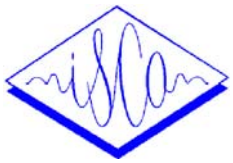

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

SECURIZED FLEXIBLE VOCABULARY VOICE MESSAGING SYSTEM ON UNIX WORKSTATION WITH ISDN CONNECTION

Philippe Renevey *Andrzej Drygajlo*

Signal Processing Laboratory, Swiss Federal Institute of Technology, CH-1015 Lausanne-Switzerland
e-mail: Philippe.Renevey@lts.de.epfl.ch, Andrzej.Drygajlo@lts.de.epfl.ch

10.21437/Eurospeech.1997-462

ABSTRACT

This paper considers applications of automatic speech recognition and speaker verification techniques in developing efficient Voice Messaging Systems for Integrated Services Digital Network (ISDN) based communication systems. The prototype demonstrator presented was developed in the framework of cooperative project which involves 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: the Advanced Communication Services (aComm), SunMicrosystems (Switzerland) and the Swiss Telecom PTT [1]. The project is supported by the Commission for Technology and Innovation (CTI). The goal of the project is to make available basic technologies for automatic speech recognition and speaker verification on multi-processor SunSPARC workstation and SwissNet (ISDN) platform to industrial partners. The developed algorithms provide the necessary tools to design and implement workstation oriented voice messaging demonstrators for telephone quality Swiss French. The speech recognition algorithms are based on speaker independent flexible vocabulary technology and speaker verification is performed by a number of techniques executed in parallel, and combined for optimal decision. The recognition results obtained validate the flexible vocabulary approach which offers the potential to build word models for any application vocabulary from a single set of phonetic sub-word units trained with the Swiss French Polyphone database.

1. INTRODUCTION

The telephone is the most popular current platform for remote speech command and control. Speech also provides a simple, easy-to-use, uniform interface for securized call management and message processing operations, and it is an important part of modern telephony applications. Therefore speech recognition and speaker verification over telephone channels are imperative.

Voice Messaging, also known as voice mail, handles the access mechanisms to record and replay digitized messages and the distribution of these messages between users. Voice messaging provides the user with the opportunity to leave a digitally recorded speech message for someone who is not available but who has a remote securized telephone access to this message. The demand for this technology for different languages is increasing dramatically, due to the opportunity of savings in operating costs and the rise of mobile cellular telephony. There is also a definite need for voice messaging services to have flexible vocabulary, speaker independent recognition, speaker identity verification, and flexible, programmable dialogue structure.

In this paper we present all these technical aspects when developing a universal, workstation oriented voice messaging system in the area of digital telecommunications for Swiss French.

2. PROTOTYPE VOICE MESSAGING SYSTEM (VMS)

The prototype voice messaging system developed in this work offers two main functionalities: posting a message, and consulting personal messages. The posting allows to choose the addressee and to record the message. The consulting allows to read the messages by the owner of the mail box. It includes a recognition of the speaker identity by means of a vocal PIN-code.

The main VMS functions are:

- *Command recognition* - permits to record a command, to recognize it, and to accomplish the associated action.
- *User identity verification* - permits to record the voice of a user, and to verify his identity using three speaker verification methods in parallel [2],[3].
- *Message recording* - permits to record a message, and to store it. This message can be retrieved later by the owner of the mail box.

The main goal of the ongoing work is to develop a general "toolkit" which allows a universal VMS to be built by associating these three mentioned above functions.

The basic hardware configuration of the prototype VMS demonstrator consists of a two-processor SunSPARC station 20, with 128 MB of RAM, connected to SwissNet (the Integrated Services Digital Networks (ISDN) in Switzerland) using ISDN BRI adapter card. The required software on the SunSPARC station includes Solaris 2.4 or later, SunISDN 1.0.2, XTL 1.1 and a modular speech/speaker recognition system adapted to real-time processing. The languages used to program the demonstrator are C++ and C, and shell scripts (tcsh). The general structure of the prototype VMS is presented in Fig. 1.

3. SPEECH RECOGNITION

At the heart of automatic speech recognition system of the demonstrator lies a set of algorithms for recognition and training.

In speech recognition the signal from a ISDN channel is processed using a Continuous Density Hidden Markov Model (CDHMM) technique where feature extraction and recognition using the Viterbi algorithm are adapted to a

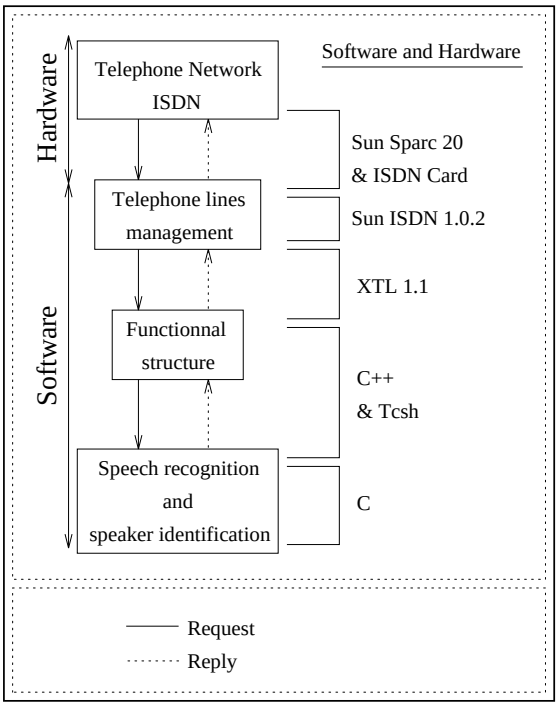

Figure 1. General structure of the VMS

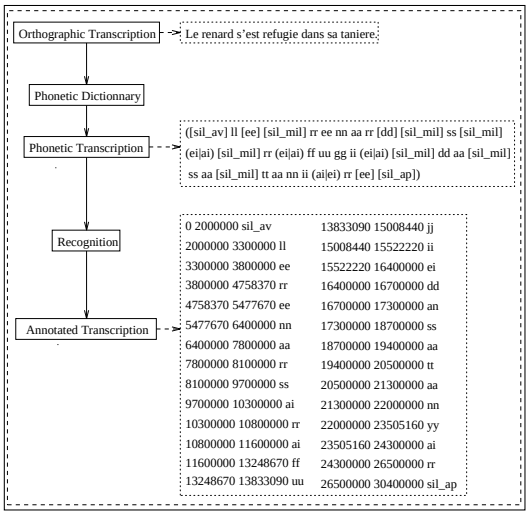

Figure 2. Automatic labeling process

Vocabulary type	Recognition rates
7 command word	98.38%
Digit recognition	97.30%
78 command word	91.95 %

Table 1. Recognition rates

real-time execution. The approach selected for this work to model command words is the speaker independent flexible vocabulary approach. It offers the potential to build word models for any speaker using Swiss French and for any application vocabulary from a single set of trained phonetic sub-word units.

The major problem of a phonetic-based approach is the need for a large telephone quality database to train, once and for all, a set of speaker-independent and vocabulary independent phoneme models. This problem was solved using the Swiss French Polyphone database [4] provided by the Swiss Telecom PTT, and massively parallel computing facilities (several multi-processor SunSPARC stations and massively parallel computer CRAY T3D (128 processors)) at the EPFL to train the statistical phonetic CDHMMs.

The CDHMM toolkit based on the Baum-Welch algorithm adapted to perform massively parallel processing was developed and used in the training and testing experiments. A total of 27671 sentences uttered by 4488 speakers were selected and labeled semi-automatically (1891 sentences) and automatically to the phoneme level (Fig. 2). Standard three-state left-to-right phoneme HMMs with five Gaussian mixtures per state, characterized by 12 Mel Frequency Cepstral Coefficients (MFCC) plus energy and their first and second derivatives were used.

In order to validate the flexible vocabulary approach for VMS, we conducted a set of experiments to compare performances of HMMs using phonemes as units. The recognition rates obtained for command words are shown in Table 1.

The reported recognition results clearly validate the flexible vocabulary approach, and show that phonetic models can be almost as performant as whole word HMMs, where the gap between recognition rates of the considered approaches is as small as 1%[5].

The VMS have been designed to work with command words. Some speakers will not utter only the command words, but place them into sentences. So these command words must be extracted from the answer to the system. For this reason a word spotting algorithm has been added to the recognition system. This word spotting recognizer uses filler models to recognize non-keyword speech. Three types of models have been tested. The first kind of filler represents words. These words models are trained on non-keyword models. With the flexible vocabulary approach, the non-keywords are not well designed and it is difficult to create accurate filler models. These models have not given good results for the word spotting.

The second type of filler was the phoneme models themselves. The problem is that these models give better recognition score for the keywords than the keyword models themselves. The results obtained with this type of models are very bad.

Finally phonetic class models have been used as filler models. The phone set used for the recognition has been divided into five classes of phonemes (fricative, labial, nasal, occlusive and voiced). For each class a model was built and trained. Then these filler models were added to the recognition grammar in order to perform the word spotting. This approach gives satisfactory results. Fig. 3 gives the grammar structure used for the word spotting.

4. SPEAKER IDENTITY VERIFICATION

The VMS offers a securized access to personal messages. At the first connection, a user has to be registered. He will be asked to repeat ten times his Personal Identification Number (PIN) and to answer several question. These utterances will be the references used in the Speaker Identity verification (SIV) process. The SIV is performed as follows: The speaker gives his PIN code. The PIN code is recognized using a HMM recognizer and its validity is tested. If the code is correct the utterance is used to per-

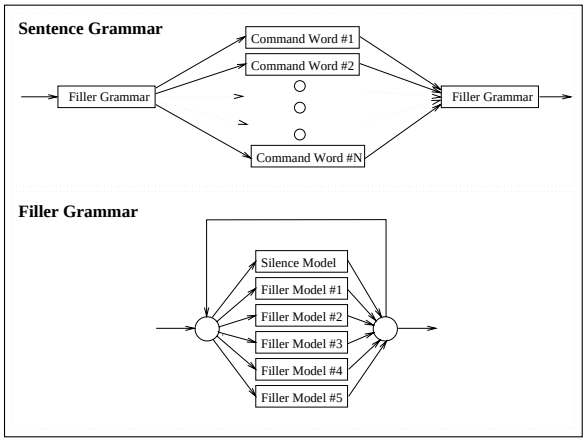

Figure 3. Word Spotting Grammar

form a text dependent SIV using three methods (described below). Three results are possible:

- *Rejected:* The access is refused.
- *Accepted:* The access is granted.
- *Doubt:* There is a doubt and a text independent verification will be performed.

In case of doubt, a question like "What is your address?" is requested to the speaker, and the answer will be used to perform a text independent SIV. This verification process is shown in Fig. 4.

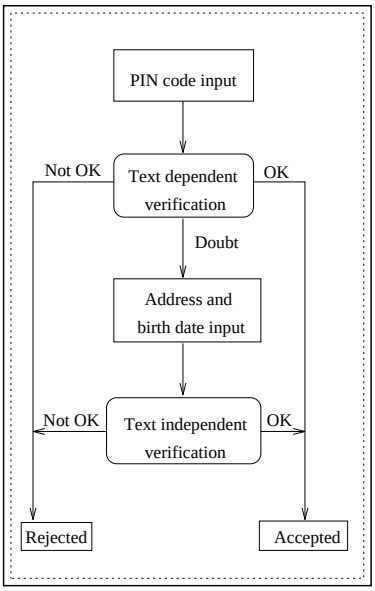

Figure 4. Speaker identity verification process

The three methods used for the SIV are:

- *HMMs:* In the text dependent SIV, digit models are built and trained with the registration PIN code utterances. For the text independent SIV, all the utterances of the registration are used to create a general speech model [6].
- *Dynamic Time Warping (DTW):* This method computes a distance between test and reference utterances. It can be used only as text dependent method [7].
- *Second order statistical measures:* A sphericity measure is computed and a distance between test and reference sequence is calculated[8].

All the three methods compare the score obtained by the test utterance for personal models and for general models (speaker independent) called "world models". If the personal models obtain the best results the identity is validated. In the other case the identity is rejected. If the personal and the general models obtain an equivalent score, there is a doubt and more information is requested to take the decision.

5. FUNCTIONAL STRUCTURE

A problem one has to cope with when designing a VMS is to describe its functional structure in a compact, precise and readable way. Some graphical formalisms have been adapted to represent the states of the VMS. One of the possible states of the VMS is presented in Fig. 5. When a flexible vocabulary approach is used, the recognition system can be easily modified, upgraded or extended. The functional structure of the prototype demonstrator was constructed keeping in mind the flexibility of its structure. Two main parts have been developed. The first part is a service independent part that manages the telephone line, the recording, and the playing of messages.The second part is service dependent. It defines the functional structure of the VMS and controls the dialogue.

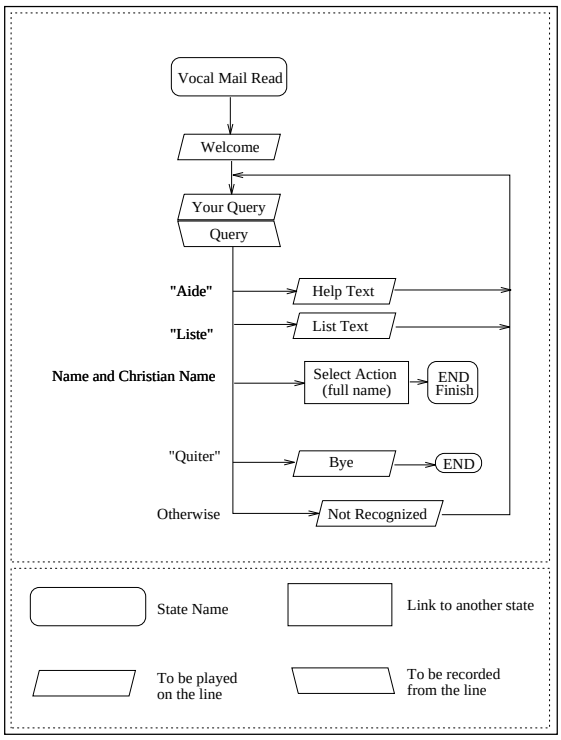

Figure 5. Graphical representation of one of the VMS states

The service independent part is written in C++. It is fixed, and do not need to be updated when constructing a new VMS with changed dialogue. It consists of a loop with the following steps:

- play request for a command

- record the answer
- call the main shell script for the recognition and wait for the answer
- play the result
- go to the first step or quit the loop.

The service dependent part of the VMS is composed using shell scripts. According to the recognition result the state of the VMS can be changed. This part works in the following way: the main script is called from the fixed part; according to the state of the VMS, select the appropriate recognition script; this script does the recognition using the grammar and the vocabulary of this state; it chooses the answer and the next state according to what is recognized; it returns to the fixed part.

These two parts of the VMS are presented in Fig. 6. To build or to modify a VMS, we have only to write or to update the shell scripts.

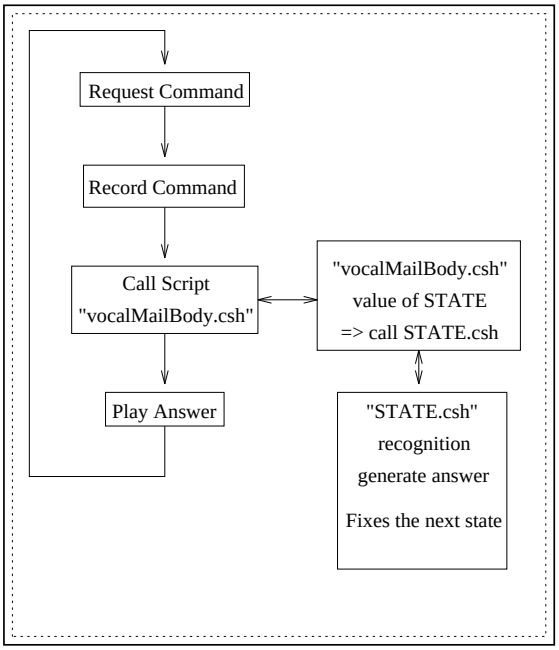

Figure 6. Functional Structure of the VMS

6. CONCLUSIONS

In this paper the complete methodology for designing and implementation of a workstation and ISDN platform oriented voice messaging system was presented. The reported recognition results clearly validate the flexible vocabulary and speaker independent recognition approach for the telephone quality Swiss French. The proposed flexible software structure allows a user to modify easily the functional structure of the prototype VMS demonstrator and to control the dialogue using only the shell scripts. The ongoing work is concentrated on a service creation tool using a Graphical User Interface (GUI) to write and modify the shell scripts.

REFERENCES

[1] A. Drygajlo, J.-L. Cochard, G. Chollet, O. Bornet, and P. Renevey, "Sun workstation and swissnet platform for speech recognition and speaker verification over the telephone", *in Workshop "Workstations and Their Applications" (SIWORK'96)*, pp. 1–4. Zurich, May 13-14, 1996.

[2] O. Bornet, G. Chollet, J-L Cochard, A. Contantinescu, and D. Genoud, "Secured access to telephone servers", *in Third IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA'96)*, pp. 41–44, Sept, 1996.

[3] D. Genoud, F. Bimbot, G. Gravier, and G. Chollet, "Combining methods to improve the phone based speaker verification decision", *in 4th International Conference on Spoken Language Processing (ICSLP'96)*, Oct. 3-6, 1996.

[4] G. Chollet, J.-L. Cochard, Ph. Langlais, and R. van Kommer, "Swiss-french polyphone: a telephone speech database to develop interactive voice servers", *in Linguistic Databases*, Gröningen, 1995.

[5] A. Constantinescu, O. Bornet, G. Caloz, and G. Chollet, "Validating different flexible vocabulary approaches on the swiss french polyphone and polyvar databases", *in 4th International Conference on Spoken Language Processing (ICSLP 96)*, pp. 2293–2296, Philadelphia, Oct. 3-6, 1996.

[6] A.E. Rosenberg, C.H. Lee, and S. Gokoen, "Connected word talker verification using whole word hidden markov models", *in ICASSP*, pp. 381–384, 1991.

[7] H. Sakoe and Chiba, "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Trans. on ASSP*, vol. 26, pp. 43–49, 1978.

[8] F. Bimbot and L. Mathan, "Second order statistical measures for text-independant speaker identification", *in ESCA*, 1994.