Building Spoken Language Systems

I. Trancoso, D. Caseiro, N. Mamede, J. Neto, L. Oliveira, A. Serralheiro, C. Viana
L2F, INESC ID Lisboa, R. Alves Redol 9, 1000-029 Lisboa, Portugal
Phone: +351-213100268, Fax: +351-213145843, e-mail: isabel.trancoso@inesc-id.pt

***Abstract* — The Spoken Language Systems Lab was formally created in 2001, bringing together the expertise of several research groups that shared a common goal: to bridge the gap between natural spoken language and the underlying semantic information, focusing on European Portuguese. This paper describes our efforts towards this long-term goal, starting by the two main areas of activity: semantic processing of multimedia contents, and multimodal dialogue systems. These strongly interdisciplinary areas integrate several core technologies developed in the lab, namely speech recognition and synthesis. This interdisciplinarity is also strongly patent in the the lab's most recent activities, such as speech-to-speech translation.**

I. Introduction

The long-term goal of the Spoken Language Systems Lab is to bridge the gap between natural spoken language and the underlying semantic information, focusing on European Portuguese. The lab was formally created in 2001, bringing together the expertise of several research groups that shared a common goal, and whose activity in speech processing dated from the mid eighties. Being a strong interdisciplinary group is one of the main strengths of this group of over 20 researchers, with backgrounds ranging from signal processing to computer science and linguistics.

The focus on European Portuguese made the lab invest stronger efforts in developing core technologies that are more language dependent. In fact, the language independent activity on speech coding, that was one of the main strengths of this group back in the eighties, is now very reduced, though new challenges are present in this area, namely in terms of Voice Over IP.

Throughout this paper, several references will be made to two core modules: AUDIMUS [1] and DIXI+ [2]. The first one is a hybrid speech recognizer that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of multilayer perceptrons (MLPs). The same recognizer is used for different complexity tasks, based on a common structure with different components: the HMM/MLP hybrid model, the lexicon and the language model. The dynamic decoder builds the search space as the composition of the corresponding three Weighted Finite-State Transducers (WFSTs).

The second module is a concatenative-based text-to-speech synthesizer, based on Festival. This framework supports several voices and two different types of unit: fixed length units (such as diphones), and variable length units. This latter data-driven approach is more suitable, by adequate design of the corpus, to a limited domain application.

This paper starts with the description of the two main areas of activity: semantic processing of multimedia contents, and multimodal dialogue systems. These strongly interdisciplinary areas integrate several core technologies developed in the lab, namely in terms of speech recognition and synthesis, and will therefore be described in comparatively more detail. The paper continues with two recent lines of activity: computer enhanced human-to-human communication and speech-to-speech translation. The final sections are devoted to activities in which we have been investing for a long time, although with more limited resources: e-inclusion, and porting to other varieties of Portuguese. The presentation will be illustrated by demos in several application domains: automatic captioning of broadcast news, domotics, and e-inclusion**.**

II. Semantic Processing of Multimedia Contents

The involvement in this type of activity was motivated by the participation in the ALERT European project [3] (2000-2002). The goal of this media watch project was to continuously monitor a TV channel, searching the news programs for stories that match the profile of a given user. The system can be tuned to automatically detect the start and end of a broadcast news show from the occurrence of the program jingles. Once the start is detected, the system automatically records, transcribes, indexes, summarizes and stores the program stories. The system then searches in all the user profiles for the ones that match the detected topics. If there is a match, an alert email is send to that user indicating the occurrence and location of one or more stories about the selected topics. This alert message enables a user to follow the links to the corresponding video clips.

A. System Architecture

The system includes three main blocks: a Capture block, responsible for capturing the broadcast news shows, a Processing block, responsible for generating all the relevant markup information, and a Service block, responsible for the user and database management interface. A simple scheme of semaphores is used to control the overall process.

In the Capture block, we have access to the list of programs and their time schedules (expected starting and ending time). The capture is done via a normal TV capture board (Pinnacle PCTV Pro), generating an MPEG-1 file. In the Processing block, the audio stream extracted from this file is processed through several stages that successively segment, transcribe and index it, compiling the resulting information into an XML file. The Service block deals with the user interface,

---

I.Trancoso, D. Caseiro, N. Mamede, J. Neto, and L. Oliveira are also with Instituto Superior Técnico, Lisboa. A. Serralheiro is also with Academia Militar. C. Viana is an invited researcher of L2F, affiliated to CLUL.

through a set of web pages and database management of user profiles and programs. Each time a new program is processed, the program database is updated and the system matches the story topics with the user profiles, generating a list of email alerts. The system has been implemented on a network of 3 ordinary PCs and has been running daily since May 2002, having been expanded to cover several channels.

RTP, the Portuguese user partner in the ALERT project, is interested on indexing every story according to the thematic thesaurus that is daily used for manual indexation. This thesaurus follows rules that are generally adopted within EBU (European Broadcast Union), and has a hierarchical structure with 22 thematic areas in the first level. Although each thematic area is subdivided into (up to) 9 lower levels, we implemented only 3 in our system, due to the sparsity of topic-labeled training data in the lower levels. This structure, complemented with geographic (places) and onomastic (names of persons and organizations) descriptors, makes our topic definition. A user can also specify a free text string. The user profile results from a combination of boolean operators on these topics. The use of this hierarchic structure makes the Portuguese system significantly different from the one developed by the other partners, and from the type of work involved in the TREC SDR Track.

Each alert email message (Fig. 1) includes the title of the news broadcast, the date, the news duration time and a percentage score indicating how well the story matched the chosen profile (e.g. **Telejornal + 2003-03-29 + 00:07:40 + 65%).** It also includes the title of the story, a link to a URL where one could find the corresponding RealVideo stream, and the topic categories that were matched.

Figure 1. Example of an alert e-mail.

B. Processing Block

Figure 2 shows a functional diagram of the Processing block, based on successive stages that transform the MPEG-1 file into an XML file containing the orthographic transcription and associated metadata.

The first stage extracts the audio file from the MPEG-1 stream downsampling it to 16kHz. In the future, we plan to integrate image processing techniques in different stages, namely for segmentation. The resulting file is then processed by a ***Jingle Detector***, which tries to select the program's start and ending time, and cut the commercial breaks

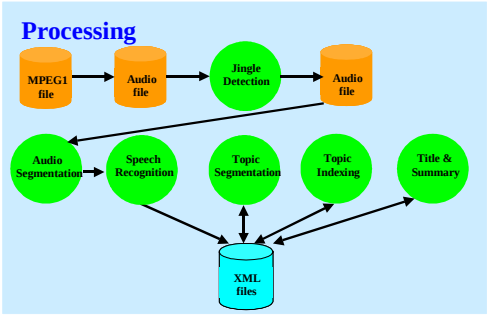

Figure 2. Functional diagram of the Processing block.

The new audio file is then fed through an ***Audio Segmentation*** module, whose output is a set of transcribable segments that are homogeneous in terms of background conditions, speaker gender and speaker cluster. The module achieved a miss ratio of 14% in terms of segment boundary detection and an insertion ratio of 18%. The error rate for tagging speech segments as non-speech is 4.4%. Background classification turned out to be a hard task, namely because there are many segments in the training material with music plus noise. The gender misclassification rate was 7.1%. Our speaker clustering method achieves a cluster purity greater than 97%. The anchor identification module achieved a deletion rate of 9%, and an insertion rate of 2%.

Each transcribable segment is then processed by the ***Speech Recognition*** module, based on AUDIMUS. For this application, its vocabulary is limited to around 60k words associated to a multi-pronunciation lexicon. The corresponding OOV rate is 1.4%. We use an interpolated 4-gram language model combining a model created from newspaper texts with a model created from broadcast news transcriptions (45 hours). For planned speech, with no background noise, high bandwidth channel, and native speech, the system achieved a word error rate (WER) of 14.8%. The average WER for all conditions was 26.5%. The fact that our audio segmentation module frequently subdivides sentences has a negative impact in terms of language model errors near incorrect sentence boundaries.

The ***Topic Segmentation*** module is based on a very simple heuristic that assumes that all new stories start with a segment spoken by the speaker identified as anchor. This module is currently being modified, as this simple heuristic does not deal with filler segments (headlines introduced by the anchor and not distinguished from the following story also by the anchor), and with multiple anchors.

For each story the ***Topic Indexing*** module generates a classification, according to the thesaurus. Since most stories are manually classified into more than one topic, multiple topics can be assigned to a story. The detection takes into account the confidence score of each decoded word, achieving a correctness of 95.3% for the higher-level topic.

Finally a module for generating a ***Summary*** is applied to each story, based on the first sentence of the anchor. This simple strategy proved satisfactory in spite of being dependent on the story segmentation process. At the end of this Processing block, the XML file contains all the relevant metadata that we were able to extract.

C. Field Trials

During the first months of operation the system was tested by a group of volunteers that reported several problems. Most were related to lack of training data for some particular topics, others were related with thesaurus limitations, vocabulary limitations and missing anchor detection.

In spite of these problems, we would like to stress the fact that having a fully operational system is a must for being able to address user needs in the future in this type of service. Our small panel of users was unanimous in finding such type of system very interesting and useful, specially since they were often too busy to watch the full broadcast and got the opportunity of watching only the most interesting parts.

III. MULTIMODAL DIALOGUE SYSTEMS

The involvement in this line of activity started in the scope of the national project DIGA and progressed more recently through the participation in the Interactive Home of the Future, at *Museu das Comunicações*. The home was built as a mean of demonstrating new telecommunications / multimedia technologies to the generic public. Our spoken dialogue system [4] gives access to a virtual butler (*Ambrósio*) that controls several devices in the master bedroom. The system combines automatic speech recognition, natural language understanding, speech synthesis and a visual interface based on a realistic animated face.

A. System Architecture

Our system is based on three main blocks (see Figure 3). Two of them are responsible for the interfaces with the user and the centralized system for controlling devices. The other block is responsible for the dialogue management.

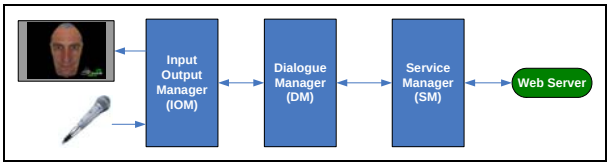

Fig. 3. Block diagram of the dialogue system.

The user interface is based on a wireless microphone available on the bedroom and the TV set, where *Ambrósio* is visualized and answers our requests through the TV speakers. The device controller interface is based on a web server, giving access to any device in the home. The system is quite generic, being able to control different domains: home environment, remote database access (weather information, bus information, stock market information), email access, etc. Access to the system may be done via microphone, telephone, GSM, PDA and web. In order to satisfy these generic goals, both interface blocks create a domain-independent level to the Dialogue Manager (DM), who does not know which type of device made the request (i.e., spoken request or click by the pen in a PDA application). The Input and Output Manager (IOM) creates this level by sending a unique XML format for each request, independent of the source, and the same principle is applied at the Service Manager (SM).

The four main blocks of the IOM are AUDIMUS, DIXI+, FACE and TM (Text Manager). The FACE block is a Java 3D implementation of an animated face with a set of visemes for Portuguese phonemes and a set of emotions and head movements. The TM block transforms the web server requests into XML format in order to access the DM. The DM determines the action requested by the user and asks the SM to execute it. The domain representation and the information extracted from the user request are handled through frames, composed by slots and rules. Slots define domain data relationships, and rules define the system behavior. The DM is based on a communication hub of the Galaxy framework that interconnects the IOM and SM with a set of blocks: Language Interpretation, Interpretation Manager, Discourse Context, Behavioral Agent, and Generation Manager. The SM provides the DM with the necessary interface with a set of heterogeneous devices grouped by domains, represented by XML structures that define the set of available services, the implementation code that will execute the services, the identification of the domain objects and the templates of the system domain utterances.

B. User Interaction

In the environment of the Interactive Home of the Future, we found different types of user: expert users, museum guides that have learnt how to use the system, and generic visitors, that want to try the system, sometimes in groups of over 30 young students. The use of a wireless microphone gives a large degree of freedom. In order to cope with this open microphone situation, we implemented a keyword detection module that detects the word *Ambrósio*, only sending commands to the recognizer that start with this keyword.

The state of the system can be visualized through the butler's face. We implemented six states, with corresponding head movements: silence (no one talking or closed mic); voice activity detection, keyword detection, service request, service execution, and system reply.

There are many different ways of accessing the system (for instance, for turning on the lights). Thus, it was necessary to consider many alternatives for the different actions and to map these new meanings to the same action in the Language Interpreter module. For AUDIMUS, this implied the modification of the vocabulary, lexicon and language model. The latter resulted from interpolating a generic model extracted from newspaper texts with a very small model generated from sentences associated to the domain.

IV. COMPUTER ENHANCED HUMAN-TO-HUMAN COMMUNICATION

The know-how gained in the media watch project enabled us to deal with related topics in the area that is now commonly known as computer enhanced human-to-human communication. Examples of applications in this area are the meeting browser and the lecture tracker. A major difference of these domains relative to broadcast news is the dominant presence of spontaneous speech. But many other issues are involved such as speaker diarization (marking the times corresponding to speaker changes, and providing speaker identification information), inclusion of other modalities, multimodal structuring, summarization and information retrieval, etc. This motivates a worldwide effort to automatically produce what is denoted by "rich transcriptions", including not only transcripts but also metadata. The accompanying metadata is useful in increasing the readability of the transcripts, not only in terms of marking speaker changes, but also in terms of disfluency recognition (marking verbal fillers such as filled pauses, discourse markers, and verbal edits) and semantic unit segmentation. The goal of the national project LECTRA, recently started, is to automatically provide enriched transcriptions of lectures.

V. SPEECH-TO-SPEECH MACHINE TRANSLATION

Automatic speech recognition has traditionally benefited of major advances thanks to the use of statistical techniques that have been also extended for the development of speech-to-speech machine translation (SSMT) systems. Under this framework, an SSMT system is built from sets of examples that must be sufficiently large and representative, usually consisting of parallel text and speech data of the source language. Our most recent research efforts have been focusing on integrated SSMT techniques, i.e., techniques that perform speech decoding and translation in the same process. We have explored the fact that AUDIMUS is able to combine finite-state models through on-the-fly composition, including models that are inferred from parallel corpora by means of different grammatical inference methods. We consider our experimental results to be very promising, specially taking into account that these are the first reported results of a speech-to-speech translation task (Portuguese-to-English) involving our language that we know of.

VI. E-INCLUSION

The application of spoken language technologies to develop aids for people with special needs has always been the concern of the group researchers. Indeed, the very first application of DIXI+'s predecessor was for children suffering from cerebral palsy. This application evolved to become *Eugénio* (the word prediction genie), a software agent operating in the Microsoft Windows environment, that suggests words which complete the text being edited, in order to speed up the writing process. The system uses statistical language models and can cope with different forms of keyboard scan and can be used together with any SAPI compliant text-to-speech synthesizer. The software is available for public download from the lab's website.

Our most recent e-inclusion project is IPSOM, a national project dealing with digital spoken books and their interfaces for visually impaired users. Our major task there has been the alignment between the text and the audio files. This has been successfully achieved by some modifications made to our WFST-based decoder. With these modifications, a 2-hour long spoken book may be aligned in a single step in much less than real time. In the scope of the above mentioned LECTRA project, the lecture transcriptions aligned the audio/video data may also potentially be of use to deaf students.

VII. OTHER VARIETIES OF PORTUGUESE

Porting our core technologies and applications to other varieties of Portuguese has been an old goal for which unfortunately we could never find the necessary funding and resources. This is regrettable, since the lab has internationally recognized work in the areas of non-native speech processing and spoken language identification, which could be important in this context. Preliminary results are promising, namely in terms of grapheme-to-phone conversion for Brazilian Portuguese.

VIII. CONCLUSIONS

Throughout this paper, no references have been made at all to an area that is very morose and time consuming – building linguistic resources for European Portuguese. This has been a very strong investment of the lab, and despite the many hours of spoken language corpora, we are still far from having resources of dimension comparable to those existing for English. Nevertheless, we believe we have played an important role, namely in building the corpora used to train the recognizers nowadays operating in the telecom domain.

Given the scope of the conference, this necessarily brief summary did not cover either natural language processing tools whose development, although more recent in our lab than spoken language tools, is nevertheless very promising. We believe that these tools will be instrumental in progressing from speech recognition to speech understanding, and from speech synthesis from text to synthesis from concepts. In short, to bridge the gap between natural spoken language and the underlying semantic information

REFERENCES

[1] H. Meinedo, D. Caseiro, J. Neto and I. Trancoso, "AUDIMUS.MEDIA a Broadcast News speech recognition system for the European Portuguese language", in *Proc. PROPOR'03*, Faro, Portugal, June 2003.

[2] S. Paulo and L. Oliveira, “Multilevel Annotation of Speech Signals Using Weighted Finite State Transducers”, in *Proc. 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, USA, 2002.

[3] I. Trancoso, J. Neto, H. Meinedo, and R. Amaral, “Evaluation of an alert system for selective dissemination of broadcast news”, in *Proc. Eurospeech´2003*, Genéve, Switzerland, Sept. 2003

[4] J. Neto, N. Mamede, R. Cassaca, and L. Oliveira, “The development of a multi-purpose Spoken Dialogue System”, in *Proc. Eurospeech´2003*, Genéve, Switzerland, Sept. 2003.