



# 4th Joint Workshop on Machine Learning and Multimodal Interaction

28-30 June 2007  
Hotel Continental  
Brno, Czech Republic

[www.mlmi07.org](http://www.mlmi07.org)



## Welcome to MLMI'07!

The fourth MLMI workshop is coming to Brno in the Czech Republic, following successful workshops in Martigny (2004), Edinburgh (2005) and Washington, DC (2006). MLMI brings together researchers from the different communities working on the common theme of advanced machine learning algorithms applied to multimodal human-human and human-computer interaction.

Revised versions of selected papers and posters presented at MLMI'07 will be published in Springer's Lecture Notes in Computer Science series, in common with the previous MLMI workshops, published as LNCS 3361, 3869, and 4299.

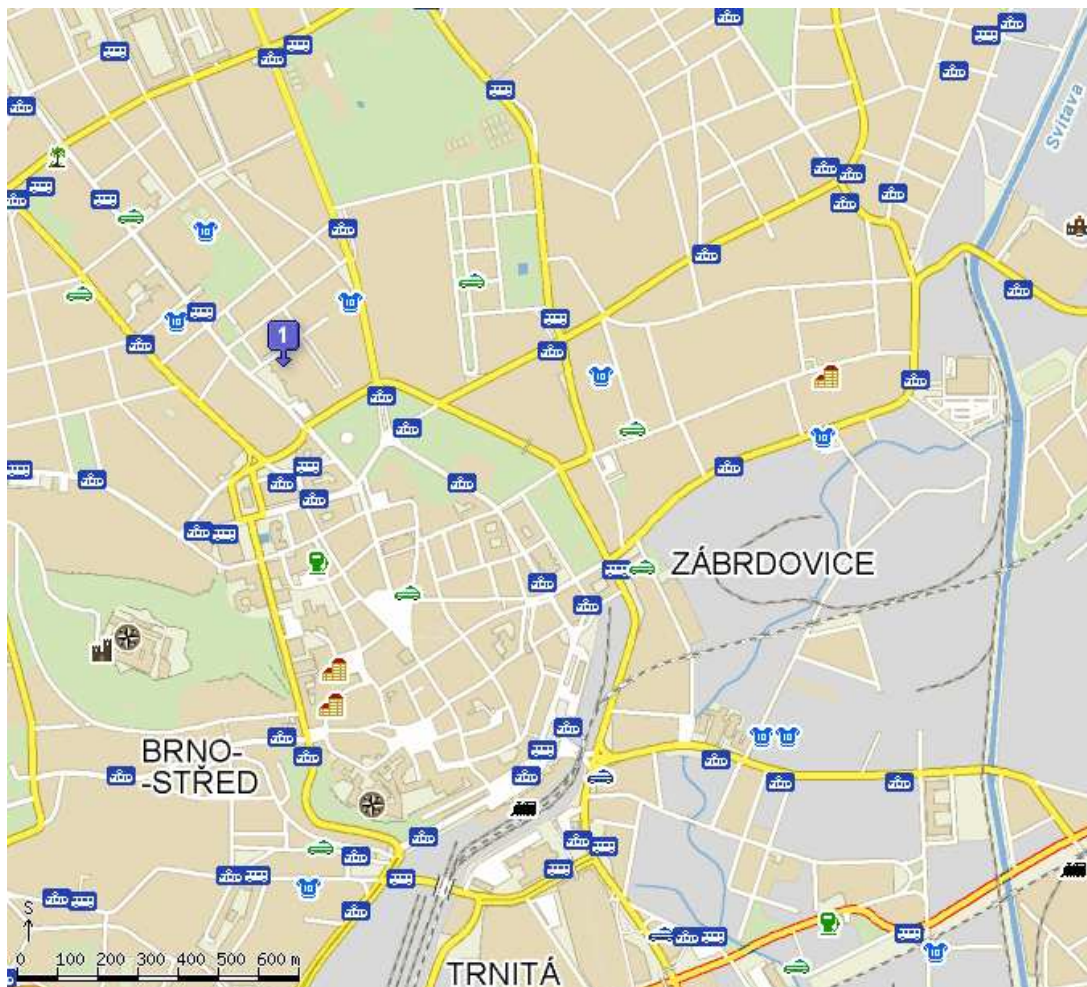
MLMI'07 is collocated with a number of events, such as the Summer school of the European Masters in Speech and Language (Brno, July 2-6), the PASCAL Speech Separation Challenge II, and the AMIDA Training Day (Brno, June 30). MLMI'07 also follows on directly from the annual conference of the Association for Computational Linguistics (Prague, June 25-27, 2007).


The MLMI'07 local organizers are members of the Faculty of Information Technology ([www.fit.vutbr.cz](http://www.fit.vutbr.cz)) at Brno University of Technology, founded in 1899.

---

## Practical Information

MLMI'07 takes place at the Hotel Continental, a modern hotel located within walking distance from the city centre. Address: Kounicova 6, 602 00 Brno; phone: +420 541 519 111; fax: +420 541 211 203; email: [info@continentalbrno.cz](mailto:info@continentalbrno.cz); web page: [www.continentalbrno.cz](http://www.continentalbrno.cz). The hotel is marked as **1** on the Brno map below (courtesy of [www.mapy.cz](http://www.mapy.cz)).



Brno has an efficient **public transport** system ([www.dpmb.cz](http://www.dpmb.cz)). To get to the hotel from the main train station (*Hlavní nadraží*,  at the bottom-centre of the map) or bus station (*Autobusové nadraží*) you can take the tram no. 1 in front of the train station, direction Reckovice, and get off at Antonínska stop. Take Antonínska St. (on the left before the stop), go 300 m uphill, and turn again left at the small park.

**Tickets** for public transportation in Brno cost 13 Kc and can be bought from tobacco shops (*Tabak*), yellow vending machines at stops, or from drivers (more expensive and exact change needed). These tickets are valid for almost all trips within Brno, and must be stamped on board.

**Sightseeing.** Brno is the second largest city in the Czech Republic and the capital of Moravia. The city is about 200 km from Prague and 130 km from Vienna. In Brno, you can visit the Old Town Hall and the St. Peter and Paul cathedral, as well as the Crypt of the Capuchin monastery, or the Spilberk castle with its great views on a sunny day. There are many baroque churches in Brno, such as the Jezuit church on Jezuitska Street (open around noon). Brno has beautiful surroundings and it is often enough to get to the terminus of a tram or bus to be able to enjoy them.

Around Brno, if you have a spare day, do not miss the Moravian Karst (*Moravský kras*, [www.cavemk.cz](http://www.cavemk.cz)) and its main cave, *Punkevní*, with its boat trip (advance booking is recommended). You can also visit the small village of Krtiny with its large baroque cathedral.

**Emergency numbers.** From mobile phones dial the standard 112, from fixed phones dial 112 or 158.

---

## Organizing Committee

Herve Bourlard, IDIAP  
Honza Cernocky, Brno University of Technology (organization co-chair)  
Andrei Popescu-Belis, University of Geneva (programme chair)  
Steve Renals, University of Edinburgh (special sessions)  
Pavel Zemcik, Brno University of Technology (organization co-chair)

## Programme Committee

Marc Al-Hames, Munich University of Technology  
Jan Alexandersson, DFKI  
Tilman Becker, DFKI  
Samy Bengio, Google  
Herve Bourlard, IDIAP  
Nick Campbell, ATR  
Jean Carletta, University of Edinburgh  
Honza Cernocky, Brno University of Technology  
John Dines, IDIAP  
Sadaoki Furui, Tokyo Institute of Technology  
John Garofolo, NIST  
Daniel Gatica-Perez, IDIAP  
Luc van Gool, ETHZ  
Thomas Hain, University of Sheffield (area chair)  
James Henderson, University of Edinburgh  
Hynek Hermansky, IDIAP  
Vaclav Hlavac, Czech Technical University Prague (area chair)  
Alejandro Jaimes, IDIAP  
Samuel Kaski, Helsinki University of Technology  
Denis Lalanne, University of Fribourg  
Yang Liu, University of Texas at Dallas (area chair)  
Stephane Marchand-Maillet, University of Geneva  
Jean-Claude Martin, LIMSI  
Helen Meng, Chinese University of Hong Kong  
Nelson Morgan, ICSI  
Ludek Muller, University of West Bohemia  
Roderick Murray-Smith, University of Glasgow  
Sharon Oviatt, OGI/OHSU (area chair)

Andrei Popescu-Belis, University of Geneva (programme chair)  
Ganesh Ramaswamy, IBM TJ Watson Research Center  
Steve Renals, University of Edinburgh  
Jan Sedivy, IBM Prague  
Elizabeth Shriberg, SRI and ICSI  
Rainer Stiefelhagen, University of Karlsruhe (area chair)  
Jean-Philippe Thiran, EPFL  
Pierre Wellner, IDIAP  
Dekai Wu, Hong Kong University of Science and Technology  
Pavel Zemcik, Brno University of Technology

## Webmasters

Jonathan Kilgour, University of Edinburgh  
Josef Zizka, Brno University of Technology

## Conference Secretary

Ms. Sylva Otahalova  
Department of Computer Graphics and Multimedia  
Faculty of Information Technology, Brno University of Technology  
Bozetechova 2, 612 66 Brno, Czech Republic  
Tel: +420 541141296 | Fax: +420 541141290 | [otahala@fit.vutbr.cz](mailto:otahala@fit.vutbr.cz)

---

## MLMI'07 Program

### THURSDAY 28 JUNE 2007

08:00-09:00 **Registration**

**09:00-09:10 Opening remarks**

PAPER SESSION 1: Features for Speech Recognition

09:10-09:40 **Using Prosodic Features in Language Models for Meetings**  
Songfang Huang and Steve Renals

09:40-10:10 **Posterior-Based Features and Distances in Template Matching for Speech Recognition**  
Guillermo Aradilla and Hervé Bourlard

10:10-10:40 **A Study of Phoneme and Grapheme Based Context-dependent ASR Systems**  
John Dines and Mathew Magimai Doss

10:40-11:10 - Coffee break

INVITED TALK 1

11:10-12:00 **How to follow a conversation without listening to the words**  
Nick Campbell (Media Information Science Laboratories, ATR, Japan)

12:00-13:30 - Lunch break

PAPER SESSION 2: Separation and Segmentation in Spoken Interaction

13:30-14:00 **Automatic Labeling Inconsistencies Detection And Correction for Sentence Unit Segmentation in Conversational Speech**  
Sebastien Cuendet, Dilek Hakkani-Tur and Elizabeth Shriberg

14:00-14:30 **Modeling Vocal Interaction for Segmentation in Meeting Recognition**  
Kornel Laskowski and Tanja Schultz

14:30-15:00 **Binaural Speech Separation Using Recurrent Timing Neural Networks for Joint F0-Localisation Estimation**  
Stuart Wrigley and Guy Brown

15:00-15:30 - Coffee break

SPECIAL SESSION: PASCAL Speech Separation Challenge II

15:30-16:00 **To Separate Speech! A System for Recognizing Simultaneous Speech**  
John McDonough, Kenichi Kumatani, Tobias Gehrig, Emilian Stoimenov, Uwe Mayer,  
Stefan Schacht, Matthias Woelfel and Dietrich Klakow

16:00-16:30 **A Microphone Array Beamforming Approach to Blind Speech Separation**  
Iain McCowan, Ivan Himawan and Mike Lincoln

POSTER SESSION 1

16:30-18:00 List of posters below

**FRIDAY 29 JUNE 2007**

PAPER SESSION 3: Image and Video Processing of Human Interaction

09:00-09:30 **Conditional Sequence Model for Context-based Recognition of Gaze Aversion**  
Louis-Philippe Morency and Trevor Darrell

09:30-10:00 **Face Recognition in Smart Rooms**  
Hazim Kemal Ekenel, Mika Fischer and Rainer Stiefelhagen

10:00-10:30 **Meeting State Recognition from Visual and Aural Labels**  
Jan Curin, Pascal Fleury, Jan Kleindienst and Robert Kessl

10:30-11:00 - Coffee break

INVITED TALK 2

11:00-11:50 **Structure and images**  
Vaclav Hlavac (Center for Machine Perception, Czech Technical University, Prague)

11:50-13:30 - Lunch break

PAPER SESSION 4: Annotation and Structuring of Spoken Input

13:30-14:00 **Term-Weighting for Summarization of Multi-Party Spoken Dialogues**  
Gabriel Murray and Steve Renals

14:00-14:30 **Automatic Annotation of Dialogue Structure from Simple User Interaction**  
Matthew Purver, John Niekrasz and Patrick Ehlen

14:30-15:00 **Computer Assisted Pattern Recognition**  
Enrique Vidal, Luis Rodriguez, Francisco Casacuberta and Ismael García-Varea

15:00-15:30 - Coffee break

PAPER SESSION 5: Meeting Browsers and their Evaluation

15:30-16:00 **An Ego-centric and Tangible Approach to Meeting Indexing and Browsing**  
Denis Lalanne, Florian Evequoz, Maurizio Rigamonti, Bruno Dumas and Rolf Ingold

16:00-16:30 **Towards an Objective Test for Meeting Browsers: the BET4TQB Pilot Experiment**  
Andrei Popescu-Belis, Philippe Baudrion, Mike Flynn and Pierre Wellner

POSTER SESSION 2

16:30-18:00 List of posters below

**SATURDAY 30 JUNE 2007**

AMIDA TRAINING DAY

	Introduction - AMI aims, objectives, overview	Steve Renals / Herve Bourlard
	The AMIDA Data Recording Environment	Mike Lincoln
	Automatic Speech Recognition in AMI	Thomas Hain
09:30-17:30	Speaker Tracking	Pavel Zemcik / Igor Potucek
	Understanding Content and Structure of Multi-modal Interaction	Tilman Becker
	Summarisation	Thomas Kleinbauer
	HCI, Application Prototyping and Evaluation	Alex Jaimes

## MLMI'07 Abstracts

### Invited Talks

#### **How to follow a conversation without listening to the words**

- **Nick Campbell**, National Institute of Information and Communications Technology & ATR Spoken Language Communication Research Labs, Japan.

This talk describes the use of nonverbal speech sounds in spoken interaction and shows how a technology can be trained to follow the flow of a conversation or dialogue without any understanding any of its propositional content. The first part of the talk describes the collection of a very large corpus of natural conversational speech, the second part outlines the major findings from its analysis, and the third part describes the development of a multimodal device for tracking participant status in such conversational dialogues.

#### **Structure and images**

- **Vaclav Hlavac**, Center for Machine Perception, Czech Technical University, Prague

Statistical pattern recognition methods have had difficulties to deal with images for several decades. The main obstacle is that a standard statistical approach cannot directly cope with the structure induced by the neighborhood relation in images. The talk will demonstrate these issues on several examples. Many researchers think that attempts to apply structural pattern recognition methods in the 1960s and 1970s led to the dead end. I like to advocate that the structural pattern recognition can be embedded into the statistical pattern recognition framework. This step has the potential to bring robustness to the structural approach. There are several possibilities in which such approach solves practically applicable tasks. A few examples from our recent research will be given, e.g., (a) optimizations on Markovian random fields applied to non-rigid matching in images or segmentation; (b) the structural construction applied to grammar-based recognition of the on-line hand written text.

### Oral Presentations

#### **Using Prosodic Features in Language Models for Meetings**

- Songfang Huang and Steve Renals

Prosody has been actively studied as an important knowledge source for speech recognition and understanding. In this paper, we are concerned with the question of exploiting prosody for language models to aid automatic speech recognition in the context of meetings. Using an automatic syllable detection algorithm, the syllable-based prosodic features are extracted to form the prosodic representation for each word. Two modeling approaches are then investigated. One is based on a factored language model, which directly uses the prosodic representation and treats it as a 'word'. Instead of direct association, the second approach provides a richer probabilistic structure within a hierarchical Bayesian framework by introducing an intermediate latent variable to represent similar prosodic patterns shared by groups of words. Four-fold cross-validation experiments on the ICSI Meeting Corpus show that exploiting prosody for language modeling can significantly reduce the perplexity, and also have marginal reductions in word error rate.

#### **Posterior-Based Features and Distances in Template Matching for Speech Recognition**

- Guillermo Aradilla and Hervé Bourlard

The use of large speech corpora in example-based approaches for speech recognition is mainly focused on increasing the number of examples. This strategy presents some difficulties because databases may not provide enough examples for some rare words. In this paper we present a different method to incorporate the information contained in such corpora in these example-based systems. A multilayer perceptron is trained on these databases to estimate speaker and task-independent phoneme posterior probabilities, which are used as speech features. By reducing the variability of features, less number of examples are needed to properly characterize a word. In this way, performance can be highly improved when limited number of examples is available. Moreover, we also study posterior-based local distances which result more effective than traditional Euclidean distance. Experiments on Phonebook database support the idea that posterior features with a proper local distance is effective in example-based approaches.

## **A Study of Phoneme and Grapheme Based Context-dependent ASR Systems**

- John Dines and Mathew Magimai Doss

In this paper we present a study of automatic speech recognition systems using context-dependent phonemes and graphemes as sub-word units based on the conventional HMM/GMM system as well as tandem system. Experimental studies conducted on three different continuous speech recognition tasks show that systems using only context-dependent graphemes can yield competitive performance on small to medium vocabulary tasks when compared to a context-dependent phoneme-based automatic speech recognition system. In particular, we demonstrate the utility of tandem features that use an MLP trained to estimate phoneme posterior probabilities in improving grapheme based recognition system performance by incorporating phonemic knowledge into the system without having to explicitly define a phonetically transcribed lexicon.

## **Automatic Labeling Inconsistencies Detection and Correction for Sentence Unit Segmentation in Conversational Speech**

- Sebastien Cuendet, Dilek Hakkani-Tur and Elizabeth Shriberg

In conversational speech, irregularities in the speech such as overlaps and disruptions make it difficult to decide what a sentence is. Thus, despite very precise guidelines on how to label conversational speech with dialog acts (DA), labeling inconsistencies are likely to appear. In this work, we present various methods to detect labeling inconsistencies in the ICSI meeting corpus. We show that by automatically detecting and removing the inconsistent examples from the training data, we significantly improve the sentence segmentation accuracy. We then manually analyze 200 of noisy examples detected by the system and observe that only 13% of them are labeling inconsistencies, while the rest are errors done by the classifier. The errors naturally cluster into 5 main classes for each of which we give hints on how the system can be improved to avoid these mistakes.

## **Modeling Vocal Interaction for Segmentation in Meeting Recognition**

- Kornel Laskowski and Tanja Schultz

Automatic segmentation is an important technology for both automatic speech recognition and automatic speech understanding. In meetings, participants typically vocalize for only a fraction of the recorded time, but standard vocal activity detection algorithms for close-talk microphones in meetings continue to treat participants independently. In this work we present a multispeaker segmentation system which models a particular aspect of human-human communication, that of vocal interaction or the interdependence between participants' on-off speech patterns. We describe our vocal interaction model, its training, and its use during vocal activity decoding. Our experiments show that this approach almost completely eliminates the problem of crosstalk, and word error rates on our development set are lower than those obtained with human-generated reference segmentation. We also observe significant performance improvements on unseen data.

## **Binaural Speech Separation Using Recurrent Timing Neural Networks for Joint F0-Localisation Estimation**

- Stuart Wrigley and Guy Brown

A speech separation system is described in which sources are represented in a joint interaural time difference-fundamental frequency (ITD-F0) cue space. Traditionally, recurrent timing neural networks (RTNNs) have been used only to extract periodicity information; in this study, this type of network is extended in two ways. Firstly, a coincidence detector layer is introduced, each node of which is tuned to a particular ITD; secondly, the RTNN is extended to become two-dimensional to allow periodicity analysis to be performed at each best-ITD. Thus, one axis of the RTNN represents F0 and the other ITD allowing sources to be segregated on the basis of their separation in ITD-F0 space. Source segregation is performed within individual frequency channels without recourse to cross-channel estimates of F0 or ITD that are commonly used in auditory scene analysis approaches. The system is evaluated on spatialised speech signals using energy-based metrics and automatic speech recognition.

## **To Separate Speech! A System for Recognizing Simultaneous Speech**

- John McDonough, Kenichi Kumatani, Tobias Gehrig, Emilian Stoimenov, Uwe Mayer, Stefan Schacht, Matthias Woelfel and Dietrich Klakow

## **A Microphone Array Beamforming Approach to Blind Speech Separation**

- Iain McCowan, Ivan Himawan and Mike Lincoln

The two most common classes of algorithms to recover mixed speech from multiple distant microphones have been microphone array beamforming and blind source separation. Of these, beamforming has perhaps been the most promising for speech recognition applications. Beamforming, however, requires knowledge of both the microphone positions and the locations of any speakers. In contrast, blind source separation techniques attempt to separate a mixture of speech and noise based only on the assumption of statistical independence between the signals, and thus have the advantage of not requiring explicit knowledge of locations. They can suffer, however, from a lack of robustness to real speech data in practical situations. This article presents a microphone array beamforming system

that does not rely on any prior knowledge of microphone or speaker geometry. The system is evaluated as a front-end to recognise overlapped sentences in the PASCAL Speech Separation Challenge II.

### **Conditional Sequence Model for Context-based Recognition of Gaze Aversion**

- Louis-Philippe Morency and Trevor Darrell

Eye gaze and gesture form key conversational grounding cues that are used extensively in face-to-face interaction among people. To accurately recognize visual feedback during interaction, people often use contextual knowledge from previous and current events to anticipate when feedback is most likely to occur. In this paper, we investigate how dialog context from an embodied conversational agent (ECA) can improve visual recognition of eye gestures. We propose a new framework for contextual recognition based on Latent-Dynamic Conditional Random Field (LDCRF) models to learn the sub-structure and external dynamics of contextual cues. Our experiments show that adding contextual information improves visual recognition of eye gestures and demonstrate that the LDCRF model for context-based recognition of gaze aversion gestures outperforms Support Vector Machines, Hidden Markov Models, and Conditional Random Fields.

### **Face Recognition in Smart Rooms**

- Hazim Kemal Ekenel, Mika Fischer and Rainer Stiefelhagen

In this paper, we present a detailed analysis of the face recognition problem in smart room environment. We first examine the well-known face recognition algorithms in order to observe how they perform on the images collected under such environments. Afterwards, we investigate two aspects of doing face recognition in smart rooms. These are: utilizing the images captured by multiple fixed cameras located in the room and handling possible registration errors due to the low resolution of the acquired face images. In addition, we also provide comparisons between frame-based and video-based face recognition and analyze the effect of frame weighting. Experimental results obtained on the CHIL database, which has been collected from different smart rooms, show that benefiting from multi-view video data and handling registration errors reduce the false identification rates significantly.

### **Meeting State Recognition from Visual and Aural Labels**

- Jan Curin, Pascal Fleury, Jan Kleindienst and Robert Kessl

In this paper we present a meeting state recognizer based on a combination of multi-modal sensor data in a smart room. Our approach is based on the training of a statistical model to use semantic cues generated by perceptual components. These perceptual components generate these cues in processing the output of one or multiple sensors. The presented recognizer is designed to work with an arbitrary combination of multi-modal input sensors. We have defined a set of states representing both meeting and non-meeting situations, and a set of features we base our classification on. Thus, we can model situations like *presentation* or *break* which are important information for many applications. We have hand-annotated a set of meeting recordings to verify our statistical classification, as appropriate multi-modal corpora are currently very sparse. We have also used several statistical classification methods for the best classification, which we validated on the hand-annotated corpus of real meeting data.

### **Term-Weighting for Summarization of Multi-Party Spoken Dialogues**

- Gabriel Murray and Steve Renals

This paper explores the issue of term-weighting in the genre of spontaneous, multi-party spoken dialogues, with the intent of using such term-weights in the creation of extractive meeting summaries. The field of text information retrieval has yielded many term-weighting techniques to import for our purposes; this paper implements and compares several of these, namely TF.IDF, Residual IDF and Gain. We propose that term-weighting for multi-party dialogues can exploit patterns in word usage among participant speakers, and introduce the SU.IDF metric as one attempt to do so. Results for all metrics are reported on both manual and automatic speech recognition (ASR) transcripts, and on both the ICSI and AMI meeting corpora.

### **Automatic Annotation of Dialogue Structure from Simple User Interaction**

- Matthew Purver, John Niekrasz and Patrick Ehlen

In previous work, we presented a method for automatic detection of action items from natural conversation. This method relies on supervised classification techniques that are trained on data annotated according to a hierarchical notion of dialogue structure; data which are expensive and time-consuming to produce. We also presented previously a meeting browser which allows users to view a set of automatically-produced action item summaries and give feedback on their accuracy. In this paper, we investigate methods of using this kind of feedback as implicit supervision, in order to bypass the costly annotation process and enable machine learning through use. We investigate, through the transformation of human annotations into hypothetical idealized user interactions, the relative utility of various modes of user interaction as well as various techniques for automatically producing training instances from interaction. We show that performance improvements are possible from interaction alone, even with interfaces that present very low cognitive load to users.

## **Computer Assisted Pattern Recognition**

- Enrique Vidal, Luis Rodriguez, Francisco Casacuberta and Ismael García-Varea

Pattern Recognition systems are not error-free. Human intervention is typically needed to verify and/or correct the result of such systems. To formalize this fact, a new framework, which integrates the human activity into the recognition process taking advantage of the user's feedback, is described. Several applications, involving Computer Assisted Speech Transcription and Multimodal Interactive Machine Translation, have recently been considered under this framework. These applications are reviewed in this paper, and some experiments, showing that the proposed framework can save significant amounts of human effort, are also presented.

## **An Ego-centric and Tangible Approach to Meeting Indexing and Browsing**

- Denis Lalanne, Florian Evequoz, Maurizio Rigamonti, Bruno Dumas and Rolf Ingold

This article presents an ego-centric approach for indexing and browsing meetings. The method considers two concepts: meetings' data alignment with personal information to enable ego-centric browsing and live intentional annotation of meetings through tangible actions to enable ego-centric indexing. The article first motivates and introduces these concepts and further presents brief states-of-the-art of the domain of tangible user interaction, of document-centric multimedia browsing, a traditional tangible object to transport information, and of personal information management. The article then presents our approach in the context of meeting and details our methods to bridge the gap between meeting data and personal information.

## **Towards an Objective Test for Meeting Browsers: the BET4TQB Pilot Experiment**

- Andrei Popescu-Belis, Philippe Baudrion, Mike Flynn and Pierre Wellner

This paper outlines the BET method for task-based evaluation of meeting browsers, based on an abstract definition of 'observations of interests' in meetings, which are empirically determined by neutral observers and then processed and ordered by evaluators. The TQB annotation-driven meeting browser was submitted to evaluation using the BET. A series of subjects attempted to answer as many meeting-related questions as possible in a fixed amount of time, and their results were measured in terms of precision and speed. Results indicate that the TQB interface is easy to understand with little prior learning and that its annotation-based search functionality is highly relevant, in particular keyword search. Two knowledge-poorer browsers appear to offer lower precision but comparable speed. These results indicate that the BET task-based evaluation method appears to be a coherent measure of browser quality.

## **Posters and Demos**

### **POSTER SESSION 1: THURSDAY 28 JUNE 2007, 16:30-18:00**

#### **Studying Multimodal Fusion and Fission Mechanisms through the Constitution of an Open Source Toolkit Allowing Rapid Creation of Multimodal Interfaces**

- Bruno Dumas, Denis Lalanne, Rolf Ingold

This poster presents HephaisTK, a project which targets the development of novel multimodal fusion and fission mechanisms through the constitution of an open-source toolkit allowing the rapid creation of multimodal interfaces. This poster hence presents the current version of an agent-based framework dedicated to the creation of multimodal interfaces; a first multimodal application has been developed using this framework and potential SMM applications are presented.

#### **Transfer Learning for Meeting-domain Tandem ASR Features**

- Joe Frankel, Ozgur Cetin, Nelson Morgan

Tandem automatic speech recognition (ASR), in which one or an ensemble of multi-layer perceptrons (MLPs) is used to provide a non-linear transform of the acoustic parameters, has become a standard technique in a number of state-of-the-art systems. In this paper, we examine the question of how to transfer MLP learning from one task to another. Experimental work is concerned with recognition of farfield speech data from the NIST RT05s evaluation.

#### **Neural Network Topologies and Bottleneck Features in Speech Recognition**

- Frantisek Grezl, Martin Karafiat, Jan Cernocky

In recent years, probabilistic features became an integral part of state-of-the-art LVCSR systems. These features are estimated probabilities of speech classes, usually phonemes, converted to suitable form for subsequent standard GMM-HMM modeling by logarithm (gaussianization) and PCA (decorrelation and dimensionality reduction). The probabilities are estimated by Artificial Neural network (ANN) usually in form of fully connected feed forward three layer Multi-Layer Perceptron (MLP). Bottle-Neck features are obtained directly from neural net. These features are

called bottle-neck, because they are obtained from five layer MLP with very narrow middle layer. Since the classification performance of the 5-layer MLP does not degrade compared to a three layer one, there was the guarantee of sufficient information for classification on the bottle neck output and thus hope for good features for GMM-HMM recognizer.

### **Automatic Decision Detection of Meeting Speech**

- Peiyun Hsueh, Jonathan Kilgour, Jean Carletta, Steve Renals, Johanna Moore

Making decisions is an indispensable aspect of meeting conversations. However, for someone who is absent from a meeting, finding out what decisions have been made from the recording of the meeting is a challenging task. AMI DecisionDetector is a system that aims to facilitate the task of reviewing decisions by performing automatic decision detection in meeting speech. Specifically, AMI DecisionDetector marks those topic segments where interlocutors have reached decisions and detects important dialogue acts pertaining to the decisions made. In this demo, we will demonstrate how AMI DecisionDetector can guide the users to find the decisions from a meeting recording and provide sufficient visual aids for them to interpret the content of those decisions.

### **Channel Compensation for Speaker Recognition**

- Valiantsina Hubeika, Lukas Burget, Pavel Matejka, Jan Cernocky

Variability in the channel and session is an important issue in text-independent speaker recognition task. To date, several techniques providing channel and session variability compensation were introduced in a number of scientific papers. Such implementation can be done in feature, model and score domain. Relatively new and powerful approach to remove channel distortion is so-called eigenchannel adaptation for Gaussian Mixture Models (GMM). The drawback of the technique is that it is not applicable to different types of classifiers, e.g. Support Vector Machines (SVM), to GMM with different number of Gaussians or in speech recognition task using Hidden Markov Models (HMM). The solution can be an approximation of the technique, eigenchannel adaptation in feature domain. Both, the original eigenchannel adaptation and eigenchannel adaptation on features using GMM are briefly introduced. Presented results are achieved using the BUT GMM system submitted for NIST SRE 2006 and on 2006 data.

### **In-Context Phone Posteriors as Complementary Features for Tandem ASR**

- Hamed Ketabdar, Herve Bourlard

In this paper, we present a method for integrating possible prior knowledge (such as phonetic and lexical knowledge), as well as acoustic context (e.g., the whole utterance) in the phone posterior estimation, and we propose to use the obtained posteriors as complementary posterior features in Tandem ASR configuration. These posteriors are estimated based on HMM state posterior probability definition (typically used in standard HMMs training). In this way, by integrating the appropriate prior knowledge and context, we enhance the estimation of phone posteriors. These new posteriors are called 'in-context' or HMM posteriors. We combine these posteriors as complementary evidences with the posteriors estimated from a Multi Layer Perceptron (MLP), and use the combined evidence as features for training and inference in Tandem configuration. This approach has improved the performance, as compared to using only MLP estimated posteriors as features in Tandem, on OGI Numbers, Conversational Telephone speech (CTS), and Wall Street Journal (WSJ) databases.

### **Czech Text-to-Sign-Speech Synthesizer**

- Zdenek Krnoul, Jakub Kanis, Milos Zelezny, Ludek Muller

For communication of aurally or speech disabled people with computer, special techniques have to be used. Using Sign Speech as communication means is one of them. To provide Sign Speech output of a computer, the textual or speech information has first to be translated to textual Sign Speech representation, then into symbolic representation. Sign Speech synthesis module then generates geometric parameterization and renders signing avatar animation. The task is a part of a project on creating Sign Speech dialogue system.

### **User-specific Training of a Music Search Engine**

- David Little, David Raffensperger, Bryan Pardo

Query-by-Humming (QBH) systems let a user find the desired song in a music database by humming or whistling its melody. Existing systems do not optimize on individual users, once deployed. We present a method to improve QBH performance with user-specific training on a deployed system. Parameters for the singer error model and note transcription are tuned using a genetic algorithm. Testing over a corpus of sung queries our preliminary tests show songs within and near the top ten songs listed.

### **Study on Correlation between ROUGE and Human Evaluation in Meeting Summarization**

- Feifan Liu, Yang Liu, Bin Li

Meeting summarization has recently gained more attention in speech and language processing. Thus finding an effective automatic evaluation metric is quite crucial to advance system development. ROUGE has been shown to correlate well with human evaluation in text summarization and widely used in various summarization tasks, such as DUC. However, in the meeting domain, current research has mainly focused on summarization approaches, most of which directly using ROUGE for performance measurement. The question of whether ROUGE also correlates well with human evaluation on meeting summaries is unclear, since the many characteristics in multiparty meeting domain (such as disfluencies, multi-speakers) may pose potential problems to ROUGE. In this paper we conduct human evaluation, adapt the ROUGE metric to account for issues in meeting domain, and carefully examine their correlation using both human and system generated summaries. We find that both R-1 and R-SU4 scores in ROUGE correlate well with human evaluation when ROUGE is adapted to the meeting style, such as removing disfluencies, using domain stop-words and speaker specific information.

### **Do Disfluencies Affect Meeting Summarization? A Pilot Study on the Impact of Disfluencies**

- Yang Liu, Feifan Liu, Bin Li, Shasha Xie

Meeting summarization and browsing can effectively help humans cope with the large amount of meeting data. Several approaches have been developed for speech summarization in the meeting domain. Most of them simply used the transcriptions as is (either human transcriptions or recognition output). However, since conversational speech (as in meetings) contains many disfluencies, we suspect they do not contribute much to the meaning of the sentences and probably have an impact on summarization performance. In this paper, we conduct a pilot study to evaluate how disfluencies affect meeting summarization.

### **Frequency Domain Linear Prediction for QMF Subbands and Applications to Audio coding**

- Petr Motlicek, Sriram Ganapathy, Hynek Hermansky

Here, we propose a high quality audio coding technique, exploiting the predictability of temporal evolution of QMF sub-band signals. The QMF filters form a perfect reconstruction tree structured filter bank with critical downsampling of the sub-band signals. Temporal envelopes in critically sampled QMF sub-bands are estimated using frequency domain linear prediction applied on relatively long time segments (1s). The sub-band residual signals, which play an important role in acquiring high quality reconstruction, are processed and the important parameters in the frequency domain are encoded. When used for coding challenging audio recordings, even with very simple quantization techniques, the proposed technique is able to reconstruct with high quality at reasonable bit rates.

### **LP-TRAP based Speech Features for ASR**

- Petr Motlicek, Hynek Hermansky

This work presents LP-TRAP speech feature extraction technique developed for Automatic Speech Recognition (ASR). The technique allows for approximating the temporal evolution of spectral envelopes in critical-band-sized sub-bands of a speech signal by the autoregressive model. By exploiting generalized autocorrelation linear predictive technique, we can control fitting the peaks and the dips of the original temporal trajectory of the sub-band signal. By using analysis windows of the order of hundred milliseconds, the procedure automatically decides how to distribute poles of the autoregressive model to best model the temporal structure within window. Recursively computed cepstral representation as well as direct temporal envelopes are used as inputs to the TANDEM-based speech recognition system evaluated on standard ASR task. We show improvements in terms of frame-error accuracies and WERs over traditional PLP or TRAP features.

### **Cross Entropy for Learning in Multimodal Streams**

- Athanasios Noulas, Ben Krose, Nikos Vlassis

In this paper we present a variation of the Cross Entropy method that can be applied on Dynamic Bayesian Networks for efficient learning of the model parameters. We demonstrate the results achieved on real world video streams using a variety of DBNs. Finally we compare this approach to the traditional EM algorithm, in terms of computational complexity, memory requirements and robustness to initialization.

### **Audiovisual Interaction in the Control of Human Overt Attention**

- Cliona Quigley, Selim Onat, Peter Koenig, Sue Harding, Martin Cooke

In every moment of awake life, we are bombarded with information from our different senses. Attentional mechanisms allow us to sort through this flood of information and process just a subset of our surroundings. Eye movements are an important measure of overt attention, and knowing where people look when presented with a certain stimulus provides a valuable insight into the processes underlying the allocation of attention. Eye-tracking studies have contributed a great deal to the current knowledge of what makes a visual stimulus salient. But what happens when other modalities come into play? When and how are individual feature channels from different modalities combined? The existing body of research into multimodal processing is based mainly on experiments with artificial stimuli that are often far from true to the statistics of our natural world. Answering the question of what a

perceptual system actually does under normal operating conditions, however, should involve the use of natural stimuli.

### **Combination of Word and Phoneme Approach for Spoken Term Detection**

- Igor Szöke

The topic of this poster is Spoken Term Detection. The goal is to provide accurate and fast technique for search for a term or keyword in a speech. The spoken term detection is based on lattices generated from word and phoneme recognizers. The term is a sequence of words (quoted query). The results are evaluated on meeting data (AMI corpus), NIST RT 05/06 eval data and on NIST STD 2006 development set.

### **Search in Meetings Using Combination of LVCSR- and Phonetic-based Spoken Term Detection**

- Igor Szöke, Michal Fapšo

This demo is a follow-up of BUT speech search demos presented at the previous Technology transfer events of AMI. On contrary to previous version, that was based solely on LVCSR lattice search and this suffered from limitation by recognition vocabulary, the version to be presented at MLMI 2007 features: combined LVCSR- and phonetic-search, multiword queries, and sorting of hits according to confidence estimated from both LVCSR and phonetic lattices. The search is integrated into JFerret meeting browser and will be demonstrated on four AMI Pilot meetings.

### **Spoken Term Detection System Based on a Combination of LVCSR and Phonetic Search**

- Igor Szöke, Michal Fapšo, Martin Karafiát, Lukáš Burget, František Grézl, Petr Schwarz, Ondřej Glembek, Pavel Matějka, Jiří Kopecký, Jan Černocký

The paper presents Brno University of Technology (BUT) system for indexing and search of speech, combining LVCSR and phonetic approach. NIST Spoken term detection evaluation data and metrics are described, as well as details and results of BUT system.

### **Combined Visual Parameterization for Automatic Lip-Reading**

- Jana Trojanova, Petr Cisar, Milos Zelezny

Experiments with various visual speech parameterizations were performed on two databases (English XMVTSDB and Czech UWB-05-HSCAVC). A new parameterization based on knowledge of human lip-reading experts was designed. It combines both shape-based and image-based methods. Results proved the contribution of the new parameterization to increasing the recognition rate.

### **Grapheme-based Spoken Term Detection in the Meetings Domain**

- Dong Wang, Joe Frankel, Simon King

In this work, we propose using context-dependent graphemes (CDG) as sub-word units for spoken term detection, in particular for out-of-vocabulary search terms. In essence, this approach moves pronunciation modelling away from the letter-to-sound rules which are used to generate phone strings, and into the Gaussian mixture models which describe the observation space. This removes the need to make potentially error-prone hard decisions at an early stage of processing. In addition, words which have multiple pronunciations have a single grapheme representation which simplifies the subsequent search. Large text corpora can be used to train long-span grapheme-based language models for use in lattice generation. These language models have words implicit within them, though given suitable smoothing can be used to support previously unseen words. We first present the results of phone and grapheme recognition, in addition to word recognition based on phone and grapheme sub-word units. We then present results on a spoken term detection (STD) task. Our preliminary results are encouraging: a grapheme-based approach offers the possibility of fast implementations, rapid adaptation to new domains, and can be trained on for languages for which resources such as large lexica are not available.

### **The Listening Room - A Speech-based Interactive Art Installation**

- Alexa Wright, Alun Evans, Mike Lincoln

In this poster we describe *The Listening Room* – an interactive art installation that incorporates a number of speech technologies. In *The Listening Room* up to three small sculptures are displayed on exhibition plinths. People entering the space are automatically tracked using webcams positioned overhead. An individual standing close to one of the sculptures triggers a disembodied voice, which will then try to engage that person in conversation. Using keywords to interpret what is said in reply, the disembodied voice will pursue a more or less meaningful dialogue that can be heard only at a particular location in the space. 'The voice' will be able to conduct conversations at up to three different locations. **POSTER SESSION 2: FRIDAY 29 JUNE 2007, 16:30-18:00**

## **Archivus: A User Performance Analysis with Speech, Keyboard and Mouse as Interaction Modalities**

- Marita Ailomaa, Agnes Lisowska

Our poster presents part of an analysis of the results of a user experiment where the task was to answer questions about what happened in meetings using Archivus, a multimodal (mouse, keyboard and speech) dialogue system with a graphical user interface. The analysis presented here focuses on users' performance on the tasks, and more specifically on the impact that the availability of different interaction modalities and the use of command-and-control actions made in natural language had on user performance.

## **Adaptable User Modeling Component for Multimodal Interaction Markup Language**

- Masahiro Araki

In this paper, we propose an adaptable user modeling component and its markup language for multimodal interactions. Our user modeling description enables a developer to define a set of user model attributes and acquire possible causal relations between them. In addition, these user model attributes can be easily used for control information for multimodal interactions.

## **Distributed Visual Sensor Network Fusion**

- Petr Chmelar, Jaroslav Zendulka

The poster deals with a framework for distributed visual sensor network metadata management system. It is assumed that data coming from many cameras is annotated using computer vision modules to produce metadata representing moving objects in their states. The data is supposed to be noisy, uncertain and some states might be missing. Firstly, here is described the spatio-temporal data cleaning using Kalman filter. Secondly, it copes with many visual sensors fusion and persistent object tracking within a large area. Thirdly, it describes the data and architecture model.

## **Multimodal Meeting Capture and Understanding with the CALO Meeting Assistant**

- Patrick Ehlen

The CALO Meeting Assistant is a multimodal meeting assistant technology that integrates speech, gestures, and multimodal data collected from multiparty interactions during meetings. Using machine learning and robust discourse processing, it provides a rich, browsable record of a meeting.

## **Gaussian Process Latent Variable Models for Human Pose Estimation**

- Carl Henrik Ek, Philip Torr, Neil Lawrence

We describe a generative approach to recover 3D human pose from image silhouettes. Our method is based on learning a shared low dimensional latent representation capable of generating both human pose and image observations through the GP-LVM. We learn a dynamical model over the latent space which allows us to disambiguate between ambiguous silhouettes by temporal consistency. The model has only two free parameters and requires no manual initialization.

## **The Hub: Real Time Data Distribution and Storage**

- Mike Flynn, Maël Guillemot, Bastien Crettol

There is a need for live meeting support and remote assistants which should also be able to retrieve historical data e.g. in case a person joins a meeting late. We present in this poster a Java-based system for live distribution of data between recognition systems and browsers: the hub. On one side, there are recognition systems (ASR, gesture recognition) that detect patterns (e.g. events, words) and produce annotations. On the other side there are browsers waiting for multimodal data. Experiments have been performed on the full 100-hours AMI meeting corpus annotation public release in order to validate the proposed scheme.

## **Hardware Acceleration of AdaBoost Classifier**

- Jiri Granat, Adam Herout, Michal Hradis, Pavel Zemčík

This paper presents a concept of hardware architecture targeting fast implementation of the AdaBoost classifier engine. Haar wavelet features are used as the weak classifiers; for evaluation of the features integral image is used. The internal part of the classifier is described in detail, as well as its interoperation with the overall image-processing system. The experiments verify the predicted performance expectations. The design concept and its implementation serves as a good starting point for further improvements and offered numerous recommendations for future constructs aiming at fast operation of the AdaBoost-based classifiers.

## **A Fully-automated Conference Recording and Webcasting System**

- Maël Guillemot, Alessandro Vinciarelli, Jean-Marc Odobez, Olivier Bornet, Olivier Masson

This work presents a system for the automatic acquisition, processing and delivery of oral presentations based on slides. The Presentation Acquisition System (PAS) has two main goals: the first is to capture oral presentations based on slides and the second is to provide users with an effective access to the presentations content. Since one of the main limits of current systems for the same tasks is the low degree of automation (operators are needed in order to make the system work), major efforts have been made to make the system fully automatic: the whole capture, processing and delivery process can be operated by pushing a single button. It is one of the first end-to-end conference recording and distribution systems. This makes the application suitable for real world applications where users do not necessarily have a technical background.

### **Automatic Camera-Selection for Meetings based on HMMs**

- Benedikt Hörnler, Marc Al-Hames, Gerhard Rigoll

Nowadays smart-meeting-rooms get more popular and so a lot of meetings and video-conferences are recorded with several cameras. One camera-view out of the available cameras can be selected for each frame to create an interesting meeting summary or to perform an online video-conference. The described system is based on Hidden-Markov-Models and fulfils this task satisfyingly.

### **Multi-Modal Interface for Information Access through Extraction and Visualization of Time-Series Information**

- Tsuneaki Kato, Mitsunori Matsushita

We often need to comprehend some trend and movement, and access a series of documents containing specific time-series information related. For example, we may wish to know the changes of gasoline prices in the last few months, or the increase in the number of cellular subscribers in the last decade. To meet these information requests, we propose a framework that extracts and visualizes given time-series information and its changes, and provides users with a multi-modal summarization and also an interactive interface for accessing that information. We emphasize the importance of changes of data during some time period rather than data points, as the unit of information extracted and represented. Based on this idea, we propose a visualization method in which qualitative and quantitative characteristics of changes of a given time-series information are plotted with textually represented comments, and a widely applicable information extraction method that regards the changes of time-series information as information primitives and extracts those for the visualization.

### **Indicative Abstractive Summaries of Meetings**

- Thomas Kleinbauer, Stephanie Becker, Tilman Becker

We present ongoing research on the generation of indicative meeting abstracts supporting quick relevance assessment of meetings, based on natural language generation techniques.

### **Analysis of the Multimodal Behaviour of Users in HCI: the Expert Viewpoint of Close Relations**

- Gilles Le Chenadec, Valérie Maffiolo, Noël Château

The aim of our research is to define an automated system to detect and characterize expressions of emotion-related states in the context of human-machine interactions. Our current work specially focuses on analysing and connecting three different forms of expressions of emotion-related states of users in a human-machine interaction. The first form concerns multimodal and visible expressions appearing during the interaction and supported by vocal, facial, gestural, postural modalities. The second form concerns verbal expressions out of the interaction but related to it, that are emotion-related comments given by users themselves and observers. Finally, the third form concerns physiological expressions appearing during the interaction. The poster will focus on the analysis of two first forms of expressions obtained in an experiment in which users were involved in an interaction with a virtual embodied agent.

### **Joint Bi-Modal Face and Speaker Authentication Using Explicit Polynomial Expansion**

- Sébastien Marcel

We propose a technique for modeling jointly audio and video streams at the feature level using an explicit polynomial expansion kernel. We concatenate audio and video frame features after the kernel expansion and then perform the classification using a linear SVM. Experiment results have been performed on a audio/visual database and compared to other techniques.

### **System for Automatic Language Identification**

- Pavel Matejka

Phonexia markets the language identifications system for which the theoretical and algorithmic foundations were laid by Speech@FIT. The experimental system is based on a combination of acoustic and phonotactic systems. The

experimental system scored excellently in NIST Language recognition evaluations 2005. The production system currently offered includes phonotactic modeling only. It consists of Hungarian phoneme recognizer and simple phonotactic language model.

### **Demonstration: Archivus - a Multimodal Dialogue System for Meeting Browsing and Retrieval**

- Miroslav Melichar, Agnes Lisowska, Pavel Cenek, Marita Ailomaa, Martin Rajman

As part of our work in the Interactive Multimodal Information Management (IM2) project, we developed a multimodal dialogue-driven interface for browsing and searching recorded and annotated meeting data in a multimedia database. The challenge lay in creating an interface that was easy to use and which smoothly blended direct manipulation and natural language interaction (both voice and keyboard based). Archivus was developed in part using the Wizard of Oz technique during extensive experiments with naive test subjects. In its current state, our Wizard's Controller Interface is quite an efficient tool for a wizard to control semantic interpretation of users multimodal input and to control dialogue aspects of the interaction in real-time. Both the Archivus system and Wizard Controller Interface will be presented during the demonstration and visitors will be invited to play the role of an Archivus user.

### **Evaluation and Comparison of Tracking Methods Using Meeting Omnidirectional Images**

- Igor Potucek, Vitezslav Beran, Stanislav Sumec, Pavel Zemčik

Visual cues, such as gesturing, looking at each other or monitoring each others facial expressions, play an important role in meetings. Such information can be used for indexing of multimedia meeting recordings. These situations are strongly focused nowadays. The omnidirectional system usage in such situations brings many advantages as portability, easy installation, large field of view, low cost etc. That is why we choose such scenarios for testing the omnidirectional system. Information about differences between omni-directional and classical images both for human presentation and tracking purposes is needed. We try to compare two different tracking methods on the various video sequences. The results of the tracking methods can help to demonstrate the benefits or drawbacks of the omni-directional system. The evaluation scheme was developed to bring us the aspects which affect the vision algorithms for detection and tracking of human bodies.

### **Object Category Recognition Using Probabilistic Fusion of Speech and Image Classifiers**

- Kate Saenko, Trevor Darrell

Multimodal scene understanding is an integral part of human-robot interaction (HRI) in situated environments. Especially useful is category-level recognition, where the the system can recognize classes of objects of scenes rather than specific instances (e.g., any chair vs. this particular chair.) Humans use multiple modalities to understand which object category is being referred to, simultaneously interpreting gesture, speech and visual appearance, and using one modality to disambiguate the information contained in the others. In this paper, we address the problem of fusing visual and acoustic information to predict object categories, when an image of the object and speech input from the user is available to the HRI system. Using probabilistic decision fusion, we show improved classification rates on a dataset containing a wide variety of object categories, compared to using either modality alone.

### **Evaluation of Automatic Video Editing**

- Stanislav Sumec, Igor Potucek

This paper describes a methodology that can be used for an evaluation of automatically generated videos. Criteria for evaluation such videos are presented. Chosen methodology combines experimental evaluation with human viewers, and synthetic experiments, which can be easily repeated. While viewers are able to evaluate an overall video quality, synthetic experiment are suitable for an evaluation of particular parts of video generator. Further, the video editing algorithm for processing of meeting data recorded with several cameras simultaneously is introduced. Some experiments performed with proposed algorithm and chosen methodology are mentioned.

### **Computer Assisted Pattern Recognition: Demonstrations**

- Enrique Vidal, Luis Rodriguez, Francisco Casacuberta, Ismael García-Varea

Demonstration related to the oral presentation.

### **Modeling Co-articulation by Visual Unit Selection in Czech Audio-Visual Speech Synthesis**

- Milos Zelezny, Zdenek Krnoul

In audio-visual speech synthesis, the effect of co-articulation has to be well solved to avoid unpleasant effects at the border of adjacent speech units. Visual Unit Selection (VUS) method was proposed as a rival candidate to commonly used Dominance Functions (DF). Results showed that VUS has better properties while at the same reaches the same level of accuracy according to the RMSE measure.

### **Local Rank Differences - Novel Features for Image Processing**

- Pavel Zemčik, Michal Hradiš, Adam Herout

One of the most important tasks in image processing and computer vision is extraction of image features. Image features serve as basic source of information for various tasks, such as segmentation, pattern matching, classification, etc. The image features used in certain image processing tasks are usually result of some compromise between the lowest "computational cost" and highest "information content" although the "information content" is in most cases evaluated only empirically. In many cases, invariance to lighting changes and possibly also to geometrical transformations is required or useful. This contribution presents novel image features that are superior to many existing ones in their invariance to lighting changes, low computational cost, and high information content. These features are described and evaluated as "weak classifiers" in AdaBoost classification method and compared to more traditional ones. Some implementation issues are also discussed including implementation in programmable hardware, such as FPGA.

---