Sun Workstation and SwissNet Platform for Speech Recognition and Speaker Verification over the Telephone*

Andrzej Drygajlo[1], Jean-Luc Cochard[2], Gérard Chollet[2], Olivier Bornet[2], Philippe Renevey[1]

[1] Signal Processing Laboratory, Swiss Federal Institute of Technology Lausanne, CH-1015 Lausanne, Switzerland, Andrzej.Drygajlo@lts.de.epfl.ch
[2] Dalle Molle Institute for Perceptive Artificial Intelligence, CH-1920 Martigny, Switzerland, Jean-Luc.Cochard@idiap.ch

1 Introduction

Voice processing systems make it possible for people to interact with computers using speech, the most natural and widely-distributed human mode of communication. Because voice processing systems will support human-machine interaction in a natural way that requires no special training, these interfaces will eventually make computer-based resources available to many new groups of users, particularly to telephone users. Therefore speech recognition and speaker verification over telephone channels is imperative.

In this paper we describe an applied research project entitled "Automatic Speech Recognition in French on Workstation with SwissNet Connection". This cooperative project involves specialists from two research institutes: the Signal Processing Laboratory (LTS) of the Swiss Federal Institute of Technology Lausanne (EPFL) and the Dalle Molle Institute for Perceptive Artificial Intelligence, Martigny (IDIAP), and three industrial partners: aComm, SunMicrosystems (Switzerland) and the Swiss Telecom PTT. The project is supported by the Commission for Technology and Innovation (CTI, formerly CERS). The goal of the project is to make available basic technologies for automatic speech recognition (ASR) and speaker verification (SV) on multi-processor SunSPARCstation 20 and SwissNet platform to industrial partners and particularly to Swiss industry for Swiss French.

2 Related Research Activities

Nowadays in Europe, the demand for Interactive Voice Response (IVR) services over the telephone is increasing dramatically. Many R&D projects are currently granted by the European Union, in the context of the 4th Framework Program, namely CAVE: "Caller Verification in Banking and Telecommunication" (Telematics Program), and M2VTS: "Multimodal Verification for Teleservices

* Published in *Proceedings of SIWORK'96: Workstations und ihre Anwendungen*, pp. 1–4, Zürich, Switzerland, May 1996

and Security Applications" (ACTS Program). For both projects, IDIAP is a technology supplier for SV components, as well as an active partner in the implementation of IVR services.

In the framework of COST (European COoperation in the field of Scientific and Technical research) allowing the coordination of national research on a European level, Switzerland is granting two actions dedicated to speech processing, namely COST 249: "Continuous Speech Recognition over the Telephone", and COST 250: "Speaker Recognition in Telephony", with a participation of LTS EPFL and IDIAP.

To fully describe the general context of the project presented in this paper, the SpeechDat project: "Speech Databases for Creation of Voice Driven Teleservices" (Telematics Program) has to be mentioned. This project addresses the fields of production, standardization, evaluation and dissemination of spoken language resources recorded over the telephone. Twelve partners, including IDIAP, are in charge of recording data in their own countries, which will cover more than ten languages.

3 Voice Messaging and Voice Response Technology

The main goal of the project presented in this paper is to design and implement workstation oriented voice messaging and voice response demonstrators based on ASR and SV technologies in the area of digital telecommunications for Swiss French.

The basic hardware configuration consists in a multi-processor Sun SPARCstation 20, with 128 MB of RAM, connected to SwissNet (the ISDN network in Switzerland). The required software on the Sun SPARCstation includes Solaris 2.4 or later, SunISDN 1.0.2, XTL 1.1 [3] and a modular speech/speaker recognition system adapted to real time processing.

3.1 Voice Processing Demonstrators

The first planned demonstrator is a voice messaging system. The caller can leave a message to somebody, or consult all the messages sent to him. This application includes ASR and SV components. The caller can also ask the system for some help or decide to quit the system.

The tasks of *posting a message* can be divided in two steps: pronouncing the name of the addressee and recording the message. The system should recognize the given name and ask him for a confirmation. If the caller confirms the recognition result, the recording step can start. Otherwise, a second try is possible or the possibility to send a message to the local operator (secretary) is proposed. For improving the recognition phase, it is possible to spell the name of the addressee.

The *consultation of personnal messages* includes a recognition of the speaker identity, by means of a vocal PIN-code. After the caller has given his personnal code, the system checks the caller's identity by comparing it to the user's reference. If granted, the caller can access his vocal mailbox. If denied, he is rejected.

The second demonstrator is an IVR server, InfoMartigny. The purpose of this server is to provide information on cultural events taking place in the city of Martigny. The system proposes a menu-based dialogue where the user interacts with the server by uttering isolated words. Word-spotting capability is also available which greatly enhances naturalness of the man-machine dialogue.

Thus, at the first level, the valid words are “cinema”, “manifestation” and “concert”, among others. At a second level for “cinema”, there is a choice between all the cinema names. Once a cinema is selected by the speaker, information concerning the movie currently played is spoken. And so on for the other commands. Like in the voice messaging demonstrator the possibility exists to ask for some help or to quit the system.

3.2 Automatic Speech Recognition

At the heart of automatic speech recognition systems lies a set of suitable algorithms for recognition and training. The main successfully implemented technology to recognize speech sounds is based on Hidden Markov Models (HMM) and its allied techniques which can be appropriately modified for different applications. In speech recognition for the project demonstrators the signal from a ISDN channel is processed using a Continuous Density HMM (CDHMM) technique. The approach selected for this project to model command words is a*flexible vocabulary* approach. It offers the potential to build word models for any application vocabulary from a single set of trained phonetic sub-word units. The price to pay for having such a versatile system is twofold. First, the size of HMM models is enlarged by a sub-word units approach. For example, if we consider the extreme values in InfoMartigny, we have the following results:

<table>
<tr><th rowspan="2">Word</th><th rowspan="2">Phonetic Transcription</th><th colspan="2">Model size (kB)</th></tr>
<tr><th>Global</th><th>Concat.</th></tr>
<tr><td>Guide</td><td>/gg/ /ii/ /dd/ [/ee/]</td><td>17</td><td>$4 \times 42 = 168$</td></tr>
<tr><td>Galerie du Manoir</td><td>/gg/ /aa/ /ll/ (/oe/|/eu/|/ee/) /rr/ /ii/ /dd/ /uu/ /mm/ /aa/ /nn/ /ww/ /aa/ /rr/ [/ee/]</td><td>96</td><td>$17 \times 42 = 714$</td></tr>
</table>

The second drawback of a phonetic-based approach is the need of a large database to train, once and for all, a set of speaker-independent and vocabulary independent sub-word models for the chosen language.

The availability of the telephone quality Swiss French Polyphone database provided by the Swiss Telecom PTT is the determining factor of this project enterprise. Another factor is the availability of sufficient computer resources at EPFL to train the statistical phonetic CDHMMs. The Polyphone database [4] contains about 12 GBytes of speech data. The use of computing facilities at EPFL (several multi-processor SunSPARCstations and massively parallel computer CRAY T3D) facilitates this realisation. It allows us to keep to the stated schedule for the adaptation and optimisation of the architecture of the developed models by multiple training as well as testing of the quality of the models. The CDHMM toolkit available at LTS and IDIAP laboratories with suitable adaptations to perform massively parallel processing is used in the training and testing experiments.

Initial tests with the vocabulary of InfoMartigny show that the flexible vocabulary approach is a very promising one if compared to the global modelling approach (see Table 1). The training of the phonetic models has been performed on a subset of the Polyphone database, containing sentence utterances. The test has been performed on a set of 5 000 utterances of InfoMartigny command words, pronounced by 3 male and 3 female speakers of the Polyvar database [4] that contains among others occurrences of the InfoMartigny vocabulary recorded through an analog telephone line.

Word	Recognition rate		Word	Recognition rate	
	Global	Concat.		Global	Concat.
`annulation`	100	100	`manifestation`	100	100
`Casino`	98.99	99.66	`message`	100	100
`cinéma`	100	96.32	`mode d'emploi`	99.66	98.66
`concert`	100	100	`Louis Moret`		100
`Corso`	100	99.66	`musée`	82.71	89.83
`exposition`	100	98.99	`précédent`	95.96	98.65
`Galerie du Manoir`	99.66	98.99	`quitter`	100	99.66
`Gianadda`	99.33	99.66	`suivant`	96.3	100
`guide`	98.99	95.62	average rate	98.3	98.54

Table 1. Comparative results of recognition rates on a small vocabulary IVR service.

3.3 Speaker Verification

This project aims at integrating a SV component that is currently developed in the framework of the projects CAVE and COST 250. The selected approach is based on a parallel processing of a speech utterance by three different algorithms: Dynamic Time Warping (DTW) [1], Sphericity [2], and HMMs . This task, in the context of this project, is clearly a technology transfer from academic research to commercial use.

4 Conclusion

The project started in September 1995, and the work on the demonstrators is still in progress. The two types of demonstrators cover a large range of possible applications of speech technology and this project has to play a key-role in a global process of technology transfer from research centers to industrial partners.

References

1. M. Homayounpour: "Vérification du locuteur: Dépendante et indépendante du texte". PhD Thesis, Université Paris-Sud, 1995.
2. F. Bimbot and L. Mathan: "Second-order statistical measures for text-independent speaker identification". in "ESCA Workshop on Automatic Speaker Recognition Identification Verification", Martigny, pp. 51-54, April 94.

3. SunSoft, Sun Microsystems Inc.: "XTL Architecture Guide". March 1994.
4. G. Chollet, J.L. Cochard, A. Constantinescu, C. Jaboulet, Ph. Langlais: "Swiss French PolyPhone and PolyVar : telephone speech databases to model inter- and intra-speaker variability". IDIAP technical rapport, 1995.

This article was processed using the LaTeX macro package with LLNCS style