

Corpus Design For Speaker Recognition Assessment

ANDREA DI CARLO, MAURO FALCONE, ANDREA PAOLONI

Abstract — We present the project of an Italian speech database collation over the Public Switched Telephone Network. The goal of this database is to create a reference material for speaker recognition assessment, as regard the Italian language. It is part of a wider Italian national project managed by the ISPT (Istituto Superiore delle Poste e Telecomunicazioni) that is the tutor of the "Italian Association of Acoustic Databases" (AIDA). After the production of a set of 6CD for speech recognition assessment and the design of two databases for speech synthesis, the AIDA4 corpus, for the assessment of speaker recognition over the telephone line, has been designed. In this paper we describe the contents of this database, the collation modalities and the technical set up of the collation. As far as possible in designing this database we consider the different kind of speaker recognition systems (text-dependent, text-independent, listening tests, etc.). The start of the collation is planned for the second half of this year, and it will take about eight months. A total of 500 speakers will take part to the realisation of this database. The database will be ownership of the ISPT, that is also the official distributor of all AIDA corpora.

Index Terms — databases, speaker verification, speaker recognition, data collation, telephonic speech.

I. INTRODUCTION

Around the word speech processing technology is being incorporated in a wide range of systems [1]. It seems new applications are being introduced daily, especially in telecommunication area. Several types of technologies are used in there applications.

Speech synthesis offers enormous potential for useful custom service. Applications such as reverse directory or customization of standard call intercept messages provides interesting solution for services which can be accessed twenty four hours per day, seven day per week.

Automatic speech recognition systems (ASR), available from multiple vendors in multiple forms, have successfully pioneered in numerous applications. This technology has improved sufficiently to provide for

satisfactory operational performance. Speech recognition plays a central role in recent medical applications in a hospital setting. The large vocabulary speech recognition system can be used such as a vocal typewriter. Numerous consumer applications, e.g. the Voice Activated Telephone (VAT) has been demonstrated. Finally a word spotting ASR, a new technology which permits a recogniser to "spot" a spoken stream of speech has been applied in operator assisted telephone call automation.

As interactive voice technology evolves, application once considered high-end move down market becoming a new segment of the voice technology market.

Automatic speech processing techniques for identification of people by their voice characteristics have a number of important potential applications. These applications include [2][3]:

- security, where the task is to verify the identity of an individual for control of access to a restricted facility (speaker verification);
- surveillance of communication channels, where the task is to identify a speaker from samples of unconstrained text (speaker recognition);
- forensic applications [4], which can involve either, a recognition and verification task, but where control over the available speech sample is often limited, and the potential number of impostor may be very large.

To provide enough speech data to develop speaker recognition technology, to provide evaluation of speaker recognition systems (both text-dependent, and text-independent), and to provide large corpus to study or to model phonetic variations, the design of a large goal oriented corpus is crucial.

II. OTHER CORPORA

For the English language several speech corpora are available to support research in speaker recognition. We review in this paragraph those distributed by Linguistic Data Consortium (LDC).

- The KING corpus [5] collected at ITT contain monologues by 51 male speakers each divided into 10 sessions per speaker of short 1 minute duration.
- The YOHO corpus [5] collected at ITT contain phrases of trip digits and other combinations by 186 speakers. More than 1900 test sessions are included in this 2 CD-ROM corpus.

The authors are with the speech group of the "Fondazione Ugo Bordonini" (FUB), Via B. Castiglione 59, 00142 Rome, Italy.

You can reach them by fax at +39.6.5480.4405 or via Email at pao90@itcaspur.caspur.it

- The SWITCHBOARD corpus [6] collected automatically at Texas Instruments contains 2500 conversations among more than 500 talkers from all over the United States.

For non English languages, the "International Coordinating Committee on Speech Databases and Speech I/O Systems Assessment" (COCOSDA) as been made at a satellite meeting of the Eurospeech Conference in Chiavari, Italy [7].

This Commette has proposed the Polyphone project [8], an international collaboration for collect a telephonic corpus. It will contain a figure of 5000 callers for each language with not more than three minutes per caller is 15.000 of speech.

This corpus can be utilized for text-independent speaker identification or verification.

For the Italian language the "Istituto Superiore delle Poste e Telecomunicazioni" (ISPT), accepting the suggestions from other research centres, has decided to create an expert committee to define corpora for the assessment and research purpose.

The Committee has started its work in October 1989 specifying the aims that it was necessary to achieve with the creation of the vocal database. The committee had to consider different necessities and that is, as just said, the requirements of evaluating apparatus for the recognition and synthesis, and the requirements of realisation of systems. Starting from these considerations for different corpus was defined [9].

- AIDA1 has been realised to solve the problems of testing and evaluating automatic voice recognition systems. It consists of digits and bisyllabe words, choose in order to cover all Italian phonetic constraints. It is distributed in 6 CD ROMs [10].
- AIDA2 and AIDA3 has been designed to solve problems arising in evaluating text-to-speech synthesis systems at both segmental and sovra-segmental level.
- AIDA4 corpus has been designed for speaker recognition and verification. Two outcomes are carried out: speaker verification for having assess to specific services, and speaker recognition for forensic applications.

A large amount of data collected from many speakers (more than 500) is forecast. The data will be also useful to characterize speaker variability, dialect variability, speech, speed, speech level and psychological factors. The corpus will contain 28 isolated words, a small dialogue (11 utterances), and a read passage (60s).

III. THE DESIGN

The goal of AIDA4 speech corpus are to support research in speaker recognition, speaker verification and related areas.

As designed to support several types of speech and language research, it have to contain various kind of

speech: isolated words, passages and a example of spontaneous speech.

Therefore we will produce a corpus constituted by a continuous passage, a spontaneous dialogue and a set of isolated words. The variability dimensions will be the personal identity, the dialectal area, the sex and the acquisition time. 500 speakers of both sexes will be involved.

In the first phase 500 speakers will repeat the acquisition after N days; then 20 (selected) speakers (10 males and 10 females) will repeat the acquisition 12 times in seven months (we foresee 12 acquisition phases in 0, 3, 7, 15, 30, 45, 60, 90, 120, 150, 180, 210-th days). People aged between 18 and 60 will be recruit in main dialectal Italian areas, individuated by linguistic experts. Acquisition will be centralised that is a central station for the material collection will be set up in Rome at ISPT laboratories.

3.1 Corpus definition

3.1.1 Isolated word session

In order to realise test comparable to the literature is necessary to have the list of digits and commands for a Speaker Verification System. Then we propose the following corpus of 28 words. Word that is a *digits* sequence like "9.7.2.3", must be produced as connected words (that is NINE SEVEN TWO THREE and not NINTYSEVEN TWENTYTHREE)

TABLE I
Isolated word list

SI	A.F.B.8
NO	9.7.2.3
GLI SCACCHI	DUE
RICHIESTA DI ACCESSO	1.E.5.D
TRE	CANCELLA
4.C.0.6	AIUTO
INVIO	SETTE
SEI	CORRETTO
ESEGUI	ERRATO
QUATTRO	AVANTI
CINQUE	INDIETRO
UNO	NOVE
OTTO	FERMA
INIZIA	ZERO

3.1.2 Dialogue session

An example of the dialogue between the system (S) and the speaker (P) follows. The answers of the speaker are only for example purpose but they will be produced in real spontaneous mode.

3.1.3 Continuous passage session

This text will be read by the speaker. It should be about sixty seconds long, depending on the utterance modalities. The speaker is free to read the number in the more natural way.

TABLE II
Example of dialog session

S	Come si chiama?
P	e.g. Carlo Rossi
S	Luogo di nascita
P	e.g. Sono nato a Roma
S	Data di nascita
P	e.g. 5 dicembre 1948
S	luogo abituale di residenza
P	e.g. Torino
S	Dove ha frequentato le scuole elementari?
P	e.g. a Bari in Puglia
S	Ripeta la sua data di nascita come sequenza di cifre
P	e.g. 0.5.1.2.1.9.4.8
S	Mi dica il numero di codice del TEST
P	e.g. Il mio codice del test e' RSIALO48N5Y908D
S	Qual'è il suo titolo di studio?
P	e.g. Sono laureato in Fisica
S	Dove lavora?
P	e.g. Sono dipendente della Fondazione Bianchi
S	E' coniugato?
P	e.g. Si
S	Ha già eseguito questa prova?
P	e.g. No
S	Arrivederci e grazie.

TABLE III
Continuous passage

Mi chiamo Mario Rossi; sono nato a Parma addì 26 maggio 1961. Se vuoi conoscere il mio numero di codice fiscale è RSS MRO 61E26 N110F. Mi sono laureato in Scienze Agrarie a Napoli nel giugno 1984 e sono dipendente, come funzionario aggregato, presso la Fondazione "SMEG" che ha sede a Venezia dal 1° ottobre 1985. Sono stato abilitato ad accedere a codesto servizio con autorizzazione numero abbonato 52341 del 18 febbraio 1992; il mio codice di accesso è GLI SCACCHI 1300. Ripeto più dettagliatamente il mio codice d'accesso GLI SCACCHI 1300.

3.2. Work planning

3.2.1 Working phases

We foresee the following working phases

- 1) alerting a sufficient number of speakers;
- 2) acquisition;
- 3) data collected verification and selection of 520 speaker data useful for the database;
- 4) CD-ROM prototyping;
- 5) CD-ROM verification;
- 6) CD-ROM production.

3.2.2 Scenario

4000 peoples will be alerted by post; 2000 speakers will produce the first repetition; 1000 speakers will

propose themselves for the second repetition and will receive some gadgets. The collected material of 500 speaker will be selected a posteriori for the final corpus realization. We assume individuation from these 500 speakers 20 people for 12 repetitions.

3.2.3 Duration

We assume that the equipment will be ready in 6 to 12 months. The work duration to carry out the corpus will be close to 12 months.

3.3 Equipment

The necessary equipment consists of :

1) a workstation for acquisition with:

- two distinct toll-free telephone numbers;
- automatic answering and "prompting" on several lines;
- signal acquisition functionalities with predefined modalities;
- interface computer system (SESAM like) in order to realise the corpus management and validation;

2) a workstation for backup, storage, and verification of data and for prototyping CD-ROMs.

The workstation designed to collect speech database on the Public Switched Telephone Network (PSTN) is based on a Personal Computer equipped with a Dialogic telephonic board. We decide for the analogue D/41 board, that is a four channel board with onboard loopstart interface. It use the Analog Expansion Bus (AEB), a four line bidirectional audio bus , to access additional call processing resources like fax or voice recognition.

This board has received the approval from the Italian Ministry of Post and Telecommunications in 1993, so that the correct link between hardware platform and the telephone line is guaranteed. This is the first reason that move our decision in this direction; the second one is that a lot of successful experiences has been gained in other Italian laboratories [11] (and in many other place all around the world) using such kind of hardware.

We are now evaluating if the collation of this database will be feasible using the standard facilities provided in the basic system configuration or if some additional hardware- software improvement is required at our principal scope. In a previous telephonic data collation of the TI isolated word vocabulary performed in FUB last year using a different system, we found that collation success is rather sensitive to the uniformity, rate and timing of the acoustic prompt.

As the different speaking styles enclosed in our database, an optimisation of prompting methodology is required. By now it is our believe that a system as the one described in [12] should be sufficient to fulfil our requirements.

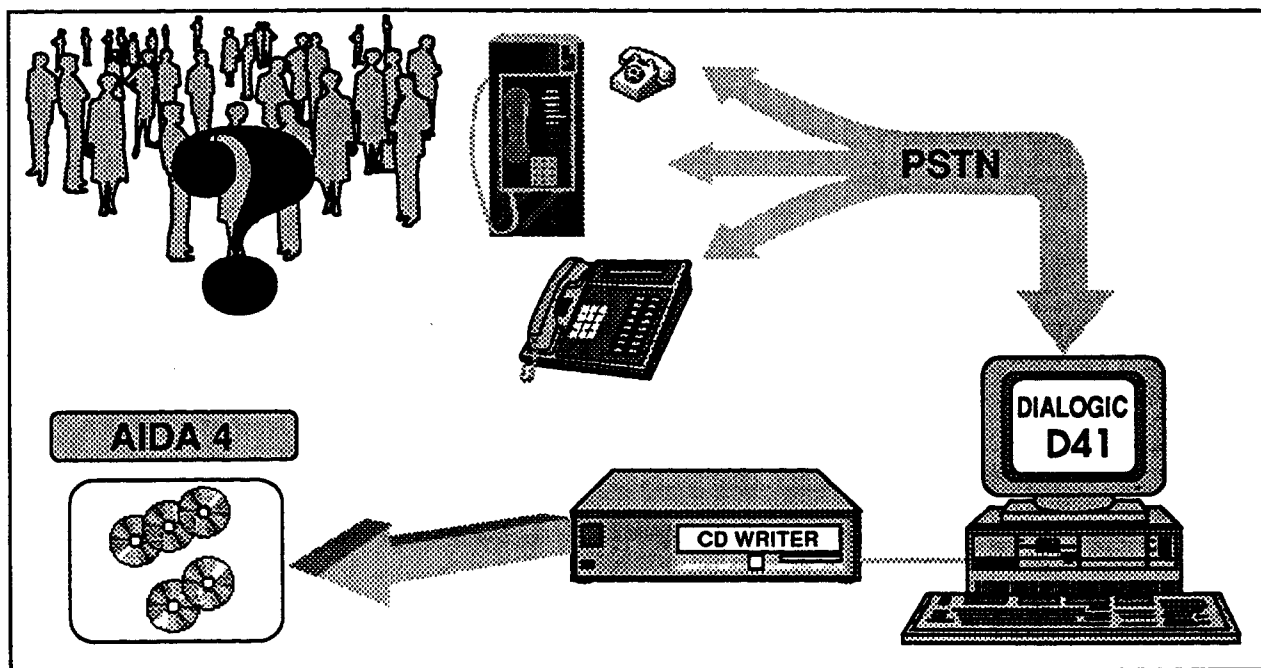


Fig. 1. Schematic blocks of the AIDA4 collation

3.4 Database size

Since 1 CD ROM = 550 Mbyte, 8 kHz sampling rate and 2 bytes per sample produce 16 kbyte per second of speech, the duration of each session is about 120 seconds, we obtain 120×16 kbyte = 2 Mbyte per speaker or 250 speaker per CD ROM on 5 CD ROMs.

TABLE IV
CD ROM distribution

CD-ROM	repetition	sex	TYPE
1°st	1°	M	250x1
2°nd	2°	M	250x1
3°rd	1°	F	250x1
4°th	2°	F	250x1
5°th	12	M+F	20x12

IV. CONCLUSION

We presented the design of an acoustic corpus for speaker recognition assessment. As it must support several types of speech and language research, it have to contain various kind of speech isolated words, read passages and spontaneous speech by 500 speakers.

Some attention is given to the linguistic coverage of the read portion. The recruitment of speakers appears an expensive phase of the work. The complete corpus will be stored on 5 CD ROMs.

The equipment have been set not only on the basis of operational requirements but in the respect of the authority recommendations on standard.

ACKNOWLEDGMENT

We would like to thank Prof. Federico Albano-Leoni, of the CIRASS Institute in Naples, that provide the text for the continuous passage, and all the members of "Italian Commission of Acoustic Database", for their helpful hints given in designing this database.

REFERENCES

- [1] A.J. Fourcin et alii, "Speech Input and Output Assessment, Multilingual Methods and Standards", Ellis Horwood Books in Information Technology, 1989.
- [2] F. Bimbot, G. Chollet, M. Falcone, "The Assessment of Speaker Recognition System", ESPRIT P.6819 (SAM-A), first year progress, report N.9
- [3] F. Bimbot, et alii, "Methodology for the Assessment of Speaker Verification Systems", these proceedings
- [4] G. Ibba, A. Paoloni, "Trends in Speaker Recognition", proc. of VERBA '90, International Conference on "Speech technologies", Rome, Italy, January 1990.
- [5] J.J. Godfrey, D. Graff, A. Martin, D. Pallet, "Public Databases for Speaker Recognition and Verification", these proceedings
- [6] J.J. Godfrey, E.C. Holliman, J.McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development", proc. ICASSP'92, pp.517-520
- [7] Proceedings of the IInd meeting of the "International Coordinating Committee on Speech Databases and Speech I/O Systems Assessment (COCOSDA)", Chiavari, Italy, September 1991.
- [8] Proceedings of the IIIrd meeting of the "International Coordinating Committee on Speech Databases and Speech I/O Systems Assessment (COCOSDA)", Banff, Canada, October 1992.
- [9] E. Laj, A. Paoloni, "AIDA: The Italian Corpora", Proceedings of the Second Workshop of 'Neural Networks for Speech Processing', Firenze, Italy, December 1992.
- [10] G. Castagneri, K. Vaggas, "The Italian National Database for Speech Recognition", Proceedings of ICSLP '90, Kobe, Japan, November, 1990.
- [11] F. Canavesio, G. Castagneri, G. Di Fabrizio, A. Massone., "TESCOS - An Integrated Workstation to Collect Large Databases on the Telephone Network", proc. of AVIOS'93, pp. 25-30
- [12] G. Castagneri, G. Di Fabrizio, A. Massone, M. Oreglia, "SIRVA - A Large Speech Database Collected on the Italian Telephone Network", proc. EUROSPEECH'93, pp. 199-201