Samy Bengio
Hervé Bourlard (Eds.)

LNCS 3361

Machine Learning for Multimodal Interaction

First International Workshop, MLMI 2004
Martigny, Switzerland, June 2004
Revised Selected Papers

Samy Bengio Hervé Bourlard (Eds.)

Machine Learning for Multimodal Interaction

First International Workshop, MLMI 2004
Martigny, Switzerland, June 21-23, 2004
Revised Selected Papers

Springer

Volume Editors

Samy Bengio
Hervé Bourlard
IDIAP Research Institute
Rue du Simplon 4, P.O. Box 592, 1920 Martigny, Switzerland
E-mail: {bengio,bourlard}@idiap.ch

Library of Congress Control Number: 2004118425

CR Subject Classification (1998): H.5.2-3, H.5, I.2.6, I.2.10, I.2, I.7, K.4, I.4

ISSN 0302-9743
ISBN 3-540-24509-X Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11384212 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science 3361

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Lecture Notes in Computer Science

For information about Vols. 1–3287

please contact your bookseller or Springer

Vol. 3412: X. Franch, D. Port (Eds.), COTS-Based Software Systems. XVI, 312 pages. 2005.

Vol. 3403: B. Ganter, R. Godin (Eds.), Formal Concept Analysis. XI, 419 pages. 2005. (Subseries LNAI).

Vol. 3398: D.-K. Baik (Ed.), Systems Modeling and Simulation: Theory and Applications. XIV, 733 pages. 2005. (Subseries LNAI).

Vol. 3397: T.G. Kim (Ed.), Artificial Intelligence and Simulation. XV, 711 pages. 2005. (Subseries LNAI).

Vol. 3391: C. Kim (Ed.), Information Networking. XVII, 936 pages. 2005.

Vol. 3388: J. Lagergren (Ed.), Comparative Genomics. VIII, 133 pages. 2005. (Subseries LNBI).

Vol. 3387: J. Cardoso, A. Sheth (Eds.), Semantic Web Services and Web Process Composition. VIII, 148 pages. 2005.

Vol. 3386: S. Vaudenay (Ed.), Public Key Cryptography - PKC 2005. IX, 436 pages. 2005.

Vol. 3385: R. Cousot (Ed.), Verification, Model Checking, and Abstract Interpretation. XII, 483 pages. 2005.

Vol. 3382: J. Odell, P. Giorgini, J.P. Müller (Eds.), Agent-Oriented Software Engineering V. X, 239 pages. 2004.

Vol. 3381: P. Vojtáš, M. Bieliková, B. Charron-Bost, O. Sýkora (Eds.), SOFSEM 2005: Theory and Practice of Computer Science. XV, 448 pages. 2005.

Vol. 3376: A. Menezes (Ed.), Topics in Cryptology – CT-RSA 2005. X, 385 pages. 2004.

Vol. 3375: M.A. Marsan, G. Bianchi, M. Listanti, M. Meo (Eds.), Quality of Service in Multiservice IP Networks. XIII, 656 pages. 2005.

Vol. 3368: L. Paletta, J.K. Tsotsos, E. Rome, G. Humphreys (Eds.), Attention and Performance in Computational Vision. VIII, 231 pages. 2005.

Vol. 3363: T. Eiter, L. Libkin (Eds.), Database Theory - ICDT 2005. XI, 413 pages. 2004.

Vol. 3362: G. Barthe, L. Burdy, M. Huisman, J.-L. Lanet, T. Muntean (Eds.), Construction and Analysis of Safe, Secure, and Interoperable Smart Devices. IX, 257 pages. 2005.

Vol. 3361: S. Bengio, H. Bourlard (Eds.), Machine Learning for Multimodal Interaction. XII, 362 pages. 2005.

Vol. 3360: S. Spaccapietra, E. Bertino, S. Jajodia, R. King, D. McLeod, M.E. Orlowska, L. Strous (Eds.), Journal on Data Semantics II. XI, 223 pages. 2004.

Vol. 3359: G. Grieser, Y. Tanaka (Eds.), Intuitive Human Interfaces for Organizing and Accessing Intellectual Assets. XIV, 257 pages. 2005. (Subseries LNAI).

Vol. 3358: J. Cao, L.T. Yang, M. Guo, F. Lau (Eds.), Parallel and Distributed Processing and Applications. XXIV, 1058 pages. 2004.

Vol. 3357: H. Handschuh, M.A. Hasan (Eds.), Selected Areas in Cryptography. XI, 354 pages. 2004.

Vol. 3356: G. Das, V.P. Gulati (Eds.), Intelligent Information Technology. XII, 428 pages. 2004.

Vol. 3355: R. Murray-Smith, R. Shorten (Eds.), Switching and Learning in Feedback Systems. X, 343 pages. 2005.

Vol. 3353: J. Hromkovič, M. Nagl, B. Westfechtel (Eds.), Graph-Theoretic Concepts in Computer Science. XI, 404 pages. 2004.

Vol. 3352: C. Blundo, S. Cimato (Eds.), Security in Communication Networks. XI, 381 pages. 2004.

Vol. 3350: M. Hermenegildo, D. Cabeza (Eds.), Practical Aspects of Declarative Languages. VIII, 269 pages. 2005.

Vol. 3349: B.M. Chapman (Ed.), Shared Memory Parallel Programming with Open MP. X, 149 pages. 2005.

Vol. 3348: A. Canteaut, K. Viswanathan (Eds.), Progress in Cryptology - INDOCRYPT 2004. XIV, 431 pages. 2004.

Vol. 3347: R.K. Ghosh, H. Mohanty (Eds.), Distributed Computing and Internet Technology. XX, 472 pages. 2004.

Vol. 3346: R.H. Bordini, M. Dastani, J. Dix, A.E.F. Seghrouchni (Eds.), Programming Multi-Agent Systems. XIV, 249 pages. 2005. (Subseries LNAI).

Vol. 3345: Y. Cai (Ed.), Ambient Intelligence for Scientific Discovery. XII, 311 pages. 2005. (Subseries LNAI).

Vol. 3344: J. Malenfant, B.M. Østvold (Eds.), Object-Oriented Technology. ECOOP 2004 Workshop Reader. VIII, 215 pages. 2004.

Vol. 3342: E. Şahin, W.M. Spears (Eds.), Swarm Robotics. IX, 175 pages. 2004.

Vol. 3341: R. Fleischer, G. Trippen (Eds.), Algorithms and Computation. XVII, 935 pages. 2004.

Vol. 3340: C.S. Calude, E. Calude, M.J. Dinneen (Eds.), Developments in Language Theory. XI, 431 pages. 2004.

Vol. 3339: G.I. Webb, X. Yu (Eds.), AI 2004: Advances in Artificial Intelligence. XXII, 1272 pages. 2004. (Subseries LNAI).

Vol. 3338: S.Z. Li, J. Lai, T. Tan, G. Feng, Y. Wang (Eds.), Advances in Biometric Person Authentication. XVIII, 699 pages. 2004.

Vol. 3337: J.M. Barreiro, F. Martin-Sanchez, V. Maojo, F. Sanz (Eds.), Biological and Medical Data Analysis. XI, 508 pages. 2004.

Vol. 3336: D. Karagiannis, U. Reimer (Eds.), Practical Aspects of Knowledge Management. X, 523 pages. 2004. (Subseries LNAI).

Vol. 3335: M. Malek, M. Reitenspieß, J. Kaiser (Eds.), Service Availability. X, 213 pages. 2005.

Vol. 3334: Z. Chen, H. Chen, Q. Miao, Y. Fu, E. Fox, E.-p. Lim (Eds.), Digital Libraries: International Collaboration and Cross-Fertilization. XX, 690 pages. 2004.

Vol. 3333: K. Aizawa, Y. Nakamura, S. Satoh (Eds.), Advances in Multimedia Information Processing - PCM 2004, Part III. XXXV, 785 pages. 2004.

Vol. 3332: K. Aizawa, Y. Nakamura, S. Satoh (Eds.), Advances in Multimedia Information Processing - PCM 2004, Part II. XXXVI, 1051 pages. 2004.

Vol. 3331: K. Aizawa, Y. Nakamura, S. Satoh (Eds.), Advances in Multimedia Information Processing - PCM 2004, Part I. XXXVI, 667 pages. 2004.

Vol. 3330: J. Akiyama, E.T. Baskoro, M. Kano (Eds.), Combinatorial Geometry and Graph Theory. VIII, 227 pages. 2005.

Vol. 3329: P.J. Lee (Ed.), Advances in Cryptology - ASIACRYPT 2004. XVI, 546 pages. 2004.

Vol. 3328: K. Lodaya, M. Mahajan (Eds.), FSTTCS 2004: Foundations of Software Technology and Theoretical Computer Science. XVI, 532 pages. 2004.

Vol. 3327: Y. Shi, W. Xu, Z. Chen (Eds.), Data Mining and Knowledge Management. XIII, 263 pages. 2004. (Subseries LNAI).

Vol. 3326: A. Sen, N. Das, S.K. Das, B.P. Sinha (Eds.), Distributed Computing - IWDC 2004. XIX, 546 pages. 2004.

Vol. 3323: G. Antoniou, H. Boley (Eds.), Rules and Rule Markup Languages for the Semantic Web. X, 215 pages. 2004.

Vol. 3322: R. Klette, J. Žunić (Eds.), Combinatorial Image Analysis. XII, 760 pages. 2004.

Vol. 3321: M.J. Maher (Ed.), Advances in Computer Science - ASIAN 2004. XII, 510 pages. 2004.

Vol. 3320: K.-M. Liew, H. Shen, S. See, W. Cai (Eds.), Parallel and Distributed Computing: Applications and Technologies. XXIV, 891 pages. 2004.

Vol. 3319: D. Amyot, A.W. Williams (Eds.), Telecommunications and beyond: Modeling and Analysis of Reactive, Distributed, and Real-Time Systems. XII, 301 pages. 2005.

Vol. 3318: E. Eskin, C. Workman (Eds.), Regulatory Genomics. VIII, 115 pages. 2005. (Subseries LNBI).

Vol. 3317: M. Domaratzki, A. Okhotin, K. Salomaa, S. Yu (Eds.), Implementation and Application of Automata. XII, 336 pages. 2005.

Vol. 3316: N.R. Pal, N.K. Kasabov, R.K. Mudi, S. Pal, S.K. Parui (Eds.), Neural Information Processing. XXX, 1368 pages. 2004.

Vol. 3315: C. Lemaître, C.A. Reyes, J.A. González (Eds.), Advances in Artificial Intelligence – IBERAMIA 2004. XX, 987 pages. 2004. (Subseries LNAI).

Vol. 3314: J. Zhang, J.-H. He, Y. Fu (Eds.), Computational and Information Science. XXIV, 1259 pages. 2004.

Vol. 3313: C. Castelluccia, H. Hartenstein, C. Paar, D. Westhoff (Eds.), Security in Ad-hoc and Sensor Networks. VIII, 231 pages. 2004.

Vol. 3312: A.J. Hu, A.K. Martin (Eds.), Formal Methods in Computer-Aided Design. XI, 445 pages. 2004.

Vol. 3311: V. Roca, F. Rousseau (Eds.), Interactive Multimedia and Next Generation Networks. XIII, 287 pages. 2004.

Vol. 3310: U.K. Wiil (Ed.), Computer Music Modeling and Retrieval. XI, 371 pages. 2005.

Vol. 3309: C.-H. Chi, K.-Y. Lam (Eds.), Content Computing. XII, 510 pages. 2004.

Vol. 3308: J. Davies, W. Schulte, M. Barnett (Eds.), Formal Methods and Software Engineering. XIII, 500 pages. 2004.

Vol. 3307: C. Bussler, S.-k. Hong, W. Jun, R. Kaschek, D.. Kinshuk, S. Krishnaswamy, S.W. Loke, D. Oberle, D. Richards, A. Sharma, Y. Sure, B. Thalheim (Eds.), Web Information Systems – WISE 2004 Workshops. XV, 277 pages. 2004.

Vol. 3306: X. Zhou, S. Su, M.P. Papazoglou, M.E. Orlowska, K.G. Jeffery (Eds.), Web Information Systems – WISE 2004. XVII, 745 pages. 2004.

Vol. 3305: P.M.A. Sloot, B. Chopard, A.G. Hoekstra (Eds.), Cellular Automata. XV, 883 pages. 2004.

Vol. 3303: J.A. López, E. Benfenati, W. Dubitzky (Eds.), Knowledge Exploration in Life Science Informatics. X, 249 pages. 2004. (Subseries LNAI).

Vol. 3302: W.-N. Chin (Ed.), Programming Languages and Systems. XIII, 453 pages. 2004.

Vol. 3300: L. Bertossi, A. Hunter, T. Schaub (Eds.), Inconsistency Tolerance. VII, 295 pages. 2005.

Vol. 3299: F. Wang (Ed.), Automated Technology for Verification and Analysis. XII, 506 pages. 2004.

Vol. 3298: S.A. McIlraith, D. Plexousakis, F. van Harmelen (Eds.), The Semantic Web – ISWC 2004. XXI, 841 pages. 2004.

Vol. 3296: L. Bougé, V.K. Prasanna (Eds.), High Performance Computing - HiPC 2004. XXV, 530 pages. 2004.

Vol. 3295: P. Markopoulos, B. Eggen, E. Aarts, J.L. Crowley (Eds.), Ambient Intelligence. XIII, 388 pages. 2004.

Vol. 3294: C.N. Dean, R.T. Boute (Eds.), Teaching Formal Methods. X, 249 pages. 2004.

Vol. 3293: C.-H. Chi, M. van Steen, C. Wills (Eds.), Web Content Caching and Distribution. IX, 283 pages. 2004.

Vol. 3292: R. Meersman, Z. Tari, A. Corsaro (Eds.), On the Move to Meaningful Internet Systems 2004: OTM 2004 Workshops. XXIII, 885 pages. 2004.

Vol. 3291: R. Meersman, Z. Tari (Eds.), On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE, Part II. XXV, 824 pages. 2004.

Vol. 3290: R. Meersman, Z. Tari (Eds.), On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE, Part I. XXV, 823 pages. 2004.

Vol. 3289: S. Wang, K. Tanaka, S. Zhou, T.W. Ling, J. Guan, D. Yang, F. Grandi, E. Mangina, I.-Y. Song, H.C. Mayr (Eds.), Conceptual Modeling for Advanced Application Domains. XXII, 692 pages. 2004.

Vol. 3288: P. Atzeni, W. Chu, H. Lu, S. Zhou, T.W. Ling (Eds.), Conceptual Modeling – ER 2004. XXI, 869 pages. 2004.

Preface

This book contains a selection of refereed papers presented at the 1st Workshop on Machine Learning for Multimodal Interaction (MLMI 2004), held at the "Centre du Parc," Martigny, Switzerland, during June 21–23, 2004. The workshop was organized and sponsored jointly by three European projects,

- AMI, Augmented Multiparty Interaction, http://www.amiproject.org
- PASCAL, Pattern Analysis, Statistical Modeling and Computational Learning, http://www.pascal-network.org
- M4, Multi-modal Meeting Manager, http://www.m4project.org

as well as the Swiss National Centre of Competence in Research (NCCR):

- IM2: Interactive Multimodal Information Management, http://www.im2.ch

MLMI 2004 was thus sponsored by the European Commission and the Swiss National Science Foundation.

Given the multiple links between the above projects and several related research areas, it was decided to organize a joint workshop bringing together researchers from the different communities working around the common theme of advanced machine learning algorithms for processing and structuring multimodal human interaction in meetings. The motivation for creating such a forum, which could be perceived as a number of papers from different research disciplines, evolved from a real need that arose from these projects and the strong motivation of their partners for such a multidisciplinary workshop. This assessment was indeed confirmed by the success of this first MLMI workshop, which attracted more than 200 participants.

The conference program featured invited talks, full papers (subject to careful peer review, by at least three reviewers), and posters (accepted on the basis of abstracts) covering a wide range of areas related to machine learning applied to multimodal interaction—and more specifically to multimodal meeting processing, as addressed by the M4, AMI and IM2 projects. These areas included:

- human-human communication modeling
- speech and visual processing
- multimodal processing, fusion and fission
- multimodal dialog modeling
- human-human interaction modeling
- multimodal data structuring and presentation
- multimedia indexing and retrieval
- meeting structure analysis
- meeting summarizing
- multimodal meeting annotation
- machine learning applied to the above

Out of the submitted full papers, about 60% were accepted for publication in this volume, after the authors were invited to take review comments and conference feedback into account.

In this book, and following the structure of the workshop, the papers were divided into the following sections:

1. HCI and Applications
2. Structuring and Interaction
3. Multimodal Processing
4. Speech Processing
5. Dialogue Management
6. Vision and Emotion

In the spirit of MLMI 2004 and its associated projects, all the oral presentations were recorded, and synchronized with additional material (such as presentation slides) and are now available, with search facilities, at: **http://mmm.idiap.ch/mlmi04/**

Based on the success of MLMI 2004, a series of MLMI workshop is now being planned, with the goal of involving a larger community, as well as a larger number of European projects working in similar or related areas. MLMI 2005 will be organized by the University of Edinburgh and held on 11–13 July 2005, also in collaboration with the NIST (US National Institute of Standards and Technology), while MLMI 2006 will probably be held in the US, probably in conjunction with a NIST evaluation.

Finally, we take this opportunity to thank our Program Committee members for an excellent job, as well as the sponsoring projects and funding agencies. We also thank all our administrative support, especially Nancy Robyr who played a key role in the management and organization of the workshop, as well as in the follow-up of all the details resulting in this book.

December 2004

Samy Bengio
Hervé Bourlard

Organization

General Chairs

Samy Bengio	IDIAP Research Institute, Switzerland
Hervé Bourlard	IDIAP Research Institute and EPFL, Switzerland

Program Committee

Jean Carletta	University of Edinburgh, UK
Daniel Gatica-Perez	IDIAP Research Institute, Switzerland
Phil Green	University of Sheffield, UK
Hynek Hermansky	IDIAP Research Institute, Switzerland
Jan Larsen	Technical University of Denmark
Nelson Morgan	ICSI, Berkeley, USA
Erkki Oja	Helsinki University of Technology, Finland
Barbara Peskin	ICSI, Berkeley, USA
Thierry Pun	University of Geneva, Switzerland
Steve Renals	University of Edinburgh, UK
John Shawe-Taylor	University of Southampton, UK
Jean-Philippe Thiran	EPFL Lausanne, Switzerland
Luc Van Gool	ETHZ Zurich, Switzerland
Pierre Wellner	IDIAP Research Institute, Switzerland
Steve Whittaker	University of Sheffield, UK

Sponsoring Projects and Institutions

Projects:

- Augmented Multiparty Interaction (AMI), http://www.amiproject.org
- Pattern Analysis, Statistical Modeling and Computational Learning (PASCAL), http://www.pascal-network.org
- Multi-modal Meeting Manager (M4), http://www.m4project.org
- Interactive Multimodal Information Management (IM2), http://www.im2.ch

Institutions:

- European Commission
- Swiss National Science Foundation, through the National Centres of Competence in Research (NCCR) program

Table of Contents

MLMI 2004

III Multimodal Processing

IV Speech Processing

Accessing Multimodal Meeting Data: Systems, Problems and Possibilities

Simon Tucker and Steve Whittaker

Department of Information Studies, University of Sheffield,
Regent Court, 211 Portobello Street, Sheffield, S1 4DP, UK
{s.tucker,s.whittaker}@shef.ac.uk

Abstract. As the amount of multimodal meetings data being recorded increases, so does the need for sophisticated mechanisms for accessing this data. This process is complicated by the different informational needs of users, as well as the range of data collected from meetings. This paper examines the current state of the art in meeting browsers. We examine both systems specifically designed for browsing multimodal meetings data and those designed to browse data collected from different environments, for example broadcast news and lectures. As a result of this analysis, we highlight potential directions for future research - semantic access, filtered presentation, limited display environments, browser evaluation and user requirements capture.

1 Introduction

Several large-scale projects (e.g. [1,2]) have examined the collection, analysis and browsing of multimodal meeting data. Here we provide an overview of browsing tools, where we refer to any post-hoc examination of meetings data (e.g. searching a meeting transcript or reviewing a particular discourse) as *browsing*. As a result of our analysis, we are also in a position to highlight potential areas of research for future meeting browsers.

Despite being an emerging field, there are a large number of browsers described in the literature, and therefore the first stage of summarising the field was to determine a suitable browser taxonomy. The scheme used in tthis paper is to classify browsers according to their focus of navigation or attention. The taxonomy is described in more detail in Section 2 and is summarised in Table 1.

The structure of this papers is as follows. We begin by discussing how meeting browsers are classified and continue by describing each browser, according to its classification. A summary of all the browsers is then given, as a result of which we highlight directions for future research.

2 A Meeting Browser Taxonomy

Since browsing of meetings data is still an emerging field, the classification system used here is necessarily preliminary, but achieves a segregation of the range of browsers described in the literature. Browsers are classified primarily by their *focus*,

S. Bengio and H. Bourlard (Eds.): MLMI 2004, LNCS 3361, pp. 1 – 11, 2005.

and secondarily by properties unique to that focus. The focus of a browser is defined to be either the main device for navigating the data, or the primary mode of presenting the meeting data to the user.

Table 1. Overview of taxonomy of meeting browsers and typical indexing elements used in each class

PERCEPTUAL	SEMANTIC
Audio	*Artefacts*
• Speaker Turns • Pause Detection • Emphasis • User determined markings	• Presented Slides • Agenda Items • Whiteboard Annotations • Notes - both personally and privately taken notes. • Documents discussed during the meeting.
• Video	• Derived Data
• Keyframes • Participant Behaviour	• ASR Transcript • Names Entities • Mode of Discourse • Emotion

Given this definition, and the range of data collected from meetings, three classes of browsers immediately present themselves. Firstly, there are browsers whose focus is largely *audio*, including both audio presentation [3] and navigation via audio [4]. Secondly, there are browsers whose focus is largely *video*; again, including both video presentation systems [5] and those where video is used for navigation [6]. The third class of browsers are focused on *artefacts* of the meetings. Meeting artefacts may be notes made during the meeting, slides presented, whiteboard annotations or documents examined in the meeting.

A fourth class of browser accounts for browsers whose focus is on *derived data forms*. Since analysis of meeting data is largely made on the nature and structure of conversation, this final class is largely concerned with browsing discourse. In this class are browsers whose focus is the automatic speech recognition (ASR) transcript and its properties [7], and those which focus on the temporal structure of discourse between participants [8].

This taxonomy is shown in Table 1 and each of the following sections describe each browser class in detail, in the order in which they were presented above. We refer to audio and video indices as *perceptual* since they focus on low-level analysis of the data. Artefacts and derived indices are referred to as *semantic* since they perform a higher-level analysis of the raw data.

2.1 Audio Browsers

This section discusses browsers whose main focus is *audio*. We separate these browsers into two main subcategories. The first subcategory consists of audio browsers with detailed visual indices; the second category is audio browsers with limited, or no visual feedback.

Both Kimber *et al.* [9] and Hindus and Schmandt [10] describe a meeting browser whose primary means of navigation is via a visual index generated from speaker segmentation. The view presented to the listener is of participant involvement in the meeting - users are able to navigate to each speaker segment and can also navigate between neighbouring speaker segments.

Degen *et al.* [3] describe an indexed audio browser designed for visually reviewing recordings made with a personal tape recorder. The tape recorders allow users to mark salient points whilst recording, the marked recordings then being digitised for review on a computer. The computer interface affords several methods of browsing the recordings. Firstly, users can randomly access any part of the recording, and can also navigate using the markings they made during the recording phase. The visual representation of the recording is of amplitude against time, displayed as a vector or colour plot. Users can also zoom in and out of this display and also have the ability to speed up playback (see the discussion surrounding SpeechSkimmer below).

A key element to these browsers is that the visual representations allow users to immediately see the structure of a meeting. This view, however, is dependent on the browsing environment allowing visual representations to be made. There are situations and devices which do not allow for this visual feedback, so that 'pure' audio browsing requires a substantially different interface.

SpeechSkimmer [4] is a system for interactive 'skimming' of recorded speech. Skimming is defined as system-controlled playback of samples of original audio. A four level skimming system is implemented, each level compressing the speech further, whilst attempting to retain salient content. The first level is unprocessed playback, the second shortens pauses, whilst the third level plays back only speech which follows significant pauses. The final level uses an emphasis detector to select salient segments of the speech to present to the listener. On top of these skimming levels is a mechanism which allows the playback speed to be altered whilst maintaining the pitch of the speaker. In this way the playback speed can be increased without a significant loss in comprehension. It should also be noted that the interface allows users to skim backwards in a recording - in this mode short segments of speech are played forwards but in reverse order.

Roy and Schmandt [11] describe a portable news reader implemented in a small, Walkman style device. The interface was designed iteratively in software, before being transferred to the hardware device. The resulting interface allowed listeners to playback a news report and to also navigate through the report using pre-defined jump locations, computed from an analysis of pause lengths in the audio. In the course of designing the device it was noted that users preferred simpler, more controlled interfaces, preferring manual skims via jumping rather than having software controlled skims. The device also implements a form of speed-up similar to that described above, with users able to select from three different playback speeds.

Because of their nature, audio browsers are largely implemented in hardware devices and so can be argued to be distinct from meeting browsers making use of multimodal data. It has been seen, however, that these browsers have overcome the

limitations of just audio and are able to provide means of browsing audio using computed indices and speed-up techniques. As a complement to this, the following section describes browsers whose primary focus is video.

2.2 Video Browsers

The following class of browsers focus on *video*. Note that whilst each of these browsers have audio and video components, the main component for presentation or navigation in each case is video.

Foote *et al.* [5] describe a simple video browser with two primary modes of navigation. Firstly the user has the ability randomly access any section of the meeting, or to jump between index points which are precomputed from properties of the audio and video. The same indexing, when converted to a continuous 'confidence' measure can also be used to control the playback speed. For example, the playback speed could be related to gesture recognition so that portions of the meeting with significant gestures are played at different speeds, and index marks are made according to these significant gestures.

Girgensohn *et al.* [6] describe video interfaces centred around the use of *keyframes*. Keyframes are static images which have been automatically selected from continuous video according to some heuristic. In the video browsing system, keyframes are chosen according to an importance score, depending on the rarity and duration of each shot. Frames are then sized according to their importance (so that keyframes of higher importance are larger) and are placed linearly on the page. The resulting interface is then similar to a comic book or Japanese Manga drawings. This method can be used to produce a single summary of a full meeting and the user can playback salient portions of the meeting by selecting keyframes, or by choosing a point on a horizontal time line.

A more complex video focused meeting browser is described by Lee *et al.* [13]. A novelty for this system is that it does not require a dedicated meeting room; instead, capture is performed by a single device, encompassing a camera which captures a panoramic video of the meeting and four microphones to record audio. A real-time interface allows meeting participants to examine audio and video during the meeting, as well as making notes during the course of the meeting. The meeting is then archived and processed in preparation for browsing.

The browsing interface has a large number of navigational options. Central to the interface is the video screen, showing both the panorama and a close-up of the currently speaking participant. Users can navigate via a number of indexes, including representations of speaker transitions and visual and audio activity. There is also the opportunity to review an automatically produced transcript of the meeting, and to navigate the meeting via this transcript. A final option for navigating the meeting is a set of automatically generated keyframes. The interface also allows the user to review any notes made during the meeting and to examine any artefacts produced from the meeting.

We note that the number of browsers in this class is relatively small, mainly because video is largely supplemented with other browsing devices and is rarely solely used as a means of navigation. Furthermore, often meeting data does not contain the salient visual events that are useful for video browsing. The browsers described above, however, have shown that there is potential for making use of video as a means of browsing meeting data, although its value is not yet determined [12].