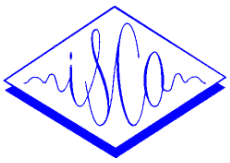

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13–16, 2003

AN OVERVIEW ON EUROPEAN PROJECTS RELATED TO SPONTANEOUS SPEECH RECOGNITION

Gerhard Rigoll

Institute for Human-Machine Communication
Munich University of Technology
D-80290 Munich, Germany
rigoll@ei.tum.de

ABSTRACT

The purpose of this paper is the presentation of a survey on activities in spontaneous speech recognition in Europe over the past 10 years. A brief historical review on the development of this topic in Europe is presented first, and then a few technical issues are addressed, that distinguish research projects on spontaneous speech recognition from other research activities in speech. Some of the major projects that have been carried out in the last years are then briefly presented and finally, the paper concludes with an outlook on the future of research in spontaneous speech recognition in Europe.

1. INTRODUCTION

During the last 30 years, substantial progress has been made in the area of speech processing. Spontaneous speech recognition is a special research area within the wide field of automatic speech recognition, that emerged as a special discipline about 10 years ago. It is quite interesting to note that actually speech is in almost any situation spontaneous, except in the scenario for read speech. It is therefore almost amazing to see that it took so long to establish this special research area. One explanation for that could be the fact that – especially since the end of the 1980´s – speech research has been very much driven by the databases that became available and very often, the release of a specific database lead to the creation of a new research branch in speech processing.
A closer look at the applications and situations that are typical for spontaneous speech recognition problems yield the following possible scenarios:

- dialogue situations
- negotiation situations
- general conversational situations
- voice controlled devices
- smart environments
- broadcast news
- interviews
- meetings

Work on dialogue systems is already going on for a long time, basically since the first commercial speech recognition systems became available at the beginning of the 80´s. It is therefore not amazing that, the first European projects within the 1st EU Framework Program that lasted from 1984 to 1987, were already heavily concerned with dialogue systems and it is thus reasonable to say that research work on spontaneous speech has been already pursued since the early 1980´s. But at that time, nobody called this research work “spontaneous speech recognition”. The spontaneous speech aspects were just implicitly given in the applications investigated during that time. Many of those earlier applications were telephone-based dialogue systems with limited vocabulary, often only involving numbers and a few keywords. One major difference to today’s projects on spontaneous speech recognition is however the fact that the former systems were mainly dealing with isolated word input while the more actual systems have continuous speech input, consisting mostly of entire spontaneously spoken sentences. Many of the typical problems in spontaneous speech, such as hesitations and interjections are of course more apparent in entire sentences rather than single words, and this may be another reason why the topic “spontaneous speech recognition” has not been created during that time but instead later, when the recognition of spontaneously spoken sentences made those problems much more obvious.
As mentioned before, mainly during the 1990´s, many databases have been created and many of them triggered a specific research area in speech processing. For all the above mentioned scenarios, databases have been

eventually created and thus the official research topic "spontaneous speech recognition" emerged in the beginning of the 1990's.

2. RESEARCH ISSUES FOR SPONTANEOUS SPEECH RECOGNITION

Before the most important projects are discussed, this brief section addresses the major research issues that come with the investigation of spontaneous speech. It is then possible to identify these issues partially in the various projects discussed thereafter. The list of research issues for spontaneous speech recognition includes topics such as

- databases for spontaneous speech
- pronunciation modeling
- modeling of hesitations and interjections
- speaking rate variability

It will be interesting to see that for the most projects listed in the following section, the first issue - namely the database aspect – is mainly addressed in these projects and all other listed research issues are rarely handled, except in a few projects that will be mentioned later. It is thus reasonable to say that in most cases, the spontaneous speech recognition aspects have been handled "implicitly" by collecting appropriate data and hoping to automatically learn and capture the effects of spontaneous speech in the statistical models resulting from training with that database. In contrast to that, a more explicit approach would have been the detailed investigation of those effects and the derivation of specific methods to solve them.

3. DESCRIPTION OF THE MAJOR EUROPEAN RESEARCH PROJECTS ON SPONTANEOUS SPEECH RECOGNITION

In the following sub-sections, some of the major European projects are listed, that either address the topic of spontaneous speech in their project description or at least cover applications and scenarios that implicitly involve spontaneous speech recognition technologies.

3.1. The Verbmobil Project

Although this was a national project sponsored by the German ministry for research and no European project in that sense that it involved the participation of several European partners, the Verbmobil project was probably among the first projects in Europe that was closely associated with spontaneous speech recognition and heavily contributed to the creation of this topic.
The basic project idea of Verbmobil has been speech-to-speech translation of dialogues between business people. The first phase of the project ran from 1993 to 1996, and the second and final phase from 1997 to 2000. The project has been sponsored with almost 60 Mio. Euro by public funding from the German research ministry and approx. 25 Mio Euro industrial funding, making it the probably largest public single speech project ever funded. In the second phase, the translation of spontaneous speech dialogues in the domains travel planning and hotel reservation have been demonstrated for German-English and German-Japanese translation with vocabularies up to 10,000 words. Many speech research groups were active within this project and some of them were quite engaged in spontaneous speech recognition phenomena. The investigated spontaneous speech recognition phenomena included topics such as e.g. prosodics, containing intonation aspects as well as variations in duration. Other investigations included the treatment of missing prepositions and other general problems related to parsing of spontaneous speech, where e.g. incorrectly recognized function words have been recovered during the parsing process. Additionally to these investigations, some important contributions to spontaneous speech recognition have been made by collecting a large database of spontaneous speech dialogues, reflecting the application scenario of appointment scheduling dialogues.

3.2. The INSPIRE Project

The INSPIRE project (Infotainment Management with Speech Interaction via Remote Microphones and Telephone Interfaces) is a European project aiming to integrate a multilingual, interactive, natural, speech dialogue-based assistant for wireless command and control of home appliances (e.g. consumer electronics). The project's emphasis is on infotainment (information and entertainment) equipment and services, whose complexity makes advanced dialogue techniques necessary. The voice-activated assistant ensures natural access to the appliances supported by it from several distributed points inside the house, and through the public telephone network and the Internet. Inside the house, spoken human-machine communication using wireless wall-mounted microphone arrays pick up the user's speech and self-powered loudspeaker modules convey the assistant's synthesized speech. The INSPIRE system enables the user to initiate natural spoken dialogues and ask for information about the current status of the appliance and/or control it, requesting assistance on its use, etc.
Spontaneous speech recognition aspects are implicitly covered by the dialogue situation of the application, but not very strongly explicitly addressed. Typical dialogues for home appliances control are relatively brief and thus typical spontaneous speech recognition phenomena are not as frequent as in applications were commands are longer and less keyword-oriented.

3.3. The CORETEX Project

The Coretex project aims at improving core speech recognition technologies, which are central to the most important applications involving voice technology, e.g. multimedia information access and automatic services over the telephone network. The first scenario has implications to spontaneous speech recognition in the area of broadcast speech transcription, where spontaneous speech recognition phenomena occur e.g. for the transcription of interviews and conversations. The latter application covers the typical dialogue situation in query and information systems, where dialogues are spontaneous but often contain brief utterances.

Some of the interesting activities in this project were investigations to port large vocabulary broadcast news (BN) recognition systems to spontaneous speech dialogue domains (in particular appointment scheduling and tourist information), where the first type of system is trained mostly on regular speech data and the second type contains all well-known spontaneous speech phenomena. The porting process has been mainly accomplished by acoustic model and language model adaptation to the new domains. It turned out that both measures were successful, where the large vocabulary coverage of the adapted BN language model contributed especially to a low OOV rate and thus resulting very low word error rates in the target domain.

3.4. The IDAS Project

The IDAS (Interactive Telephone-based Directory Assistance Services) project addresses the challenging problem of automating the provision of directory assistance services to the public over the telephone network. Its primary target is to demonstrate the applicability of very large vocabulary speech recognition and speech dialog technologies in the development of cost-effective and user-friendly applications for automated (without the intervention of human operators) and interactive telephone-based directory assistance services. The challenges of this project cover the following issues: Very large vocabularies (several millions entries) with some very similar or even non-discernible entries. Need for a very effective dialog to achieve the (highly required) disambiguation and to narrow down the search space.

The project's contribution to spontaneous speech recognition lies mostly in the design of sophisticated dialogues to keep the required speech input quite short and thus to avoid too many typical spontaneous speech phenomena that would have a tendency to occur more frequently in longer utterances.

3.5. The NESPOLE Project

NEgotiating through SPOken Language in E-commerce) aims at providing a system capable of supporting advanced needs in e-commerce and e-service by resorting to automatic speech-to-speech translation. It does not only address accuracy of translation, but extends investigation to the ability of two humans to communicate ideas, concepts, thoughts and to jointly solve problems. The project's achievements are demonstrated by means of two showcases. The first one supports multilingual negotiations and discussion between a tourist information/service provider and a customer who wants to organize a trip exploring all available possibilities, including travel, accommodation, attractions and recreation, cultural events, dining, etc.. In a second showcase, the tourist scenario is enlarged. In addition, a second, completely different domain is addressed, consisting of a video help-desk for technical support in trouble shooting and repair. Four languages are considered: English, German, French and Italian.

In contrast to other projects, this project does address spontaneous speech recognition issues directly in its major goals. Some of the investigated phenomena contain ill formed speech utterances with hesitations (um, hmm, etc.), repetitions (so I, I, I guess, what I was saying...), false starts (how about we meet on Tue.. um.. on Wednesday...), fragments often containing two or more concepts (... no, Tuesday doesn't work for me...how about...Wednesday morning...Wednesday the twelfth) and noisy recording environments containing e.g. coughs, laughter, telephone rings, door slams, etc..

3.6. The CATCH-2004 Project

The objective of the CATCH-project (Converse in AThens, Cologne and Helsinki) is to develop a multilingual, conversational system with a novel unifying architecture across devices and services. The system will provide pervasive access to multiple applications and sources of information available to citizens from public and private service providers by supporting multiple client devices, and by using multiple input modalities. Client devices are kiosks, telephones (standard and wireless) and smart wireless devices. Applications include access to information over the Internet, travel and city information/services, phone-directories and completion of transactions.

The contributions to spontaneous speech recognition are also here more implicit, and may offer some add-on to other activities in this area by noting that spontaneous speech has been addressed here especially in the context of wireless access to information services, involving a speech data collection of spontaneous speech over wireless channels. Another contribution is the quite

extensive investigation of spontaneous speech involving a relatively strong natural language understanding component, using statistical NLU techniques, making this a "spontaneous speech understanding" activity.

3.7. The ALERT Project

The ALERT system uses advanced speech recognition technology and video processing techniques in order to process large broadcast speech archives and multimedia information resources for the purpose of extracting specific information from such databases and inform selected customers about its contents. The objective of the ALERT project is to develop an intelligent software system that automatically scans multimedia data like TV or radio broadcasts for the presence of specific topics and that alerts users whenever topics of their interest are detected.
As in many typical broadcast news transcription projects, spontaneous speech recognition aspects are covered here implicitly by considering all parts and aspects of broadcast news, thus including also interviews, conversations and other items more oriented towards spontaneous situations. An interesting contribution of this project was the investigation on the effectiveness of audio-visual tools for segmentation in spontaneous speech, where e.g. hesitations and interjections could lead to undesired segment boundary hypotheses which could be resolved by visual information, exploiting the video track of the news.

3.8. The M4 Project

The abbreviation M4 stands for MultiModal Meeting Manager. The basic idea of this project is the analysis and transcription of meetings recorded in so-called smart meeting rooms, equipped with multimodal sensors. Thus, the overall objective of the M4 project is the construction of a demonstration system to enable structuring, browsing and querying of an archive of automatically analyzed meetings.. For each meeting, audio, video, textual, and (possibly) interaction information will be available. Audio information will come from close talking and distant microphones, as well as binaural recordings. Video information will come from multiple cameras. While the video and audio information will form several streams of data generated during the meeting, the textual information, the agenda, discussion papers, text of slides will be pre-generated and can be used to guide the automatic structuring of the meeting. The interaction stream consists of any information that can help in analyzing events within the meeting, for example, mouse tracking from a PC-based presentation or laser pointing information.
Without any doubt, meeting transcription is an area where the aspect of conversational speech is heavily involved, since large parts of meetings consist of conversations between two or more people. Similarly to the ALERT project, the M4 project also offers the potential to tackle spontaneous speech effects by means of multimodal and multi-stream information processing, e.g. by resolving false segmentation boundaries with support of the visual action or the facial expression that has been simultaneously observed in the video channel.

3.9. The ERMIS Project

The ERMIS project (Emotionally Rich Man-machine Intelligent System) is not a typical spontaneous speech recognition project, but deserves some attention and thus is listed as final project in this survey, because it investigates the emotional state of a person during speaking. This emotional state is a typical effect of spontaneous speech, and it is quite obvious that the current emotional state of a person has a strong influence on his current speaking performance in typical spontaneous speech situations, as listed on the first page of this paper. Activities within that project that point into this direction include e.g. the development of a module which performs emotional speech analysis and feature extraction, the development of a module which performs facial expression analysis, and the integration of the extracted features, performing multi-sensory feature analysis and recognition of the user's emotional state.

4. SUMMARY AND CONCLUSION

This paper gave an overview on some European projects in the area of spontaneous speech recognition. Some of these projects do not address this topic directly, but rather more implicitly, by focusing on scenarios that are directly related to spontaneous speech phenomena, mostly dialogue scenarios, but partially also other systems, such as broadcast and meeting transcription, and systems analyzing the emotional state of persons. One can conclude that spontaneous speech recognition has been identified and picked up – at least partially - by the European research community and that it would be desirable to further strengthen the activities in this important research area in Europe in order to keep up with the international standards in this field.

5. REFERENCES

[1] http://verbmobil.dfki.de/
[2] http://inf2.pira.co.uk/factsheets/inform/hlt/inspire.html
[3] http://coretex.itc.it/
[4] http://www.tik.ee.ethz.ch/~idas/
[5] http://nespole.itc.it/
[6] http://www.catch2004.org
[7] http://alert.uni-duisburg.de
[8] http://www.dcs.shef.ac.uk/spandh/projects/m4/
[9] http://www.image.ntua.gr/ermis/