IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

SECURED VOCAL ACCESS TO TELEPHONE SERVERS

O. Bornet [a] G. Chollet [a b] J.-L. Cochard [a]
A. Constantinescu [a] D. Genoud [a]

IDIAP–RR 96-04

SEPTEMBER 1996

PUBLISHED IN
Third IEEE Workshop on Interactive Voice Technology for Telecommunications Applications

Dalle Molle Institute for Perceptive Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] Dalle Molle Institute for Perceptive Artificial Intelligence PO Box 592 CH-1920 Martigny, Switzerland
[b] CNRS URA820 École Nationale Supérieure des Télécommunications F-75634 Paris, France

IDIAP Research Report 96-04

Secured vocal access to telephone servers

O. Bornet G. Chollet J.-L. Cochard A. Constantinescu D. Genoud

September 1996

published in
Third IEEE Workshop on Interactive Voice Technology for Telecommunications Applications

Abstract. A number of applications of man-machine interaction over the telephone requires a combination of speech recognition and speaker verification. This paper describes current work carried out at IDIAP in the framework of national and European projects. A generic Interactive Voice Server (IVS) is described by means of a graphical formalism. It includes speech recognition based on speaker independent flexible vocabulary technology and speaker verification performed by a number of techniques executed in parallel, and combined for optimal decision.

1 Introduction

A number of applications of man-machine interaction over the telephone requires a combination of speech recognition and speaker verification. This paper describes technology developments and evaluation carried out at IDIAP in the framework of national (ATTACKS, CERS, SwFPolyphone, SwGPolyphone) and European (SpeechDat, CAVE, M2VTS, COST 249, COST 250) projects. This effort resulted in the installation of a number of demonstrators running on Sun-Sparc workstations connected to the Swiss ISDN network (SwissNet). They are being used to analyse the user acceptance of new telephone services and to enable iterative improvement of the technology.

A basic effort was devoted since 1994 to database collection and annotation [4]. 5,000 Swiss-French speakers were recorded over the telephone according to the COCOSDA Polyphone protocol (SwFPolyphone database). This database is representative for inter-speaker variability. Intra-speaker variability is captured in a companion database called PolyVar. 50 Swiss French speakers were requested to make Polyphone-like recordings with several days between recordings. 3,500 PolyVar calls are available now. The goal is to have inter- and intra-speaker variability databases of similar sizes.

Section 2 describes the generic demonstrator of an Interactive Voice Server (IVS), connected to ISDN (Integrated Services Digital Networks), integrating speech recognition and speaker verification technologies. Section 3 gives details on the experimental validation of the 'flexible vocabulary' approach used for speech recognition. In section 4 combination of techniques for optimal speaker verification are given. The paper is concluded with a discussion on further developments and future applications of this work.

2 InfoMartigny, a generic IVS

InfoMartigny is a prototype of a system able to provide touristic and cultural information about Martigny area. It is considered as a good textbook case to assess technology currently available at IDIAP and to work on integration aspects of speech recognition and speaker verification components.

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 needed software on the Sun SPARCstation includes Solaris 2.4 or later, SunISDN 1.0.2, XTL 1.1 [16] and some tools from HTK [10] adapted to real time speech recognition.

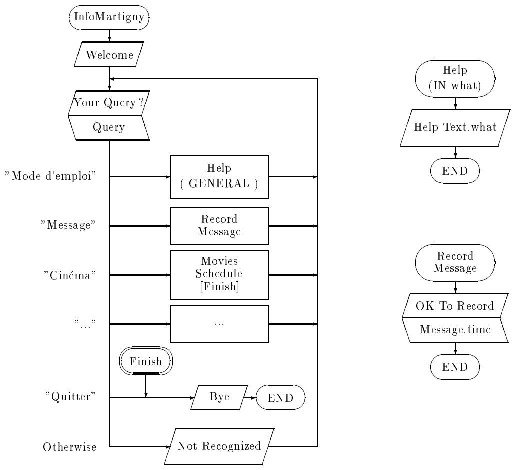

Figure 1: A graphical description of procedures Main, Help and Record.

A problem one has to cope with when designing an IVS is to describe all the possible dialogue scenarios in a compact, precise and readable way. Some graphical formalisms have been adopted for this task [17, 12, 13]. A new one is used here to illustrate interleaving of interaction and processing steps within InfoMartigny (see [3] for a complete description of the formalism).

Figure 1 contains three procedures definitions, one for a procedure InfoMartigny (main procedure), with an entry point labeled InfoMartigny, a second for a Help procedure that accepts one 'input' argument and simply plays a pre-recorded message (denoted by a parallelogram leaning over right), and a third for recording a message that plays a vocal prompt and records a user message.

Executing a procedure can be understood as moving a pointer inside the procedure graph from the entry point to an exit point labeled with END. If an 'end alone' exit is reached, the execution resumes at the exit arrow of the procedure call, denoted by a rectangle containing the same label. If the exit is a 'end "something"', like END Finish in Figure 2(a), the execution resumes at the exception point labeled Finish, inside the graph of the

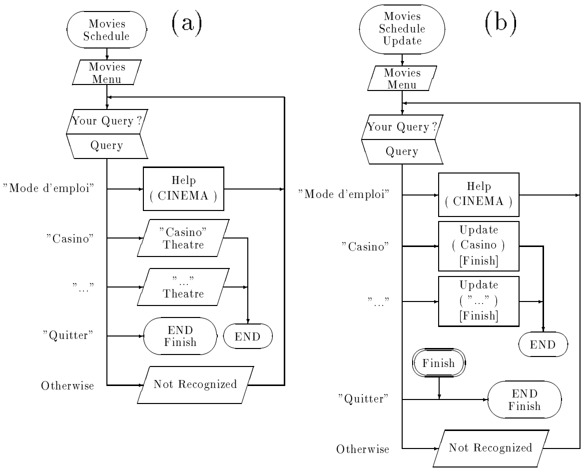

Figure 2: Interaction in the theatre sub-menu (a) to get information, (b) to update information.

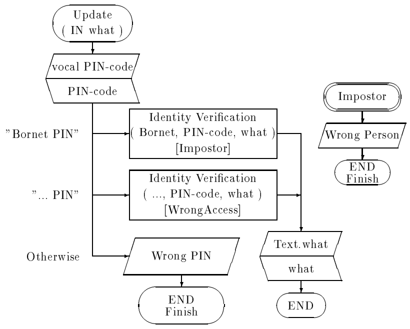

Figure 3: Generic secure access to perform update of information.

procedure call.

Speech recognition is denoted by a case selection (see all Figures). It is initiated by a voice prompt, combined with recording of a user request. Below the list of possible keywords with their dedicated actions, an Otherwise case can be fired if recognition rate is below a certain threshold.

Figure 2(b) is replacing Figure 2(a) in a second IVS allowing some granted speakers to update information of some topics. As one specific speaker can have modification permission for one topic, Speaker Verification (SV) only applies after selection of a particular topic. Thus, the Update procedure has one argument, the topic for which verification has to be achieved (see Fig. 3). Identity Verification as described in §4, can either be successful and therefore the user has the right to store new information, or it can fail and then the procedure raises the Impostor exception. Another error case occurs when speech recognition of the PIN code fails, raising a Finish exception at the outer level.

3 Speech Recognition for IVS

The speech recognition module of a versatile IVS is based on subword units, which are most successfully implemented as Hidden Markov Models (HMMs) [14]. Such 'flexible vocabulary' approach allows the modeling of any word or phrase the IVS has to recognize. The obvious advantage of this technique lies in the reusability of such segments, which need to be trained only once and can then be easily used to create new vocabulary entries for recognizers. This represents a considerable advantage compared to the use of rigid whole word models, which imply new recording sessions for any new vocabulary entry.

3.1 Phonemes as Subword Units

The most common subword units used nowadays are phonemes. In order to validate the flexible vocabulary approach for IVS, we conducted a set of experiments, on speaker independent single word and connected words recognition. They are described in more details in [5]. The scope of this study was to compare advantages and performances of HMMs using words or phonemes as units.

The modeling of these two families of units was performed with HTK [10]. The created left-right HMMs had mainly three emitting states per phoneme, while the number of Gaussian mixtures per state and stream, which represent the observation probability, varied from about five (for words), to approximately twenty (for phonemes). As parametrisation, we used 12 MFCCs and energy, together with their first and second derivatives.

3.2 Training Data

Training of models was performed on parts of the SwFPolyphone (inter-speaker variability) and PolyVar (intra-speaker variability) telephone speech databases. For training of subword models, a total of 2250 sentences uttered by 231 speakers (9–10 sentences per speaker) were selected from

Table 1: Comparative recognition rates.

Validation	on connected digits	on 17 isolated words
Whole word HMMs	97.52%	98.13%
Phoneme HMMs	95.44%	97.10%

SwFPolyphone, labeled semi-automatically to the phoneme level, and thereafter used to train phoneme HMMs.

As reference for the proposed validation, we created two sets of whole word HMMs:

- The French digits from zero to nine, including the words for the hash (#) and star ($\star$) symbols, where the models were trained on the utterances of 471 speakers from SwFPolyphone, each pronouncing one sequence of six digits (and symbols).
- A vocabulary of 17 application words from PolyVar for which the training material consisted of recordings of 10 speakers (5 female, 5 male), where each of them repeats all 17 words 30 times.

3.3 Evaluation Data

For evaluation purpose, the flexible vocabulary approach and the whole word models were tested for two tasks on the same sets of utterances.

1. For *connected words recognition*, 497 sequences of connected French digits including hash and star, where uttered by male and female SwFPolyphone speakers (one sequence of six digits per speaker).
2. For *isolated word recognition*, the mentioned 17 words were pronounced by 103 PolyVar speakers between 1 and 26 times, resulting in a total of 5321 isolated words. This test is supposed to give also an idea about the robustness of the evaluated subword units, since SwFPolyphone and PolyVar calls were recorded on different recording platforms.

3.4 Validation Results

The reported recognition results (see Table 1) clearly validate the possible use of the flexible vocabulary approach for IVS.

The Recognition rate of context independent phonemes as subword units approaches that of whole word HMMs to 1% for isolated words and 2% for connected words. Moreover, the created phoneme models appear sufficiently robust, given the fact that training and evaluation were carried out on two differently recorded telephone databases.

4 Speaker Verification Module

In order to update information for the different topics of InfoMartigny, a secured access has to be implemented by means of a Speaker Verification (SV) module.

Before the secured access can be used, each granted speaker has to complete an enrolment phase. During this phase, he is asked to pronounce his name, first name, address, all the digits from 0 to 9 sequentially, and 5 times his 7 digits personal identification number (PIN code). Speech recognition is performed on all the digits sequences in order to time-label them.

During the access phase, the speaker pronounces only once his PIN code. A verification process is a two steps task performed sequentially (see Fig. 3). Firstly, the digits sequence is recognized using a HMM-based speaker independent speech recognizer (see §3). Secondly, the speech sequence is compared to the speaker references corresponding to the recognized PIN code. Depending on the similarity between the references and the incoming sequence the speaker access is granted or denied.

4.1 Methods Used

Experiences in speaker verification show that errors made by verification algorithms are not similar for every speaker, therefore three text dependent verification methods are used. All these methods take as input a set of LPC cepstral coefficients with delta and delta-delta coefficients. The first method is a *Dynamic Time Warping* algorithm (DTW), consisting mainly in a dynamic comparison between a reference and a test matrix. The algorithm computes a distance between the test and reference patterns [11]. The second method is based on *Second Order Statistical Method* (SOSM), using a comparison between a reference and a test covariance matrix. A symmetrical measure, called sphericity measure is obtained [2].The last method is based on *Hidden Markov Models* (HMM). Two types of HMM [15] are created for each digit (0 to 9): (1) *a world model* which is speaker independent and is used to normalize the speaker score; (2) *a speaker model*, which uses parameters of the world model as initial parameters, and then re-estimates them with the speaker data. All models have the same HMM left-right structure with one mixture per state. Each model has one state per phoneme and one state per phoneme transition.

At the access time, for each digit issued by the speaker, the log likelihood ratio (LLR) is computed as follows: $LLR_{sw} = \log(L_s) - \log(L_w)$, L_s, L_w being likelihood of the speaker and world models respectively.

To improve the global performance of this SV module, decisions given by each method (DTW, SOSM, HMM) are combined by means of a weighted majority test [8]. Practically, each method provides a numerical score and a weighting factor. This factor is computed as a normalized distance between a threshold assigned to each method or each speaker, and the current method score. This factor can be understood as the confidence degree of the decision.

4.2 Database Used and Results

The Polycode database [9] used for speaker verification validation is composed of 42 speakers recorded over a telephone line in several sessions. During one session, each speaker has to say, among other sentences in French, 5 times his own 7 digits PIN code and 4 times 10 digits sequences (all the digits from 0 to 9, in different order for each sequence).

Polycode database was split up into several subdatabases in order to train and test the SV module (see [8] for detailed explanations). Training and testing sessions have been realized on 20 speakers with respectively 100 utterances for training (5×20), and 800 true accesses and around 11,000 impostor accesses for testing.

Table 2 gives false acceptance (FA) and false rejection (FR) rates for each method performing alone. It also shows that rates improve significantly by combining separate scores.

Table 2: Comparison of performance of isolated methods and combined decision.

Method	FR% (800 tests)	FA% (11,000 tests)
SOSM	12.12	9.41
DTW	3.52	7.12
HMM L/R	4.78	3.85
Combined Decision	3.12	4.53

5 Discussion

With the rapid spread of mobile telephony and the foreseen availability of portable computers, the use of a vocal interface for distant applications becomes more and more mandatory. Such an interface should allow the user to speak in a natural manner. Simple IVS react on the detection of specific vocabulary items ('word spotting') and ignore any other speech utterances and noises. This is usually realised by training the recognizer on samples of the specific items uttered by many (~1000) speakers. The costs associated with the collection, validation and processing of such application dependent databases have limited the use of this technology. The *flexible vocabulary* approach reduces these costs. The model for a new vocabulary entry is generated by concatenating subword models. Subwords are often chosen *a priori* (by experts), but, in the future, it is hoped that such units could be discovered automatically [7]. The discovery of a set of language independent units could even be envisaged [6]. The models for such units could be adapted using a limited language and application specific database.

For some applications (banking, teleshoping, access to sensitive or private information, ...) the identity of the user must be verified. The enrolment of new users could be a sensitive issue in some cases. Maximal security requires that the new user performs enrolment under supervision of an authorised human operator. Current technology using text dependent, text prompted and text independent techniques could achieve any level of security (very limited false acceptance) at the cost of some frustration from the customer [1]. Some IVS could identify the speaker (or the speaker type) so that speaker dependent recognition models could be used. The origin of the call could help to take this decision. Telephone users (and specially those using mobile sets) could use an automatic assistant (Majordome) to access personal information. In such case, the assistant must authentify the caller.

In this context, availability of large and similar corpora of speech samples from different countries, in different languages, as recorded by the SpeechDat project (EU Telematics program), is a very important milestone that clearly gives new opportunities for IVS, based on a flexible vocabulary approach, to be deployed across Europe.

References

[1] F. Bimbot and G. Chollet. Assessment of speaker verification systems. In *Handbook of Spoken Language Resources and Assessment.* EAGLES, 1995.

[2] Frédéric Bimbot and Luc Mathan. Second-order statistical measures for text-independent speaker identification. In *ESCA Workshop on Automatic Speaker Recognition, Identification, Verification*, pages 51–54, 1994.

[3] Olivier Bornet and Jean-Luc Cochard. Graphical formalism for ivs dialogue description. Communication 96–04, IDIAP, PO Box 592, CH-1920 Martigny, August 1996.

[4] G. Chollet, J.-L. Cochard, A. Constantinescu, C. Jaboulet, and Ph. Langlais. Swiss french Polyphone and PolyVar: telephone speech databases to model inter- and intraspeaker variability. Research Report 96–01, IDIAP, PO Box 592, CH-1920 Martigny, 1996.

[5] A Constantinescu and G Chollet. Swiss Polyphone and PolyVar: Building databases for speech recognition and speaker verification. In *3rd Slovenian-German and 2nd SDRV Workshop, Speech and Image Understanding, Ljubljana*, Ljubljana, April 1996.

[6] D. Constantinescu and G. Chollet. Towards language independent recognition of telephone speech: The future generation of voice servers. Thesis proposal, IDIAP, 1996.

[7] S. Deligne, F. Yvon, and F. Bimbot. Variable-length sequence matching for phonetic transcription using joint multigrams. In *EUROSPEECH*, 1995.

[8] D. Genoud, F. Bimbot, and G. Chollet. Combining methods to improve speaker verification decision. In *ICSLP 96, International Conference on Speech and Language Processing*, Philadelphia, USA, October 1996.

[9] Dominique Genoud and Gérard Chollet. Polycode, a verification database. Technical report, IDIAP, CH-1920 Martigny, 1995.

[10] Cambridge University Speech Group. *HTK Hidden Markov Model Toolkit.* Entropic Research Laboratories Inc., Cambridge, December 1993.

[11] Sakoe H. and Chiba. Dynamic programing algorithm optimization for spoken word recognition. *IEEE Trans. on ASSP*, 26(1):43–49, 1978.

[12] Yevgeny K. Ludociv and Valery G. Sibirtsen. Intelligent answering machine-secretary. In *EUROSPEECH*, volume 1, pages 277–280, 1995.

[13] Y. Niimi and Y. Kobayashi. Modeling dialogue control strategies to relieve speech recognition errors. In *EUROSPEECH*, volume 2, pages 1177–1180, 1995.

[14] L Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition.* Prentice-Hall signal processing series, 1993.

[15] A.E. Rosenberg, C.H. Lee, and S. Gokoen. Connected word talker verification using whole word hidden markov model. In *ICASSP-91*, pages 381–384, 1991.

[16] Xtl architecture guide. Technical report, SunSoft, Sun Microsystems Inc., March 1994.

[17] B. L. Zeigler and B. Mazor. Dialog design for speech-interactive automation system. In *IVTTA*, pages 113–116, 1994.