SD-TEAM: Interactive Learning, Self-Evaluation and Multimodal Technologies for Multidomain Spoken Dialog Systems

María Inés Torres[1], *Eduardo Lleida*[2], *Emilio Sanchis*[3], *Ricardo de Córdoba*[4], *Javier Macías-Guarasa*[5]

[1]Pattern Recognitiona & Speech Technologies Group. University of the Basque Country. Spain
[2]Communication Technologies Group. University of Zaragoza, Spain
[3]Pattern Recognition and Artifitial Intelligence Group. Polytechnic University of Valencia, Spain
[4]Speech Technology Group. Polytechnic University of Madrid, Spain
[5]Intelligent spaces and Transport Group. University of Alcalá de Henares, Spain

http://www:sd-team.es

Abstract

Speech technology currently supports the development of dialogue systems that function in limited domains for which they were trained and in conditions for which they were designed, that is, specific acoustic conditions, speakers etc. The international scientific community has made significant efforts in exploring methods for adaptation to different acoustic contexts, tasks and types of user. However, further work is needed to produce multimodal spoken dialogue systems capable of exploiting interactivity to learn online in order to improve their performance.

The goal is to produce flexible and dynamic multimodal, interactive systems based on spoken communication, capable of detecting automatically their operating conditions and especially of learning from user interactions and experience through evaluating their own performance. Such ?living? systems will evolve continuously and without supervision until user satisfaction is achieved. Special attention will be paid to those groups of users for which adaptation and personalisation is essential: amongst others, people with disabilities which lead to communication difficulties (hearing loss, dysfluent speech, ...), mobility problems and non-native users.

In this context, the SD-TEAM Project aims to advance the development of technologies for interactive learning and evaluation. In addition, it will develop flexible distributed architectures that allow synergistic interaction between processing modules from a variety of dialogue systems designed for distinct tasks, user groups, acoustic conditions, etc. These technologies will be demonstrated via multimodal dialogue systems to access to services from home and to access to unstructured information, based on the multi-domain systems developed in the previous project TIN2005-08660-C04.

1. Introduction

Technological advances have resulted in many devices with which we interact and that have transformed our everyday life. Within speech technology, the topic of the current proposal, of special relevance is the study of robust, intuitive and easy to use spoken dialogue systems as a means of human-computer interaction. In recent decades, significant advances in automatic speech recognition have been achieved, which has in turn led to an increased demand for voice-based interactive systems which are able to handle more complex tasks. Such is the case for spoken dialogue systems, which are currently in a very early stage of development and still not integrated into commercial products in widespread use within society. Similar remarks can be applied to other spoken language processing systems such as automatic translation, summarisation and information extraction.

These advances can and should allow for the entry of large swathes of the population into the *Information Society*, amongst which one can mention the disabled or those who, due to their age, have missed out on contact with this type of technology. However, in spite of efforts to the contrary, rapid technological growth has led to the marketing of products which sometimes do not meet expectations of people. Consequently, more robust and versatile systems are needed that go much further than current artefacts, which can be characterised as *fixed* as a result of their initial training. What is needed are systems that learn throughout their lifetime through interactions with users, with error-detection mechanisms and adaptation to novel situations and environments, which would allow increases in robustness, usability and maintainability. Such systems will have advanced sensory capabilities to permit the acquisition of relevant information from their environment, and which allow them to plan their behaviour in response to user demands [1]. Another questionable notion is the idea of the computer being the centre of the human-machine interaction. Rather, the human should be the centre of attention. Under this new view, the role of the computer is limited to that of an observer of interpersonal interactions, contributing helpful information such as speech transcription, automatic translation, topic and speaker changes and the like [2]. On the other hand, it is useful to think of the user as another system module, a view which introduces new scientific challenges [3]. Some of these involve the use of user feedback to reduce system error, adaptation and evolution to different users, environmental conditions and tasks. At the same time, the recognition that human interactions are inherently multimodal should permit improvements in the general usability of systems and hence their acceptance amongst end users.

Significant scientific advances will be required to handle the technological challenges implied these new ideas. Tackling these problems is the purpose of the SD-TEAM project described here. SD-TEAM is a natural successor of the EDECAN project [4] developed by the proposed group of scientists. The EDECAN project attempted to go beyond classical dialogue systems to allow them to perform robustly in the face of changes in acoustic conditions due both to the environment and the user, to minimise the effort required in system redesign for

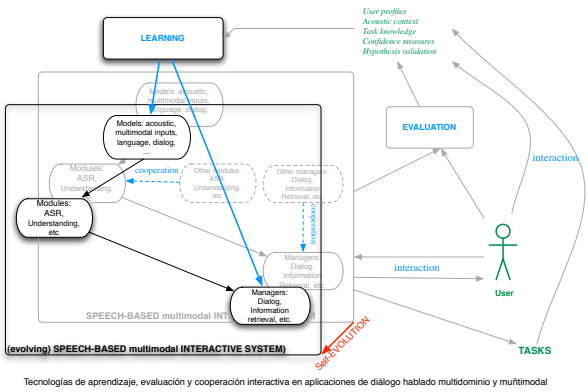

Figure 1: *Illustration of the goals of the SD-TEAM Project: interactive, voice-based systems capable of unsupervised evolution via the development of technologies which support cooperation between different modules, operating environments, evaluation techniques and interactive learning.*

new applications, and to customise the system for individual users. SD-TEAM project attempts to go further in converting *closed* dialogue systems into dynamic and flexible interactive systems with a clear capacity for self-evolution via learning, evaluation and cooperation during normal usage. These new *living* systems will be based on information and experience gained in earlier projects, but equally will be capable of adaptation, interaction with users and other systems, and, most important of all, able to learn online in order to reconfigure themselves . For this, a flexible distributed architecture (being developed from that proposed in the EDECAN project) permits cooperation amongst different subsystems for understanding, recognition, handling of dialogue gestures etc, developed for different environments, as well as natural multimodal user interaction, both of which are based on autonomous operation the will allow for constant, unsupervised evolution (see Fig. 1). Evolution is based on objective functions aimed both at increasing system performance and user satisfaction. A good summary of the overall goal of the SD-TEAM project can be found in objective 2 (*Smarter machines, better services*) of the Information and Communication Technologies theme of the 7th Framework Programme of the EU: *research must lead to systems more capable of detecting what is going on in their environment and learning, reasoning and interacting with people in a more natural way [...] Instead of forcing users to understand how machines function, it should be the systems which learn to work and interact better with us*. This aspiration is even more relevant for disabled users. Paradoxically, it is precisely this section of the population who would benefit most of such technological advances, but for whom the use of such systems is out of reach if their communication abilities are reduced by disabilities such as dysfluencies or hearing impairment. At the very least, these systems should evolve to a level of performance useful for this sector of society, as is proposed in the SD-TEAM project.

2. Objectives

Over the years, dialogue systems have been developed which *know what they know and don?t know what they don?t know*. Such systems are static and force the user to adapt to their needs. On the other hand, systems exist which are increasingly *organic*, capable of learning, *growing*, reconfiguring and self-repairing, leading to greater robustness in the face of a wide range of operating conditions. Consequently, it is essential to study in depth the technological and methodological elements that will allow dialogue systems to approach the organic model, in order to produce artefacts which automatically evolve to a level where they guarantee user satisfaction. Amongst potential users are those groups for which a static system simply would not function. These include users whose disability leads to communication problems (dysfluent speech, hearing loss, ...), those with limited mobility, and non-native users. Project SD-TEAM will improve the capabilities of voiced-based interactive multimodal systems using advances in the development of self-evaluation and interactive learning technologies.

2.1. Scientific objectives

1. Concerning interactive learning:
   - To explore efficient mechanisms for interactive learning: methods of incremental adaptive learning capable of using information contained in each interaction with the user to allow automatic and dynamic evolution of the system itself.
   - To explore efficient techniques for environment detection: speaker localization, identification of the user, acoustic conditions and semantic domain, etc. To develop robust methods for adaptation to variable environments, application changes and user customisation (multilingualism, non-native speech, dysfluent speech, emotional speech, etc).
2. Concerning interaction and cooperation between modules, systems and the user:
   - To develop approaches which consider user interaction as an additional module or element of the system that co-operates with the rest of the system. To explore methods that permit direct use of information from the interaction in order to improve the performance of highly-complex systems with imperfect behaviour (e.g. at the level of audio classification, transcription, etc.) by validating or updating hypotheses, asking new questions, etc.
   - To develop likewise approaches that can handle the interaction between systems of differing complexity, communicating and cooperating to achieve given objectives.
   - To develop distributed architectures that integrate user interaction, inter-module interaction and adaptive learning as well as multimodal information processing.
3. Concerning learning through self-evaulation:
   - To explore methods for automatic extraction of information relevant to the interaction process: signal segmentation, event identification/classification, information extraction, etc.
   - To define measures of system self-evaluation, confidence and system performance at all levels and incorporation of these into the learning process.
4. Horizontal objectives related to multimodality and e-inclusion:

- To explore efficient methods for integrated multimodal input processing: voice, touch screen, video etc. for understanding and/or learning. To develop methods and techniques for multimodal information generation: voice, multilingual text, sign language, handwriting, etc.
- To use techniques developed in SD-TEAM to improve access for disabled people.

2.2. Technical objectives

- To improve significantly the technological capabilities of the consortium (large-vocabulary recognisers, dialogue systems with wider application domains, processing of out-of-vocabulary items, coupled and decoupled architectures, integration of user interaction, etc.). Development of a flexible platform based on distributed architectures for the integration of modules involved in a voice-based multimodal interactive system capable of learning and dynamic evolution using information obtained during interaction with the user.
- To construct a prototype demonstrator to illustrate the scientific results and technology of the project. Dynamic dialogue systems applied to multiple domains starting from prototypes from the EDECAN project will be integrated with interactive systems to access unstructured speech (audio from TV programmes). Accessibility for disabled people is foreseen.

3. Partners and local objectives

The groups in this proposal have a long history of working together in the development of automatic speech recognition and spoken dialogue systems. The varied origins and specialisms of each group make collaboration especially attractive. The Communication Technology group from the University of Zaragoza (UZ) come from the field of signal and communication theory, fundamental for the development of robustness in recognition systems, and having recently introduced speech technology to aid people with disabilities and speech production difficulties. The Pattern Recognition and Artificial Intelligence group from the Polytechnic University of Valencia (UPV) and the Pattern Recognition and Speech Technology group from the University of the Basque Country (EHU) possess essential knowledge concerning model learning, allowing the consortium to deepen the development of methods for learning from examples, language modelling, understanding and dialogue as well as, in the case of EHU, the development of limited-domain translation systems. The group from UPV also work in natural language processing, in particular in morphosyntactic labelling, word-sense disambiguation, named-entity recognition and their application in information extraction and question answering. The Speech Technology group from the Polytechnic University of Madrid (UPM) have extensive experience in the design and evaluation of person-machine dialogue systems based on speech technology, with powerful systems for speech recognition and understanding and high-quality text to voice conversion. Finally, the Intelligent Spaces and Transport group at the University of Alcal (UAH) have wide experience in the positioning of intelligent mobile agents with multiple cameras as well as multimicrophone speech processing. The objective of the SD-TEAM project is to improve the capabilities of advanced voice-based multimodal interactive systems through the development of interactive learning and self-evaluation technologies. The project will also develop flexible distributed architectures that allow subsystems created for different environments to work together. SD-TEAM will lead to significant advances in the technological capacity of the consortium which will be demonstrated through prototypes for multiple-domain dialogue systems and in systems for audio information extraction, with special attention to accessibility for disabled users with problems of communication or mobility. To fulfil its objectives, SD-TEAM relies on collaborative work with the network of project partners. Each team has specialised expertise (as detailed in the lab descriptions) which complements that of other partners and will be needed to develop the distinct modules that make up complex voice-based interactive systems (and in which each group brings experience and solutions from different perpectives). This diversity in viewpoints will enrich proposed solutions to scientific problems which arise during the project. All the subprojects share the general scientific and technological objectives of the project, alongside those involved in the design and implementation of the distributed architecture and in the constructions and evaluation of demonstrators. In addition, each subproject has its own objectives:

3.0.1. SD-TEAM-EHU: objectives

- incorporation of multilingual input-output (Spanish-Basque-English) into voice-based interactive systems.
- development of robust language identification systems based on multilingual acoustic models
- translation model inference using finite-state transducers. Integration of systems for ASR and automatic translation from voice in limited domains. Maninly into a Spanish/Basque context.
- research into the development of hierarchical, cooperative language models and their application in recognition, understanding, dialogue management and automatic translation
- incorporation of the user in the design of dialogue systems; management of multimodality in search and models of understanding
- significant growth in the technical capacity of the group in recognition, dialogue and voice translation

3.0.2. SD-TEAM-UZ:objectives

- Robustness in adverse acoustic environments.
- Description of acoustic scenario: acoustic segmentation, identification of acoustic events and acoustic environment, speaker identification, i.e. audio indexing.
- Lexical Robustness: detection and learning lexical features from the speaker, which is fundamental for impaired or non-native speaking users.
- Development of algorithms for obtaining robust acoustic confidence measures for self evaluation and assessment purposes.
- Cooperation of multimodal information sources (audio-visual, tactile, etc.) to increase the robustness in spoken dialogue systems.
- Help to handicaped users: developing support systems for oral communication interfaces based on voice inputs/outputs.

- Architectures for distributed dialogue systems: making advances to manage errors, multilingual interaction, self and inter-pair evaluation and assessment of every system module and the continuous and autonomous adaptation

3.0.3. SD-TEAM-UPV: objectives

- Codification of the multimodal input information and its integration in the system models.
- Detection and classification of different semantic contexts.
- Interaction with the user: personalization.
- Dynamic learning in dialog systems: adaptation of the lexicon, the semantics and the dialog manager.
- Learning with samples that contain errors, rejecting or incorporating them to the models.
- Dynamic learning of the dialog manager by means of its use. Definition of success parameters to be used in the self-evaluation and self-learning process.
- Development of methodologies for accessing unstructured information (voice or text) using speech (Information retrieval and Question Answering). Detection of Named Entities and keywords. Development of interaction methods with the user.
- Study of cooperation techniques between different system modules: homogeneous (e.g., between different speech recognizers) or heterogeneous (e.g., between a recognizer and a dialog manager).

3.0.4. SD-TEAM-UPM:objectives

- Inclusion of multimodal inputs (speech and tactile screen). Inclusion of multimodal outputs: speech (including emotional speech), screen, and sign language translation.
- Technologies for environment detection in adaptive learning: speaker / user identification and language identification.
- Dialog management with dynamic information: dynamic generation of LM, vocabularies, etc.
- Automatic learning of dialog management (BNs architecture from labeled dialogs)
- Technologies for the collection of relevant information for the interaction: audio indexing, topic recognition, emotion recognition, especially the detection of anger.
- Personalization of dialog management: the system learns in an unsupervised way the user preferences, proposing solutions to the user wishes with the smallest possible number of interactions.

3.0.5. SD-TEAM-UAH:objectives

- Robust systems for multimodal detection, localization, tracking and pose estimation of multiple users in intelligent environments: using microphone arrays and multiple cameras, and applying audio-visual sensor fusion strategies.
- Speech enhancement techniques: based in binaural and microphone array speech processing and adaptation strategies based in simulation of reverberant environments.
- Audio-visual sensor fusion strategies for speaker identification and emotional state classification tasks.

4. Acknowledgements

The authors would like to thank the Spanish Innovation and Science Minister for funding this project under grant TIN-2008-068-C05.

5. References

[1] IEEE Transactions on Systems, Man and Cybernetics, Vol. 35, No. 1,Jan, 2005. Special Issue on Ambient Intelligence.

[2] CHIL project: ?Computers in the Human Interaction Loop?. Proyecto integrado VI programa marco de la UE: http://chil.server.de/servlet/is/101

[3] Multimodal Interaction in Pattern Recognition and Computer Vision: Project funded by the Spanish Science Minister under special program Consolider-Ingenio 2007. http://miprcv.iti.es/

[4] EDECAN: Sistema de diálogo multidominio con adaptación al contexto acústico y de aplicación. Project funded by the Spanish Science Minister: http://www.edecan.es