HIFI-AV: An Audio-visual Corpus for Spoken Language Human-Machine Dialogue Research in Spanish

Fernando Fernández-Martínez[1], Juan Manuel Lucas-Cuesta[1], Roberto Barra Chicote[1], Javier Ferreiros[1], Javier Macías-Guarasa[2]

[1]Universidad Politécnica de Madrid,
E.T.S.I. de Telecomunicación, Ciudad Universitaria s/n, 28040, Madrid, Spain,
[2]Universidad de Alcalá, Escuela Politécnica Superior,
Carretera Madrid-Barcelona, Km 33,600, 28871, Alcalá de Henares (Madrid), Spain
[1]{ffm,juanmak,barra,jfl}@die.upm.es, [2]macias@depeca.uah.es

Abstract

In this paper, we describe a new multi-purpose audio-visual database on the context of speech interfaces for controlling household electronic devices. The database comprises speech and video recordings of 19 speakers interacting with a HIFI audio box by means of a spoken dialogue system. Dialogue management is based on Bayesian Networks and the system is provided with contextual information handling strategies. Each speaker was requested to fulfil different sets of specific goals following predefined scenarios, according to both different complexity levels and degrees of freedom or initiative allowed to the user. Due to a careful design and its size, the recorded database allows comprehensive studies on speech recognition, speech understanding, dialogue modeling and management, microphone array based speech processing, and both speech and video-based acoustic source localisation. The database has been labelled for quality and efficiency studies on dialogue performance. The whole database has been validated through both objective and subjective tests.

1. Introduction

This paper describes the generalities of the HIFI-AV corpus. The central idea of the corpus is providing relevant audio-visual data to a broad range of different research areas related to speech and video processing in intelligent environments, specifically addressing a spoken dialogue task to control a HIFI audio equipment. In this sense, this database differs from others such as the AV16.3 corpus (Lathoud et al., 2004) (not dealing with the actual speech content), the CHIL audiovisual corpus (Mostefa et al., 2008) and the AMI project corpus (Carletta, 2007) (both containing audiovisual information of lectures and natural meetings).

Summarising, the HIFI-AV corpus was designed to fulfil the following objectives:

- Allow the evaluation of speech understanding and spoken dialogue modules in a home device control scenario, more specifically for controlling a HIFI audio system.
- Allow the evaluation of microphone array based speech processing tasks.
- Allow the evaluation and fine tuning of the multichannel audio acquisition systems used in the EDECÁN project (EDECAN, 2006) demonstration room at the Speech Technology Group site in the Universidad Politécnica de Madrid, shown in figure 1.
- Allow the performance evaluation of the acoustic modules: mainly speaker localisation and tracking, recognition and beam-forming.
- Allow the performance evaluation of the artificial vision modules: mainly speaker localisation, tracking and identification.
- Allow the evaluation of audio-visual sensor fusion techniques, mainly oriented to multimodal speaker localisation, tracking and identification.

2. General description

HIFI-AV is a spontaneous speech database composed of 190 human-machine spoken dialogues, uttered by 19 Speakers (12 males and 7 females). The actual speech length is around 115 minutes and every speaker speaks for around 7 minutes on average.

The application domain of the HIFI-AV corpus is controlling a HIFI system (Sharp model CDC410) using the voice (handling playback of Cd's, tapes, use of the radio, recording operations, etc.).

2.1. The Spoken Dialogue System

The dialogue manager of our system is based on Bayesian networks and it is provided with a set of domain independent dialogue strategies for handling contextual information (e.g. the dialogue history). These strategies provide the ability to deal with dialogue phenomena such as: ellipsis, anaphora or deixis.

As an alternative to classical dialogue solutions (e.g. finite state automata or FSMs, script based or plan based systems, and so on) Bayesian Networks based dialogue approaches (Fernández-Martínez, 2008) allow a greater flexibility and naturalness thanks to a more convenient definition of dialogue as the interaction with an inference system. Regarding our application of BNs to dialogue modeling and management we can highlight:

- **The BNs we have used have been automatically obtained from training data** (Fernández-Martínez et al., 2009). Automatic learning algorithms favour portability and scalability across domains.

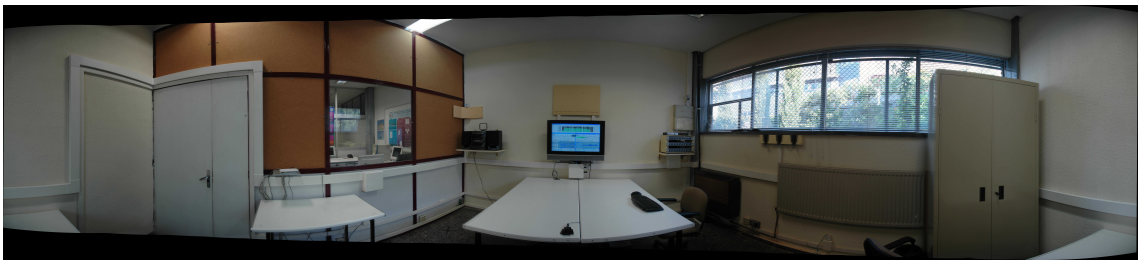

Figure 1: Panoramic view of the recording room.

- The BNs based inference system enables **a better identification of the dialogue goals** (i.e. actions or activities that the system can perform) **from the concepts provided by the user** (i.e. semantic information) and **consistently with the dialogue context**.
- BNs allow an **analysis of congruence** between the dialogue goals that the system thinks the user has requested and all data collected during the interaction. Thanks to this analysis, the system can **determine the dialogue flow** and react according to the logic of the application domain (e.g. doing the work required by the user or requesting him the information necessary to do so). In particular, it is possible to automatically detect **which concepts are necessary** (available or not), **erroneous** or **optional** regarding the inferred goals. The DM makes the decision on how to continue the dialogue using all the available information. In this way, dialogue can be directed towards the production of messages requesting the missing items, clarifying the erroneous and ignoring the optional ones. This helps to fulfil the dialogue goals in a more agile and efficient way avoiding unnecessarily lengthy dialogues.
- The BNs enable **a true mixed initiative dialogue modelling** that allows the response of the system to be flexible as there is no predefined goal or data sequence which the user has to follow. The user is free to decide the set of dialogue goals which he wants the system to offer him. This flexibility is double-edged since, besides allowing the user to decide the goals at the beginning of the interaction, the system also lets the user to jump or switch to different goals, even without having completed the previous ones. Furthermore, the user can respond with more data than requested in a question or even with data regarding a goal different to the target goal decided by the system.
- Thanks to the **negotiation process** between the user and the system, based on the Forward and Backward inference procedures (Meng et al., 2003), the system is capable of responding to complex issues (e.g. when the user provides less information than is strictly necessary to meet the required dialogue goals) and **to assist or guide the user** towards achieving his dialogue goals driving the dialogue in an efficient manner, minimising the number of questions or queries and making maximum use of available information in the context of dialogue.

2.2. On the use of contextual information

In relation to the possible use of contextual information for dialogue management:

- Some dialogue phenomena (i.e. ellipsis, anaphora and deixis) can mean a **loss of crucial information**. That information loss may hinder or difficult the proper outcome of the dialogue. However, we have solved that problem through the proper use of the dialogue context information.
- Thanks to the BNs based dialogue modeling, and through the negotiation process mentioned above, the dialogue system is able to resolve situations where the user provides an incomplete or inaccurate information in relation to the inferred goals from the ongoing utterance corresponding to the intention of the user. Nonetheless, we have provided appropriate **disambiguation procedures** for the immediate retrieval of information from **the dialogue context**. These procedures allow to exploit the contextual information to improve the robustness of dialogue and producing coherent answers from:
  - the history of the ongoing dialogue ($\sim$short term),
  - the state of the system,
  - the specific knowledge of the application domain,
  - and the history of dialogue ($\sim$long term).
- Regarding the latter case, we would like to point out the evidence attenuation mechanism that we have included for keeping the dialogue history permanently updated. As a result of that mechanism, the contents of the history are constantly renewed removing the information that becomes too old.

The convenience and appropriateness of the designed strategies for the use of contextual information have already been proved. (Fernández-Martínez et al., 2008) presents an evaluation of the developed prototype with real users aimed at measuring their true significance.

2.3. Evaluation scenarios

For the recording sessions, a list of 45 scenarios was prepared. Every scenario addressed different goals to be fulfilled by the users, according to the capabilities of the HIFI system being used.

Table 1: *Dialogue example showing the interaction of the user with the system.*

Turn (U:user; S:system)	**Details**
...	-
U: "Volume"	The user does not specify any "volume" value.
S: "What do you want to do with the volume?"	The system identifies the "volume" value as "missing" and prompts the user about it.
U: "Play track number five"	Now, the user is not interested in modifying the volume.
S: "Track number five is playing, would you like to do something with the volume?"	Though decreasing due to an evidence attenuation procedure, the previous evidence level of the "volume" parameter (recovered from the Ongoing Dialogue History) is still significant enough so that the corresponding goal, e.g. "setting the volume", is positively inferred; consequently the system continues trying to solve it.
U: "Play track number seven"	New evidence decrease.
S: "Track number seven playing, what would you like me to do with the volume?"	Still trying.
U: "Track number nine"	After several turns the evidence level of the "volume" parameter falls below a predefined threshold, therefore the system decides to remove it from the history.
S: "Track number nine now playing"	The system stops prompting the user about the volume since only the "track selection" goal is inferred as active.
...	-

All scenarios were classified according to the following nomenclature:

- 23 **Basic** scenarios, in which the user must pursue one major goal (e.g. *"The user should try to stop the CD currently being played"*).
- 19 **Advanced** scenarios: combining more than one goal, to allow the system to demonstrate advanced interaction capabilities (e.g. *"The user should try to play a particular track without referring to the specific disk the track belongs to"*).
- 3 **Free** scenarios: with no restriction for the interaction except that the user was suggested to mainly focused on one of the three main devices the Hifi system was equipped with (i.e. a Cd player, a tape player and a radio).

Table 1 contains an example of a typical dialogue between the user and the system. More details regarding those capabilities are detailed in (Fernández-Martínez, 2008) and (Fernández-Martínez et al., 2005).
Each speaker was requested to fulfil 3 basic, 6 advanced and 1 free scenarios. To address research tasks on speaker localisation and tracking, users were requested to interact with the system in given positions of the recording room and facing different orientations (for basic and advanced scenarios) or to freely move around the room (in the free scenarios).

3. Recording equipment and setup

The recording process was controlled by an operator in a room next to the recording scenario. The operator controlled the recording application and had the chance to talk to the speaker at any time.
Recording equipment was able to synchronously record 24 audio channels and 3 firewire video camera signals. Figure 2 shows the general architecture of the recording setup.

3.1. Audio Recording Hardware

Every channel's audio is sampled at 48 kHz with 24 bits/sample. The 24 audio channels are devoted to recording audio streams from different microphones/sources (close talking, lapel, linear and harmonically-spaced linear array, table top, output from the system (HIFI and text-to-speech modules) and binaural audio from a Bruel&Kjaer manikin).

3.2. Video recording hardware

In order to maximise the visible area in the demonstration room and to reduce the large amount of data to store or transmit, we have decided to use three low-cost firewire cameras. Each of the cameras is connected to a dedicated GNU/Linux workstation running as a video server. Figure 4 shows some of the images captured by the cameras at resolution 640x480 and 30 frames per second.

4. Database content

The main data and information which has been collected and generated during the recording and the labelling stages are:

- **Audio data**: every utterance has been conveniently labelled for each particular speaker, position, orientation, and scenario. For VAD on the close talk speech

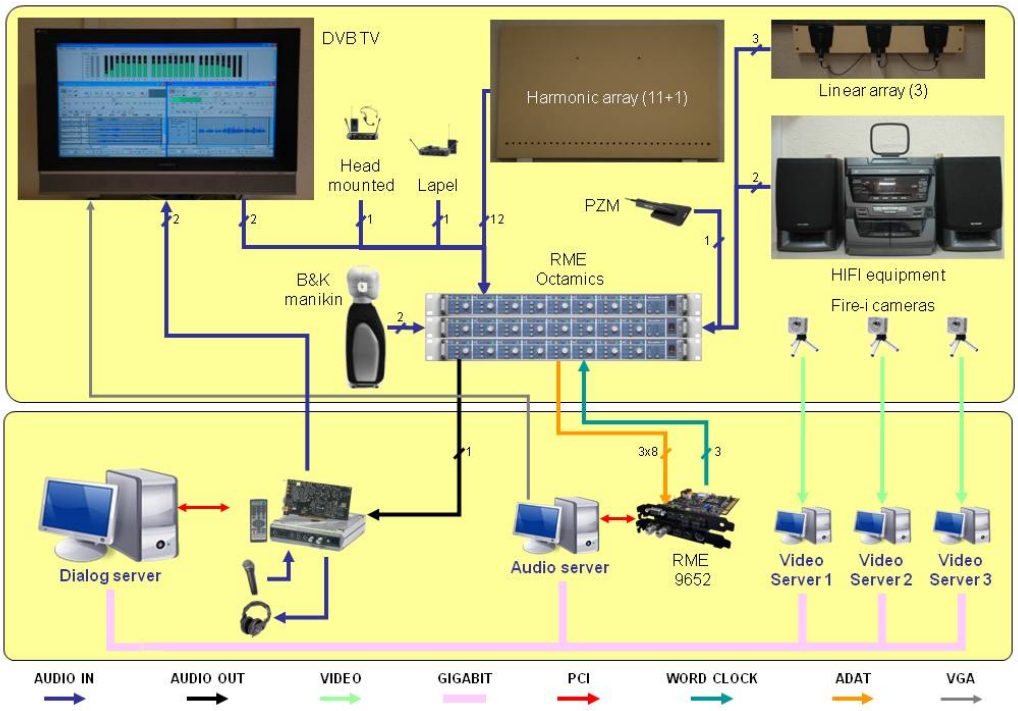

Figure 2: Schematic of the recording setup.

stream the Qualcomm-ICSI-OGI front end has been used (Adami et al., 2002).

- **Video data**: the video corresponding to every dialogue has also been recorded.

- **References and dictionary**: HIFI-AV has been manually transcribed. Additionally, automatic references (i.e. spoken sentences as recognised by the system) are also provided for each audio file. The dictionary in the HIFI-AV task is composed of 419 entries.

- **Geometrical information on speech sources**: speaker "mouth" positions and orientations are provided for evaluating speaker localisation and tracking algorithms, with data following the format used in the CHIL 2006 evaluation campaign for reference files in the Acoustic Person Tracking Task (Stiefelhagen, 2006).

- **Speaker information**: speakers' identity information is also provided.

- **Dialogue related information**: the dialogue management system used in the recordings automatically generated a large amount of information to be able to make precise evaluations on every used module (Fernández-Martínez, 2008). More precisely the following logging information has been generated:

  – **General Log information**: the general log files include all the relevant information of the dialogue process:

    * Preamble: showing general information on the dialogue context (date, time, scenario, etc.).
    * Dialogue turn sections: Providing information on the speech recognition results (including confidence values at word and sentence levels), understanding module information (the set of concepts parsed by the system, with a final confidence figure), and dialogue information (the inferred dialogue goals at any time with a probability figure).
    * Actuator module information: the sequence of actions to be performed by the HIFI.
    * Response generation information: the sequence of answer concepts used to generate a text message which is given to the text to speech (TTS) module.
    * TTS output information.
    * Dialogue context information: information regarding the dialogue status (i.e. all the variables and values describing the current status of every device in the HIFI equipment) and the dialogue history (i.e. all the relevant variables and values kept in the dialogue manager history) before and after each turn.

  – **HIFI equipment status information**: including all the variables+values describing the current status of every device in the HIFI equipment.

  – **Status files of the ongoing dialogue history**: including all the relevant variables+values kept in the ongoing dialogue history.

– **Status files of the dialogue history**: including all the relevant variables+values kept in the dialogue history.

5. Evaluation

Several research tasks have already been approached and evaluated using the HIFI-AV corpus, namely the performance of the speech recognition module, while others are currently being addressed, such as speaker localisation and tracking, or the evaluation of the whole spoken dialogue system through a set of automatically collected metrics related to both dialogue quality and efficiency.

5.1. Speech recognition performance

As we stated in Section 2., HIFI-AV comprises 190 full dialogues, uttered by 19 speakers. The labelling process showed that the database contains 1844 different sentences, with a number of reference words (i.e. correctly labelled words) of 6243.

The availability of labelled data allowed us to evaluate the performance of the automatic speech recognition engine. We have measured the Word Error Rate of the system, measured as the percentage of substitutions, insertions, and deletions on the recognition hypotheses when compared with the labelled references. Table 2 shows the average WER of each scenario type over the different speakers.

Table 2: *Speech recognition performance (measured as Word Error Rate).*

Scenario Type	Number of sentences	Reference words	WER (%)	Confidence interval
Basic	356	1010	28.51	2.78
Advanced	1065	3568	34.14	1.56
Free	423	1668	42.99	2.38
ALL	1844	6243	35.61	1.19

As could be expected, the better performance takes place during the simplest scenarios, in which the number of interactions remains relatively low, and the tasks to be performed are easy to ask for, whereas the WER is worse for more complex scenarios, in which the degree of freedom allowed to the users is greater than in other scenarios.

We also have analysed how the different Spanish accents and regions of origin of the different speakers (namely, Madrid, Aragón, Valencia, and Basque Country), affected the recognition performance. The main results of this analysis are shown on Table 3. The WER results presented in the table have been averaged among the three scenario types.

These results may look a bit surprising, providing that both Aragón and Madrid are regions in which Spanish is the only official language, whereas in Valencia and in the Basque Country there are co-official languages (Valencian, and Basque, respectively). Therefore, the worst results (in terms of speech recognition accuracy) may have been expected to take place during the interactions of those speakers from regions with several official languages, and with quite stronger accents. However, a deeper analysis of the interactions of each speaker showed that the percentage of Out-Of-Vocabulary words (OOV) that the speakers of each region used, was unexpectedly higher in the case of Valencia and, particularly, Aragón (see Table 4), not only in absolute terms (see column 3), but also considering the relative ratio between the number of OOVs, and the total number of reference words (column 4). This was a decisive reason for the lower performance that the system reaches in both cases, in comparison to Madrid and Basque Country.

Table 3: *Speech recognition performance according to dialectal regions.*

Dialectal region	Number of sentences	Reference words	WER (%)	Confidence interval
Madrid	152	513	16.37	3.20
Aragón	513	1927	47.79	2.23
Valencia	770	2617	34.77	1.82
Basque C.	409	1186	25.97	2.50

Table 4: *Distribution of Out-Of-Vocabulary words according to dialectal regions.*

Dialectal region	Number of OOVs	OOV percentage	Ratio OOV / Ref. words (%)
Madrid	7	1.26	1.36
Aragón	267	48.11	13.86
Valencia	232	41.80	8.87
Basque C.	49	8.83	4.13
ALL	555	100.00	8.89

5.2. Evaluation of the Spoken Dialogue System

As many labelling sources are available, we can evaluate the performance of complex spoken dialogue systems using the information of the database. In this regard, a similar study to that presented in (Fernández-Martínez et al., 2008) is being addressed based on different dialogue quality and efficiency metrics that were automatically collected for each scenario by the spoken dialogue system.

5.2.1. Objective Evaluation

As part of that study, we have measured the **percentage of contextual turns** as the fraction of dialogue turns in which some of the strategies are successfully applied. Logically, any piece of information that is essential for the resolution of the dialogue but can not be recovered from the dialogue context must be requested to the user. Therefore, and in connection with the above referenced metric, we have also measured the **percentage of system requests** which is limited by the contextual capabilities of the system. The results presented in Figure 4 for both metrics endorse the valuable role of the contextual information handling strategies regarding the dialogue management. Specifically, we can conclude that more than half of the turns rest on this type of information. In other words, without the contextual capabilities provided, the number of system requests would increase considerably (i.e. would double at least).

Figure 3: Sample images captured by the cameras.

As another interesting issue, we could try to assess the relevance of the contextual information handling strategies in terms of **dialogue efficiency**, a metric related to dialogue fluency which could be measured as the average number of actions executed per turn. On one hand, higher values of contextual turns should be associated with lower values of system requests. On the other, lower values of system requests should mean better turn efficiencies. Dialogue efficiency results have been presented in Table 5 for each type of scenario. We have also included a column reporting the difference between both metrics: contextual and request turns. For instance, by comparing the results corresponding to the Basic and Advanced scenarios, it is clear that there is a strong dependency between efficiency and the number of requests formulated by the system. Advanced and Free scenarios are pretty similar in this regard (i.e. 1.75 and 1.71 respectively), nonetheless it can be observed a slightly worse behaviour for the latter which could be due to the worse performance of the speech recogniser (see Table 2, particularly due to a higher number of OOVs).
Advanced and Free scenarios allow the user to better exploit the contextual information handling strategies, thus resulting in a lower amount of requests and therefore, speeding up the dialogue making it more fluent and flexible (e.g. users are free to make more use of ellipsis).

Table 5: *Dialogue efficiency and its connection with the percentage of contextual turns and system requests.*

Scenario Type	**Dialogue Efficiency** (# actions/turn)	**Contextual turns - System requests** (%)
Basic	1.48	14.22
Advanced	1.75	36.19
Free	1.71	28.88
ALL	1.69	30.61

5.2.2. Subjective Evaluation

As in (Fernández-Martínez et al., 2008), in order to get subjective ratings of the spoken dialogue system we conducted user satisfaction surveys. Every user was requested to rate task or scenario success after each scenario on a 1 (worst) to 5 (best) scale (see Figure 5). An average user satisfaction rate of 4.25 shows the goodness of our Bayesian Networks based dialogue modeling approach (Fernández-Martínez et al., 2005). Subjectively, the system's response has been assessed very positively by users as it can be deduced from the presented results. In particular, the individualised analysis conducted for each type of scenario puts the "free" scenarios as the highest-rated throughout the assessment process. This is undoubtedly a result of particular importance because the complexity of the "free" scenarios is maximum. These are scenarios without any restriction in which the initiative of the user reaches its top.

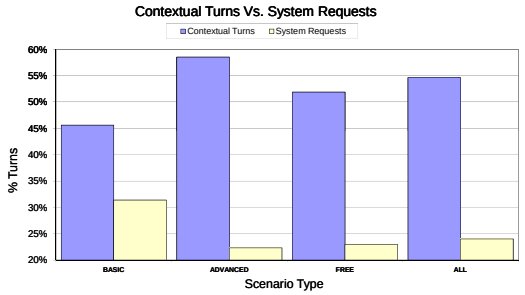

Figure 4: Percentage of contextual turns and system requests.

6. Conclusions and Future Work

In this paper we have presented a new multi-purpose audio-visual database on the context of speech interfaces for controlling household electronic devices (i.e. a Hifi system). The available labelling sources, including different dialogue quality and efficiency metrics automatically collected by the spoken dialogue system, have allowed the evaluation of several performance features. Particularly, as a result of a similar study to that presented in (Fernández-Martínez et al., 2008) it has been proved that a more natural, flexible and robust dialogue is possible thanks to the suggested BN based dialogue modelling approach. This is supported by a good user satisfaction rate and the obtained results for the collected metrics. In this regard, the strategies for handling contextual information have been proved to be essential,

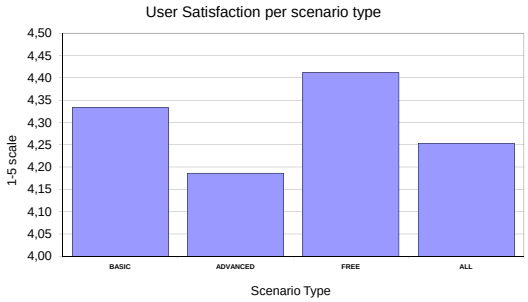

Figure 5: Average user satisfaction for each type of scenario.

saving a significant amount of system's requests, and thus speeding up the dialogue.

Audio and Video-based speaker localisation and tracking are also being addressed with strategies such as the ones described in (Castro, 2007) and (Marron et al., 2009). Our main interest in the near future is the evaluation of audio-visual sensor fusion techniques on this task.

7. Acknowledgements

This work has been supported by ROBONAUTA (DPI2007-66846-c02-02) and SD-TEAM (TIN2008-06856-C05-03).

8. References

Andre Adami, Lukas Burget, Stephane Dupont, Hari Garudadri, Frantisek Grezl, Hynek Hermansky, Pratibha Jain, Sachin Kajarekar, Nelson Morgan, and Sunil Sivadas. 2002. Qualcomm-icsi-ogi features for asr. In *Proc. ICSLP*, pages 4–7.

Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

Carlos Castro. 2007. Speaker localization techniques in reverberant acoustic environments, master thesis.

EDECAN. 2006. EDECAN project. `http://www.edecan.es/en/index.html`.

F. Fernández-Martínez, J. Ferreiros, V. Sama, J.M. Montero, R. San Segundo, J. Macas, and R. Garca. 2005. Speech interface for controlling an hi-fi audio system based on a bayesian belief networks approach for dialog modeling. In *Eurospeech*, pages 3421–3424, Lisboa (Portugal).

F. Fernández-Martínez, J. Blazquez, J. Ferreiros, R. Barra, J. Macias-Guarasa, and J.M. Lucas-Cuesta. 2008. Evaluation of a spoken dialogue system for controlling a hifi audio system. pages 137 –140, dec.

F. Fernández-Martínez, J. Ferreiros, R. Cordoba, J. M. Montero, R. San-Segundo, and J. M. Pardo. 2009. A bayesian networks approach for dialog modeling: The fusion bn. In *ICASSP '09: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4789–4792, Washington, DC, USA. IEEE Computer Society.

F. Fernández-Martínez. 2008. *Análisis, diseño y aplicación de modelos de diálogo flexibles, contextuales y dinámicos basados en Redes Bayesianas*. Ph.D. thesis, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain, Octubre.

Guillaume Lathoud, Jean-Marc Odobez, and Daniel Gatica-Perez. 2004. AV16.3: an audio-Visual Corpus for Speaker Localization and Tracking. Idiap-RR Idiap-RR-28-2004, IDIAP, Martigny, Switzerland. Published in "Proceedings of the 2004 MLMI Workshop".

M. Marron, D. Pizarro, J.C. Garcia, A. Marcos, R. Jalvo, and M. Mazo. 2009. Multi-agent 3d tracking in intelligent spaces with a single extended particle filter. pages 305 –310, aug.

H.M. Meng, C. Wai, and R. Pieraccini. 2003. The use of belief networks for mixed-initiative dialog modeling. *Speech and Audio Processing, IEEE Transactions on*, 11(6):757 – 773, nov.

D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Cristoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet. 2008. The chil audiovisual corpus for lecture and meeting analysis inside smart rooms. *International Journal of Language Resources and Evaluation. Special issue on Multimodal Corpora for Modelling Human Multimodal Behavior*, 41(3-4):389–407. ISSN 1574-020X.

Rainer et al. Stiefelhagen. 2006. The clear 2006 evaluation.