86	28.69	0.18	63.26
4 – 5	3,785,525	117.4	0.047	1.14	11.12	0.08	87.66	7.73	27.81	0.21	64.25
5 – 6	2,638,330	118.2	0.042	1.04	10.38	0.09	88.49	7.33	26.97	0.24	65.45
6 – 7	1,512,204	121.0	0.022	1.03	10.52	0.09	88.36	7.26	27.05	0.25	65.44
7 – 8	887,684	123.3	0.014	1.23	11.03	0.10	87.64	7.85	27.39	0.23	64.53
8 – 9	410,533	124.2	0.021	1.07	10.59	0.09	88.25	7.36	26.85	0.27	65.52
9 – 10	303,744	131.5	0.001	0.94	10.51	0.07	88.48	6.96	27.42	0.19	65.43
10 – 11	133,526	136.4	0.000	0.89	10.40	0.07	88.64	6.90	27.64	0.26	65.20
11 – 12	71,767	130.0	0.010	0.92	8.87	0.10	90.11	6.66	24.78	0.31	68.25

Table 25: Results of collecting web-data according to log likelihood ratio value

The 3-gram overlap in R_2, which shows the overlap between the web-data and B, is enormous and indicates that a general search will just retrieve a lot of B material. The factor seven difference (between 3 and 4-grams) in the amount of overlap in R_1 is especially interesting as it relates to the likelihood ratio directly and suggests that it may actually be easier find in-domain material by searching for 3-grams.

Note that in many cases it is possible to achieve more significant perplexity reductions by collecting more web-data. For example, it is possible to achieve a perplexity of 115 by collecting 7.5M words for the log likelihood range 2 – 3. The final perplexity obtained by interpolating all the models of table 25 using the weights in column 4 was 109.7. Curiously, the interpolation weights tell a different story from the perplexity numbers: the weights decrease steadily as the likelihood ratio increases giving no indication of the "sweet spot" mentioned above.

2.3.7 Speech decoder - Juicer

In order to provide a platform for large vocabulary speech recognition research, the Juicer speech recognition decoder has been under development at IDIAP, with further input from partner institutions. Juicer is a decoder for HMM-based large vocabulary speech recognition that uses a weighted finite state transducer (WFST) representation of the search space. The major advantages of the WFST-based approach is the decoupling of the decoding network from the decoding engine, as well as a common representation of the various ASR knowledge sources allowing standardised techniques to be used for constructing a complete, optimised search space. These characteristics allow straightforward incorporation of new capabilities, such as decoding with a custom grammar or non-standard lexical constraints, without requiring modification of the decoding engine. Presently, Juicer is ready for (beta) release to AMI partners. The Juicer package consists of a number of command line utilities: the Juicer decoder itself, along with a number of tools and scripts that are used to combine the various ASR knowledge sources (language model, pronunciation dictionary, acoustic models) into a single, optimised WFST that is input to the decoder. The following functionalists are currently supported:

Language modelling Simple word loop, word-pair and N-gram language models (subject to memory restrictions, see below)

Acoustic modelling Monophone, word-internal n-phone, cross-word triphone, HTK MMF file format

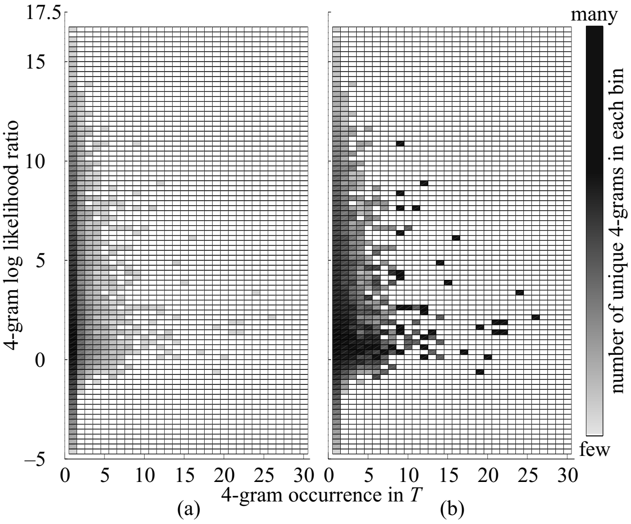


Figure 3: A 3D histogram of the number of unique 4-grams with a certain count in T and have a log likelihood ratio within a certain range: (a) is the unnormalised histogram showing that the vast majority of 4-grams occur once; (b) has the histogram peaks in each column normalised to the same height to make the more frequent 4-grams visible.

Dictionary Multiple pronunciation with pronunciation probabilities

Decoder Fixed WFST search with beam-search pruning, histogram pruning, lattice generation (only in AT&T format), word and model level output with times

Aside from providing greater compatibility with the current set of models and tools used in the AMI ASR system (based around the HTK speech recognition software and HDecode LVCSR decoder), future development of juicer is aimed at improving the performance with respect to speed and memory efficiency. In particular, early benchmarking of Juicer has revealed a major performance restriction in terms of memory requirements for the composition of large static networks, such as is required for the AMI system. In light of this, the current solution is to use heavily pruned language models which greatly reduce network size (resulting in an increase to WER). Table 26 shows the performance of Juicer on two tasks, for the first of which language model pruning was not necessary and the second where heavy pruning was required. We see that in the first task the accuracy exceeds that of the benchmark decoder whereas in the second accuracy has has to be sacrificed.

There is an on-going effort to overcome the memory limitation caused by static transducer composition. Instead of composing a unified transducer statically, the transducer is partially constructed on-the-fly. Such an approaches relies on splitting the unified transducer into two parts. One part comprises an optimal composition of the lexicon transducer and the context-dependent transducer, which is built statically before decoding. Another part comprises the language model, which is combined with the static part dynamically during decoding. This can lead to performance loss when beam-search pruning is employed during decoding, since the dynamic composition leads to a sub-optimal transducer composition.

2.3.8 Recognition Experiments using the MC-WSJ-AV corpus

Two beamforming approaches applied to the microphone array processing ASR front end were studied. The first approach relies on knowledge of the recording environment and array geometry (fixed

Task		Language model	WER	
			Juicer	Benchmark
WSJ	5k	bigram, unpruned	19.3	20.3
	20k	bigram, unpruned	18.7	21.2
RT05s	P1	trigram, pruned	43.6	41.1
	P2	trigram, pruned	34.5	33.1

Table 26: Performance of Juicer and benchmark systems on tasks: Wall Street Journal (WSJ) and NIST Rich Transcription 2005 Spring Meeting Recognition Evaluation (RT05s). The benchmark decoders for these tasks were respectively *HVite* and *HDecode*, both developed at Cambridge University Engineering Department.

Channel	No adaptation	Channel adaptation	Speaker Dependant adaptation
Headset	14.8	14.0	12.3
Lapel	26.3	20.2	18.0
Fixed Beamformer	48.6	35.6	28.1
Blind Beamformer	55.2	36.5	31.6
SDM	87.6	73.3	66.5

Table 27: % Word error rates for the 5k closed vocabulary task using stationary speaker MC-WSJ-AV data from the UEDIN room.

beamformer), while the second is a completely automatic system, requiring no knowledge of the speaker location or microphone placement (blind beamformer). The blind beamformer was used as as ASR front-end for the multiple distant microphone (MDM) condition in the Spring 2005 NIST Rich Transcription evaluation.

Recognition experiments were carried out on the UEDIN component of the corpus using data from headset and lapel microphones, two beamforming techniques and a single distant microphone. The recognition results showed that, when channel adaptation is applied to the acoustic models, the array techniques provide recognition accuracies far superior to those obtained using a single distant microphone. Also, it has been observed that if information about the room layout is available, this can be used to estimate the beamformer filters providing small improvements in accuracy over blind estimation of the filters. The results of these studies are shown in Table 27 .

Currently, the integration of audio-visual person tracking with microphone array ASR front end is under investigationinvestigated using the IDIAP component the corpus. This study intends to compare the ASR performance of audio only localisation estimates with the output of audio-visual tracker. The audio observations are based on sector-based localisation algorithm in which the space is divided into sectors and each sector is characterised by sector activity measure. High sector activity indicates the likelihood of the active speaker, from which the candidate 3-D estimates are derived. Video observations are based on shape and structure of human head.

2.4 Conclusion and summary

Work on automatic speech recognition in the AMI project was described in the previous sections. In particular we have discussed:

- Our evaluation framework
  Here we have shown how internal and external (public) evaluations of our systems are conducted.

- Data used in experiments
  Here we have defined the corpora and the sets used in training and testing for general and specific tasks.

- A complete ASR system
  Here we have discussed system architecture as well as individual system components and we have evaluated these components on our standard test sets.

- Research results on specific aspects.
  Here we have discussed subjects that are being evaluated internally in the current development cycle and have not yet found their way into the complete system.

Given the results and our experiences with the above we conclude that

- We have defined an evaluation framework that is generic, flexible, comparable, and that allows us to conduct research and development in a stable environment.

- We have built a system that is very competitive and performs exceptionally well on AMI data.

- We can focus our attention on extending our work to the full AMI corpus and the specific problems to be faced there.

- we have a research infrastructure that allows all sites to work on small subtasks without the need to build large and labour-intensive systems.

Our research results also give a better understanding of many interesting questions such as the distant microphone problem, the segmentation problem the use of language, or the presence of many accents. Our investigations highlight where major improvements in performance can be obtained. Finally it is worth mentioning that all of this was achieved by 6 research sites in Europe and the U.S in very close collaboration.

3 Speech related tasks

3.1 Keyword spotting

3.1.1 Overview

Keyword spotting (KWS) is an important technique for fast access to relevant information in meetings. We compare 3 approaches to KWS: acoustic keyword spotting, spotting in word lattices generated by large vocabulary continuous speech recognition (LVCSR) and a hybrid approach making use of phoneme lattices generated by a phoneme recognizer. Systems are compared on carefully defined test data extracted from ICSI meeting database. The advantages and drawbacks of different approaches are discussed. For acoustic KWS, we also propose a posterior masking algorithm to speed-up acoustic keyword spotting.

3.1.2 Introduction, Data and Pre-processors

The goal for Keyword Spotting (KWS) is to find the keyword and its in speech data including its position. All three mentioned techniques are based on a comparison of two likelihoods: 1) that of the keyword and 2) likelihood of a background model (Fig. 4).

KWS systems were evaluated on data carefully selected from ICSI meeting database (17 hours). Four keyword sets were defined:

- set **Test 17** contains 17 most frequent words with 33 pronunciation variants in total.
- set **Test 1** with 2310 words and 3514 variants contains rare words appearing just 1× in the test set.
- set **Test 5** with 4104 and 6537 variants contains rare words appearing at most 5×.
- set **Test 10** with 4710 words and 7567 variants covering rare words appearing at most 10×.

All systems were evaluated using standard Figure-of-Merit (FOM) measure defined by NIST.

We used 2 pre-processors (or "speech recognizers"). Our **phoneme recognizer** [137] (TRAPNN-LCRC) is based on temporal patterns (TRAPs) and neural networks (NN) with split left and right contexts (hence the name LC-RC). It produces a matrix of phoneme-state posterior probabilities with 3 states per phoneme. Theses posterior matrices are used as features for acoustic KWS and for phoneme lattices generation.

The Word (LVCSR) lattices are generated using **standard LVCSR GMM/HMM** based system. Both systems for lattices are trained on 10h of ICSI meetings, acoustic KWS system is trained on 40h of ICSI meetings.

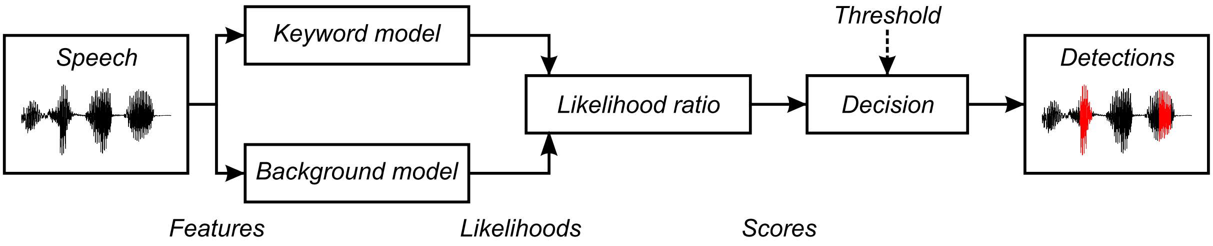


Figure 4: General scheme of KWS.

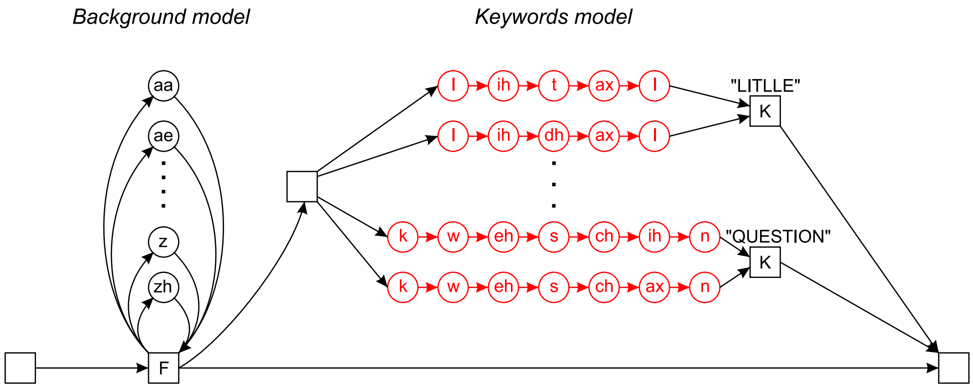


Figure 5: Acoustic KWS.

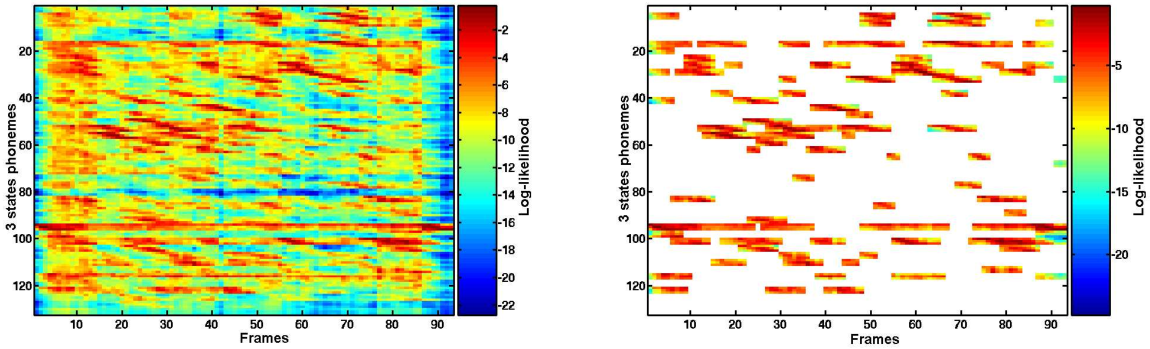


Figure 6: Original and masked phoneme-state posterior matrix.

3.1.3 Comparison of KWS approaches

In the **acoustic approach**, the models of keywords are assembled from phoneme models and run against a background model (a simple phoneme loop) according to Fig. 5. The difference of 2 log likelihoods at the outputs of these models forms the score. In acoustic KWS, it is advantageous to pre-generate the phoneme-state posteriors. The actual decoding is then very fast. We have further accelerated the decoding by pruning the phoneme-state posterior matrices (Fig. 6) by "masking" them using phoneme lattices discussed below. Using this approach, the decoding runs about 0.01×RT on an average Pentium IV machine.

The **KWS in LVCSR lattices** "greps" the keywords in lattices generated by LVCSR system. The confidence of each keyword is given by the difference of log-likelihood of the path on which the keyword lays and log-likelihood of the optimal path.

The **KWS in phoneme lattices** is a hybrid approach. First, phoneme lattices are generated. This is in fact equivalent to narrowing the acoustic search space. The phonetic form of keyword is then "grepped" in such lattices and the confidence of keywords is given by the acoustic likelihoods of individual phonemes again normalized by the optimal path in the lattice.

Table 28 presents results of the mentioned approached on the four test-sets. The Eurospeech paper [148] contains detailed description of different systems, features, comparison of neural networks and GMMs in acoustic KWS etc.

Test set	Acoustic	Word lattice	Phoneme lattice
Test 17	64.46	**66.95**	60.03
Test 10	72.49	66.37	64.1
Test 5	74.11	64.71	65.0
Test 1	**74.95**	61.33	69.3

Table 28: Comparison of FOM [%] of three KWS approaches on different test-sets.

3.1.4 Discussion and future work

The results summarized in Table 28 confirmed our previous assumptions about the advantages and drawbacks of different approaches:

- The **LVCSR-KWS** is fast (lattices can be efficiently indexed) and accurate, however only for common words. We see a clear degradation of performance for sets Test1/5/10. We should take into account that less common words (such as technical terms and proper names) carry most of the information and are likely to be searched by the users. LVCSR-KWS has therefore to be completed by a method unconstrained by recognition vocabulary.

- **acoustic KWS** is relatively precise (the precision increases with the length of the keyword) and any word can be searched provided its phonetic form can be estimated. This approach is ideal for on-line KWS in remote meeting assistants, but even with the mentioned high speed of 0.01×RT, it is not suitable for browsing *huge* archives, as it needs to process all the acoustic (or at least posterior probabilities) data.

- **phoneme lattice KWS** is a reasonable compromise in terms of accuracy and speed. Currently, our work on indexing phoneme lattices using tri-phoneme sequences is advancing and preliminary results show good accuracy/speed trade-off for rare words.

Our future work will concentrate on:

- improving the core technologies: in phoneme recognition, using advanced architectures defined in [137], in LVCSR, using the state-of-the-art AMI LVCSR system [64].

- definition of test set on AMI data which is now available including the transcriptions.

- efficient indexation and retrieval from word and phonetic lattices (this work is extending to WP5).

- enhancing the accuracy of acoustic and phonetic KWS by including basic semantics in the queries (such as specifying if the string searched is a noun, adjective, etc.) – preliminary work done by AMI trainee Gaurav Pandey [116] showed promising results.

- unified indexing of LVCSR and phonetic lattices and integration of the three approaches

- upgrade of our demo integrated with JFerret meeting browser to support also phoneme-lattice search.

3.2 Language identification

3.2.1 Overview

The task of language identification (LID) is to detect the language a particular speech segment was spoken. This technology is used for example for routing calls in call-centers and for emergency numbers.

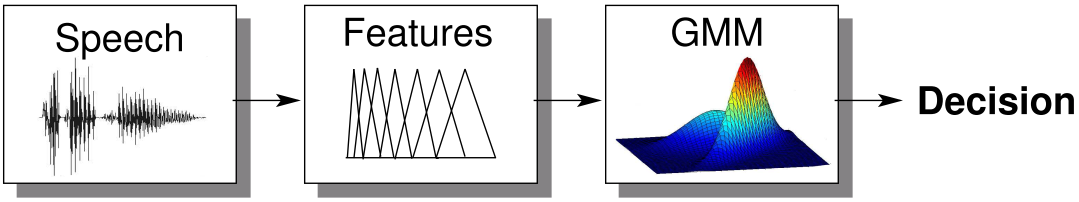


Figure 7: Acoustic LID.

Within AMI, the LID can be used for example to detect the language of back-channel speech in meetings and routing it to appropriate speech recognizer. Security is another main application domain for LID.

The group at BUT has developed two systems for LID: acoustic and phonotactic. They were thoroughly tested on data from NIST 2003 LRE evaluations. BUT took also part in 2005 evaluations, where it recorded an important success.

3.2.2 Evaluations and data

National Institute of Standard and Technology (NIST) regularly organizes language recognition evaluations, the goals of which are to establish a current baseline of performance capability for language and dialect recognition of conversational telephone speech using uniform evaluation procedure. Task is the detection of a given target language or dialect. Given a test segment of speech and a target or dialect, the system must determine whether or not the speech is from the target language or dialect.

The test segments in NIST 2005 LRE evaluations contained three nominal durations of speech: 3 seconds, 10 seconds, and 30 seconds from a set of 7 languages and two dialects (English-American, English-Indian, Hindi, Japanese, Korea, Mandarin-Mainland, Mandarin-Taiwan, Spanish and Tamil). Actual speech durations varied but were constrained to be within the ranges of 2-4 seconds, 7-13 seconds, and 25-35 seconds of actual speech contained in segments, respectively. The performance is evaluated separately for test segments of each duration, considering each system is a language *detector* rather than recognizer. A standard detection error trade-off (DET) curve is evaluated as a plot of probability of false alarms against the probability of misses with the detection threshold as parameter and equal priors for target and non-target languages. Equal error rate (EER) is the point where these probabilities are equal. The total EER of the whole LID system is the average of language-dependent EERs.

Training data *Phoneme recognizers* were trained on Hungarian, Russian and Czech SpeechDat-East. *Phonotactic language models* and *acoustic models* were trained on the CallFriend database containing telephone speech of 15 different languages or dialects (each contains 20 complete half-hour conversations) and OGI multilingual and OGI 22 languages corpora. Only seven target languages were used for building models, other languages were used for training out-of-set model. *Development Data* comes from NIST 1996 and 2003 LRE plus 40 additional segments from OGI Foreign Accented English database (Hindi part) to compensate for the lack of English accented by Indian.

3.2.3 LID systems

Acoustic LID (Fig. 7) determines the language directly on the basis of features derived from the speech signal. This approach can for example well separate between French and English - in the former, the nasal cavity is more frequently open which is directly translated into speech features. For the NIST evaluation, BUT researchers improved the existing technologies by adding discriminative training of acoustic models [23].

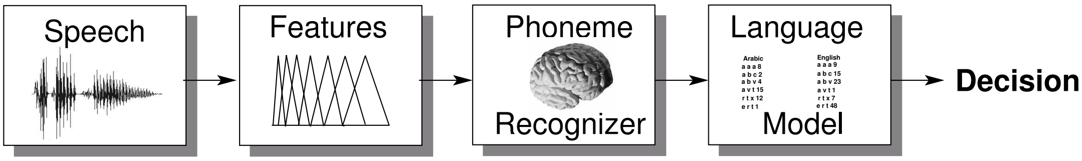


Figure 8: Phonotactic LID.

System EER [%]	Duration		
	30sec	10 sec	3 sec
PRLM+lattice	9.83	14.93	23.44
PRLM+lattice+anti.m.	8.49	13.82	22.98
PPRLM+lattice+anti.m.	7.30	11.38	19.46
GMM-MMI 256	4.63	8.61	17.23
Fusion	**3.09**	**6.54**	**14.14**

Table 29: EER for different system for NIST LRE 2005 for 30sec condition

In **Phonotactic LID** (Fig. 8) speech is first transcribed by phoneme recognizer into strings or graphs (lattices) of phonemes. On these, "language" models are trained to capture statistics of couples and triples of phonemes. In this way, German and English can be for example separated based on different statistics of "und" and "and". Speech@FIT group pioneered the use of so called "anti-models" for this task [98]. It is also advantageous not to use only phoneme strings but phoneme lattices. This system is denoted as PRLM (Phoneme recognizer followed by language model) or PPRLM (Parallel connection of phoneme recognizers followed by language models).

3.2.4 Results on NIST 2006 LRE data

Table 29 shows the results of separate systems for LRE 2005 for all three test conditions. The presented PRLM system is based on Hungarian phoneme recognizer and is the best out of our 3 PRLM systems. On 30s segments, PRLM with trigram language model derived from lattices and tested using lattices performs with EER=9.83%. Using antimodels resulted in relative improvement of 14%. Fusion of three phonotactic systems (PPRLM) performed with EER=7.30% which is also almost 14% relative improvement.

For GMM model, only segments labeled 'speech' by Czech phoneme recognizer were used. We used 256-component GMM trained under Maximum Mutual Information framework. In our previous work this system proved its superiority over the state-of-the-art highly dimensional (2048-component) GMM trained under conventional Maximum Likelihood framework. The EER of GMM-MMI 256 system is 4.63%.

Fusion of GMM-MMI and PRLM system gives 33% relative improvement over the best separate system and the final EER reaches 3.09%. Unfortunately, this is a post-evaluation results – due to bad fusion weights, the EER of our submitted 30 sec system was 5.01% (even worse than GMM-MMI 256 itself). The second and third columns in Table 29 show results for 10 sec and 3 sec durations (here, the table contains the evaluation results).

3.2.5 Discussion and future work

Among systems from 13 academic and industrial labs present in 2005 evaluations, BUT system scored the 1st in 10s and 3s test segment durations and was the 2nd in the 30s category. Especially the discriminative training brought substantial improvement over the standard ML scheme. In the phonotactic approach,

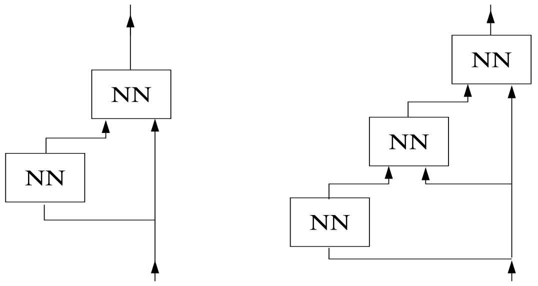


Figure 9: Tandem architectures of NNs.

good performance obtained on NIST 2003 did not fully generalize and we are currently analyzing the results and performing post-evaluation runs of our system. In future, we plan to apply at least some of our successful technologies in other domains such as speaker verification, which could be also advantageously used in off-line and on-line technologies to process meetings.

3.3 Phoneme recognition

Phoneme recognition plays very important role in speech processing and is the underlying technology of many "application" algorithms. Phoneme strings are basic representation for automatic language recognition and it is proved that language recognition results are highly correlated with phoneme recognition results [99]. Phoneme posteriors are useful representation for acoustic keyword search, they contain enough information to distinguish among all words and they are small enough to store compared for example to the size of posteriors from context dependent Gaussian Mixture Models (GMM) [148]. Another usable representation for acoustic keyword search are phoneme lattices that can be also generated by phoneme recognizer. Phoneme recognition can also improve speaker recognition.

In 2005, we have conducted an exhaustive study of different phoneme recognition architectures with the aim to compare multiple hierarchical structures of neural networks for phoneme recognition and to emphasize approaches to reach lower recognition error rates. The results were obtained on standard TIMIT database.

The experimental setups and results are discussed in detail in [137]; the main improvement was achieved by 2 approaches:

1. use of **split temporal context (STC)** architectures, where the temporal trajectory of a feature is divided into several parts, each is processed independently by a neural net (NN) and another NN serves as a merger.

2. use of **hierarchical structures of NNs** (Fig. 9) where one NN pre-estimates posteriors of target classes (phonemes or phoneme-states) and the second network is fed by these posteriors and by the original features.

Table 30 summarizes the results of our system on TIMIT database in terms of phoneme error rate (PER) for different time and/or frequency split architectures and Table 31 shows the improvement while using the hierarchical structures of NNs (Fig. 9) on the left-block net from a two-block STC architecture.

3.3.1 Discussion

We have carried out several experiments concerning architectures of neural nets used for phoneme recognition. The main message should be "adding the most of knowledge about what we want to recognize in all levels (features, output, architecture) is necessary to obtain good results". We have compared TRAPs

system	1 state	3 states
3 band TRAPs	29.24	25.78
5 band TRAPs	-	24.84
STC - 2 blocks	28.47	24.41
STC - 5 blocks	-	**23.44**
2 x 2	-	24.06

Table 30: PER for different time and/or frequency split architectures.

# nets	1	2	3	3 ext.
PER (%)	31.64	30.63	30.33	**29.66**

Table 31: Tandem of neural networks

and split temporal context (STC) systems and concluded the later offer better results. We have also experimented with tandem-NN architectures. Preliminary results show that using one net to "focus" another net on features is advantageous, though this approach needs more experiments.

At the end, we have tuned the five-block STC system by increasing the sizes of neural nets and modifying the training algorithm. The resulting PER was 21.48% which is a very competitive result. In order to be able to compare to similar published works, phoneme classification task was run (running our system with fixed phoneme boundaries); the resulting classification error rate was 17.19%.

4 Localization and Tracking

4.1 Objectives

The main target of the localization and tracking subgroup is to provide information about persons visible in the video sequence. Within the AMI context, this information, comprising data like the number of visible persons, person identifiers and head positions, serves as a basic input for various subsequent meeting analysis tasks like pose estimation as well as the recognition of identities, meeting gestures, emotions or actions. The methods employed for person detection and tracking have to be robust against real-world conditions present in meetings, like variations in object appearance and pose due to natural unrestricted motion and changing lighting conditions, and the presence of multiple self-occluding objects. Many of the above issues still represent research challenges. At the same time, the methods have to be efficient in computational terms, given their extensive use by later recognition tasks.

In this subgroup a number of novel statistical models and algorithms to deal with the above issues have been investigated and implemented. Specific research issues involved the devising of reliable yet tractable object models, the design of statistical methods for visual data fusion, both from multiple sensors (cameras) and multiple visual modalities (appearance, shape, motion), and the development of efficient mechanisms for inference in models that involve multiple people.

To compare the tracking methods based on an objective quality measure, further efforts had to be spent for the definition of a common evaluation scheme.

4.2 Sidecorpus AV16.7

The AV16.7.ami corpus was specifically collected as spoke data in the AMI project to evaluate localization and tracking algorithms. The corpus, consisting of 16 sequences, each with a duration in the range of 1-4 minutes, was collected at IDIAP, and includes sequences ranging between one and four people. The sequences depict phenomena that are both of particular challenge for tracking methods (e.g. person occlusion, camera occlusion by people passing, partial views of head backs when people sit) and occurring in reality in the context of meetings (e.g. sitting down, discussing around the table, etc.). The corpus was recorded according to a predefined agenda (i.e., the order in which people would enter the room, pass, sit, etc), but the behavior of the subjects is otherwise natural (e.g. people were not asked to look at the camera, or to stand adopting a specific body pose). In addition, one full meeting from the hub corpus (IS1008b) was also been added to the evaluation corpus. This meeting was selected due to the fact that it presented higher dynamics compared to other hub meetings, and was thus more challenging for tracking algorithms.

The evaluation corpus (e.g. the AV16.7.ami and the meeting from the hub corpus) was manually annotated at IDIAP for head and mouth locations using the annotation software described at [90], and developed in the context of the project.

4.3 Evaluation scheme for Tracking in AMI

4.3.1 Introduction

Since a number of tracking algorithms for AMI meeting scenarios is developed at several institutes, there is a certain necessity to agree on a common scheme to evaluate the performance of the different approaches. In the following paragraph a fundamental concept based on [139] for such a scheme is introduced, defining how to evaluate multiple object tracking for unknown configurations.

4.3.2 Coverage test

In order to determine the quality of a tracking result for a single object, we introduce two shape-independent measures, indicating if a ground truth object is being tracked and which $\mathcal{E}_i$ is connected to which $\mathcal{GT}_j$:

$$\begin{array}{ll} \text{Recall} & \alpha_{i,j} = \frac{|\mathcal{E}_i \cap \mathcal{GT}_j|}{|\mathcal{GT}_j|} \\ \text{Precision} & \beta_{i,j} = \frac{|\mathcal{E}_i \cap \mathcal{GT}_j|}{|\mathcal{E}_i|} \end{array}$$

While the first measurement (recall) represents the ratio of the ground truth area, which is covered by the estimate, the precision embodies the ratio of the estimate area covered by the ground truth. As it can be shown very easily, both α and β must be high to obtain good tracking results. For this reason, a coverage test using the F-measure [129]

$$F_{i,j} = \frac{2\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j} + \beta_{i,j}} \tag{5}$$

has to be passed, returning only a high value if $\alpha_{i,j}$ and $\beta_{i,j}$ are high. This test is considered to be passed, if $F_{i,j}$ exceeds a fixed threshold t_c and thus determines, that $\mathcal{GT}_j$ is being tracked by $\mathcal{E}_i$.

4.3.3 Configuration test

To facilitate the explanations in the following sections some definitions will be introduced at first. In this document labeled tracking targets are denoted as ground truth objects $\mathcal{GT}$, tracker outputs are referred to as estimates $\mathcal{E}$. The output of a tracking approach is considered to be correct, if and only if one $\mathcal{GT}$ (resp. $\mathcal{E}$) is tracking exactly one $\mathcal{GT}$ (resp. $\mathcal{E}$). In the following sections there will be defined what kind of errors arise and how they can be detected.

4.3.4 Configuration error measures

In this context, configuration means the number, the location and the size of all objects in a frame of the scenario. According to the above definition of a correct tracker output, a configuration error occurs if the size or the location of a certain $\mathcal{E}_i$ and its related $\mathcal{GT}_j$ do not match. To identify all types of errors that may occur, 4 configuration measures are introduced:

a) **Measure** FP - False positive. There is an $\mathcal{E}$ indicating an object, where no $\mathcal{GT}$ is.

b) **Measure** FN - False negative. A $\mathcal{GT}$ is not tracked by an $\mathcal{E}$.

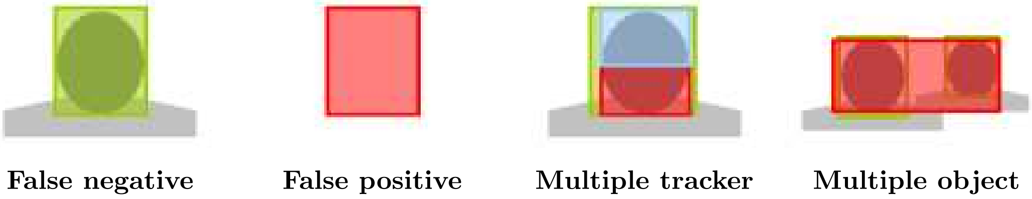
c) **Measure** MT - Multiple trackers. More than one $\mathcal{E}$ is associated with only one $\mathcal{GT}$. In order to obtain the subjective impression of a human spectator each excess $\mathcal{E}$ is counted as a MT error.

d) **Measure** MO - Multiple objects. More than one $\mathcal{GT}$ is associated with only one $\mathcal{GT}$. Again a MO error is assigned for each excess $\mathcal{E}$.

For each of these errors above an example is depicted in Fig. 10, where the ground truth is marked with green, the estimates with red resp. blue colored boxes.

4.3.5 Occlusion handling

Situations with occlusion will be treated in a special manner, since MO or MT errors might occur although the estimates are correctly placed. For this reason ground truth labels are enlarged by an additional flag occ_j indicating an occlusion in the image data. This flag is defined for each object and



False negative **False positive** **Multiple tracker** **Multiple object**

Figure 10: Example for the configuration errors

is set to one, if the ratio of the ground truth area from object j, which is covered by the ground truth object k, exceeds a certain threshold t_o.

$$occ_j = \begin{cases} 1, & \exists \mathcal{GT}_k s.t. |\mathcal{GT}_j \cap \mathcal{GT}_k| > t_o \\ 0, & otherwise \end{cases} \tag{6}$$

For all situations with a set occlusion flag there will be no evaluation of any error, i.e. none of the error measurement scores introduced above is increased and thus no ground truth data has to be available for these frames.

4.3.6 Configuration evaluation procedure

To enable a performance evaluation of different tracking approaches evaluated on diverse data sets, all those measurements presented above have to be normalized by both the number of ground truth objects $N^t_{\mathcal{GT}}$ per frame and the number of frames n as listed in the structure chart below. Since there may occur frames with no $\mathcal{GT}$ labeled at all, normalizing by simply $N^t_{\mathcal{GT}}$ would fail and thus the denominator was chosen to $max(N^t_{\mathcal{GT}}, 1)$ to avoid a division by zero for $N^t_{\mathcal{GT}} = 0$.

For an easy comparison of tracking algorithms a quality measure $\overline{ME}$ is computed from the error measurements. Since the human impression does not consider one of the error types much more severe than other ones, again the F-measure is used to compute the quality measure.

Structure chart for the configuration evaluation procedure

- calculate $F_{i,j}$ for each $\mathcal{E}_i$ combined with each $\mathcal{GT}_j$
- if $F_{i,j} > t_c$
  - if $\mathcal{GT}_j$ not already mapped: map $\mathcal{GT}_j \rightarrow \mathcal{E}_i$
    else increment MO

  else increment FP
- if $F_{i,j} > t_c$
  - if $\mathcal{E}_i$ not already mapped: map $\mathcal{E}_i \rightarrow \mathcal{GT}_j$
    else increment MT

  else increment FN
- report $\overline{FP}$, $\overline{FN}$, $\overline{MT}$ and $\overline{MO}$

$$\overline{FP} = \frac{FP}{n}\sum_{t=0}^{n}\frac{1}{max(N_{\mathcal{GT}}^{t},1)} \quad , \quad \overline{FN} = \frac{FN}{n}\sum_{t=0}^{n}\frac{1}{max(N_{\mathcal{GT}}^{t},1)}$$

$$\overline{MT} = \frac{MT}{n}\sum_{t=0}^{n}\frac{1}{max(N_{\mathcal{GT}}^{t},1)} \quad , \quad \overline{MO} = \frac{MO}{n}\sum_{t=0}^{n}\frac{1}{max(N_{\mathcal{GT}}^{t},1)}$$

- compute $\overline{ME} = \frac{4\overline{FN}\,\overline{FP}\,\overline{MT}\,\overline{MO}}{\overline{FN}+\overline{FP}+\overline{MT}+\overline{MO}}$

4.3.7 Identification test

In the field of tracking, identification means that a particular $\mathcal{E}$ tracks exactly one $\mathcal{GT}$ over its entire lifetime and thus correctly identifies this ground truth object. Among several methods to associate identities that could be considered, each with its assets and drawbacks, an approach based on a "majority rule" was chosen to represent the identification associations. Thus a $\mathcal{GT}_j$ is said to be identified by that $\mathcal{E}_i$ which tracks object j most of the time, and vice versa $\mathcal{E}_i$ identifies that $\mathcal{GT}_j$ where it spent most of the time.

4.3.8 Identification error measures

Examining tracking scenarios there arise two different types of identification failures. The first type occurs, when one estimate i suddenly stops tracking ground truth object j and another estimate k continues tracking this ground truth object. The second error type results from swapping the ground truth paths, i.e. an estimate i initially tracks $\mathcal{GT}_j$ and after a while changes to track $\mathcal{GT}_k$. To detect all these identification errors, the measures listed below are introduced:

a) **Measure** FIT - Falsely identified tracker. A $\mathcal{E}_i$ which passed the coverage test for $\mathcal{GT}_j$ is different to that identifying this ground truth object before.

b) **Measure** FIO - Falsely identified object. A $\mathcal{GT}_j$ which passed the coverage test for $\mathcal{E}_i$ has not been the identified object in the frame before.

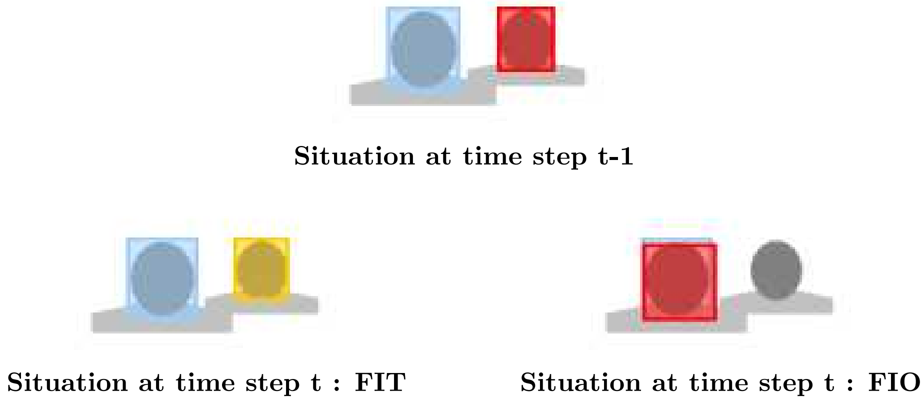
Since these measurements only report changes in associations of $\mathcal{E}$s and $\mathcal{GT}$s, a purity measure is introduced to evaluate the degree of consistency to associations between a $\mathcal{E}$ and a $\mathcal{GT}$.

a) **Measure** OP - Object purity. If $\mathcal{GT}_j$ is the ground truth object which has been identified by $\mathcal{E}_i$ for most of the time, then OP is the ratio of frames that $\mathcal{GT}_j$ is correctly identified by $\mathcal{E}_i$ $(n_{i,j})$ to the overall number of frames (n_j) $\mathcal{GT}_j$ exists.

Again the errors mentioned above are visualized in the example (Fig. 11) below, where the each box describes an estimate.

4.3.9 Identification evaluation procedure

Similar to the configuration evaluation procedure again all measurements have to be normalized by the number of ground truth objects $N_{\mathcal{GT}}^{t}$ per frame and the number of frames n as listed in the structure chart



Situation at time step t-1

Situation at time step t : FIT

Situation at time step t : FIO

Figure 11: Example for the identification errors

below. For the identification task it is difficult to create only one value indicating the performance of the algorithm, thus all three measures should be reported to get an idea of the quality of the identification capability of an approach.

Structure chart for the identification evaluation procedure

- if $\mathcal{GT}_{j,t} \rightarrow \mathcal{E}_{i,t}$
  - if $\mathcal{GT}_{j,t-1} \rightarrow \mathcal{E}_{k,t-1}$ increment FIT
  - if $\mathcal{GT}_{j,t-1}$ not mapped before increment FIO
- report $\overline{FIT}$, $\overline{FIO}$, $\overline{OP}$

$$\begin{aligned}
\overline{FIT} &= \frac{FIT}{n}\sum_{t=0}^{n}\frac{1}{max(N_{\mathcal{GT}}^{t},1)}, \\
\overline{FIO} &= \frac{FIO}{n}\sum_{t=0}^{n}\frac{1}{max(N_{\mathcal{GT}}^{t},1)}, \\
\overline{OP} &= \frac{1}{N_{\mathcal{GT}}}\sum_{j=0}^{N_{\mathcal{GT}}}\frac{n_{i,j}}{n_j}
\end{aligned}$$

4.3.10 Training and Evaluation Video Set

To get comparable evaluation results for the tracking algorithms developed by the different partners in AMI we will define a common video set for the evaluation. This video set should contain as much of the challenges which have led to the acquisition of the special side-corpus AV16.7-ami, thus the following sets have been defined for the evaluation, which may only be used for the evaluation task itself and not e.g. for tuning parameters:

Eval I : Sequences from the side corpus AV16.7ami (2, 3, 9, 12, 14)
Eval II : Sequence from the AMI core corpus (1008b)
Eval III : Sequences from the side corpus AV16.7ami (1, 8, 13, 16)

Since each of the specified sequences consists of three avi-files (left, right and central camera view) on which our algorithms will be evaluated, this material offers a total amount of approximately 1.5 h of video data for the evaluation of tracking modules. For the deliverable only results for Eval I and Eval II have to be reported. Below you can find the weblinks to get the video sequences:

AMI core corpus : http://mmm.idiap.ch/private/AMIzone/idiapHub.html
AV16.7ami : ftp://mmm.idiap.ch/private/ami/906401383/

All measurement errors introduced above will be reported according to this video evaluation set. The video material is fully annotated using different annotation rates depending on the level of dynamics of the person in the sequence. The advantage of this proceeding is a reduction of the effort in annotating parts (especially easy parts like seated people) while giving more "annotation resolution" on parts that are more interesting for tracking (e.g. somebody leaving). For this reason videos will be annotated based on three different levels of accuracy:

- Slow (1 frame/ 5 seconds) - people seated or standing for several minutes
- Middle (1 frame/ 1 second) - people standing for one minute or so max
- Fast (2 frames/ 1 second) - people entering/seating/standing up/moving to white board

The annotation data can be found at ftp://mmm.idiap.ch/private/ami/906401383/. To derive the annotation resolution please refer to the frame number explicitly given in the files. All other video material from the AMI corpus (both main and side corpus) - except the evaluation test set mentioned above - is free to be used for training the detectors and modules of the invented tracking algorithms.

4.4 Evaluation Tools

In order to facilitate a joint evaluation in the scope of AMI tracking technologies, a common MATLAB based evaluation tool has been developed and spread among all partners (also downloadable at http://www.idiap.ch/ smith/AMItrack.html). For simplifying the usage of this tool each tracking algorithm has to provide the output in the same way, i.e. a head bounding box is generated enclosing each tracked object. This result has to be stored for the evaluation tool in a simple ASCII-file according to the following file format:

```
frame [frame number]
    object [identifier]    <tab>    [head bounding box]
    object [identifier]    <tab>    [head bounding box]
```

In this file format description all expressions in brackets have to be replaced by the real numbers. For each frame, first provide the frame number (the results and ground truths must cover the same set of frame numbers), followed by the object parameters. Object parameters include a unique identifier and the location of the object in the image. The identifiers need not (and should not necessarily) match between the ground truth and tracking results, but they should be consistent within each. For each frame, provide the object parameters of every object present (in the results or the ground truth). If there are no ground truths or estimates present, just provide the frame number. Objects must be represented by bounding boxes (in both tracking and ground truth). The bounding boxes are defined by four numbers, (x,y,w/2,h/2). The point (x,y) indicates the location of the center of the bounding box,

w/2 is the distance from the center to one of the vertical edges (or half-width), and h/2 is the distance from the center to one of the horizontal edges (or half-height). All coordinates have to be referenced to the top left image origin.
The evaluation tool described in the paragraph above was re-implemented in C and further extended by some additional features resulting in the multi-object tracking evaluator. The big advantage of this evaluation tool is the possibility of displaying the annotated as well as the tracked boxes in the video and graphs showing the history of the different error types. The second advantage is the ability of reading several source data formats, enabling a more general usage not limited to a strict file format any longer. The AMI Tracking Evaluator with a graphical user interface on Windows platforms has been developed based on the requirements described in Sec. 4.3 and also exists as a console version for scripting purposes. The evaluator is able to import data from text as well as XML files. The txt-files can be one of the following formats:

Format 1
frame [frame number]
    object [identifier]   <tab>   [head bounding box]
    object [identifier]   <tab>   [head bounding box]

Format 2
image[frameID].* objectID minXPos minYPos maxXPos maxYPos

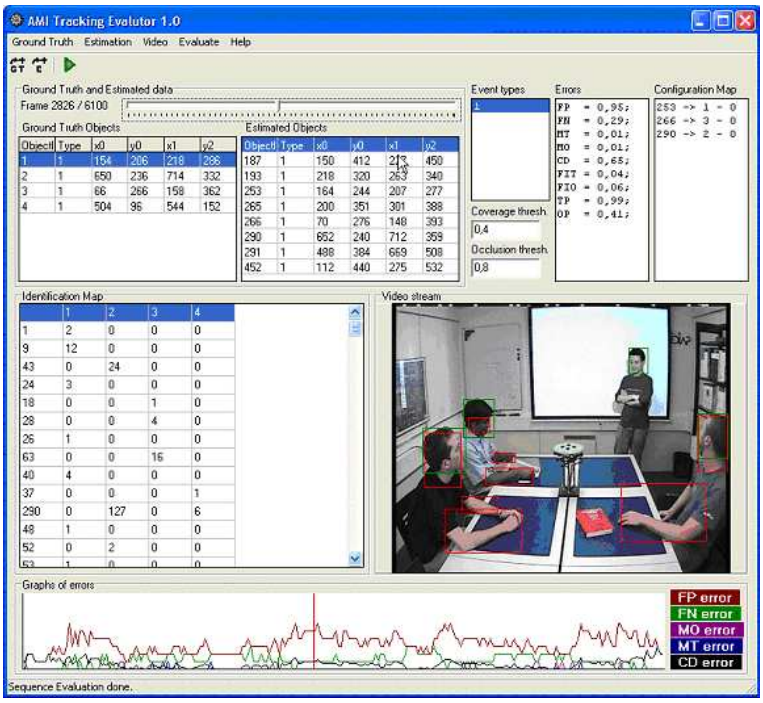
Format 3
frameID objectID visibility minXPos minYPos maxXPos maxYPos

The evaluation tool converts all defined formats. Two data sets are necessary at least. The first one, representing the annotated data referred to as the ground truth ($\mathcal{GT}$) objects, and a second one, describing the output of an image-based tracking system referred to as the estimations ($\mathcal{E}$). The output of the evaluation tool are error values defined and described in Sec. 4.3 (cf. [135]). The configuration errors can be evaluated for a particular frame and the identification errors for the entire sequence. Error values are displayed in an interactive graph, which allows finding trouble frames effectively. The most important and controlling data set is the ground truth. According to this set, the numbers and amount of frames in the sequence are initialized and also consequent browsing is allowed only through frames occurring in the ground truth set.

4.5 Tracking approaches

4.5.1 Face detection

The goal of face detection is to determine whether or not there are any faces in the image and, if present, their location. It is the crucial first step of any application that involves face processing systems including face recognition, face tracking, pose estimation or expression recognition. Thus, accurate and fast human face detection is the key to a successful operation. Recently, the IDIAP Research Institute has developed a face detection system which detects, in real-time, multiple upright frontal faces in complex backgrounds. This frontal face detection system is based on the cascade paradigm of [164] and also on the use of Modified Census Transform (MCT) features [44]. MCT belongs to the family of Local Binary Pattern (LBP) features [113]. By contrast to the Haar-like features used by Viola and Jones, LBP features are invariant to illumination and summarizes the local structure of the image. Like most face detection systems, the face detector scans the input image at many scales. The conventional approach is to compute a pyramid of sub-sampled images like Rowley et al. [130]. A fixed scale sub-window $\mathbf{x}$ is then scanned across each of these images and sent to the cascade. The objective of a cascade is to eliminate as



Ground Truth and Estimated data

Frame 2826 / 6100

Ground Truth Objects

ObjectI	Type	x0	y0	x1	y2
1	1	154	206	218	286
2	1	650	236	714	332
3	1	66	266	158	362
4	1	504	96	544	152

Estimated Objects

ObjectI	Type	x0	y0	x1	y2
187	1	150	412	213	450
193	1	218	320	263	340
253	1	164	244	207	277
265	1	200	351	301	388
266	1	70	276	148	393
290	1	652	240	712	359
291	1	488	384	669	508
452	1	112	440	275	532

Event types

1

Coverage thresh.

0,4

Occlusion thresh.

0,8

Errors

```
FP  = 0,95;
FN  = 0,29;
MT  = 0,01;
MO  = 0,01;
CD  = 0,65;
FIT = 0,04;
FIO = 0,06;
TP  = 0,99;
OP  = 0,41;
```

Configuration Map

```
253 -> 1 - 0
266 -> 3 - 0
290 -> 2 - 0
```

Identification Map

	1	2	3	4
1	2	0	0	0
9	12	0	0	0
43	0	24	0	0
24	3	0	0	0
18	0	0	1	0
28	0	0	4	0
26	1	0	0	0
63	0	0	16	0
40	4	0	0	0
37	0	0	0	1
290	0	127	0	6
48	1	0	0	0
52	0	2	0	0
53	1	0	0	0

Sequence Evaluation done.

Figure 12: AMI Tracking Evaluator GUI

many negative examples as possible at the earliest stage possible: simple classifiers are used to reject the majority of the sub-windows while detecting almost all of the positive instances. More complex classifiers can then focus on a reduced number of sub-windows. Thus, the construction of a cascade of classifiers reduces the computation time. The local binary pattern (LBP) operator is a non-parametric 3x3 kernel which summarizes the local spacial structure of an image. It was first introduced by Ojala et al. [113] who showed the high discriminative power of this operator for texture classification. At a given pixel position, LBP is defined as an ordered set of binary comparisons of pixel intensities between the center pixel and its eight surrounding pixels. Due to its texture discriminative property and its very low computational cost, LBP is becoming very popular in pattern recognition. Recently, LBP has been applied for instance to face detection [78] and face recognition [175, 1].

4.5.2 Multi-Person Tracking based on Dynamic Bayesian Networks

Our model, which was adapted from previous work on tracking the head pose multiple people outdoor scenes as described in [140], uses a multi-person tracking approach based on a hybrid Dynamic Bayesian Network that simultaneously infers the number of people in the scene, and their body and head locations in a joint state-space formulation that is amenable for person interaction modeling. The state of the system at any given time contains a varying number of interacting person models. These person models move and interact according to a dynamical model and a Markov Random Field (MRF) based interaction model. The interaction model is biased against trackers overlapping (which helps prevent multiple trackers following a single person).

A person is modeled by two bounding boxes, one corresponding to the body and one corresponding to the head. The body bounding box is defined by its image coordinates, its eccentricity, and its height. A prior on the body parameters is learned from training data and used to enforce valid configurations.

The head is modeled in a similar manner, with the additional parameter of rotation angle.

Our model is capable of automatically adjusting the number of people in the scene (by adding and removing them from the state) and exploits a global observation model to this end. The global observation model is combined with individual observations (used to localize the head) in the overall observation model. The global observation model consists of binary and color measurements defined pixel-wise over the entire image. Because these measurements are global, their cost does not vary with the number of objects in the scene. The binary observations predict the multi-object configuration using an adaptive background subtraction scheme which separates the image into foreground and background pixels. A binary observation model is trained using switching GMM's (one for every possible number of people in the scene) on features that measure the overlap of the predicted body locations with foreground and background pixels. These features are defined in the precision-recall space of the foreground and background. When observed values match these features well (hopefully as a consequence of good tracking) a high response is given. A global color model, also defined pixel-wise, is used to maintain object identity.

The individual head observations also make use of the background subtraction. A head silhouette model is constructed from the training set by averaging the binary patches taken from known head locations. These observations yield a high response when the head bounding box is configured to lay over a head-shaped area of the binary image.

Inference on the filtering distribution in our model is done by trans-dimensional Markov Chain Monte Carlo (MCMC) sampling. Trans-dimensional MCMC has the following advantages: 1) because it can change the dimension of the state, it can easily add or remove people from the scene, 2) it can efficiently search high dimensional state spaces compared to other particle filters, and 3) it can help solve the problem of normalizing the likelihoods of various objects by decomposing move types. Briefly, our sampling method works as follows. First, a proposed configuration is generated by first selecting a move type chosen from the following: *birth* of a new object or *rebirth* of a dead object, *death* of an existing object, *swap* of two object identities, update of the *body* parameters for a particular person, update of the *head* parameters of a particular person. After the state has been modified according to the move type, the likelihood of the proposed configuration is computed from the observation model. The proposal is then accepted with a probability proportional to the ratio of its likelihood to the likelihood of the previous state (so usually only good proposals are accepted). If the proposal is accepted, it is added to the Markov Chain. If not, the previous state is added. After a sufficiently long chain is produced, a MAP estimate is computed.

4.5.3 Binaural audio-visual localisation and tracking

Since the binaural audio-visual localisation and tracking algorithm takes a significantly different approach to conventional, statistical trackers, it's output is not compatible with the common evaluation scheme. It has been developed within a psychophysically plausible framework in which a single object is the tracking focus; it has been defined that this focus is positioned over a single pixel. Hence, the system cannot have more than one tracking estimate $\mathcal{E}$ as expected by the common evaluation scheme; furthermore, there is no concept, in our system, of the estimate $\mathcal{E}$ having two-dimensional extent (required for metrics $\alpha_{i,j}$, $\beta_{i,j}$ and $F_{i,j}$). Thus, an alternative evaluation metric was used based on the tracking pixel error over the duration of a number of representative sequences. The ideal system would maintain the tracking focus over the centre of the to-be-tracked object.

The ground truth position of the to-be-tracked object in the meeting recordings was established by manual transcription of the video data. The transcriber used a mouse pointer to follow the centre of the target participant's face during movie playback. The playback was slowed to approximately 8 frames per second to allow periods of rapid motion to be accurately annotated.

4.5.4 Omnidirectional image processing

Within the context of AMI research was also concerned in omni-directional image processing methods, which render the image in suitable form for further use. These methods are concerned on the fast and precise image stabilization and image transformation into the proper perspective view without significant distortions. This pre-processing is needed for presentation of omnidirectional images to a human in suitable form and mainly for further processing as tracking human body parts for activity evaluation. The developed methods for image stabilization were published on the Fifth IASTED International Conference on Visualization, Imaging & Image Processing in September 2005.

4.5.5 Gabor wavelet based face detector

A framework incorporating various methods into a single tracking system for human body parts and evaluation of different method combinations has been invented. We have developed and implemented various types of methods for hand and head detection, tracking and also algorithms for face detection. The basic method used for image segmentation is the skin color detection. Skin colored blobs are further described by ellipses, which contain information about the rotation angle and size of the main major and minor axes. A generalized skin color model is used and tested which does not need a manual initialization. Further, the Gaussian skin color model with manual initialization is used to achieve better results which are needed for the face detection. The face detection is applied only in the detected skin colored areas for increasing the speed of the whole algorithm. The face detector is based on the weak classifier compound of a Gabor wavelet and a decision tree. Its output determines a ”probability” that the input image is a face. We constructed a strong classifier as a linear combination of several weak classifiers issued from AdaBoost algorithm (Viola&Jones) and Gabor wavelets. The face detector is trained on normalized face images (24x24 pixels), where simple rectangle image/facial features are replaced by more complex Gabor wavelets.

4.5.6 KLT based tracking

A further tracking system consists of the single object tracker based on a public domain KLT tracker (Kanade-Lucas-Tomasi feature tracker). This tracker implementation provides accurate and stable single object tracking and is capable to accurately and stable track single object over relatively long image sequences despite object pose changes, fast movement and nontrivial background. However, this is only a tracker, so an object detector is needed to place the tracker on a relevant initial position. For this purpose, we use some combinations of the ellipse fitting, skin color model and/or Gabor Wavelet Networks face detector. We are testing the combination of the feature tracker with skin color model and background subtraction to improve tracker results. Improvement achieved by these methods will be evaluated to figure out if it is reasonable to use these supporting methods according to their computational cost. The main advantages of such trackers are the ability to track permanently the foreground object even if another tracked object passes behind and it also provides means to maintain object tracking continuity during partial and full occlusions possibility. Further it provides accurate and stable information about object movement velocity and direction which are used for purposes like speaker identification and gesture recognition in intelligent video editing.

4.5.7 Probabilistic Active Shape Tracking

The basic idea of this method is to utilize a factored sampling technique called ICondensation [[73], [74]] to generate several hypotheses of a possible location for the tracked object. Due to the non-rigid appearance of these objects - the human heads - especially for a varying perspective, this method is combined with a flexible Active Shape Model ([34]) of the head, which will be used for comparing each hypothesis with the observations derived from the true image data. Roughly our approach works like follows:

At first a sample-set consisting of a fixed number of hypotheses is generated. Each hypothesis in this set represents the position, scale, rotation angle and silhouette of a possible head. For the initialization of the hypotheses the data received by a skin color detector is binarized and clustered. Randomly skin colored regions (clusters) are chosen and their euclidean properties (translation, scale and rotation) serve as initial values for the hypotheses.

In the next step each element of this sample-set is predicted by a linear dynamical model with constant velocity resulting in a deterministic drift diffused by additive Gaussian noise. Then for each of the hypotheses a weight, representing the probability for a head described by the corresponding hypothesis, is computed by an Active Model based measurement. During this measure, the shape parameters are adapted to the available image data and thus the whole hypothesis is optimized towards a better representation of the image itself. A quality score describing the degree of the shape fitness to the image data is used to update the hypothesis' weight. Finally the sample-set is propagated for the subsequent frame by choosing some of the hypotheses from the old sample-set, each relating to its weight. After this procedure some of the old elements will be lost, while others may appear more than only one time in our new set. Additionally some new hypotheses are generated again by the skin color detection.

Thus our algorithm provides a stable trajectory also in very cluttered environments with non-rigid object shapes. As an important advantageous aspect of this approach, we need only a few samples for tracking people stable, resulting in a very time efficient algorithm, while comparable methods require at least between 100 and 1000 hypotheses.

4.5.8 Tracking architecture for method fusing

Video object trackers can be used in many scenarios. Within the AMI-context alone valuable information could be obtained from tracking faces, heads, hands and people as a whole, but also objects like notebooks, chairs or pencils and pens. For each tracking-task a different type of tracker would be created. From beforehand, a researcher does not want to be restricted to only one type of tracker, yet at the same time, there are many aspects of trackers that do not have to be created twice. Therefore, TNO has been investigating the common ground that is needed for ranges of types of trackers. We have consolidated this work in a tracking architecture. This architecture is being used to create a system that integrates different types of trackers, on the different camera's that were used in AMI to create one model of the scene in the meeting room. The architecture now consists of about 20 base classes. What follows is an outline of the most important classes, and their purpose.

The system perceives the scene through different Sensors, and more specific, CameraSensors. Although video object tracking mainly deals with camera's, there are still different types of camera's, and many different possible camera-settings, such as resolution and frame rate, but also pixel-type, angle of opening and coordinates for position and orientation. This class is responsible for giving access to the actual frame buffer. It also supports low-level features for drawing in the video buffer and flood-search algorithms.

For each sensor a TrackerManager is responsible for administering all trackers that work on that sensor. This class takes care of a small history of information about trackers and makes sure that all trackers have access to new frame buffers. One trackermanager can administer many types of trackers at the same time. This makes it possible to have for example face-trackers, hand-trackers and people-trackers working at the same time.

All these different types of tracker share one common base, the BaseTracker. This class stores information about the position of the tracker, it's rotation and gives access to the underlying representation of what it is actually tracking. This could be anything, and depends on the specific trackers. At this moment we have experience with BlobTrackers, TemplateTrackers, HistogramTrackers and OpenCV-Featuretrackers. It is abvious that all these different trackers have a different internal representation of what they are tracking. This is typically stored in a subclass of a BaseTemplate. This means we have a BasePixelTemplate, a BaseHistogramTemplate, etc.

The next important collection of classes is the BaseFactory and it's derivatives. A BaseFactory is responsible for starting new trackers. This can be depending on motion (BaseMotionTriggeredFactory) or on user-request (BaseCustomStartFactory) but also on detection of a specific type of object (BaseFaceDetectionFactory).

The final class to describe is the BaseScene. This class is responsible for combining information from different sensors into one coherent world view. This is the main entry point for an application to request information about the current state of affairs in the system. Currently we have several different scenes. There are four scenes that deal with different views from one camera: in the flat view nothing is changed in the coordinate-system. In the top-down view it is possible to correct for the effect that objects from straight above are seen differently as objects seen from under an angle. In the side view objects are assumed to move only horizontally, and in the perspective view a correction for distance and camera angle on the scene is applied. Current work focuses on a different aspect, namely the MultCameraScene. A tracking architecture was developed with features that enabled reusability of many core features of (classes of) trackers. A start was made with building a system for integrating different types of trackers, but this system is not ready yet for evaluation on tracking-results.

4.6 Results

4.6.1 Face detection

Comparing face detection methods is a difficult task, even though they are evaluated on the same databases. Indeed, only a few researchers give their definition of *what is a correctly detected face*, such as [48], or use some error measures, like the one introduced by Jesorsky et al. [77]. Moreover, the correct detections and false alarms are usually counted by hand. In this work, a detection is considered as correct if the bounding box contains the eyes and the mouth without too much background (see Fig. 13).

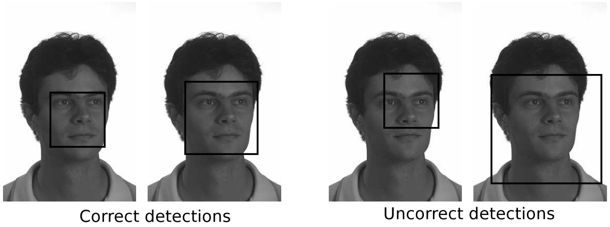


Figure 13: Examples of correct and incorrect detections.

Performance Evaluation on a Benchmark Database Table 32 provides the detection rate and the number of false alarms of the IDIAP frontal face detector on the CMU Frontal Test Set. The proposed system achieves a good detection (84.6%) but still with too many false alarms. Indeed, even though the CMU Frontal Test Set contains only frontal faces, some of them can be slightly rotated in-plane or out-of-plane. Therefore, our frontal face detector is missing those faces.

Table 32: Results obtained by the IDIAP frontal face detector on the CMU Frontal Test Set.

Detection Rate	Number of False Alarms
84.6%	435

Evaluation on AMI Data We don't have currently any numerical results to provide on AMI data. However, we already performed several tests on meeting recordings (Fig. 14). These tests have shown that the frontal face detector is performing well without any specific tunings.

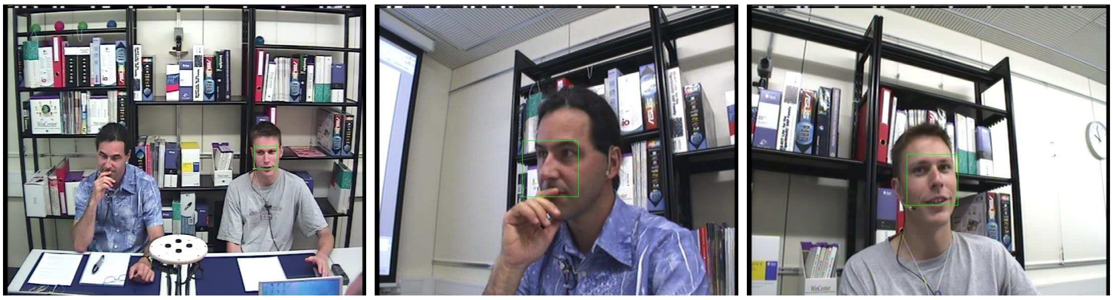


Figure 14: Examples of face localization on AMI data. From left to right: full view (faces are small) and close-up (faces are very often not frontal).

The original videos can be found at `http://www.idiap.ch/~marcel/en/facedemos.php#demo3`.

A Torch Machine Vision Package for Face Localisation We developed a machine vision package for **Torch** (a well-known open source machine learning library) called **Torch vision**. **Torch vision** [2] provides basic image processing and feature extraction algorithms but also a general modular framework for *face detection*.

Demonstrations We developed several face detection demonstration systems publicly available at `http://www.idiap.ch/~marcel/en/facedemos.php`. These systems includes (1) a real-time frontal face detector using a video camera, (2) an on-line frontal face detector which offers the possibility to upload an image and to save the result, (3) a frontal face detector using video files.

4.6.2 Multi-Person Tracking based on Dynamic Bayesian Networks

The results for the MCMC tracking method developed at IDIAP (described in Sec. 4.5.2) appear in Tab. 33. These results should be seen as a first trial of our method on the data. This dry-run has highlighted a number of issues that could be refined in order to improve performance. In particular, the head tracking in our model relies heavily on good quality body tracking, which suffered from the lack of annotated data to train from. Additionally, the binary observations perform poorly for situations in which object size can vary dramatically. This occurred frequently due to the close proximity of the cameras to the meeting participants.

Some of the data is indeed quite challenging. The size of heads varied greatly within the data set. Some scenes existed where virtually the entire field of view of the camera was occluded by the back of a head. Cases occurred quite often in which a head was only partially in the scene, or worse, only a portion of the back of their head (as seen in Figure 16). Generally, our model performed well when the meeting participants were further in the field of view of the camera and did not linger at the edge of the scene (example in Figure 15). Merged foreground blobs of meeting participants caused problems for our model. Normally, this can be overcome by placing a strong prior on the size of the body, but it was necessary to relax this prior in order to accomidate the variation in size of a person between sitting and standing position. Our model uses identity swapping to better match object color and recover from occlusion, but this proved to be difficult in some instances. Typical errors included: FN errors caused by not tracking

[2] `http://www.idiap.ch/~marcel/en/torch3/introduction.php`

partial heads as seen in Figure 16, FN errors caused by heads of extremely large size in close proximity to the camera, MT and MO errors caused by meeting participants in close proximity for a long duration, FIT and FIO errors caused by people entering and exiting the scene from or standing in close proximity to each other.

Table 33: Tracking Results - Multi-Person Tracking based on Dynamic Bayesian Networks

Eval	Seq	F-meas	**FN**	**FP**	**MT**	**MO**	**CD**	$\overline{\textbf{FN}}$	$\overline{\textbf{FP}}$	$\overline{\textbf{MT}}$	$\overline{\textbf{MO}}$	$\overline{\textbf{CD}}$	**FIT**	**FIO**	$\overline{\textbf{FIT}}$	$\overline{\textbf{FIO}}$	$\overline{\textbf{TP}}$	$\overline{\textbf{OP}}$
I	seq02L	0.84	1	1	0	0	0	0.02	0.02	0	0	0	0	0	0	0	0.50	0.88
I	seq02R	0.84	7	22	0	0	15	0.14	0.44	0	0	0.31	0	0	0	0	0.20	0.67
I	seq03L	0.79	92	10	0	0	-82	0.44	0.05	0	0	0.46	0	0	0	0	0.08	0.02
I	seq03R	0.70	77	15	0	0	-62	0.37	0.07	0	0	0.41	0	8	0	0.04	0.63	0.21
I	seq09L	0.76	6	11	2	0	6	0.04	0.11	0.02	0	0.08	20	41	0.16	0.40	0.65	0.49
I	seq09R	0.18	20	6	0	0	-13	0.27	0.08	0	0	0.21	0	8	0	0.11	0.75	0.22
I	seq12L	0.75	112	8	0	0	-33	0.38	0.07	0	0	0.43	18	37	0.09	0.24	0.75	0.43
I	seq12R	0.19	121	0	0	0	-73	0.70	0	0	0	0.70	0	2	0	0.01	1	0.04
I	seq14L	0.50	141	28	6	0	-40	0.48	0.11	0.03	0	0.41	9	11	0.04	0.05	0.76	0.40
I	seq14R	0.77	89	23	11	0	-34	0.32	0.06	0.01	0	0.32	32	41	0.21	0.27	0.65	0.37
III	seq01L	0.82	4	2	0	0	-2	0.06	0.03	0	0	0.06	0	1	0	0.02	0.64	0.74
III	seq01R	0.85	3	16	2	0	15	0.05	0.25	0.03	0	0.23	0	7	0	0.11	0.54	0.73
III	seq08L	0.69	15	15	10	0	6	0.08	0.08	0.06	0	0.09	4	15	0.03	0.09	0.80	0.56
III	seq08R	0.20	132	3	0	0	-69	0.71	0.03	0	0	0.74	0	3	0	0.03	0.83	0.02
III	seq13L	0.70	94	14	43	0	-11	0.38	0.09	0.17	0	0.17	62	63	0.37	0.28	0. 69	0.51
III	seq13R	0.11	73	10	33	0	-14	0.44	0.06	0.23	0	0.36	1	49	0.01	0.33	0.83	0.31
III	seq16L	0.85	43	1	0	0	-22	0.22	0.01	0	0	0.22	3	6	0.02	0.04	0.84	0.56
III	seq16R	0.53	13	0	0	0	-8	0.08	0	0	0	0.08	9	22	0.05	0.18	0.95	0.46

4.6.3 KLT based tracking

Results obtained for KLT based tracking are shown in Tab. 34.

Table 34: Tracking Results- KLT based tracking

Eval	Seq	F-meas	**FN**	**FP**	**MT**	**MO**	**CD**	$\overline{\textbf{FN}}$	$\overline{\textbf{FP}}$	$\overline{\textbf{MT}}$	$\overline{\textbf{MO}}$	$\overline{\textbf{CD}}$	**FIT**	**FIO**	$\overline{\textbf{FIT}}$	$\overline{\textbf{FIO}}$	$\overline{\textbf{TP}}$	$\overline{\textbf{OP}}$
I	seq02L	-	21	0	0	0	-	1.00	0.00	0.00	0.00	1.00	0	0	0.00	0.00	0.00	0.00
I	seq02R	-	7	2	0	0	-	0.33	0.10	0.00	0.00	0.33	0	2	0.00	0.00	1.00	1.00
I	seq03L	-	25	54	0	0	-	0.27	0.57	0.00	0.00	0.84	0	2	0.00	0.02	1.00	1.00
I	seq03R	-	35	51	0	0	-	0.48	0.33	0.00	0.00	0.72	1	7	0.01	0.07	1.00	0.92
I	seq09L	-	24	31	0	0	-	0.21	0.31	0.00	0.00	0.24	1	13	0.01	0.11	1.00	0.50
I	seq09R	-	36	23	0	0	-	1.00	0.66	0.00	0.01	0.97	0	0	0.00	0.00	0.00	0.00
I	seq12L	-	125	119	0	0	-	0.98	0.82	0.00	0.00	0.56	0	2	0.00	0.01	1.00	1.00
I	seq12R	-	59	150	0	1	-	0.31	0.71	0.00	0.01	0.46	0	21	0.00	0.10	0.99	1.00
I	seq14L	-	143	133	0	0	-	0.51	0.57	0.00	0.00	0.36	1	15	0.00	0.08	1.00	0.85
I	seq14R	-	200	126	0	0	-	0.63	0.41	0.00	0.00	0.31	0	19	0.00	0.05	1.00	0.91
III	seq01L	-	18	14	0	0	-	0.95	0.74	0.00	0.00	0.31	0	1	0.00	0.05	1.00	1.00
III	seq01R	-	26	24	0	0	-	0.90	0.83	0.00	0.00	0.34	0	3	0.00	0.10	1.00	1.00
III	seq08L	-	4	91	0	0	-	0.02	0.53	0.00	0.00	0.52	0	4	0.00	0.02	1.00	1.00
III	seq08R	-	90	15	0	0	-	0.11	0.68	0.00	0.00	0.61	1	3	0.01	0.02	1.00	0.83
III	seq13L	-	8	117	2	1	-	0.04	0.55	0.01	0.01	0.55	6	5	0.04	0.03	1.00	0.87
III	seq13R	-	126	66	0	0	-	0.96	0.52	0.00	0.00	0.69	0	3	0.00	0.02	1.00	0.86
III	seq16L	-	22	16	0	1	-	0.11	0.07	0.00	0.00	0.10	1	9	0.00	0.04	0.98	0.77
III	seq16R	-	25	70	0	0	-	0.29	0.73	0.00	0.00	0.65	0	5	0.00	0.05	1.00	0.90

4.6.4 Probabilistic Active Shape Tracking

Results obtained for Probabilistic Active Shape Tracking are shown in Tab. 35. These results have been created in a first dry-run on the AMI side corpus. In Fig. 17 frames of AMI16.7 sequence No. 2 are shown till the person disappears. As there can be seen, the tracker keeps constantly and precisely on the position of the persons' head.

Table 35: Tracking Results - Probabilistic Active Shape Tracking

Eval	Seq	F-meas	**FN**	**FP**	**MT**	**MO**	**CD**	$\overline{\textbf{FN}}$	$\overline{\textbf{FP}}$	$\overline{\textbf{MT}}$	$\overline{\textbf{MO}}$	$\overline{\textbf{CD}}$	**FIT**	**FIO**	$\overline{\textbf{FIT}}$	$\overline{\textbf{FIO}}$	$\overline{\textbf{TP}}$	$\overline{\textbf{OP}}$
I	seq02L	0.84	1	0	0	0	-1	0.02	0.00	0.00	0.00	0.02	0	0	0.00	0.00	1.00	0.88
I	seq02R	0.73	10	8	0	0	-2	0.20	0.16	0.00	0.00	0.16	0	0	0.00	0.00	0.58	0.52
I	seq03L	0.72	41	23	0	0	-18	0.20	0.11	0.00	0.00	0.14	0	0	0.00	0.00	0.70	0.56
I	seq03L	0.65	52	28	0	0	-24	0.25	0.13	0.00	0.00	0.12	0	0	0.00	0.00	0.66	0.51
I	seq01L	0.45	18	15	0	0	-3	0.28	0.23	0.00	0.00	0.08	0	0	0.00	0.00	0.06	0.05
I	seq01R	0.49	25	17	0	0	-8	0.39	0.27	0.00	0.00	0.13	0	0	0.00	0.00	0.19	0.14

4.6.5 Binaural audio-visual localisation and tracking

Single participant The system was evaluated using a recording of a single participant who moved around the room uttering a short phrase at 10 degree intervals. Fig. 18(a) shows the ground truth and the audio-visual tracker positions for the frames in which the participant was visible. It is evident that the system tracks the participant with good accuracy; indeed, the mean absolute error per frame across the entire sequence was only 13.4 pixels — much less than the width of a face (26 to 46 pixels depending on the distance from the camera). It should be noted that the larger errors at the end of the sequence are due to part of the participant's body and/or face being beyond the edge of the frame. In this situation the face detector fails to identify the partially occluded face, and hence the automatic tracker uses the center of the remaining visible areas of the body whereas the ground truth shows the position of the visible portion of the face.

Multiple participants The two multi-participant scenarios were similar to the single participant scenario. In addition to the target participant, two seated participants were present at approximately video column positions 220 and 555 (i.e., to the left and right of the frame). In the first recording, the seated participants remained silent; in the second recording, they were instructed to converse naturally. Fig. 18(b) shows the groundtruth and the audio-visual tracker positions for the multi-participant scenario in which there was no background speech and Fig. 18(c) shows the audio-visual tracking performance with a speech background. The mean absolute error per frame across the entire sequence was 18.7 pixels and 26.4 pixels respectively. Tracking performance across the three audio-visual scenarios is good with few gross errors. As expected, the accuracy of the system degrades, albeit fairly gracefully, with increased meeting complexity (more participants, background speech). However, even in the situation with two confounding visual and auditory sources, the audio-visual tracking system successfully follows the target participant and exhibits a mean error equivalent to the lowest bound on head width. The system's robustness can be illustrated in the multi-participant scenario in which the seated participants are conversing. As the currently silent target participant approaches the seated subject on the right of the frame the seated subject becomes the A-V object which is to be tracked. However, the system corrects this once the target participant begins to speak (see Fig. 18(c), time labels 'B' and 'C').

A further contributing factor to the mean tracking error is the behavior of the oculomotor model. As described in previous reports, this is based upon a velocity-matching framework which includes visual cortex delays but does not include prediction. Thus, when a tracked object stops, the oculomotor model continues for a period of time before slowing. This is manifested by a short overshoot in the A-V tracking behavior which is followed by a small correction in the opposite direction (e.g., Fig. 18(a) at time label ‘A’).

4.7 Conclusion and summary

We developed a face detection system which detects, in real-time, multiple upright frontal faces. We evaluated the performances on a benchmark database but not yet on AMI data. Several results have been obtained such as a machine vision package for **Torch** that implements a face detection system, or demonstrations systems. We plan to investigate non-frontal face detection in order to handle faces under different views (in-plane and out-plane rotated). Trackers based an a variety of different technologies have been investigated and initial results have been presented. These methods comprise trackers for multiple-person scenarios, techniques exploiting multimodal features (AV-Tracker) as well as view independent tracking frameworks like the Probabilistic Active Shape Tracking. At the moment the presented results indicate the potential of our methods, future work will also provide even lower error rates by tuning the parameters of the trackers.

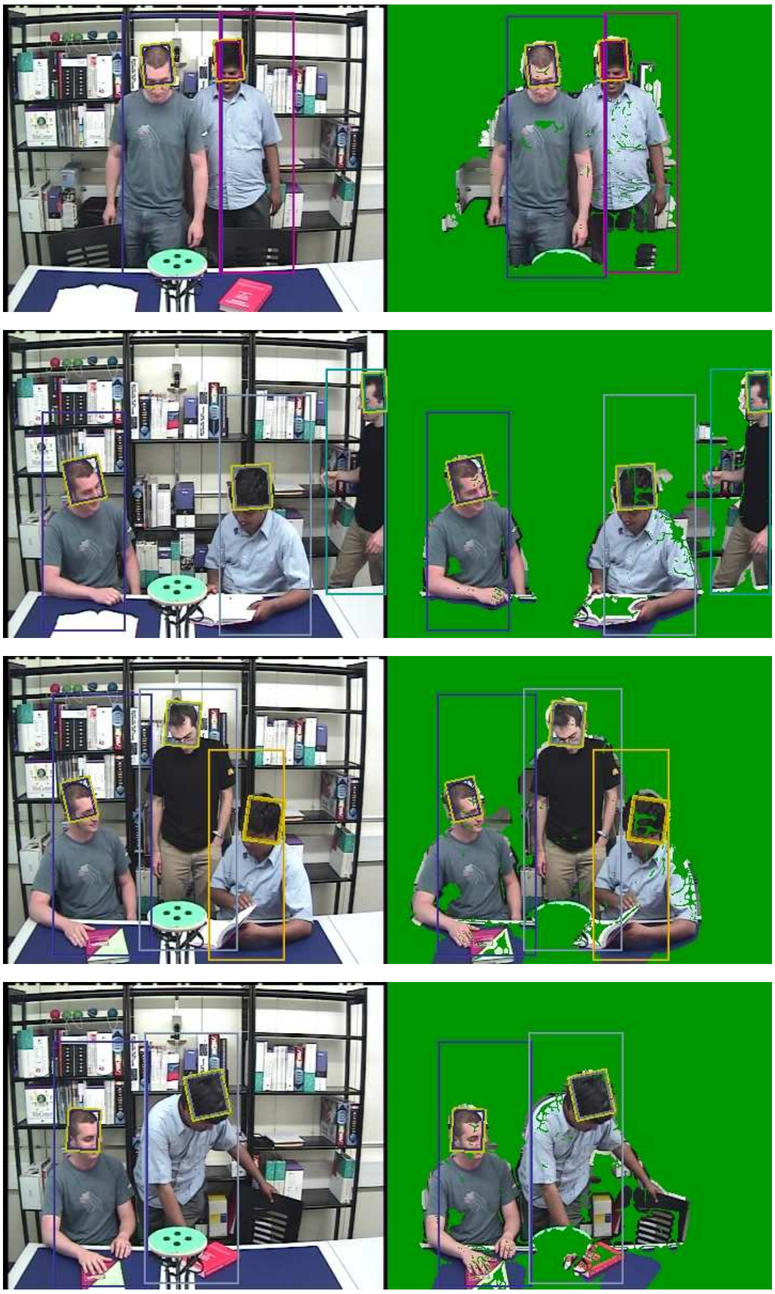


Figure 15: *Seq12L* frames 405, 843, 1693, 2253. Left: Orignal image with tracking results from MCMC method developed at IDIAP. Right: Background subtracted image with tracking results from MCMC method developed at IDIAP.

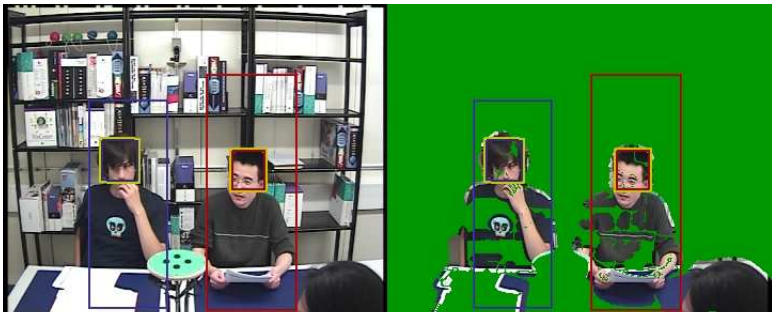


Figure 16: *Seq14L* Left: Orignal image with tracking results from MCMC method developed at IDIAP. Right: Background subtracted image with tracking results from MCMC method developed at IDIAP. Our model has failed to recognize the head of a third participant in the lower-left hand corner. Building a head model capable fo recognizing all three of the heads in the scene a challenging task.

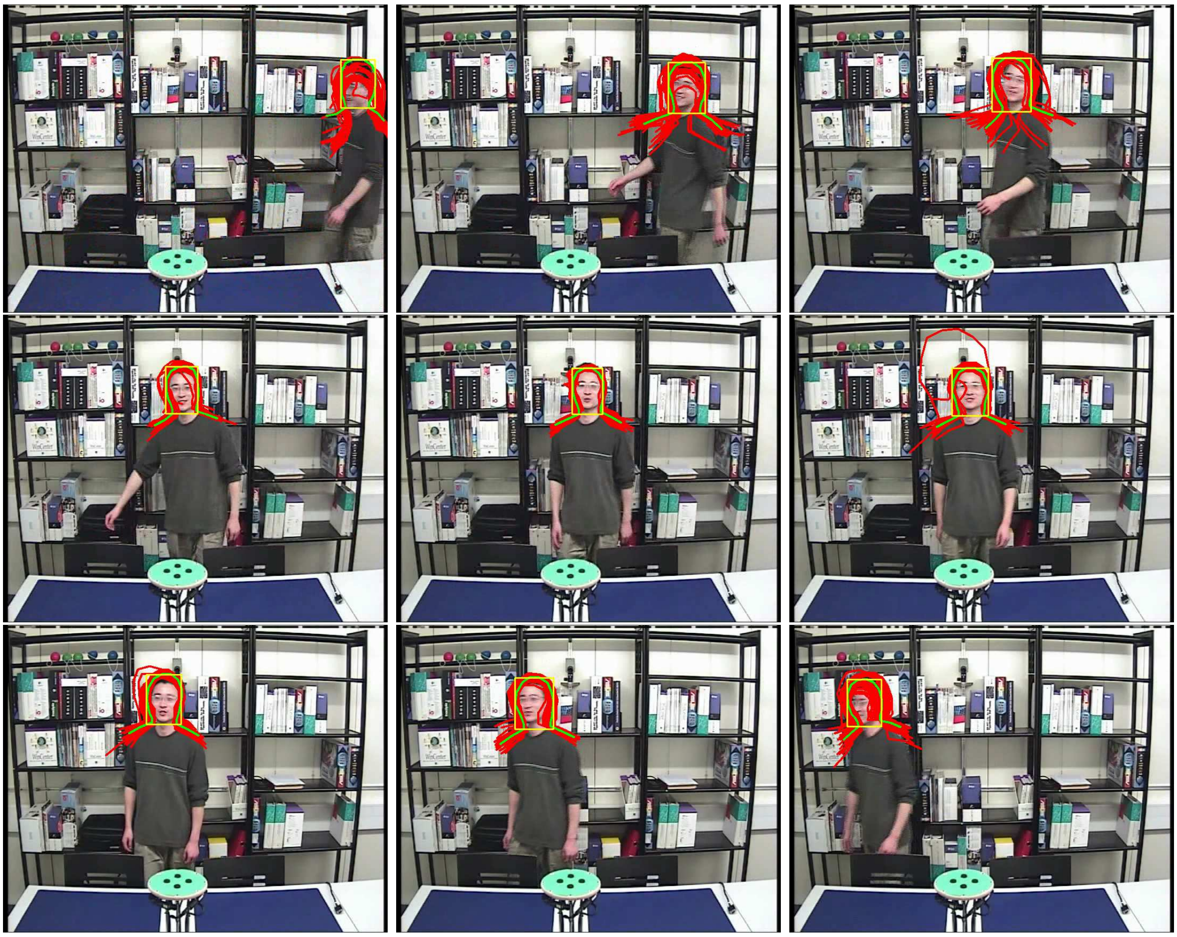


Figure 17: Seq02L - Frames 12, 24, 36, 48, 60, 140, 152, 164. The boundingbox, marked by the yellow rectangle, shows the position of the head, which is tracked by the active shapes marked with red and green color.

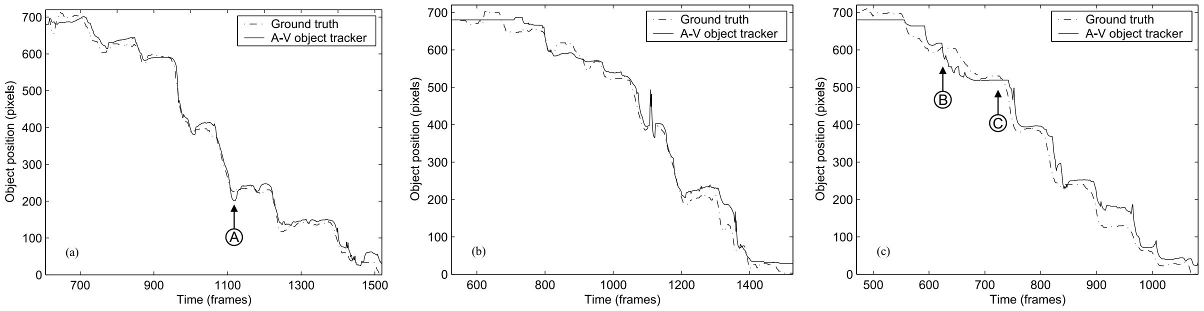


Figure 18: A-V tracking position over time compared with manually transcribed ground truth position for varying scenario complexity. (a) Single participant scenario; (b) Multi-participant scenario (silent background); (c) Multi-participant scenario (speech background). Overshoot of the oculomotor model when the target becomes stationary after a period of rapid movement is apparent at time ‘A’. At time ‘B’ the target participant becomes silent and the system incorrectly tracks a non-target participant due to their proximity. However, once the target participant resumes speaking, the tracking system corrects itself (time ‘C’). See text for details.

5 Actions and Gestures

The research group on Action and Gesture Recognition aims to identify relevant features and to develop methods for automatic recognition of important actions and gestures within the AMI domain. Actions differ from gestures in the temporal aspect. Gestures (e.g. pointing) are bound in time, whereas actions (e.g. writing) can, in general, last for any given period of time. Here we do not distinguish between the two.

5.1 Objectives

The main objectives are the evaluation of features and methods for automatic recognition of relevant actions and gestures within the AMI domain. The focus will be on important actions and gestures which occur frequently within the design meetings. Important here means that the action or gesture has a meaning within the semantic analysis of the meeting.

5.2 Relevant actions and gestures within AMI

Within AMI, we focus on gestures and actions that have a communicative function. A list of relevant gestures, cf. Table 36 has been defined in WP3 of the AMI consortium [126].

Hand	Body	Head
pointing	leaning forward	nodding
writing	leaning backward	shaking
voting	leaning with head on hand	
scratching/touching	reposition	
handling objects	standing up	
beats	sitting down	
inconic/metaphoric		
writing on whiteboard		
wiping the whiteboard		

Table 36: Relevant action and gestures within AMI.

A total of 12 randomly selected Scripted Meetings have been annotated and the number of occurrences for each of the gestures mentioned in Table 36 has been determined. The results can be found in Table 37. We only used data from Camera 1 and 2, so none of the presentations at the whiteboard have been taken into account.

Gestures that occurred only a few times, and gestures that have no clear communicative function (handling objects, scratching/touching, leaning) have been removed from the list. Beats, iconics and metaphorics have been combined in one class: speech supporting gestures (SSG). The list of gestures to be recognized is now reduced to writing, speech supporting gestures, nodding, shaking, standing up and sitting down.

Our aim is to recognize these gestures from the input data. The first step is to extract features from these data. We used Posio [121], a model-based pose estimation program, to extract for each person the 2D location of the head and hands, a set of 9 3D joint locations, and a set of 10 joint angles. These features were further augmented to obtain velocity and acceleration for the 2D and 3D locations and angular velocity and acceleration for the 3D joint angles.

Next, we describe how to segment these feature streams and evaluate the segmentation in Section 5.3. In Section 5.4 different recognition methods are evaluated.

Gesture	Occurrence	Gesture	Occurrence
Pointing	2	Nodding	> 100
Writing	58	Shaking	23
Voting	0	Leaning forward	16
Handling objects	47	Leaning backward	27
Scratching/touching	49	Leaning on table	25
Beats	> 100	Leaning with head on hand	41
Iconics/Metaphorics	25	Reposition	6
Writing on whiteboard	0	Standing up	12
Wiping whiteboard	0	Sitting down	12

Table 37: Occurrences of the different relevant gestures in one hour of randomly chosen annotated AMI video data.

Also special attention is paid to a particular type of head gesture, namely *negative signals* signals a negative response to a yes-no question; usually characterized by a head shake).

Negative signals in the AMI corpus can be subtle gestures involving little head movement. For direct feature extraction for this type of gesture deformable templates based on Active Shape Models [35, 15] are not well suited (although they may be useful for adding contextual information about the location of body parts for supporting other techniques). Active Appearance Models [12] are more promising, but making generic, as opposed to person-specific, models of heads at differing viewpoints is challenging. With these considerations in mind we have chosen to focus on our second approach.

Inspired by the success of methods used to automatically detect and describe of salient points in a still image (e.g [94]), our aim is to develop methods to automatically detect and describe salient motions in a video stream. Preliminary research as been conducted in methods based on optical flow, cf. [70].

5.3 Evaluation of different segmentation methods

In Section an evaluation of the segmentation is presented.

5.3.1 Evaluation method

For the segmentation two methods were compared: Bayes Information Criterion (BIC) and the Activity Measure (AM) approach. For BIC, different values for the window size and λ penalty for inserting a boundary were evaluated. For AM, we evaluated the use of different boundary types: minima, maxima, zero-crossings and threshold crossing. Also different combinations of possible input features were evaluated. More detailed information can be found in [65].

The performance measure p used for evaluating the resulting segmentation is given by

$$p = \frac{(m/2 - i)}{(m + d)/2)}$$

where m is the number of correctly matched boundaries i is the number of insertions (placing a boundary where there should be no boundary) and d the number of deletions (not placing a boundary where there should be a boundary). A placement of a boundary is considered exact when the difference is less than 7 frames (0.28s) from the actual boundary. Here an actual boundary means a manually placed boundary in the annotated video stream.

5.3.2 Evaluation results

For each gesture the best segmentation method, including the best values for the parameters of the method, is presented in Table 38.

Gesture	Method	Feature	Parameters
Writing	AM	2D location head	Types: threshold, maxima, zero-crossing
SSG	BIC	2D location left and right hand	Window size: 13 λ penalty: 2
Nodding	BIC	2D velocity head	Window size: 24 λ penalty: 3
Shaking	BIC	2D velocity head	Window size: 16 λ penalty: 5
Standing up	AM	3D joint angle left and right shoulder	Types: minima, maxima
Sitting down	AM	3D joint angle left and right shoulder	Types: threshold, maxima

Table 38: The best segmentation method, including the best values for the parameters of the method

The segmentation performance for the above methods and settings is given in Table 39. An explanation of the low performance, mainly due to insertions, is that writing, SGG's, nodding and shaking gestures contain repeated patterns, for instance nodding is a repetition of nods. This implies that each starting (end) point of such a pattern can be classified as a boundary. Every frame which is the start of a nod can be classified as, and is the start of, a nodding gesture. Hence there is a high chance that the segmentation algorithm will put a boundary at this frame. Moreover the SSG gesture class is a container for a very diverse set of gestures, such as beats, iconic and metaphoric gestures, and the segmentation algorithm is not able to find a general underlying principle for segmenting the video stream into SSG parts without taking the underlying gestures into account.

Gesture	Match	Insertion	Deletion	Performance
Writing	95	24	63	29.75%
SSG	1224	550	12	10.03%
Nodding	678	226	84	29.96%
Shaking	120	52	0	13.33%
Standing up	14	3	6	40.00%
Sitting down	15	5	5	25.00%

Table 39: The segmentation performance for the methods of Table 38.

Moreover the input features are generated by the Posio system, that generates rather noisy features which influence the segmentation performance in a negative way.

One possible solution for the insertion problem could be to neglect insertions because of the repetition in gestures. This will be the focus of future research in gesture segmentation. Some preliminary results have already been obtained and can be found in Table 40. It should be remarked that generated segmentation boundaries which are outside the annotated gesture boundaries are neglected and thus *not* counted as errors. One of the problems to be investigated is whether these segmented parts within gestures coincide with meaningful gesture parts. Hence future research into gesture part segmentation is needed.

Gesture	Match	Deletion	Performance
Writing	156	2	98.73%
SSG	1214	22	98.22%
Nodding	687	75	90.16%
Shaking	166	4	96.67%
Standing up	20	0	100.00%
Sitting down	20	0	100.00%

Table 40: The segmentation performance for gesture parts.

5.3.3 Conclusions

The main conclusion is that automatic segmentation of gestures is still a challenging problem and the approach taken does not give good segmentation performance for whole gestures, mainly due to the intrinsic structure of the gestures under consideration and the noise in the input features. An alternative approach could be to look at gesture parts and develop segmentation algorithms for these.

5.4 Evaluation of different recognition methods

Given the above conclusions on the segmentation, we will evaluate the recognition methods *not* on the automatically segmented video stream but on the manually segmented video stream, generated by the annotators.

5.4.1 Evaluation method

Due to the temporal aspect in gestures we focus on classifying gestures using different HMM methods. The optimal parameters for number of states, topology and type of HMM (discrete or continuous) are determined. In Table 41 the optimal parameters of the models are summarized. The parameters have been varied over a range of values, and a threshold was used to determine whether a sample was a gesture or not. The ratio between true positives and false positives for a fixed number of samples has been used as a selection criterium. The nodding and shaking gestures are left out of the table since the feature data turned out to be too noisy. This can be explained by the simplistic approach that has been taken to

Gesture	Feature set	# Clusters	States	Topology
Writing				
Garbage model 1	Hand Polar velocities	20	10	Fully connected
Garbage model 2	Hand Polar velocities	20	4	Left Right
Gesture model	Hand Polar velocities	20	10	Left Right
SSG				
Garbage model	Cartesian velocities	30	8	Fully connected
Gesture model	Polar velocities	30	18	Fully connected
Standing up				
Garbage model	Polar velocities	30	8	Fully connected

Table 41: The used garbage and gesture HMM models and their parameters. These parameters are based on the performance on a validation set.

extract the features. Furthermore, the three remaining gestures use a garbage model. A garbage model is a model which tries to classify which parts are not (part of) a gesture. Then the parts remaining can be classified using the recognition model for the gesture under consideration. These garbage models act as an extra filter that reduces the number of false positives. For writing we used two garbage models instead of one because two garbage models gave better filtering performance. In fact the standing up gesture is determined entirely by a garbage model.

When gesture parts are segmented, a streamline approach is taken to find the gestures in a feature stream. In this approach, a sliding and expanding window is used. The window is aligned with the segmented boundaries, which makes the approach more efficient. The size of the window is determined by the minimum and maximum length of the annotated gestures for the class. We used the same parameters as in the manually segmented gestures, cf. Table 41.

For the detection of *negative signals* a different evaluation approach was taken. The evaluation is based on meeting TS3005a of the AMI corpus. This meeting contains 15 examples of negative signals. Each example was checked by eye before any automatic results were generated and assigned a subjective difficulty score with values ranging between 1 and 3, where 1 indicates head shakes with large-amplitude motion, 2 indicates medium-amplitude head shakes and 3 indicates head shakes with low amplitude motion, or no head shake at all. For each gesture we have selected a clip of the video 6 seconds long starting 3 seconds before the start of the gesture. We assume that the duration of all gestures is 1 second. Any time the gesture detector fires during the gesture we count a true positive and any time it fires outside the duration of the gesture we count a false positive.

5.4.2 Evaluation results

The classification performance, determined by the F-measure for the different gestures can be found in Table 42. The last column gives the performance for the streamwise approach.

Gesture	F-measure Segmented	F-measure Streamwise
Writing	63%	38%
SSG	85%	51%
Standing up	100%	66%

Table 42: The recognition performance, F-measure, for the different gestures on the manually segmented video stream (2nd column) and the streamwise approach (3th column).

The results of the *negative signal* detector are given in terms of the precision of the detector within the clip (the number of true positives divided by the total number of detections). Note that with 6 second clips and 1 second duration gestures a gesture detector that fires at random times would on average score a precision of 0.1667 and measures of recall are not very informative. The performance of the detector is listed in Table 43.

The average precision of the simple negative signal detector is 0.2. This result, which is not significantly different from a random detector, reflects both the limitations of the detector and the difficulty factor of the gestures (the gestures were very difficult to see in many of the examples with a difficulty score of 3). The precision scores were slightly better for 2 gestures with a difficulty score of 1.

5.4.3 Conclusions

All gesture recognition methods gave reasonable performance on the manually segmented video stream. Especially Standing up (recognition performance 100%) and Speech Supporting Gestures (SSG) (recognition performance 85%) can be recognized very accurately. Recognizing Standing up is considered easy,

File name	Event number	Difficulty Score	Precision
TS3005a.ID	002	2	0.2
TS3005a.ID	108	2	0.2
TS3005a.ID	118	3	0.0
TS3005b.UI	089	1	0.5
TS3005b.UI	115	2	0.0
TS3005b.UI	390	1	0.3
TS3005a.ME	027	3	0.3
TS3005a.ME	073	3	0.0
TS3005a.ME	087	3	0.2
TS3005a.ME	101	3	0.3
TS3005a.ME	283	3	0.2
TS3005a.ME	295	3	0.2
TS3005a.ME	343	3	0.3
TS3005a.ME	349	2	0.4
TS3005a.ME	351	3	0.2

Table 43: The performance results for the simple negative signal detector.

but SSG is a container for different type of gestures, but still it can be recognized very well. It should be noted that this score is on manually segmented video data, due to the bad performance of the considered segmentation methods. If one evaluates the models on streamwise video data then the performance drops significantly, cf. Table 42.

A novel approach was the introduction of the so called “garbage model” which is a model for the parts where there is no relevant gesture in the video data. This garbage model is used as a filter to reduce the number of false positives. The use of this garbage model increased the recognition performance. One of the reasons for this increase of performance could be that the gestures under consideration are sparse, hence there is a lot of data for training the garbage model.

The results of the negative signal detector are not significantly different from a random detector. This reflects both limitations of the detector and the difficulty of recognizing *negative signals* such as shaking. The problem of detecting head movement was also observed in the gesture recognition approach.

5.5 Summary and conclusion

The evaluated segmentation methods do not give good segmentation performance, cf. Table 39, for whole gestures, mainly due to the intrinsic structure of the considered gestures. An alternative approach could be define meaningful gesture parts, relevant features for detecting gesture parts and develop algorithms for gesture part segmentation.

The evaluated gesture recognition methods, all HMM based, gave reasonable results, cf. Table 42. Especially the introduction of so-called “garbage model” increased the performance. On streamwise data the performance dropped significantly.

Detecting gestures such as shaking and nodding and negative signals is still a challenging problem that will require methods capable of detecting very subtle head movements.

6 Affective Computing in Meetings Scenarios

6.1 Annotation of Emotions

6.1.1 Objectives

In the AMI project more than 100 hours of meeting recordings are being collected and annotated. One of the annotation levels involves the 'emotional state' of the participants. Interest and engagement levels of participants may signal hotspots in the meeting where important issues are being discussed that either enthuse the participants or that are highly controversial. The development of an annotation scheme for emotion involves many issues. The most important of these are probably the following questions: What types of emotions actually occur in the AMI recordings? How can we annotate effectively with the smallest effort? How reliably can these emotions be annotated? In a number of trials we have been experimenting with emotion schemes and procedures to develop a suitable emotion scheme.

6.1.2 Selection of emotion labels

It was expected that the AMI meetings would not contain many highly emotional episodes. To establish the kinds of emotions or affective dimensions one might expect in meetings we asked people (33 participants) to select 20 'emotion' terms that they thought would be frequently perceived in a meeting. The participants in this survey were presented with a list of about 200 emotion labels compiled from various sources. Participants could also suggest other emotions. The top 20 list (words that were mentioned most) were: bored (mentioned 23 times), confident, interested, attentive, serious, joking, friendly, curious, cheerful, at-ease, amused, relaxed, nervous, frustrated, decisive, uninterested, impatient, confused, agreeable, annoyed (mentioned by 9 participants). This list already shows that the terms that are being mentioned are not all emotions, strictly speaking. Is curiosity an emotion? What to think of respectful, dominant or tired, which are other terms in the top 80 list. Of course this depends on one's definition of emotion.

Besides labels for affective dimensions, it appeared from the reactions of the participants to the survey as well as from looking at the first recordings of AMI data, that other phenomena play an important role as well. Clearly, from the point of view of the relevance for meeting browsing and other techniques for building up memories of what happened in a meeting, it is obvious that what is relevant about what goes on in people's minds is not only what they 'felt' about what was being said in the emotional meaning of the word, but also whether they were surprised by the things that were said, certain, skeptical or how clear or confusing certain issues were presented.

In the first trial, people were asked to annotate 20 minutes of meeting data (involving one participant in a meeting) using the procedure defined by Cowie et al. (2000) using the FeelTrace tools from Belfast. This involves a continuous labeling of the emotional content on a plane involving two dimensions: arousal and valence. The annotation guidelines for this trial can be found on the AMI annotation group pages (/urlhttp://wiki.idiap.ch/ami/AmiAnnotGroups).

After analyzing the annotations in many different ways (e.g., absolute coordinate distances, direction of change relations, overlap in quadrants) it was concluded that the agreement between annotators was too low, which is probably due to several reasons. One of them is that the annotators were asked to annotate the data real-time on a single pass through the data. This inevitably leads to delays, false starts, etcetera. In addition, most of the observable changes in the mental states of the participants in the meeting do not directly relate to emotional dimensions. Annotators are noticing these changes and try to accommodate these some way or another in the annotation task. Although a more intensive training of the annotators might improve on the results, we felt that the results called for a change in the procedure altogether.

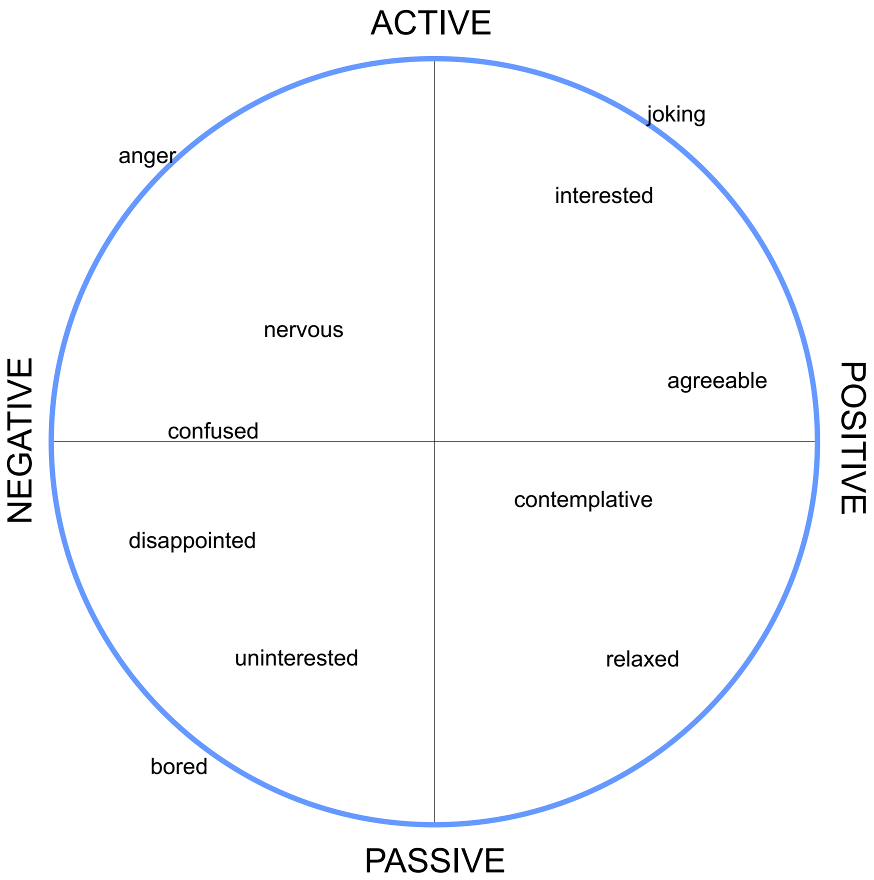


Figure 19: FeelTrace activation-evaluation space with meeting specific emotion labels

6.1.3 Second emotion annotation trial

On the basis of the results in the first trial we decided, firstly, to discretize the procedure and to provide a labeling framework that is more relevant to the mental state changes that occur in the meeting data, and secondly, to develop a training procedure using examples tailored to the domain.

In the new scheme the annotators are asked to segment the videos into 'emotion episodes', stretches of video in which a particular emotion is found to arise and disappear. For each of the fragments the annotator is required to indicate the intensity (on a 3 point scale) and the valence (positive/negative/NA) of the emotion, together with a label. The annotator can choose from a set of about 10 carefully selected emotion terms with a description, derived from the survey and the initial investigation of the data. Furthermore, annotators can provide terms of their own if they find another term more appropriate. A training procedure has been set up involving instructions on how to use the annotation tools and what the annotation procedure involves, providing several examples of when and where to segment a video and what labels to assign. These examples were taken from the meeting data in the AMI corpus. The annotation guidelines for this trial can be found on the AMI annotation group pages (/urlhttp://wiki.idiap.ch/ami/AmiAnnotGroups).

6.1.4 Results

The annotations exhibit two different patterns. Some segments can be characterized as long stretches of nothing interspersed with short relevant events. In such segments, the long stretches of nothing are annotated as a succession of segments labeled neutral or attentive. Other segments show in a short time more alternation of short segments with varied labeling. In both situation, a large confusion between attentive and neutral was observed. This might be because attending to what the others are saying is more or less the neutral state for participants in a meeting.

Based on this, we calculated derived agreement measures on the data. The first is agreement on 'something/nothing'. Taking 'neutral' and 'attentive' as 'nothing' and all others as 'something' yields kappa values between 0.55 and 0.7. The second is agreement on the evaluation dimension (rated positive/negative/NA). For the segments that have already been annotated this lies around 0.7.

From interviews with the annotators it became clear that there are important differences between annotators with respect to the signals they pay the most attention to. Some pay more attention to the facial expressions and the gestures and other to what is being said. This has effects on the annotation, both on the segmentation and the labels. So, in the process of annotation the first few meetings with 'emotional labels' we noticed that most of the labels we use relate to meta-cognitive functions. As we remarked above, many of the nonverbal expressions, even expressions typically association with emotions according to the literature (such as the six universal facial expressions associated with the six basic emotions by Ekman; see Ekman & Friesen, 1974, for instance) that we use are not directly expressing an emotion. We often found them in other contexts as well. Though we have not yet made a systematic analysis of the correlations between the behaviors correlated with the mental state annotations (as the annotations are still in a trial stage), we have started to look at the relations between facial expressions and other behaviors and the communicative actions participants take.

6.2 Detection of Emotions from Vision

Based on the closeup videos the research in AMI strives to estimate the participants' emotion from the information of head- and body pose, gestures and facial expressions. Therefore, the development and enhancement of the corresponding algorithms is crucial for emotion recognition by visual input. A description of activities can be found in the corresponding section. Independently, works are going on to analyze facial expressions. Very recent investigations are based on an application of the AdaBoost (Y. Freund, R.E. Schapire, 1996) algorithm applied on 2-dimensional Haar- and Gabor-Wavelet coefficients

(T.S. Lee, 1996), for localization of frontal faces and eyes (P. Viola, M. Jones 2002), as well as for classification of facial expressions (M. Pantic, L. Rothkranz, 200). Furthermore, an approach based on Active Appearance Models (T. Cootes, 2001) is implemented and investigated in its application to head pose estimation and facial expression analysis. Even though this method shows high requirements to computational performance of the applied hardware, the expected results should argue for this approach. Final results are available at the beginning of February 2006 and will be presented within the subsequent WP4 plenary meeting.

7 Identification

7.1 Objectives

Face recognition is a general topic that includes both face identification and face authentication (also called verification). On one hand, face authentication is concerned with validating a claimed identity based on the image of a face, and either accepting or rejecting the identity claim (one-to-one matching). On the other hand, the goal of face identification is to identify a person based on the image of a face. This face image has to be compared with all the registered persons (one-to-many matching).

The problem of face recognition has been addressed by different researchers using various approaches. These approaches can be divided into *discriminant* approaches and *generative* approaches. A *discriminant approach* takes a binary/multi-class decision and considers the whole input for this purpose. Such *holistic* approaches are using the original gray-scale face image or its projection onto a Principal Component subspace (referred to as PCA or Eigenfaces) or Linear Discriminant subspace (referred to as LDA or Fisherfaces) as input of a discriminant classifier such as Multi-Layer Perceptrons (MLPs), Support Vector Machines (SVMs) or simply a metric.

Recently, it has been shown that **generative approaches** such as Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) were more robust to automatic face localization than the above discriminant methods. A generative approach computes the likelihood of an observation (a holistic representation of the face image) or a set of observations (local observations of particular facial features) given a client model and compares it to the corresponding likelihood given an impostor model.

During recent international competitions on face authentication [104], it has been shown that the discriminant approaches perform very well on manually localized faces. Unfortunately, these methods are not robust to automatic face localization (imprecision in translation, scale and rotation) and their performance degrades. On the opposite, generative approaches emerged as the most robust methods using automatic face localization. This is our main motivation for developing generative algorithms [25, 24]. We proposed to train generative models, such as Gaussian mixture models (GMMs), one-dimensional hidden Markov models (1D-HMMs) and pseudo two-dimensional hidden Markov models (P2D-HMMs), using maximum a posteriori (MAP) training instead of the traditionally used maximum likelihood (ML) criterion. We experimentally demonstrated the superiority of this approach over other training schemes. The main motivation for the use of MAP training is the ability of this algorithm to estimate robust model parameters when there is only a few training images available.

7.2 Evaluation method

Currently, we are evaluating the algorithms on a face verification task using a well-known benchmark database.

7.2.1 Benchmark Database

The BANCA database [3] was designed in order to test multi-modal identity authentication with various acquisition devices (2 cameras and 2 microphones) and under several scenarios (controlled, degraded and adverse). For 5 different languages (English, French, German, Italian and Spanish), video and speech data were collected for 52 subjects (26 males and 26 females), i.e. a total of 260 subjects. Each language - and gender - specific population was itself subdivided into 2 groups of 13 subjects (denoted $g1$ and $g2$). Each subject participated to 12 recording sessions, each of these sessions containing 2 records: 1 true *client access* (T) and 1 informed [4] *impostor attack* (I). For the image part of the database, there is 5 shots per record. The 12 sessions were separated into 3 different scenarios.

[3] `http://www.ee.surrey.ac.uk/banca`

[4] The actual speaker knew the text that the claimed identity speaker was supposed to utter.

7.2.2 Performance Evaluation

The authentication decision is then reached as follows. Given a threshold τ, the claim is accepted when $\Lambda^*(X,Y) \geq \tau$, and is rejected when $\Lambda^*(X,Y) < \tau$. This threshold is chosen to optimize a given criterion such as the Equal Error Rate (EER), i.e when $FAR = FRR$ (Fig. 20).

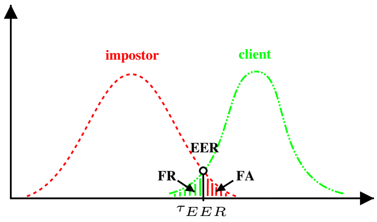


Figure 20: Illustration of typical errors of a biometric system.

FRR is the False Rejection Rate (when the system rejects a client), FAR is the False Acceptance Rate (when the system accepts an impostor), $HTER$ is the Half Total Error Rate (an unique measure given by $HTER = \frac{FRR+FAR}{2}$).

7.3 Results

7.3.1 Performance Evaluation on Benchmark Databases

We present results (in terms of HTER), on the BANCA database, obtained by several IDIAP generative models and baseline state-of-the-art systems (Table 44) with an automatic face localisation. In order to provide an unbiased evaluation of the performance, the decision threshold has to be chosen a priori (not optimize on the test set itself). Thus, we determine the threshold τ on the development set which minimizes the EER criterion.

System	Protocol			
	Mc	Ud	Ua	P
PCA	22.4	29.7	33.7	29.0
LDA/NC (from [132])	22.6	25.4	27.1	25.2
SVM (from [132])	19.7	30.4	33.2	27.8
GMM	**9.5**	**21.0**	**24.8**	**19.5**
1D-HMM	**13.8**	**25.9**	**23.4**	**21.7**
P2D-HMM	∗ **6.5**	∗ **15.9**	∗ **14.7**	∗ **14.7**

Table 44: HTER performance for different experiment protocols

Note that the result table presented here contains performance figures for the two best systems reported in [132]. The first system is based on combination of Linear Discriminant Analysis and Normalized Correlation (LDA/NC), while the second system is based on a Support Vector Machine (SVM) classifier. Like the PCA based system, these LDA/NC and SVM systems are holistic in nature.

Results show that generative models are providing better results than discriminant models. The best results are achieved by P2D-HMM. However, it should be noted that P2-HMM are also much slower (18.80 seconds to process 5 images) than GMM (only 0.24 seconds to process the same images).

7.3.2 A Torch Machine Vision Package for Face Recognition

We developed a machine vision package for **Torch** (a well-known open source machine learning library) called **Torch vision**. **Torch vision** [5] provides basic image processing and feature extraction algorithms but also several modules for *face recognition*.

7.4 Conclusion and summary

We developed robust-to-localisation generative models for face recognition. In order to obtain preliminary results, these algorithms have been evaluated on a face verification task using a well-known benchmark database.

We will apply these generative models to the AMI domain (face identification task) on AMI data (meeting room recordings).

[5] `http://www.idiap.ch/~marcel/en/torch3/introduction.php`

8 Speaker segmentation and clustering

8.1 Objectives

The objective of this work is to be able to segment, cluster and recognize the speakers in a meeting based on their speech. Speaker information can be included in the meeting browser so that the used will have a better understanding of what is going on and will have a better context of the contents (such as the transcripts).

Several technical approaches exists, some of which are

Channel energy Using the acoustic energy of individual speaker's microphones,

Voice characteristics Using the spectral, phonetic and idiolectical content od the speech signal,

Source localization Using directional information obtained from microphone arrays.

Within the work carried out for AMI at TNO, we have worked on the approach of using the acoustic contents of the microphone signal to segment and cluster speakers. This extends our earlier work on speaker recognition (for telephone speech) and speaker segmentation/clustering (for broadcast news).

In the AMI WP4 meeting in Prague the opportunity to participate in the NIST Rich Transcription Evaluation on Meeting Data in the spring of 2005 (RT05s) was discussed. This evaluation contained a *speaker diarization* (who spoke when?) track, which seemed to be an ideal task to work on. The RT05s evaluation data contained AMI meeting recording data.

8.2 Evaluation method

The evaluation took place in the typical NIST speech technology evaluation series. An evaluation period (two weeks) is defined. At the beginning of this period, the evaluation data is sent to participants, who can run their systems in order to perform a *task* specified by NIST. At the end of the period the results are sent to NIST, who then scores the results. After a little while, these results are disseminated to the participants and the reference files (the true 'answers' for the task). Before the evaluation period commenses, NIST defines and distributes 'development data' which is similar in structure and character to the evaluation data, and contains the reference truth as well. With this materials the participants can develop and tune their systems.

The evaluation measure for speaker diarization is the speaker diarization error rate (SDE). This measure is basically the fraction of time attributed to the wrong speaker compared to the spoken time. The precise definition is quite complicated because the measure has to incorporate periods in time where there is more than one speaker speaking. Note that the *absolute* identity of the speakers is not required, the speakers found by the system can be identified by arbitrary names, e.g., 'a,' 'b,' 'c,' etc.

Because the denominator in the SDE is formed by spoken time, it is quite essential to keep the 'false alarm' time low. This means, that the system must incorporate a good speech activity detection (SAD) algorithm.

8.3 Results

The evaluation set contained 10 meetings in total, two meetings each from five different sources. One meeting source was AMI, for which the two meetings were obtained from both the Edinburgh and IDIAP location. TNO participated in both the Speech Activity Detection task and the Speaker dizarization task. The results are shown in table 45.

The results on the bottom line is the performance obtained in the RT05s evaluation. Although these are the results submitted for the primary *multiple distant microphones* condition we actually only used the information from one central distant microphone.

Table 45: SAD and speaker diarization results, in % error, for non-overlapping speaker segments. The last three columns show SDE results where the input to the clustering system is either our own SAD (primary evaluation system), ICSI's SAD and perfect SAD (post-evaluation).

Test	SAD error	SDE		
		SAD input from:		
		TNO	ICSI	perfect
AMI dev	10.3	35.7		45.9
RT04s − CMU	2.8	35.4		31.9
RT05s	5.0	35.1	37.1	32.3

8.4 Conclusion and summary

We have extended our broadcast news speaker segmentation/clustering system to operate on meeting data. We obtained very competative results in the NIST RT05s evaluation for speech activity detection (the lowest error rate reported) and our speaker diazrization system performed satisfactorily, given the technology we used. The best system, that of ICSI/SRI (ICSI also an AMI partner) showed about half the SDE reported here. We hope to participate in next year's speaker diarization task, and to perhaps use information from multiple microphone sources, and integrate efforts with other AMI groups.

9 Focus of Attention analysis

9.1 Objectives

The objective of this part is to study tasks related to the focus-of-attention (FOA) of meeting participants, where we decided to restrict the definition of the focus-of-attention of people to the spatial locus defined by the person's gaze[6]. Using this definition, the identified research tracks related to the FOA were the following:

- the first track is concerned with the **recognition** of the FOA. More precisely, given recorded meeting data streams, can we identify at each instant the FOA of people ? Thus, the research direction for this task is the study and development of gaze estimation algorithms, or, as a surrogate, of head orientation estimation algorithms. Tasks 1 and 2 below are in this direction.
- in the second track, the objective is **to identify the role played by the FOA** in the dynamics of meeting (e.g. can we predict the current speaker from the FOA of all participant). Answering such questions will be useful to understand the relationship between the FOA and other cues (such as speaker turns) as well as to more precisely identify the interactions between participants (e.g. by contributing to the recognition of the higher level dialog acts), which in turn could translate into better FOA recognition algorithms. The investigated tasks 3 and 4 belong to this objective.

We are currently working in both directions. The next section will describe the tasks that have been investigated up to now, as well as the databases and protocols employed to evaluate these tasks. Section 9.3 will summarize the methods employed to address the tasks, along with the obtained results and comments. Finally, Section 9.4 will draw the main conclusions and present future AMI investigations.

9.2 Evaluation method

9.2.1 Tasks

During the past year, the following tasks have been studied and evaluated in AMI:

- **FOA.T1 task**: head pose and head tracking.
  One first step towards determining a person's FOA consists of estimating its gaze direction. Then, from the geometry of the room (object, cameras) and the location of meeting participants, the FOA can normally be estimated. However, as estimating gaze is difficult (and requires very close-up views of people to assess the position of the pupil in the eye globe), we have developped, as an approximation, algorithms for tracking the head and estimate its pose.
- **FOA.T2 task**: recognition of focus-of-attention.
  In this task, the emphasis is on the recognition of a finite set $\mathcal{F}$ of specific FOA loci. One approach to this problem, that we have first studied, corresponds to the mapping of head orientations to FOA labels. However, other strategies that model people interactions and fuse multiple cues will be investigated in the upcoming year.
- **FOA.T3 task/experiment**: perception of head orientation in a Virtual Environement.
  This task consists of assessing how accurately people perceive gaze direction.
- **FOA.T4 task/experiment**: identifying speaker amongst meeting participants.
  In this experiment we investigate whether observers use knowledge about differences in head orientation behavior between speakers and listeners by asking them to identify the speaker in a four-person setting.

[6] This restriction was necessary to avoid ambiguity in definition/annotation, for instance if the FOA would be defined as some internal mental state.

9.2.2 Databases

D1: IHPD, the IDIAP Head Pose Database (available at `http://mmm.idiap.ch/HeadPoseDatabase/`)

purpose : the purpose of this database is to allow numerical evaluation of both tasks **T1** and **T2**. In most research works on head pose estimation, algorithms are assessed by visual inspection on few sequences. However, in view of the limitations of visual evaluation, and the inaccuracy obtained by manually labeling head pose in real videos, we decided to record a video database with head pose ground truth produced by a flock-of-birds device. At the same time, as the database is also annotated with the discrete FOA of participants, we will be able to evaluate the impact of having the true vs an estimated head pose on the FOA recognition.

description : the database has the following characteristics:

- content: the database comprises 8 meetings of 4 people (duration of meetings ranged from 7 to 14 minutes), recorded in IDIAP's smart meeting room. The scenario of the meeting was to discuss statements displayed on the projection screen. Due to technological constraints we were able to capture the head ground truth of only two participants (the left and right person in Fig. 21), using 3D magnetic sensors attached to the head.
- head pose annotation: the head pose configuration with respect to the camera was ground truthed. This pose is defined by three Euler angles (α, β, γ) which parameterize the decomposition of the rotation matrix of the head configuration with respect to the camera frame. Among the possible decompositions, we have selected the one whose rotation axes are rigidly attached to the head to report and comment the results. With this choice, we have: α denotes the pan angle, a left/right head rotation; β denotes the tilt angle, an up/down head rotation; and finally, γ, the roll, represents a left/right "head on shoulder" head rotation.
- foa annotation: for a given person ('left' and 'right' in Fig. 21), the set of potential focus is composed of the other participants, the slide-screen, the table, and an additional label (unfocused) when none of the previous could apply. In the IHPD meetings, the whiteboard was not used. As a person can not focus on himself/herself, the set of focus is thus different from person to person. For instance, for the left person, we have: $\mathcal{F} = \{right\_person, organizer1, organizer2, slide\_screen, table, unfocus\}$. The guidance for the annotation were the same as for the annotation of the AMI meetings (see below).

D2: AMI database: a subset of the AMI meetings were annotated with FOA ground truth.

purpose: the FOA annotation will be used as ground truth to evaluate FOA recognition algorithms involving non-verbal people interaction modeling and cue fusion (something that can not be done with the IHPD, as several pieces of information are missing there), and as data input for higher cognitive tasks (e.g. addressee detection, cf workpackeg WP5).

description: the FOA annotation of the AMI meetings is specifically characterized by the following elements:

- content: these are the standard AMI meetings, which features four persons with specific roles (e.g. project manager, marketing expert) involved in the design of a new remote control, during the course of four meetings.
- amount: 12 meetings from the IDIAP smart meeting room (SMR), 1 meeting from TNO and 1 meeting from the Edimburg SMR. The choice made by IDIAP and University of Twente of the IDIAP SMR as main source for annotation is motivated by the presence of medium range cameras which simultaneously allows for a better annotation of the FOA (than long range or close-up cameras), and for conducting FOA recognition experiments.

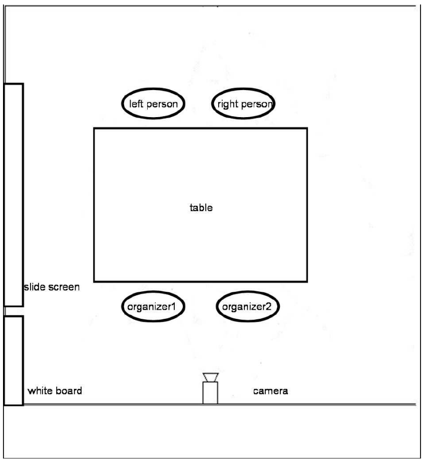


Figure 21: FOA: the set $\mathcal{F}$ of potential FOA comprises: other participants, the table, the slidescreen, the whiteboard, and an unfocus label when none of the previous applies.

- annotation: the FOA of each participant was annotated using specific guidelines [112] (available on the AMI wiki pages). With respect to the original plan (cf deliverable D4.1), the set of FOA loci has been reduced to 8, mainly to reduce annotation costs. They are: the participants, Px, the table, the slide-screen and whiteboard, and an unfocus label (see Fig. 21).

D3: Conversation corpus

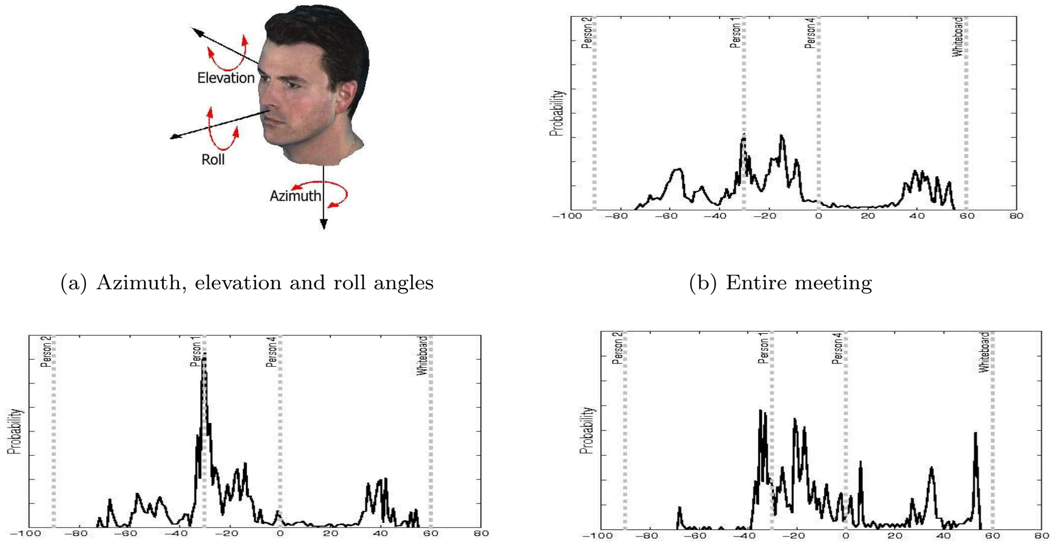
purpose : We have collected a corpus of conversations in which the head-orientations of the participants were tracked by flock-of-bird sensors. This allows us to study with more precision the differences in head-orientations for the various roles. Some of the findings are reported below.

description : Three meetings with a total duration of 21 minutes were recorded in the IDIAP (Institut Dalle Molle d'Intelligence Artificielle Perceptive) smart meeting room in Martigny, Switzerland (Figure 25(a)). Each meeting consisted of debates about three issues. These were presented to the participants on the whiteboard. The four meeting participants were sitting two-by-two, at opposite sides of the table. Three cameras and an overhead microphone were used to record the audio and video. The head positions and orientations of all meeting participants were tracked using electromagnetic sensors at a rate of 50 Hz. The Flock of Bird sensors we used, Ascension Technology 6DFOB, have an orientation accuracy of 0.5°. Each sensor is only a small box and when mounted on top of a participant's head it does not cause any distraction during the meeting.

	Meeting 1	Meeting 2	Meeting 3	Total
Samples	28148	13078	11333	52559
Turns	214	85	92	391

Table 46: Number of samples and turns per meeting

We analyzed head orientation and video data to discover possible biases due to incorrect mounting of the Flock sensor on the head. We corrected the orientation data for these biases, which were all within the {-10°, 10°} interval. For each participant, an azimuth orientation angle (Figure 22(a)) of 0° corresponds to looking straight forward and looking to the right corresponds to a positive



(a) Azimuth, elevation and roll angles

(b) Entire meeting

(c) When Person 1 is speaking

(d) When Person 3 is speaking

Figure 22: Distribution of azimuth angles for Person 3 in meeting 3

rotation. The speech was transcribed manually from the audio recordings allowing us to determine who was speaking at any time. Head orientation data and speaker data were time aligned and all occurrences with non-speech or with speech overlap were removed from the data set.

The analysis of this corpus gave the following findings:
As all the participants are located within the same elevation-roll plane, the azimuth angle is the most informative rotation to distinguish the different focus of attention targets. If one plots the azimuth angle distribution for each of the persons over a whole meeting one can see that the peaks in the orientations of the head correspond more or less with items of interest. An example of such a plot is displayed in Figure 22(b) where the distribution of azimuth angles of Person 3 is given. The locations of the others and the center of the whiteboard are indicated with dotted lines. The seating arrangement of the participants is presented in Figure 25(b). The graph shows that the four peak areas correspond more or less with the three other participants and the whiteboard.

The fact that the correspondence is not exact has several explanations. First of all, we solely used head orientation and ignored the head position. By leaning forward or backward, the relative positions between the persons change. Also, part of the gaze direction is constituted by eye orientation. We expect this to be the main reason for the fact that the orientations towards the two people sitting at the opposite side of the table tends to be a bit in between the persons.

Figure 22(c) correlates the orientation of the head with information about the person who is speaking. It shows the distribution of azimuth orientation angles of Person 3, when Person 1 is speaking. The figure shows a clear indication for the expected correlations between head orientation of listeners towards the current speaker. The highest peak reveals that Person 3 is directing his head mostly towards Person 1, when Person 1 is speaking. We obtained similar graphs, for all the other speaker-listener combinations. Generally the highest peak in these plots corresponds to the location of the speaker. Results of a quantitative analysis of where speakers and listeners rotate their heads towards, are given in Table 47. We defined a person as being looked at by another person if the head orientation of the latter was within a range of {-15°, 15°} from the angle between them, as calculated from the mean position of the head during a meeting.

	Looks at		
Role	Speaker	Listeners	Other
Speaker	N.A.	88.32%	11.68%
Listener	46.39%	42.65%	10.97%

Table 47: Percentage of time that speakers and listeners look at the others

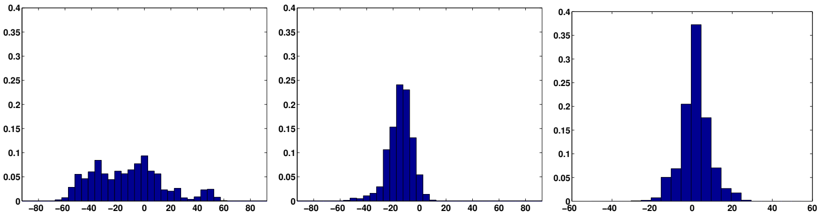


Figure 23: Histograms of pan, tilt and roll values of the evaluation data in task 1.

The table shows that listeners' heads are oriented towards the speaker 46.39% of time and to one of the two other listeners in 42.65%. [160] found for a comparable setting with four persons that listeners gazed at the speaker in 62.4% of the time, while speakers gazed at listeners in approximately 55% (as calculated from the data by the authors). It appears that his findings, when using gaze, are indeed in line with ours. In part, the differences can be explained by the criteria that was used to determine who was looked at. [160] used an eye-tracker and considered all gazes within the face as eye gaze. But is should be granted that differences in the conversational setting, the task, or between individuals will have an influence as well. A bigger difference between our results and those of [160] is that in his work listeners gazed 7.3 times more at the speaker than at the other listeners. Although the amount of time that listeners look at the speaker (46.39%) is significantly higher ($t(52558) = 128.86$; $p < 0.0001$) than the time spent looking at each of the other two listeners (21.33%), the difference is only of factor 2.

If we take a closer look at how many heads were oriented towards a listener or the speaker we obtain the results from Table 48. This table shows that in 13.88% of the time three people have their heads oriented towards the speaker, compared to only 1.98% for listeners. The table shows clearly that a speaker is generally being looked at by more people than a listener.

	Looked at by number of persons			
Role	3	2	1	0
Speaker	13.88%	28.56%	35.56%	22.00%
Listener	1.98%	12.92%	40.73%	44.38%

Table 48: Percentages of time that a certain number of meeting participants had their heads oriented to either a speaker or a listener

We can conclude that there are differences between speakers and listeners with respect to the head orientations in the azimuth plane that indicate the focus of attention similar to the findings in earlier studies that looked at gaze.

9.2.3 Protocols

The following protocols were used to evaluate the algorithms of tasks **T1-T4**.

Task 1: head pose estimation

data: we used the IHPD database here. Amongts the 16 recorded people, we used half of the database (8 people) as training set to learn the pose dynamic model and the half remaining as test set to evaluate the tracking algorithms. In addition, from the 8 meetings of the test set, we selected 1 minute of

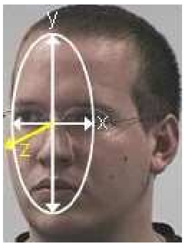


Figure 24: Illustration of the pointing vector.

recording (1500 video frames) for evaluation data. This decision was made to save machine computation time, as we use a quite slow matlab implementation, and to easily allow for different parameterization testing. Figure 23 shows the distribution of the pan (or azimuth), tilt (or elevation), and roll values on the evaluation data. Because of the scenario used to record data, people often have negative pan values corresponding to looking at the projection screen. But the pan values range from -60 to 60 degree. Tilt values range from -60 to 15 degrees and roll value from -30 to 30 degrees.
performance measures: four error measures are used. The three first measures are the errors in pan, tilt and roll angle, i.e. the absolute difference between the pan, tilt and roll of the ground truth (GT) and the tracker estimation. Also, as a head pose defines a vector in the 3D space, the vector indicating where the head is pointing at (cf Figure 24), the angle between the 3D pointing vectors defined by the head pose GT and the pose estimated by the tracker can be used as pose estimation error measure. This vector depends only on the head pan and tilt values (given the selected representation), and is directly related to the FOA recognition.

Task 2: FOA recognition
data: for these first experiments on FOA recognition, we exploited the IHPD database. This will allow us to compare the difference between using either the head pose ground truth or the estimated head pose to infer the FOA of people. Experiments on FOA recognition are done separately for the left and right person (see Fig. 21). Thus, for each seating position, we have 8 sequences. We adopt a leave-one-out protocol, where for each sequence, the parameters of the recognizer that is applied to this sequence are learned on the 7 other sequences.
performance measures: two different types of measures are used.

- *frame-based recognition rate*: this corresponds to the percentage of frames in the video whose estimated FOA match the ground truth label. While this measure is interesting from the pure recognition point of view, it emphasis the events that are longer (i.e. when someone is continuously focused) and may not reflect whether we are able to capture shorter focus of people, which might be important in understanding meeting dynamics and human interaction.

- *event-based recall/precision*: we are given two sequences of FOA events, : the recognized sequence of FOA, $R = \{R_i\}_{i=1..N_R}$, with each event R_i characterized by its label l_i (one of the target FOA), and its duration interval $[b_i, e_i]$, with starting time b_i and ending time e_i; similarly, we have the ground truth sequence $G = \{G_j\}_{j=1..N_G}$ with similar definitions. To compare the 2 sequences, we first need to align the two sequences. This was done using an adapted string alignment procedure[7]. Given this alignement, we can then compute for each event $l \in \mathcal{F}$, the *recall* ρ and *precision* π measures of that event, defined as:

$$\forall l \in \mathcal{F}, \rho(l) = \frac{N_{matched}(l)}{N_G(l)} \text{ and } \pi(l) = \frac{N_{matched}(l)}{N_R(l)} \tag{7}$$

[7] Indeed, we have here to take time into account when doing the alignment: a recognized event $R_i = (l_i, b_i, e_i)$ can only be said to match a ground truth event $G_j = (l_j, b_j, e_j)$ if not only the labels are corresponding (i.e. $l_i = l_j$), but the times at which the events occur are intersecting a minimum (i.e. $[b_i, e_i] \cap [b_j, e_j] \neq \emptyset$).

where $N_{matched}(l)$ represents the number of events l in the recognized sequence that match the same event type in the ground truth after the alignment, $N_R(l)$ denotes the number of occurence of event l in the recognition sequence, and $N_G(l)$ denotes the number of occurence of l in the ground truth. Qualitatively, the recall of l indicates the percentage of true looks at the FOA l that were recognized, while the precision indicates the percentage of looks at l that were recognized that indeed corresponds to the ground truth. To obtain a composite value[8], we use the F *measure*, defined as the harmonic mean of the precision and recall, i.e.

$$\frac{1}{F_{meas}(l)} = \frac{1}{2}(\frac{1}{\rho(l)} + \frac{1}{\pi(l)}) \tag{8}$$

Finally, the performance for a given person FOA sequence are computed through averaging:

$$\rho = \frac{\sum_{l \in \mathcal{F}} \rho(l)}{|\mathcal{F}|} \quad , \quad \pi = \frac{\sum_{l \in \mathcal{F}} \pi(l)}{|\mathcal{F}|} \quad (\text{but } \frac{1}{F_{meas}} = \frac{1}{2}(\frac{1}{\rho} + \frac{1}{\pi})) \tag{9}$$

Finally, the performance measures for the whole database are obtained through averaging of the individual recall, precision and F measures.

Task 3 and 4: perception of head orientation in a Virtual Environement (VE); identifying speaker amongst meeting participants

To evaluate how observers can identify the speaker amongst meeting participants, we have used a virtual environment in which the meeting room is replicated in 3D and the participants are replaced by avatars. The data collected from the converstational corpus serves to animate the head movements of the avatars. If observers are aware of the systematics in head orientation behavior for speakers and listeners, one should expect them to be able to correctly identify the speaker in a significant number of cases.

The use of a virtual environment (VE) has certain advantages for this kind of research. Within perception research, it is of major importance to have a good control over the stimulus. With the traditional tools that are used for research into conversational behavior such as video, it is difficult to focus on one single modality (for example speech, facial expression, gesture or gaze) while ignoring all others. This limitation makes it hard to study in detail how humans interpret the effect of a specific modality on the conversation. In our research, we address this limitation by using a 3D virtual environment, as suggested by [147]. This allows full control over all stimuli which makes it an appropriate tool for research into human perception [93]. As VE we use the virtual meeting room (VMR, Figure 25(b)), described in [124] to manipulate head orientations while keeping all other modalities unchanged. More details about the specific setup are given with the results.

9.3 Results

In the following, we present the results we obtained for the different tasks. We first start by describing the different algorithms tested.

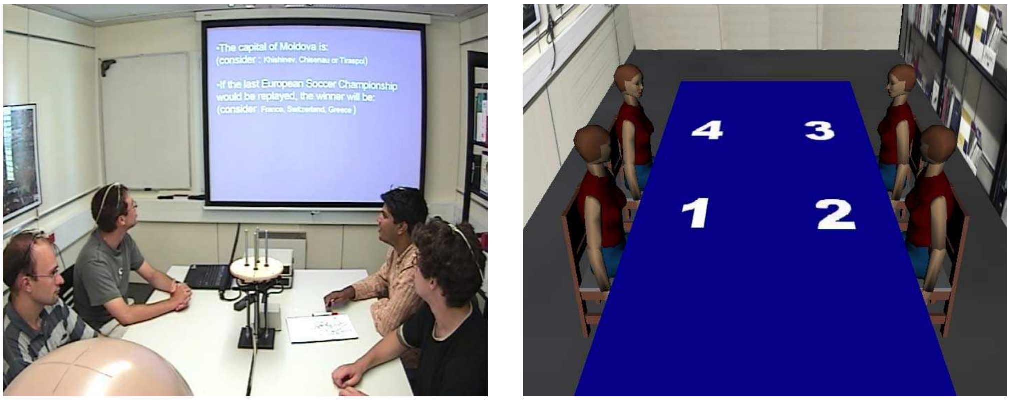
9.3.1 Task 1: results on head pose estimation

Algorithms:

During the 2nd year of AMI, we continued the work pesented in the D4.1 deliverable. That is, we formulate the coupled problems of head tracking and head pose estimation in a Bayesian filtering framework, which is then solved through sampling techniques. In this paragraph, we sumarize the main points of our approach. More details can be found in references [8, 9].

The Bayesian formulation of the tracking problem is well known. Denoting by X_t the hidden state

[8] In many circumstances, increasing the recall tends to decrease the precision, and vice-versa.



(a) (b)

Figure 25: (a) Real meeting setting and (b) Virtual Meeting Room setting

representing the object configuration at time t, and by Y_t the observation extracted from the image, the objective is to estimate the filtering distribution $p(X_t|Y_{1:t})$ of X_t given all the observations $Y_{1:t} = (Y_1 \dots Y_t)$ up to the current time. This can be done through a recursive equation, which can be approximated through sampling techniques (or particle filters PF) in the case of non-linear and non-Gaussian models. The basic idea behind PF consists of representing the filtering distribution using a weighted set of samples $\{X_t^n, w_t^n\}_{n=1}^{N_s}$, and updating this representation as new data arrives. That is, given the particle set at the previous time step $\{X_{t-1}^n, w_{t-1}^n\}$, configurations at the current time step are drawn from a proposal distribution $q(X_t) = \sum_n w_{t-1}^n p(X_t|X_{t-1}^n)$. The weights are then computed as $w_t^n \propto p(Y_t|X_t^n)$.

Four elements are important in defining a PF:

1. the *state space*, which defines the elements we are looking for.
2. the *dynamical model* $p(X_t|X_{t-1})$ defines the temporal evolution of the state.
3. the *observation likelihood* $p(Y_t|X_t)$ measures the adequacy between the observation and the state. This is an essential term, where data fusion occurs, and whose modeling accuracy can greatly benefit from additional discrete variables in the state space.
4. the *sampling mechanism* places new samples as close as possible to regions of high likelihood.

These elements along with our model are described in the next paragraphs.

State space: The state contains both discrete and continuous variables. More precisely, the state $X = (S, \gamma, l)$ is the conjunction of a discrete index $l = (\theta, k)$ which labels an element of the set of head pose models e_k^θ, while both the discrete variable γ and the continuous variable $S = (x, y, s^x, s^y)$ parameterize the transform $\mathcal{T}_{(S,\gamma)}$ which characterizes the image object configuration. (x, y) denotes the position of the object, (s^x, s^y) denote the object width and height scales, and γ the in-plane rotation.

Dynamical model: The graphical model in Figure 26 describes the dependencies between our variables. from which the equation of the process can be defined. The chosen model, learned from training sequences, allows to set some prior on the head eccentricity, as well as to take into account the head rotation dynamic.

The observation model: $p(Y|X)$, where the observation Y are composed of texture and color observations (Y^{text}, Y^{col}), was defined as follows :

$$p(Y|X = (S, \gamma, l)) = p_{text}(Y^{text}(S, \gamma)|l)p_{col}(Y^{col}(S, \gamma)|l), \tag{10}$$

where the texture likelihood p_{text} and the color likelihood p_{col} were learned from the Pointing database [38]. The parameters (S, γ) allow to extract an image patch, on which the features are computed, while

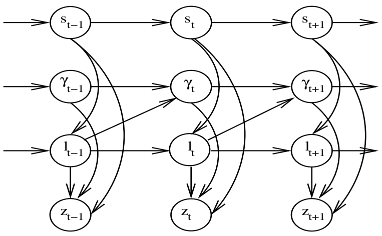


Figure 26: Mixed State Graphical Model.

	pointing vector			pan			tilt			roll		
	mean	std	med	mean	std	med	mean	std	med	mean	std	med
MSPF	22.5	12.5	20.1	10.0	9.6	7.8	19.4	12.7	17.5	11.5	9.9	8.8
RBPF	20.3	11.3	18.2	9.10	8.6	7.0	17.6	12.2	15.8	10.1	9.9	7.5

Table 49: Mean, standard deviation and median of errors on the different angles.

the examplar index l allows to select the appropriate appearence model.
The sampling mechanism: here we studied two different approaches. The first one, denoted MSPF, was a plain particle filter. The second approach, RBPF, the Rao-Blackwellisation, consists of applying the standard PF algorithm over the tracking variables S and γ while applying an exact filtering step over the exemplar variable l. The method theoretically results in a reduced estimation variance, as well as a reduction of the number of samples.

Results

Table 49 shows the pose errors for the two methods over the test set. Overall, given the small head size, and the fact that none of the head in the test set were used for appearence training, the results are quite good, with a majority of head pan errors smaller than 10 degrees. However, these results hide a large discrepency between individuals, as the mean errors for each person of the test set show (Fig. 27). This variance depends mainly on whether the tracked person resembles one of the person of the training set used to learn the appearance model. The table also shows that the errors in pan and roll are smaller than the errors in tilt. This is due to the fact that, even in a perceptive point of view, discriminating between head tilts is more difficult than discriminating between head pan or head roll [17]. Finally, as can be seen, the errors are smaller for the RBPF than for the MSPF, though not being statistically significant. This improvment is mainly due to a better exploration of the configuration space of the head poses with the RBPF, as illustrated in Figure 28 which displays sample tracking results of one person of the test set. The first row presents the results from the MSPF while the second row shows those of the RBPF for the same time instants. Because of a sudden head turn, the MSPF lags behind in the exploration of the head pose configuration space, to the contrary of the RBPF approach which nicely follows the head pose.

9.3.2 Task 2: results on FOA recognition

The goals of the experiments on FOA recognition are to evaluate 1) whether the head-pose information is sufficient to infer the visual FOA of the participants 2) the degradation in recognition when using the automatically estimated head-poses (cf task 1) instead of the ground truth head poses.

Algorithms:

For a given participant location in the room, we used a simple FOA recognition algorithm. Let us denote by $\vec{v}$ the head pointing vector of a participant (cf Figure 24), and by f a focus class. For each focus different than the unfocused one, we model the class-conditional likelihood as a gaussian:

$$p(\vec{v}|f) = \mathcal{N}(\vec{v}; \mu_f, \Sigma_f) \tag{11}$$

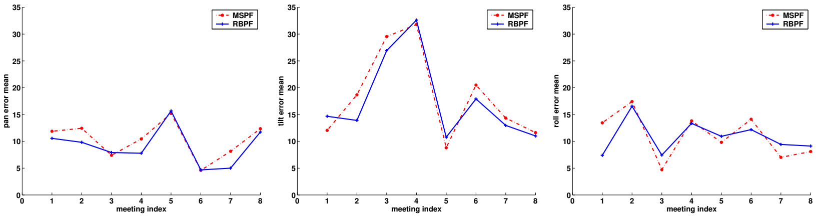


Figure 27: Pan, tilt, and roll errors over individual participants.

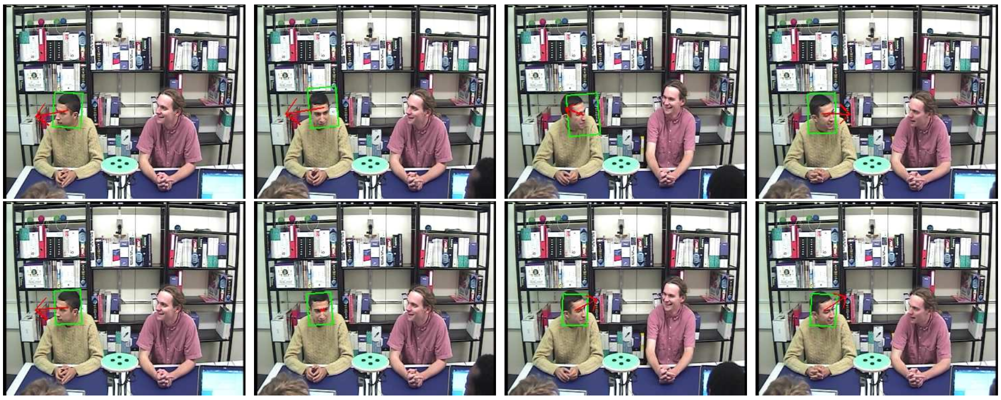


Figure 28: Sample of tracking failure for MSPF. First row : MSPF; Second column: RBPF.

where μ_f and Σ_f represent the mean and (full) covariance matrices that are learned from the data in the training set. Besides, the probability density function (pdf) of the unfocus label is modelled as a uniform density over the pointing vector range. Then, for recognition, we used a the Maximum A Posteriori (MAP) principle, that is:

$$\widehat{f}_t = \arg\max_{f \in \mathcal{F}} p(f|\vec{v}_t) = \arg\max_{f \in \mathcal{F}} \frac{p(\vec{v}_t|f)p(f)}{p(\vec{v}_t)} = \arg\max_{f \in \mathcal{F}} p(\vec{v}_t|f) \tag{12}$$

As can be seen, this scheme (denoted as GMM in the result section) does not exploit any temporal smoothing. This was done by using a Hidden Markov Model (HMM) as sequence modeling. In this case, in addition to the class-conditional pdf, the transition matrix $p(f_t|f_{t-1})$ is learned from the training data and used in the Viterbi decoding algorithm.

Results

Two sets of experiments need to be distinguised: those where we use the ground-truth $\vec{v}^{gt}$ as input data both in the training and test phase, and those where we use the output of the head-pose tracker (cf Task 1).

With ground truth head pointing vector:

Tables 50 and 51 provide the classification results we obtain. Firstly, we can remark that these results are not as high as one could expect, despite the use of ground truth data: around 68% and 46% frame recognition rate for left and right person respectively, and 70% and 62% in event recognition. They are below the numbers reported in other work ([143]): around 88%. There are several explanations for these results and differences. The two main ones are: firstly, in [143], the analysis of focus was restricted to participants (and thus involved less labels) and these four participants were equally distributed around a round table, making it almost necessary to accompany gaze shifts with head pose shifts. This was not the case in our setting. This illustrates that the correlation between gaze and head-pose is very dependent on

test set	1	2	3	4	5	6	7	8	average
frame based recognition rate	70.6	69	76.2	68.4	76.7	69	70.5	42.8	67.9
event recall	0.76	0.72	0.78	0.55	0.66	0.83	0.71	0.55	0.70
event precision	0.63	0.72	0.60	0.52	0.59	0.55	0.62	0.57	0.60
event Fmeas	0.69	0.72	0.68	0.53	0.62	0.66	0.67	0.56	0.64

Table 50: performance measures for each of the left persons of the test set (GMM algorithm, GT data).

test set	1	2	3	4	5	6	7	8	average
frame based recognition rate	33.2	28.8	44.9	57.4	58.5	40.3	45.5	62.8	46.4
event recall	0.46	0.30	0.78	0.62	0.74	0.69	0.66	0.73	0.62
event precision	0.65	0.60	0.42	0.54	0.49	0.37	0.47	0.45	0.50
event Fmeas	0.54	0.40	0.55	0.58	0.59	0.48	0.55	0.55	0.53

Table 51: performance measures for each of the right persons of the test set (GMM algorithm, GT data).

target FOA configurations. Secondly, less importantly, we do not currently use any individual adaptation modeling at recognition time (i.e. the model learned from other meetings is directly applied on the test meeting). This should possibly improve the results.

By comparing the results depending on the position in the meeting room, we get a striking idea of the impact of the FOA configurations on the results: while good results are achieved for the left person, we get a high degradation (20% decrease in frame based recognition, around 10% in event recall and precision), due for a large part to the ambiguity between two foci for the right person: the slide screen, and the left person. This ambiguity can be easily identified by looking at the the confusion matrices, tables 52 and 53, and is illustrated in the two images of Figure 30. Such ambiguous situations are unvoidable in practice, and can only be resolved using contextual information.

Table 54 provides the recognition results for the left location when the HMM approach is employed. We can observe a degradation of the results due to oversmoothing. A similar decrease can be observed for people sitting in the right position. There are currently investigating two possible explanations for this degradation: first, the HMM, through the transition matrix, introduce some prior on each class, while in the GMM case, we assumed a uniform prior for all classes. Secondly, the pdf of the unfocus class is modeled as a gaussian in the HMM case, thereas it is modeled as a uniform pdf in the GMM approach.

With estimated head pointing vector:

Table 55 presents the results we obtain when using the estimated head-pose output. As can be seen, despite the noisy measurements, the loss of performance is not dramatic (less than changing position in the meeting room): around 8% for the frame recognition rate, and 7% for the event recognition rate. Indeed, the main drop comes from the precision, which means that, alltogether, the degradation comes essentially in the form of more shorter erroneous FOA detection. Overall, we still have the finding that when a person is correctly tracked, the degradation is smaller[9], whereas when the tracker is in difficulty (example of person/set 1), the results can become random. As a way of understanding the large recognition variability among individuals, we plotted in Figure 29 the focus of individual people. As can be seen in this plot, there is a quite large overlap between targets, depending on participants. Similarly, by comparing ground truth data with estimated ones, we observe the larger variance along the tilt angle that was identified in Task 1.

9.3.3 Task 3: perception of head orientation in a Virtual Environement (VE)

Approach

Stimuli In the VMR, an avatar was positioned at the left side of the table, either on the chair in front

[9] Note that it is still affected by the other results, since they are used to train the recognition model.

foa	right person	organizer 1	organizer 2	table	slide screen	unfocus	deletion
right person	70.5	7.9	11.1	0.6	0.2	2.5	7.2
organizer 1	0.2	82.8	1.6	0.2	0.2	0.9	14.2
organizer 2	0	1.2	73.9	0.71	0.8	2.0	21.3
table	1.2	12.8	25.6	25.1	1.3	4.4	29.6
slide screen	0	0	0.8	1.3	95.2	0.5	2.1
unfocus	2.1	11.2	25.2	1.8	1.8	30.9	27.1

Table 52: event based confusion matrix (left person, GT data).

foa	left person	organizer 1	organizer 2	table	slide screen	unfocus	deletion
left person	50.9	1.5	0.5	1.2	23.4	6.1	16.3
organizer 1	2.5	67.3	4.5	0	4.6	4.6	16.5
organizer 2	0.6	1.6	84	0	1.3	0.3	12.1
table	4.3	16.9	9.8	26.4	7.3	8.1	27.1
slide screen	0.1	1.1	3.1	0.4	81.3	0.6	13.2
unfocus	0	11.2	15.7	2.5	4.3	48.5	17.8

Table 53: event based confusion matrix (right person, GT data).

(*C1*) or at the back (*C4*) (Figure 31). A number of balls was placed at eye height for the avatar and at a distance corresponding to 1.5 meters away from the avatar. To ensure good depth estimation, each ball was placed on a stick that intersected with the table. For enhanced discrimination, the balls were numbered and were alternately red and green. For *C4*, the balls were placed in the range {-30°, 60°} from the avatar, where 0° corresponds to looking straight ahead. For *C1*, this range was mirrored. We used three values for the angular distance between the balls: 15°, 22.5° and 30°. For the given angular range of 90° this amounts to 7, 5 and 4 balls respectively. The eyes of the avatar were fixed and pointed straight ahead. The viewpoint was placed at a height of 3.0 meters and 3.5 meters away from the center of the table, at an angle of 30° downwards. The avatar was placed 0.85 meter left of the center of the table and 0.5 meter to the front or back for the *C1* and *C4* condition respectively.

Observers were seated in front of a 19” TFT screen that was placed on a desk in an office environment. On this screen the VMR was displayed together with a button panel. The head of the avatar measured approximately 2.5 by 3 centimeters on the screen.

Procedure Each observer was asked to complete a session, each consisting of 6 session parts. Each session part was assigned a different combination of conditions for avatar position (*C1* or *C4*) and angular ball distance (15°, 22.5°and 30°). The sessions started either with the three parts of *C1*, or with three parts of *C4*. The order of angular distance between balls in the first three parts and the last three parts was identical. The total number of session types was 12, 6 starting with *C1* and 6 starting with *C4*. Within a session part, for each position a ball was presented exactly once and the order within each part was randomized.

The observers were asked to predict at which ball the avatar was looking. The selection was made by pressing a button with the corresponding number. Then the experiment proceeded with the next sample. A total of 32 samples were judged by each observer. There was no time limit and no breaks were needed as the experiment never lasted more than 3 minutes. Observers could view their progress in the experiment but did not receive any feedback on their judgement.

Participants A total of 36 persons (2 women and 34 men) participated in the experiment. The participants were students and employees of our department in the age range between 22 and 59. None of the participants was a trained observer.

Results

test set	1	2	3	4	5	6	7	8	average
frame based recognition rate	54.9	56.3	42.6	47.3	65.3	61.1	69.3	61.2	57.3
event recall	0.63	0.57	0.60	0.53	0.60	0.64	0.59	0.6	0.59
event precision	0.56	0.62	0.55	0.45	0.55	0.55	0.60	0.53	0.55
event Fmeas	0.59	0.59	0.57	0.48	0.58	0.59	0.59	0.56	0.57

Table 54: performance measures for each of the left persons of the test set (HMM algorithm, GT data).

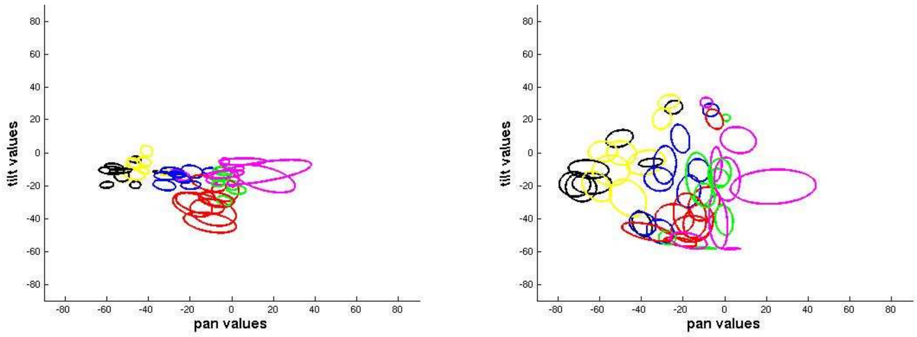


Figure 29: Focus of right persons. Each ellipses displays the focus of one persone for one target. Colors: black: left person. Dark blue: organizer 1. Green: organizer 2. Red: table. Yellow: slide screen. Soft blue: unfocused. We plotted the ellipses that are at a Mahalanobis distance of 0.5 of the mean. Left: using ground truth data. Right: using estimated data.

Each of the 12 session types was completed 3 times, which resulted in a total of 1152 judged samples. The performance scores for each of the conditions are summarized in Table 56.

As a first observation, we found no significant difference for the location of the avatar. Next, we compared the results of the first half of the session with the results of the second half for each observer, to ensure no learning effects occurred. A paired t-test revealed no significant difference between the two, suggesting no learning effect.

The results further indicate that decreasing the angular distance between the balls increases the judgement error. In Experiment II, observers must be able to correctly assess who is being looked at. The minimum azimuth angle between two persons is 30°, the angle between two persons at one side of the table as seen from a person at the other side. Our results indicate that discrimination in this situation is possible with an accuracy of 97.57%, which is sufficiently accurate.

9.3.4 Task 4: identifying speaker amongst meeting participants

Given the differences in head orientation behavior, we expect the observers to have some clue about who is the speaker when being shown the set of azimuth angles of participants' head orientation on avatars. [115] identify three looking regimes: *convergence* (there is one person attracting the others' gaze more than any of the others), *dyad-link* (the situation where two people look at each other) and *divergence* (gaze patterns that do not match the other two). We expect the best speaker identification performance for the convergence regime.

Approach:

Stimuli We used the VMR with the same viewpoint as in Task 3. We removed the balls and placed an identical avatar on each chair. Similar task 3, only the azimuth head angles have been varied. The setting now corresponds to the setting in the recorded meetings, with the distances between all participants and the whiteboard properly scaled.

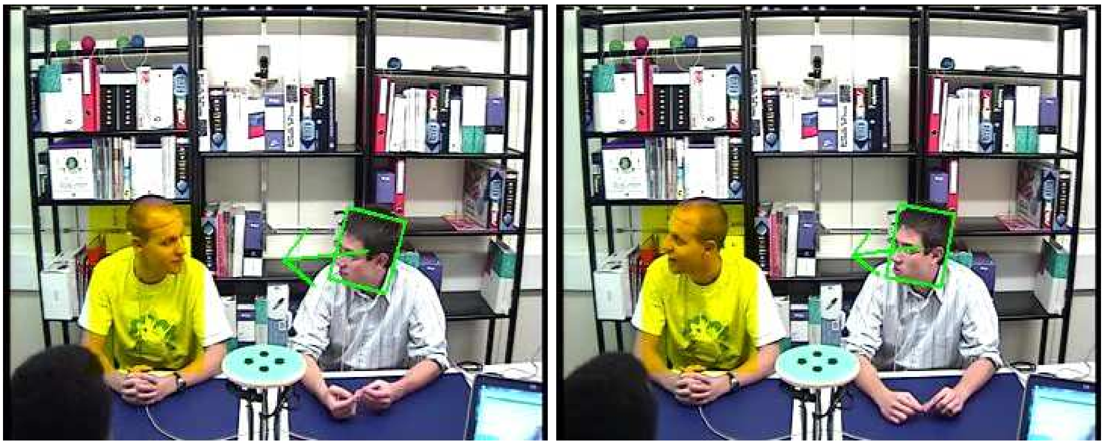


Figure 30: Ambiguity in focus: despite the high visual similarity of the head pose of the right person, the two focus are different (left image: left_person: right image: slide_screen). Resolving such cases can only be done by using context (speaking status, other's people gaze, slide activity etc).

test set	1	2	3	4	5	6	7	8	average
frame based recognition rate	11.9	48.3	42.7	40.5	51.7	30.7	45.2	28.4	37.4
event recall	0.26	0.30	0.67	0.51	0.70	0.64	0.59	0.60	0.54
event precision	0.08	0.36	0.30	0.40	0.31	0.166	0.35	0.19	0.27
event Fmeas	0.13	0.33	0.41	0.45	0.43	0.26	0.44	0.29	0.34

Table 55: performance measures for each of the right persons of the test set (GMM algorithm, estimated data).

We used two stimuli types: stills and animations. In the still condition, we provided the observers with a static scene. In this scene, the heads of the meeting participants were oriented in the azimuth plane in accordance with a specific time stamp in a specific meeting. For each sample, randomly chosen from a meeting, there was exactly one speaker. The observers had to identify the person who they thought was the speaker.

Animations of complete turns provide more context and display the dynamics of head orientations during a turn, with typical differences in speaker and listener behavior. In the animation condition, we displayed the head orientations of the meeting participants during an entire meeting turn which was derived from the speaker annotation of the data. The speaker turns, randomly chosen from a meeting, varied in length between half a second and 25 seconds. The animation was played with the same speed as the recorded data. Again, observers had to identify the speaker.

Procedure Each observer completed 4 session parts of 20 samples each. The samples of the first two parts were taken from Meeting 1, the third part from Meeting 2 and the last part from Meeting 3. There

Condition	15°	22.5°	30°	Average
C1	75.00%	85.56%	97.92%	84.03%
C4	74.21%	87.78%	97.22%	84.20%
Average	74.60%	86.67%	97.57%	84.11%

Table 56: Performance scores for ball identification in all conditions

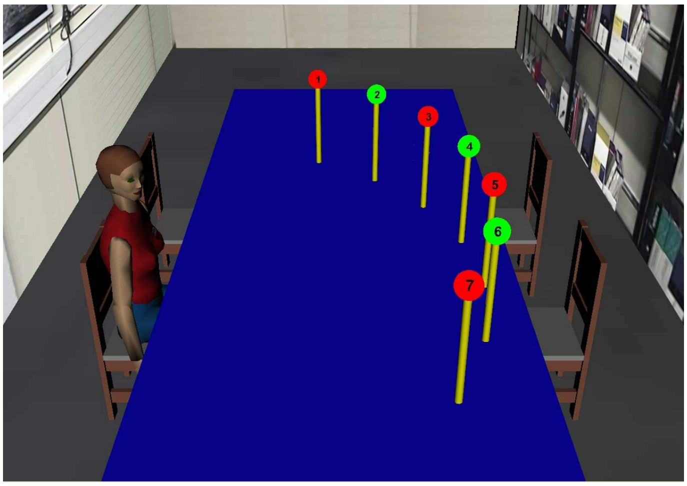


Figure 31: The VMR in condition C1, with an angular ball distance of 15°

was no time constraint imposed. In the animation condition, observers could replay the animation as a whole.

In accordance with task 3, the observers were asked to press the button with the number that corresponded with the number on the table before the speaker (see Figure 25). After pressing the button, the experiment advanced to the next sample. A forced-choice methodology was abandoned by introducing a 'no idea' button to prevent participants from conveying indifference to the task [123].

The mean durations of the experiment were approximately 9 and 21 minutes for the still and animation conditions respectively.

Participants A total of 40 persons (6 women, 34 men) took part in the experiment. The participants were students and employees of our department in age between 21 and 48. Observers were presented with either the still condition or the animation condition. Some of the participants also took part in task3 experiment.

Results and discussion:

Both conditions were completed 20 times and consisted of 80 samples, resulting in 1600 judged samples per condition. The results are shown in Table 57. Samples where observers identified no speaker but instead used the 'no idea' button have not been taken into account. This button was used for 148 (9.25%) and 40 samples (2.5%) in the still and animation condition respectively.

Condition	Part 1	Part 2	Part 3	Part 4	Average
Still	44.13%	44.69%	37.88%	38.32%	41.25%
Animation	45.52%	42.52%	35.62%	49.37%	43.27%

Table 57: Speaker identification performance in still and animation condition

First we checked for learning effects by comparing the scores of session part 1 and 2, both containing samples taken from meeting 1. A paired samples t-test showed no significant improvement of part 2 over part 1, in neither still nor animation condition.

The baseline for performance is 25%, the expected outcome when no *a priori* probabilities are known. It is clear that the results are better than random. However, the overall percentage of correct guesses (slightly over 40%) is rather low.

If we look at the differences between the two conditions, we notice only a small difference in performance. We found no significant difference between the still and animation condition. This suggests, somewhat surprisingly, that the extra information present in the animation condition does not improve identification results.

To test our hypothesis that observers' performance would be the best in the convergence regimes, we used all the samples from the still condition and calculated the scores for convergence regimes *Convergence-n*. These are the situations where there is one person attracting the others' gaze more than any of the others. The number n indicates how many persons look at this person. We looked at the *Convergence-2* and *Convergence-3* regimes, *Convergence-1* occurred only 9 times. The results are shown in Table 58.

Condition	Occurrence	Performance score	Most looked at is guessed	Most looked at is speaker
Convergence-2	865	36.88%	51.10%	42.89%
Convergence-3	317	55.84%	75.39%	70.03%

Table 58: Performance for Convergence-2 and Convergence-3 regime

The table shows that Convergence-3 regimes scored much better (55.84%) than the Convergence-2 regimes (36.88%). To find out if observers indeed used the differences in head orientation behavior as a means to predict the speaker we calculated how often the person where most heads were oriented to actually was identified as the speaker. Table 58 shows that observers identified the person that was looked at most in 51.10% and 75.39% for the Convergence-2 and Convergence-3 regime respectively. This reveals that observers indeed seem to think, or at least applied the systematic, that speakers are generally being looked at more than a listener. These results confirm our expectations that humans apply knowledge about systematic differences in head orientation behavior between speakers and listeners.

9.4 Conclusion and summary

In this section, we have presented the work that is pursued in AMI on the FOA analysis. The main results are:

- head-pose estimation: we presented a methodology for jointly tracking the head and estimating its head pose. We obtained an average error of around 10 degrees in pan angle, and 18 degrees in tilt angle. We showed that there was an important variation of results among individuals, depending on their resemblence with people in the appearence training set.

- FOA recognition: we have shown that, for the IDIAP SMR configuration, the recognition of the FOA purely from the individual head-pose pointing vector (pan and tilt) achieves 46% and 68%, depending on the person position in the SMR, clearly demonstrating the impact of position on the recognition. In addition, we have shown that the use of estimated head-pose rather than ground truth readings were degrading the results not so strongly (respectively 8% and 4% in frame resp. event based recognition rate). We have also shown that there was a large variation amongst individuals, which directly calls for adaptation approaches like Maximum A Posteriori techniques for FOA recognition.

- Speaker prediction from head-pose patterns: in a thorough study on the role of FOA in meeting converstations, we showed through the use of a Virtual Environement display that people are indeed using the gaze/head pose of participants to assess who is speaking. This results demonstrate that humans apply knowledge about systematic differences in head orientation behavior between speakers and listeners.

In the 3rd year of AMI, we will continue the work on FOA recognition and analysis. We aim at proposing better computational model to represent multimodal, non-verbal human interaction involving FOA in meetings, and explore the impact of contextual information (other cues, other people pose, other scene events) on our understanding of people visual focus of attention.

10 Conclusion

WP4 is concerned with the automatic recognition from audio, video, and combined audio-video streams, with an emphasis on developing models and algorithms to combine modalities. In this report we described the implementation and evaluations of ported and developed algorithms on common AMI data. Seven main tasks have been identified for WP4:

Baseline speech recognition system: The automatic speech recognition subgroup is concerned with the development of a speech recognition system for the use on AMI data. The developed system has been evaluated in international evaluations of ASR systems for meeting transcription, conducted by the U.S. National Institute for Standard and Technology (NIST). AMI has successfully competed in the NIST RT05s STT evaluations yielding very competitive results on both conference meeting and lecture room transcription.

Event spotting: Acoustic event (mainly keyword) spotting (KWS) in meetings has the goal to find all occurrences of entered word in a meeting and sort them according to confidences. This allows for Google-like browsing of meetings using acoustics. It also has the goal to verify if a word really occurred in a particular meeting, which is linked to WP5 summarisation work. The AMI KWS systems were evaluated on data from the ICSI meeting database. All systems were evaluated using standard Figure-of-Merit (FOM) measures defined by NIST and showed very promising results. Advantages and differences of the different AMI systems have been shown and evaluated in detail.

Person identification, segmentation, and clustering: AMI developed robust-to-localisation generative models for face recognition. The algorithms have been evaluated yielding very good results on a face verification task using the well-known BANCA benchmark database. For the speaker segmentation and clustering the TNO broadcast news speaker segmentation/clustering system has been expanded to operate on meeting data. The AMI speaker diarisation system performed satisfactorily in a NIST evaluation. For the speech activity detection AMI obtained very competitive results in the NIST RT05s evaluation (the lowest error rate reported).

Emotion recognition: AMI performed two major studies about emotions in meetings. These studies investigated important questions about emotions in meetings: What types of emotions actually occur in the recordings? How can these emotions be effectively annotated and how reliably are the annotations. Furthermore AMI developed algorithms for the detection of emotions in meeting scenarios from the facial expression, gestures, as well as head- and body pose. The algorithms have been preliminary evaluated with very promising results. Final evaluations are currently performed.

Localisation and Tracking: AMI developed a face detection system which detects, in real-time, multiple upright frontal faces. The performances has been evaluated on a benchmark database. Trackers based an a variety of different technologies have been investigated and AMI results have been presented. These methods comprise trackers for multiple-person scenarios, techniques exploiting multimodal features (AV-Tracker) as well as view independent tracking frameworks like the Probabilistic Active Shape Tracking.

Gestures and actions: Algorithms to automatic recognise relevant actions and gestures have been developed. A special focus has been set on gestures which occur frequently within the AMI meetings. All gesture recognition methods gave reasonable performance on manually segmented video streams. Yet segmentation of gestures in AMI is still a very challenging task: the evaluated segmentation methods do not give good segmentation performance. Detecting gestures such as shaking and nodding and negative signals is still a very challenging problem that will require methods capable of detecting very subtle head movements. However AMI made a substantial process for the automatic segmentation of gestures in meeting scenarios.

Focus of attention: AMI has achieved good results for the automatic recognition of the focus of attention in meetings on both a standard head pose database and a sub-set of the AMI meetings. The results are good enough to perform different tasks, and can measure with the state-of-the-art.

In this deliverable D4.2 we described several implemented and ported algorithms and methods for each of the seven tasks. The output of the described procedures can then be used as input for WP5. Common evaluation schemes on AMI data have been established. This allows to compare different approaches to a problem on the common AMI data set. Furthermore the common evaluation schemes guarantee common interfaces among the involved partners, and a common, stringent output of the different recognisers. Therefore WP5 has defined inputs from WP4 - independent of the actual used algorithm. Finally the common interfaces allow the fusion of several algorithms to a larger system.

Several algorithms have been evaluated in international evaluations, conducted e.g. from NIST. Developed AMI systems for a wide range of audio-visual problems performed very successfully in these evaluations. This shows that AMI has developed a range of state-of-the-art algorithms for audio, visual, and audio-visual problems.

Given these very good results, AMI is now going to further improve the algorithms and methods, as well as development towards real-time processing of the huge amount of audio-visual data.

References

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *Proc. 8th European Conference on Computer Vision (ECCV), Prague, Czech Republic*, pages 469–481, 2004.

[2] Marc Al-Hames, Alfred Dieleman, Daniel Gatica-Perze, Stephan Reiter, Steve Renals, Gerhard Rigoll, and Dong Zhang. Multimodal integration for meeting group action segmentation and recognition. In *Proceedings MLMI'05, Edingburgh, Scotland*, 2005.

[3] Stuart M. Anstis, John W. Mayhew, and Tania Morley. The perception of where a face or television 'portrait' is looking. *American Journal of Psychology*, 82(4):474–489, 1969.

[4] Michael Argyle and Mark Cook. *Gaze and mutual gaze.* Cambridge University Press, London, United Kingdom, 1976.

[5] Michael Argyle and Janet Dean. Eye-contact, distance and affiliation. *Sociometry*, 28(3):289–304, 1965.

[6] Michael Argyle, Roger Ingham, Florisse Alkema, and Margaret McCallin. The different functions of gaze. *Semiotica*, 7:19–32, 1973.

[7] Michael Argyle, Mansur Lalljee, and Mark Cook. The effects of visibility on interaction in a dyad. *Human Relations*, 21:3–17, 1968.

[8] Sileye O. Ba and Jean Marc Odobez. Evaluation of head pose tracking algorithm in indoor environments. In *International Conference on Multimedia & Expo, ICME 2005, Amsterdam*, 2005.

[9] Sileye O. Ba and Jean Marc Odobez. A rao-blackwellized mixed state particle filter for head pose tracking. In *ACM-ICMI Workshop on Multi-modal Multi-party Meeting Processing (MMMP), Trento Italy*, pages 9–16, 2005.

[10] Jeremy N. Bailenson, Andrew C. Beall, and Jim J. Blascovich. Gaze and task performance in shared virtual environments. *The journal of visualisation and computer animation*, 13:313–320, 2002.

[11] Jeremy N. Bailenson, Andrew C. Beall, and Matthew Turk. Transformed social interaction: Decoupling representation from behavior and form in collaborative virtual environments. *Presence*, 13(4):428–441, 2004.

[12] S. Baker, I. Matthews, J. Xiao, R. Gross, T. Kanade, and T. Ishikawa. Real-time non-rigid driver head tracking for driver mental state estimation. In *Proceedings of the 11th World Congress on Intelligent Transportation Systems*, 2004.

[13] I. Bakx, K. van Turnhout, and J. Terken. Facial orientation during multi-party interaction with information kiosks. In *Proc. of INTERACT*, 2003.

[14] R. F. Bales and S. P. Cohen. *SYMLOG: A System for the Multiple Level Observation of Groups.* The Free Press, NY, 1979.

[15] A. Baumberg and D. Hogg. An efficient method for contour tracking using active shape models. In *Proceedings of the IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, 1994.

[16] A.W. Black, P. Taylor, and R. Caley. (2004). The Festival Speech Synthesis System, Version 1.95beta. CSTR, University of Edinburgh, Edinburgh.

[17] L. Brown and Y. Tian. A study of coarse head pose estimation. *IEEE Workshop on Motion and Video Computing*, Dec 2002.

[18] I. Bulyko, M. Ostendorf, and A. Stolcke. Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures. in Proc HLT'03.

[19] I. Bulyko, M. Ostendorf, and A. Stolcke. Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In *Proc HLT'03*, 2003.

[20] I. Bulyko, M. Ostendorf, and A. Stolcke. Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In *Proc HLT*, 2003.

[21] S. Burger, V. MacLaren, and H. Yu. (2002). The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style. In Proc. ICSLP'2002.

[22] L. Burget. (2004), Combination of Speech Features Using Smoothed Heteroscedastic Linear Discriminant Analysis. in Proc. ICSLP'04, Jeju island, KR, 2004, p. 4.

[23] L. Burget, P. Matějka, and J. Černocký. "Discriminative training techniques for acoustic language identification", accepted to ICASSP, Toulouse, France, 2006.

[24] F. Cardinaux, C. Sanderson, and S. Bengio. Face verification using adapted generative models. In *The 6th International Conference on Automatic Face and Gesture Recognition*, Seoul, Korea, 2004. IEEE.

[25] F. Cardinaux, C. Sanderson, and S. Marcel. Comparison of MLP and GMM classifiers for face verification on XM2VTS. In *Proceedings of the International Conference on Audio- and Video-Based Biometric Person Authentication*. Springer-Verlag, 2003.

[26] J. Carletta, S. Ashby, S. Bourban, M. Guillemot, M. Kronenthal, G. Lathoud, M. Lincoln, I. McCowan, T. Hain, W. Kraaij, W. Post, J. Kadlec, P. Wellner, M. Flynn, and D. Reidsma. The AMI meeting corpus. In *Proc MLMI*, 2005.

[27] J. Carletta, S. Ashby, S. Bourban, M. Guillemot M. Kronenthal, G. Lathoud, M. Lincoln, I. McCowan, T. Hain, W. Kraaij, W. Post, J. Kadlec, P. Wellner, M. Flynn, and D. Reidsma. (2005). The AMI Meeting Corpus. Submitted to MLMI'05.

[28] Jean Carletta, Simon Garrod, and Heidi Fraser-Krauss. Placement of authority and communication pattern in workplace groups: the consequences for innovation. *Small Group Research*, 29(5):531–559, 1998.

[29] Jean C. Carletta, Anne H. Anderson, and Rachel McEwan. The effects of multimedia communication technology on non-collocated teams: a case study. *Ergonomics*, 43:1237–1251, 2000.

[30] Jean C Carletta, Anne H. Anderson, and Garrod S. Seeing eye to eye: an account of grounding and understanding in work groups. *Bulletin of the Japanese cognitive sciences*, 9(1):1–20, 2002.

[31] Justine Cassell and Kristinn R. Thórisson. The power of a nod and a glance: envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(4):519–538, 1999.

[32] Marvin G. Cline. The perception of where a person is looking. *American Journal of Psychology*, 80(1):41–50, 1967.

[33] R. Alex Colburn, Michael F. Cohen, and Steven M. Drucker. The role of eye gaze in avatar mediated conversational interfaces. Technical Report MSR-TR-2000-81, Microsoft Research, 2000.

[34] T. Cootes and C. Taylor. Statistical models of appearance for computer vision. Technical report, 2004.

[35] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):35–59, 1995.

[36] H. Cox, R. Zeskind, and I. Kooij. (1986). Practical supergain. IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-34(3):393–397.

[37] H. Cox, R. Zeskind, and M. Owen. (1987). Robust adaptive beamforming. IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-35(10):1365–1376.

[38] Head Pose Database. Prima-pointing head pose database. www-prima.inrialpes.fr/Pointing04/data-face.html.

[39] A. Dielmann and S. Renals. Multistream dynamic Bayesian network for meeting segmentation. *Lecture Notes in Computer Science*, 3361:76–86, 2005.

[40] Starkey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292, 1972.

[41] Starkey Duncan and George Niederehe. On signalling that it's your turn to speak. *Journal of Experimental Social Psychology*, 10:234–247, 1974.

[42] Ralph V. Exline. Explorations in the process of person perception: visual interaction in relation to competition, sex, and need for affiliation. *Journal of personality*, 31:1–20, 1963.

[43] S. Fitt. (2000). Documentation and user guide to UNISYN lexicon and post-lexical rules, Tech. Rep., Centre for Speech Technology Research, Edinburgh.

[44] B. Fröba and A. Ernst. Face detection with the modified census transform. In *IEEE Conference on Automatic Face and Gesture Recognition (AFGR)*, 2004.

[45] M.J.F. Gales and P.C. Woodland. (1996). Mean and Variance Adaptation within the MLLR Framework. *Computer Speech & Language*, Vol. 10, pp. 249–264.

[46] J.S. Garafolo, C.D. Laprun, M. Michel, V.M. Stanford, and E. Tabassi. (2004). In Proc. 4th Intl. Conf. on Language Resources and Evaluation (LREC'04).

[47] Maia Garau, Mel Slater, Simon Bee, and Martina A. Sasse. The impact of eye gaze on communication using humanoid avatars. In *Proceedings of the conference on human factors in computing systems (CHI'01)*, pages 309–316, Seattle, WA, 2001.

[48] Christophe Garcia and Manolis Delakis. Convolutional face finder: A neural architecture for fast and robust face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11), 2004.

[49] J.L. Gauvain and C. Lee. (1994). MAP estimation for multivariate Gaussian mixture observation of Markov Chains, IEEE Tr. Speech& Audio Processing, 2, pp. 291-298.

[50] D.M. Gavrilla. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

[51] James J. Gibson and Anne D. Pick. Perception of another person's looking behavior. *American Journal of Psychology*, 76(3):386–394, 1963.

[52] E. Goffman. *Forms of Talk*. University of Pennsylvania Press, Philadelphia, 1981.

[53] Erving Goffman. On face-work: An analysis of ritual elements in social interaction. *Psychiatry*, 18:213–231, 1955.

[54] Erving Goffman. *Behaviour in Public Places, Notes on the Social Organization of Gatherings*. The Free Press, Glencoe, IL, 1963.

[55] C. Goodwin. *Conversational Organization: Interaction Between Speakers and Hearers*. NY:Academic Press, 1981.

[56] Charles Goodwin. *Conversational organization: Interaction between speakers and hearers*. Academic Press, New York, 1981.

[57] D. Gopher. *The Blackwell dictionary of Cognitive Psychology, chapter Attention*. Basil Blackwell Inc., 1990.

[58] B. J. Grosz and C. L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, Vol. 12:175–204, 1986.

[59] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals. (2005). The 2005 AMI System for the Transcription of Speech in Meetings. In Proc. of the NIST RT05s workshop, Edinburgh.

[60] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals. The 2005 AMI system for the transcription of speech in meetings. In *Proc. Rich Transcription 2005 Spring Meeting Recognition Evaluation*, 2005.

[61] T. Hain, J. Dines, G. Garau, M. Karafiat, D. Moore, V. Wan, R. Ordelman, and S. Renals. (2005). Transcription of Conference Room Meetings: an Investigation. in Proc. Interspeeh 2005.

[62] T. Hain, P. Woodland, T. Niesler, and E. Whittaker. (1999). The 1998 HTK system for transcription of conversational telephone speech. Proc. IEEE ICASSP, 1999.

[63] Thomas Hain. Implicit modelling of pronunciation variation in automatic speech recognition. *SPCOMM*, 46(2):171–188, 2005.

[64] Thomas Hain, Lukas Burget, John Dines, Giulia Garau, Martin Karafiat, Mike Lincoln, Iain McCowan, Darren Moore, Vincent Wan, Rolland Ordelman, and Steve Renals. “The 2005 AMI System for the Transcription of Speech in Meetings”, in *Proc. NIST Rich Transcription 2005 Spring Meeting Recognition Evaluation*, Edinburgh, July 2005.

[65] N. Hassink and M. Schopman. Gesture recognition in a meeting environment. Master’s thesis, University of Twente, Department of Computer Science, HMI group, 2006.

[66] H. Hermansky. (1990). Perceptual Linear Predictive (PLP) analysis of speech. Acoustical Society of America, 87(4):1738–1752.

[67] Dirk Heylen. Challenges ahead. Head movements and other social acts in conversation. In *Proceedings of the Social Presence Cues for Virtual Humanoids workshop at the conference of Artificial Intelligence and the Simulation of Behaviour (AISB)*, Hatfield, United Kingdom, 2005.

[68] Dirk Heylen, Ivo van Es, Betsy van Dijk, and Anton Nijholt. Experimenting with the gaze of a conversational agent. In *Proceedings of the Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*, pages 93–99, Copenhagen, Denmark, 2002.

[69] Jon Hindmarsh, Mike Fraser, Christian Heath, Steve Benford, and Chris Greenhalgh. Fragmented interaction: establishing mutual orientation in virtual environments. In *Proceedings of the conference on Computer supported cooperative work (CSCW'98)*, pages 217–226, Seattle, WA, 1998. ACM Press.

[70] B.K.P Horn and B.G. Schunk. Determining optical flow. AI Memo 572, MIT, 1980.

[71] Eric Horvitz, Carl Kadie, Tim Paek, and David Hovel. Models of attention in computing and communication. from principles to applications. 2004.

[72] X. Huang, S.Z. Li, and Y. Wang. Shape localization based on statistical method using extended local binary pattern. In *Proc. Third International Conference on Image and Graphics (ICIG), Hong Kong, China*, pages 184–187, 2004.

[73] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. In *International Journal of Computer Vision*, volume 29(1), pages 5–28, 1998.

[74] M. Isard and A. Blake. ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework. *Lecture Notes in Computer Science*, 1406:893–908, 1998.

[75] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. (2003). The ICSI Meeting Corpus. ICASSP'03, Hong Kong.

[76] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The ICSI meeting corpus. In *Proc. ICASSP*, 2003.

[77] Oliver Jesorsky, Klaus J. Kirchberg, and Robert W. Frischholz. Robust face detection using the hausdorff distance. In *Proceedings of the Third International Conference on Audio- and Video-Based Biometric Person Authentication*, pages 90–95, 2001.

[78] H. Jin, Q. Liu, H. Lu, and X. Tong. Face detection using improved LBP under bayesian framework. In *Proc. Third International Conference on Image and Graphics (ICIG), Hong Kong, China*, pages 306–309, 2004.

[79] N. Jovanovic and R. op den Akker. Towards automatic addressee identification in multi-party dialogues. In *5th SIGdial Workshop on Discourse and Dialogue*, pages 89–92, 2004.

[80] N. Jovanovic, R. op den Akker, and N. Nijholt. A corpus for studying addressing behavior in face-to-face meetings. In *6th SIGdial Workshop on Discourse and Dialogue. Lisbon, Portugal*, 2005.

[81] M. Katzenmaier, R. Stiefelhagen, and T. Schultz. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proc. of ICMI*, 2004.

[82] Adam Kendon. Some functions of gaze direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[83] Youngjun Kim, Randall W. Hill, and David R. Traum. Controlling the focus of perceptual attention in embodied conversational agents. In *AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 1097–1098. ACM Press, New York, NY, USA, 2005.

[84] Chris L. Kleinke. Gaze and eye contact: a research review. *Psychological Bulletin*, 100(1):78–100, 1986.

[85] B. Klimt and Y. Yang. (2004). Introducing the Enron Corpus, Second Conference on Email and Anti-Spam, CEAS 2004.

[86] C. H. Knapp and G. C. Carter. (1976). The generalized correlation method for estimation of time delay/ IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-24:320–327, August 1976.

[87] Kunibert Krüger and Bärbel Hückstedt. Die beurteilung von blickrichtungen. *Zeitschrift fur experimentelle und angewandte psychologie*, 16:452–472, 1969.

[88] N. Kumar. (1997), Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition.PhD thesis, John Hopkins University, Baltimore.

[89] Stephen R.H. Langton. The mutual influence of gaze and head orientation in the analysis of social attention direction. *Quarterly Journal of Experimental Psychology*, 53A(3):825–845, 2000.

[90] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez. Av16.3: an audio-visual corpus for speaker localization and tracking. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, 2004.

[91] C.J. Leggetter and P.C. Woodland. (1995).Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs. Computer, Speech and Language, , Vol. 9, pp. 171–186.

[92] Gene H. Lerner. Selecting next speaker: the context-sensitive operation of a context-free organization. *Language in Society*, 32:177–201, 1998.

[93] Jack M. Loomis, Jim J. Blascovich, and Andrew C. Beall. Immersive virtual environment technology as a basic research tool in psychology. *Behavior Research Methods, Instruments and Computers*, 31:557–564, 1999.

[94] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91 – 110, 2004.

[95] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[96] L. Mangu, E. Brill, and A. Stolcke. Finding consensus among words: Lattice-based word error minimization. In *Proc. Eurospeech'99*, pages 495–498, 1999.

[97] Kinya Maruyama and Mitsuo Endo. The effect of face orientation upon apparent direction of gaze. *Tohoku Psychologica Folia*, 42(1–4):126–138, 1983.

[98] P. Matějka, L. Burget, P. Schwarz, and J. Černocký. "Use of anti-models to further improve state-of-the-art PRLM Language Recognition System", accepted to ICASSP 2006, Toulouse, France, 2006.

[99] P. Matějka, P. Schwarz, J. Černocký, and P. Chytil. "Phonotactic Language Identification using High Quality Phoneme Recognition" in *Proc. Eurospeech 2005*, Lisabon, Portugal, Sept. 2005.

[100] Evelyn Z. McClave. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32:855–878, 2000.

[101] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard. Modeling human interaction in meetings. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP 03)*, 2003.

[102] I. McCowan, D. Gatica-Perez, S. Bengio, and G. Lathoud. Automatic analysis of multimodal group actions in meetings. Technical Report RR. 03-27, IDIAP, Martigny, 2003.

[103] Ian McCowan, D. Gatica-Perez, Bengio, D. Moore, and H. Bourlard. Towards computer understanding of human interactions. *Proceedings of EUSAI 2003, LNCS 2875, (E. Aarts et al. ed.)*, pages 235–251, 2003.

[104] Kieron Messer, Josef Kittler, Mohammad Sadeghi, Miroslav Hamouz, Alexey Kostyn, Sebastien Marcel, Samy Bengio, Fabien Cardinaux, Conrad Sanderson, Norman Poh, Yann Rodriguez, Jacek Czyz, and al. Face authentication test on the BANCA database. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, Cambridge, August 23-26 2004.

[105] D. Messerschmitt, D. Hedberg, A. Haoui C. Cole, and P. Winship. (1989). Digital voice echo canceller with a TMS32020. Appl. Rep. SPRA129, Texas Instruments.

[106] J. Moore, M. Kronenthal, and S. Ashby. Guidelines for AMI speech transcriptions. Technical report, IDIAP, Univ. of Edinburgh, February 2005.

[107] Yukiko I. Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. Towards a model of face-to-face grounding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'03)*, pages 553–561, Sapporo, Japan, 2003.

[108] Gerhard S. Nielsen. *Studies in self-confrontation*. Howard Allen, Cleveland, OH, 1962.

[109] NIST. Spring 2005 (RT05S) rich transcription meeting recognition evaluation plan. http://www.nist.gov/speech/tests/rt/rt2005/spring, 2005.

[110] David G. Novick, Brian Hansen, and Karen Ward. Coordinating turn-taking with gaze. In *Proceedings of the international conference on Spoken Language Processing (ICSLP'96)*, volume 3, pages 1888–1891, Philadelphia, PA, 1996.

[111] Brid OĆonaill and Steve Whittaker. Chapter 6, characterizing, predicting and measuring video-mediated communication: a conversational approach. In Kathleen E. Finn, Abigail J. Sellen, and Sylvia B. Wilbur, editors, *Video-Mediated Communication (Computers, Cognition, and Work)*, pages 107–131. Lea, 1997.

[112] Jean-Marc Odobez. Focus of attention coding guidelines. Technical Report 2, IDIAP-COM, January 2006.

[113] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29, 1996.

[114] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine intelligence*, 24:971–987, 2002.

[115] Kazuhiro Otsuka, Yoshinao Takemae, Junji Yamato, and Hiroshi Murase. A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In *Proceedings of International Conference on Multimodal Interface (ICMI'05)*, pages 191–198, Trento, Italy, October 2005.

[116] Gaurav Pandey. *Keyword spotting on continuous speech data using semantic categories*, AMI-Traineeship Project report, Brno University of Technology, July, 2005.

[117] Derrick Parkhurst, Klinton Law, and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42:107–123, 2002.

[118] David I. Perrett and Nathan J. Emery. Understanding the intention of others from visual signals. *Current psychology of cognition*, 13:683–694, 1994.

[119] T. Pfau and D.P. W. Ellis. (2001). Hidden markov model based speech activity detection for the ICSI meeting project. Eurospeech'01.

[120] Isabella Poggi, Catherine Pelachaud, and Fiorella de Rosis. Eye communication in a conversational 3D synthetic agent. *European Journal on Artificial Intelligence*, 13(3):169–181, 2000.

[121] R. Poppe, D. Heylen, A. Nijholt, and M. Poel. Towards real-time body pose estimation for presenters in meeting environments. In *Proceedings of the International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision'2005 (WSCG'2005)*, pages 41–44, 2005.

[122] Daniel Povey and Philip C. Woodland. Minimum phone error and i-smoothing for improved discriminative training. In *In Proc. ICASSP'02*, 2002.

[123] John J. Ray. Acquiescence and problems with forced-choice scales. *Journal of Social Psychology*, 130(3):397–399, 1990.

[124] Dennis Reidsma, Rieks op den Akker, Rutger Rienks, Ronald Poppe, Anton Nijholt, Dirk Heylen, and Job Zwiers. Virtual meeting rooms: from observation to simulation. In *Proceedings of the Third International Workshop on Social Intelligence Design (SID'05)*, Stanford, CA, 2005.

[125] Dennis Reidsma, Rutger Rienks, and Natasa Jovanovich. Meeting modeling in the context of multimodal communication. In *Proceedings of the workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI'04)*, pages 22–35, Martigny, France, 2004.

[126] S. Reiter. Annotation scheme for gestures and individual actions. Annotation guideline, AMI consortium, 2004.

[127] Stephan Reiter and Gerhard Rigoll. Multimodal meeting event recognition fusing three different types of recognition techiques. Joint AMI/PASCAL/IM2/M4 Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI), Martigny, June 2004.

[128] Stephan Reiter and Gerhard Rigoll. Multimodal meeting analysis by segmentation and classification of meeting events based on a higher level semantic approach. In *Proceedings of the 30th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, USA, March 2005.

[129] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 1979.

[130] Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[131] Spring 2004 (RT04S). Rich Transcription Meeting Recognition Evaluation Plan. NIST, US. Available at `http://www.nist.gov/speech`.

[132] M. Sadeghi, J. Kittler, A. Kostin, and K. Messer. A comparative study of automatic face verification algorithms on the banca database. In *4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*, pages 35–43, Guilford, UK, 2003.

[133] Noam Sagiv and Shlomo Bentin. Structural encoding of human and schematic faces: holistic and part-based processes. *Journal of Cognitive Neuroscience*, 13(7):937–951, 2001.

[134] Emanuel A. Schegloff. Sequencing in conversational openings. *American Anthropologist*, 70(6):1075–1095, 1968.

[135] S. Schreiber and D. Gatica-Perez. Evaluation scheme for tracking in ami. Technical report, 2006.

[136] T. Schultz, A. Waibel, M. Bett, F. Metze, Y. Pan, K. Ries, T. Schaaf, H. Soltau, M. Westphal, H. Yu, and K. Zechner. (2001). The ISL Meeting Room System. In Proc. of the Workshop on Hands-Free Speech Communication (HSC-2001), Kyoto.

[137] Petr Schwarz, Pavel Matějka, and Jan Černocký. "Hierarchical structures of neural networks for phoneme recognition", accepted to ICASSP 2006, Toulouse, 2006.

[138] Abigail J. Sellen. Speech patterns in video-mediated conversations. In *Proceedings of the conference on Human factors in computing systems (CHI'92)*, pages 49–59, Monterey, CA, 1992.

[139] K. Smith, S. Ba, J. Odobez, and D. Gatica-Perez. Evaluating multi-object tracking. volume Workshop on Empirical Evaluation Methods in Computer Vision (EEMCV), San Diego, CA, USA, June 2005.

[140] K. Smith, S. Ba, J.M. Odobez, and D. Gatica-Perez. Multi-person wander-visual-focus-of-attention tracking. Technical report, 2005.

[141] K. Smith, S. Ba, J.M. Odobez, and D. Gatica-Perez. Multi-person wander-visual-focus-of-attention tracking. Technical Report 80, IDIAP-RR, submitted to IEEE Conference on Computer Vision and Pattern Recognition (CVPR), November 2005.

[142] R. Stiefelhagen, J. Yang, and A Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, Vol.13, No. 4, 2002.

[143] R. Stiefelhagen and J. Zhu. Head orientation and gaze direction in meetings. *Conference on Human Factors in Computing Systems, Minneapolis, Minnesota, USA*, 2002.

[144] Rainer Stiefelhagen. Tracking focus of attention in meetings. In *Proceedings of the IEEE International Conference on Multimodal Interfaces (ICMI'02)*, pages 273–280, Pittsburgh, PA, 2002.

[145] Rainer Stiefelhagen and Jie Zhu. Head orientation and gaze direction in meetings. In *Extended abstracts on Human factors in computing systems (CHI'02)*, pages 858–859, Minneapolis, MN, 2002.

[146] A. Stolcke, C. Wooters, N. Mirghafori, T. Pirinen, I. Bulyko, D. Gelbart, M. Graciarena, S. Otterson, B. Peskin, and M. Ostendorf. (2004). Progress in Meeting Recognition: The ICSI-SRI-UW Spring 2004 Evaluation System. NIST RT04 Workshop.

[147] Lawrence A. Symons, Kang Lee, Caroline C. Cedrone, and Mayu Nishimura. What are you looking at? Acuity for triadic eye gaze. *Journal of general psychology*, 131(4):451–469, 2004.

[148] Igor Szöke, Petr Schwarz, Pavel Matějka, Lukáš Burget, Martin Karafiát, Michal Fapšo, and Jan Černocký. "Comparison of Keyword Spotting Approaches for Informal Continuous Speech", in *Proc. Eurospeech 2005*, Lisabon, Portugal, Sept. 2005.

[149] California. The SRI Language Modelling Toolkit (SRILM).
`http://www.speech.sri.com/projects/srilm`, SRI international.

[150] Carlo Tomasi and Takeo Kanade. Detection and tracking of point features. 1991.

[151] The Hidden Markov Model Toolkit. (HTK). `http://htk.eng.cam.ac.uk`, Cambridge University, UK.

[152] D. Traum and J. Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings AAMAS02*, pages 15–19, 2002.

[153] Ivo van Es, Dirk Heylen, Betsy van Dijk, and Anton Nijholt. Gaze behavior of talking faces makes a difference. In *Extended abstracts on Human factors in computing systems (CHI'02)*, pages 734–735, Minneapolis, MN, 2002.

[154] Jeroen van Rest and Jurgen den Hartog. An architecture for dedicated real-time tracker development and management. In *Joint AMI/Pascal workshop*, 2004.

[155] K. van Turnhout, J. Terken, I. Bakx, and B. Eggen. Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *Proc. of ICMI*, 2005.

[156] R. Vertegaal. *Look who's talking to whom. Mediating Joint Attention in Multiparty Communication and Collaboration.* PhD thesis, University of Twente, 1998.

[157] R. Vertegaal. Attentive user interfaces. *Communications of the ACM*, Vol. 46(3):33–36, 2003.

[158] Roel Vertegaal. *Look Who's Talking to Whom.* PhD thesis, University of Twente, Enschede, The Netherlands, 1998.

[159] Roel Vertegaal. The GAZE groupware system: mediating joint attention in multiparty communication and collaboration. In *Proceedings of the conference on Human factors in computing systems (CHI'99)*, pages 294–301, Pittsburgh, PA, 1999.

[160] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. Why conversational agents should catch the eye. In *Extended abstracts on Human factors in computing systems (CHI'00)*, pages 257–258, The Hague, The Netherlands, 2000.

[161] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *Proceedings of the conference on Human factors in computing systems (CHI'02)*, pages 301–308, Seattle, WA, 2002.

[162] Roel Vertegaal, Gerrit C. van der Veer, and Harro Vons. Effects of gaze on multiparty mediated communication. In *Proceedings of Graphics Interface 2000*, pages 95–102, Montreal, Canada, 2000.

[163] Ian Vine. Judgement of direction of gaze: an interpretation of discrepant results. *British Journal of Social and Clinical Psychology*, 10:320–331, 1971.

[164] Paul Viola and Michael J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[165] Mario L. Von Cranach and Johann H. Ellgring. Chapter 10, problems in the recognitions of gaze direction. In Mario L. Von Cranach and Ian Vine, editors, *Social communications and movement: studies of interaction and expression in man and chimpanzee*, pages 419–443. Academic Press, London, 1973.

[166] V. Wan and T. Hain. Strategies for language model web-data collection. In *Proc. ICASSP*, 2006.

[167] Rita M. Weisbrod. Looking behavior in a discussion group. Technical report, Cornell University, Ithaca, New York, 1965. Term Paper submitted for Psychology 546 under the direction of Professor Longabaugh.

[168] Hugh R. Wilson, Frances Wilkinson, Li-Ming Lin, and Maja Castillo. Perception of head orientation. *Vision Research*, 40(5):459–472, 2000.

[169] P.C. Woodland, M.J.F. Gales, D. Pye, and S.J. Young. (1997). Broadcast News Transcription using HTK. In *Proc. ICASSP'97*, pp. 719-722, Munich.

[170] S. Wrigley, G. Brown, V. Wan, and S. Renals (2005). Speech and crosstalk detection in multichannel audio. IEEE Trans. Speech& Audio Proc., 13(1):84–91.

[171] V. H. Yngve. On getting a word in edgewise. *Papers from the sixth regional meeting of the Chicago Linguistics Society, Chicago: Chicago Linguistics Society.*, 1970.

[172] Victor H. Yngve. On getting a word in edgewise. *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pages 567–577, 1970.

[173] R. Zabih and J. Woodfill. A non-parametric approach to visual correspondence. *IEEE Transactions on Pattern Analysis and Machine intelligence*, 1996.

[174] Dong Zhang, Daniel Gatica-Perez, Samy Begio, Iain McCowan, and Guillaume Lathoud. Modeling individual and group actions in meetings: a two-layer hmm framework. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Workshop on Event Mining in Video (CVPR-EVENT)*, Washington DC, July 2004.

[175] G. Zhang, X. Huang, S.Z. Li, Y. Wang, and X. Wu. Boosting local binary pattern (LBP)-based face recognition. In *Proc. Advances in Biometric Person Authentication: 5th Chinese Conference on Biometric Recognition, SINOBIOMETRICS 2004Guangzhou, China*, pages 179–186, 2004.