THE "SIVA" SPEECH DATABASE FOR SPEAKER VERIFICATION: DESCRIPTION AND EVALUATION

Mauro **F***alcone,* ***A****lessandra Gallo*

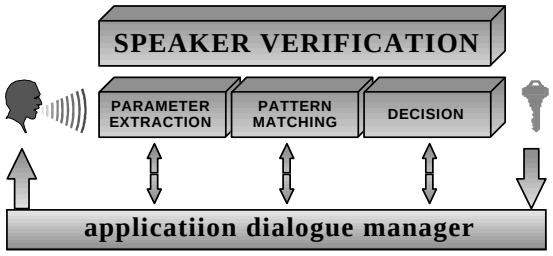
Fondazione Ugo Bordoni
Via Baldassarre Castiglione 59, 00142 Roma

ABSTRACT

The description and characterization of the Italian speech database SIVA is given. After a brief review of the available corpora designed for speaker verification task, we introduce the "Speaker Identification and Verification Archives: SIVA", a database that consists actually of more than two thousands calls, collected over the public switched telephone network. A detailed description of speech material, a proposal for an acoustic characterization, and the performances obtained using a speaker verification reference system are presented and discussed herein after.

1. INTRODUCTION

In the last years the importance of corpora for speech technology has been definitively recognized in basic science, as well as in research and development limits. There are now several organizations that distribute speech and lexicon databases: the "Linguistic Data Consortium" (LDC) from the University of Pennsylvania, is the most famous one and distributes almost every American public database; the "European Language Resources Association" (ELRA), founded last year, is going to be the reference point for European countries.



Fig*ure* **1:** The standard speaker verification system may be divided in three main blocks. In order the speech parametrisation block; the pattern matching algorithm block, and last the decision strategy block. These blocks may be controlled by a dialogue manager that controls the input and output of the whole system.

Speaker verification (SV) concerns the problem of verifying if a given utterance has been pronounced by the declared authorized speaker or not. The simplest scheme to represent a speaker verification system is shown in figure 1. Authorized speakers are commonly called "users", while speakers that try to force the system, i.e. to mimic another person's voice, are called "impostors". The taxonomy and the exact terminology of speaker recognition (by speaker recognition we indicate every possible task including verification, identification, monitoring, etc.) may be quite difficult, and in any case there is not a definitive agreement on it. A detailed description is given in [1]. The present work only concerns speaker verification problem, in its common understanding, i.e. as previously defined.

2. SPEECH DATABASES FOR SPEAKER VERIFICATION

Generally speaking a speech database for speaker verification assessment and evaluation should contain many repetitions of the users' voice, and few (at least only one) repetitions of the impostors' voice. It should also contain many impostors' voice, and (only for practical reasons) a limited number of users' voice. In addition the speakers' population should be balanced in gender, age, social extraction, etc. Of course, these are only broad guidelines following a general purpose approach [2]; specific tasks and solutions may require designs of ad hoc databases.

2.1. Av*ai*l*a*b*le s*peec**h** *da*t*a*b*ases* f*or* SV

As speaker recognition has been considered until today just a marginal field of speech technology, there are few public databases on this topic. Nowadays there is an increasing interest in SV, from both service providers and end-users. It is for this reason that in the last few years we had some databases [3] made for speaker verification goals. Here it is a list of the available ones, i.e. the databases utilized in the most important experiments.

TI*M***IT (***and* **NTI***M***IT).** Certainly this is the most famous database. Even if it was designed for speech recognition, it has been widely used also in speaker recognition [4]. Its telephonic version NTIMIT, has a detailed technical description. This is the only case of acoustic description, that we know, and it is devoted to describe the transformation of the original database in a telephone quality speech database.

KIN*G***.** It is the first database designed for speaker verification. It is also famous for the "great-divide", an effect related to some variations in the acquisition instruments. The effect is described in term of system performance, and not in relation to the characteristics of the speech signal (that is of course a more reasonable and interesting description). It contains monologues by 51 male speakers each divided into 10 sessions per speaker of short 1 minute duration.

YOHO. It contains a large scale, high quality speech corpus to support text dependent speaker authentication research. The data was collected under a US Government contract. It contains "combination lock" phrases of triple digits and other combinations by 186 speakers. In all there are 553 enrollment sessions, and 1380 trial sessions, with a nominal time interval of three days between sessions.

SPIDRE. This is a subset of the SWITCHBOARD collection, selected for speaker identification research, and with special attention to telephone instrument variation. It contains training and testing data for experiments in closed or open set recognition or verification. Combining the two sides of the conversations also permits speaker change detection, or speaker monitoring experiments.

POLYVAR. This is the first database designed to evaluate the intra speaker variability. It is also the first non English database. It has been collected in Switzerland, over the telephone line. The language is French.

POLYPHONE. It is the most ambitious and wide project. It consists of a 5000 speakers' voice, collected over the telephone line. It is a multi lingual project, so it will be available in several languages (including Italian). It was designed for speaker recognition purpose, but it will be very useful for speaker verification studies too.

COST250. The European project COST250 "Speaker Recognition in Telephony" promoted the realization of a speech database. IDIAP and EPFL, two Swiss institutes that take part in the project, manage the acquisition system and the collection. Each partner (about 13 countries) which joined the project, contributed with 10 speakers (5 male plus 5 female) calling 10 times, for a total of 100 calls per partner. The collection is just finished in May 1996. The database consists of digits and short sentences in English, and free speech in native mother tongue. It is, of course, a telephonic speech database.

2.2. Italian databases

There are only few public speech databases in Italian language: EUROM0, EUROM1 and AIDA. They are "high quality" speech databases designed for speech recognition aims. Other Italian collections are known: e.g. SIRVA, POLYPHONE and others realized in the European project SPEECHDAT. Finally there are databases collected by companies that produced commercial systems or services as DRAGON, IBM, etc. Usually these are reserved and so there are few information about them.

Our speech group was looking for speech material to provide enough data to develop speaker recognition technology, to provide evaluation of speaker recognition systems (both text-dependent, and text-independent), and to provide large curpus to study or to model phonetic variations. We did not find, among the available Italian databases, enough material that satisfies these requirements, so we started an internal project for collecting a large speech database over the public telephone network. A pilot collection was gathered in 1994 [5[, then the project was renewed and it is still a running activity of our group.

3. SIVA DESCRIPTION

The SIVA database consists of four speaker categories: male users, female users, male impostors, female impostors. Speakers were formerly advised via mail, and they were suggested to read carefully the information and the instructions before making the call. About 500 speakers were recruited using a company specialized in selection of population samples. The others were volunteers contacted by our institute.

Speakers access the recording system calling a 'toll free' number. An automatic answering system guides them along the three sessions that complete a recording. In the first session a list of 28 words (including digits and some commands) is recorded using a standard enumerated prompt. The second session is a simple unidirectional dialogue (the caller answers to the questions put by the system) where personal information is asked (name, age, etc.). At length, in the third session, the speaker is asked to read a continuos passage of phonetically balanced text that resembles a short "curriculum vitae". The physical system consists of a personal computer with housed a telephone interface board (a Dialogic D41/IT), and a software program written ad hoc for the present collection. The signal is a standard 8kHz sampled signal, coded using the American 8 bits mu-law format.

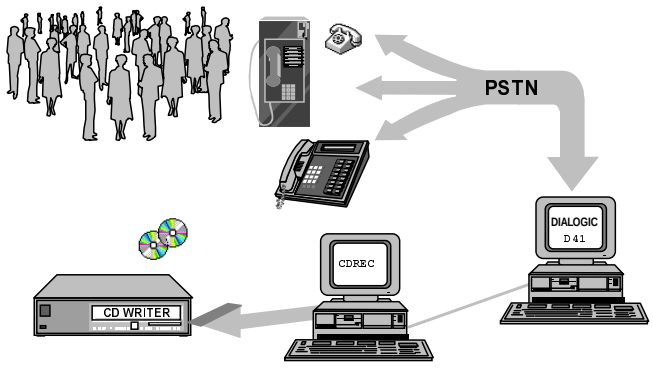


Figure 2: The SIVA database has been collected over the Public Switched Telephone Network (PSTN), using several types of telephonic hand sets. Selected material will be available on CD-ROM. The distribution policy is not yet defined.

3.1. The pilot collection (SIVA the Muser)

A first pilot collection started in spring '94 and ended in autumn of the same year. In that period we recorded the male users. All the acquired speech material has been checked by an operator who selected the error free calls. The result was a first CD that contains 18 repetitions of 20 male speakers calling from the entire national territory. This first database has already been used in several experiments, and distributed to the most important Italian laboratories working on speech technology.

3.2. Status of the SIVA database

At the beginning of 1996, after the experience of the pilot acquisition, we started the collection necessary to complete the SIVA database. Only two slight variations have been effected: the

silence-head and the silence-tail of each utterance have been augmented to prevent undesired truncation; in the second session some questions have been modified because they looked too intimate and so we supposed that many people would not answer. Up to now received calls are:

- MU: male users 18 speakers, 20 repetitions
- FU: female users 16 speakers, 26 repetitions
- MI: male impostors
  - ◊ 189 speakers, 2 repetitions
  - ◊ 128 speakers, 1 repetition
- FI: female impostors
  - ◊ 213 speakers, 2 repetitions
  - ◊ 107 speakers, 1 repetition.

3.3. F*u*t*ure ac*tiviti*es*

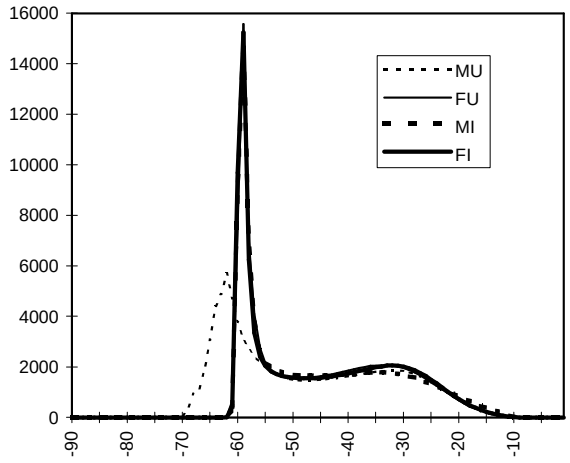
First of all we intend to complete SIVA reaching the initial goal of 500 male impostors, and 500 female impostors. In the mean time we are evaluating the possibility to create a special set of speakers with a high familiar tie (e.g. brother vs. brother, father vs. son, etc.), and another one of disguised voices.

4. SIVA ACOUSTIC CHARACTERIZATION

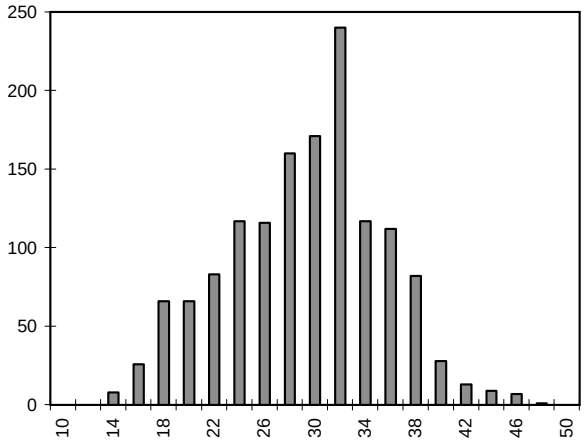
A speech database, may be characterized under a pure acoustical point of view. In fact it is a collection of speech *s*i*g*n*als*, and these *s*i*g*n*als* may be characterized objectively, i.e. without human decision by simple and well defined measures and algorithms. Definition and standardization of acoustical measures in speech, are generally available only for telephonic speech. Many of these may easily be translated to other kind of speech signal, but the main problem remains: which measures must be performed; using which instruments or algorithms; how the results should be grouped and reported; how to create a 'standard report' that will be easy to use and undertake a 'familiar look'. The answer to these questions is not a trivial task, and a definitive and comprehensive definition must be validate by the appropriate international commissions and institutes. In a former work [6] we addresses these problems, and we gave some indications on how to approach it. Here we utilize again this approach, and we show that a simple subset on measures can say a lot about the acoustical characteristic of the speech material that constitute a database.

4.1. Tim*e do*m*ain anal*y*s*i*s*

Let us consider the simplest measure in the time domain: energy. We compute the energy on the short term period, using a 256 points window (32ms). Then we graph the histograms of this measures for the four classes of speaker (see Figure 3). The position of peaks in the left side is related to silence energy level, the amplitude is related to the relative amount of silence per call. The difference between the pilot collection (MU) and the following ones, which have longer head and tail, is clearly evident as the peak for MU is about 6k, while for FU, MI and FI the peaks are near 15k. No relevant differences are present among the peaks on the right side. These are related to the speech level, and speech total duration per call, that look quite stable along the database recording.



Fig*ure* **3:** The histograms of energy distribution of MU (male users), FU (female users), MI (male impostors), FI (female impostors). The abscissa dimension is dB, the ordinate is total count of segment at the given energy of the represented class.



Fig*ure* **4**: Distribution of the Signal to Noise Ratio (SNR) of the whole SIVA database (about 2k calls). The abscissa is the SNR value in dB, the ordinate is total counts.

Using the same measure we can evaluate a second characteristic of the speech signal: the signal to noise ratio (SNR), that is an indicator of the quality of the transmission channel and of the background noise. The material under study has a mean SNR around 32dB, that is a typical value for telephonic speech. The

distribution is uni-modal, i.e. it has only one peak, so at a first glance the overall speech quality is reasonable.

4.2. Frequency domain analysis

Also for frequency domain analysis we choice the simplest measure: the mean power spectrum. The results are shown in figure 5. The shape for users and impostors of the same gender are quite similar. On the contrary we find two differences between male and female mean power spectrum. The shape for the female power spectrum under 400Hz is much more rippled than the male one: this is probably due to the female higher pitch and its convolution with the spectrum shape. It is also evident a clear difference between male and female spectrum shape in the zone of 500Hz and 3000Hz, where the two shapes assume opposite inclination. We can not give a ultimate justification to this fact, it is probably originated by the different formant values of male and female voices. Generally speaking the power spectrums look regular and comparable in both shape (according to the former notes) and amplitude; so we can say that no spectral anomalies are found in the database.

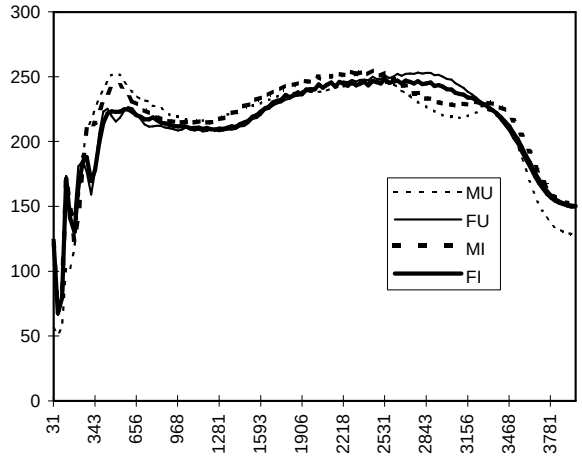


Figure 5: The mean power spectrum of MU (male users), FU (female users), MI (male impostors), FI (female impostors). The abscissa dimension is Hertz, the ordinate is dBov*10.

4.3. The reference system

The use of a reference system is recommended to check and to validate speech material. We used a text independent speaker verification algorithm based on the similarity measure of covariance matrices, similat to the one proposed in [4]. We only use the read passage (continuous speech) of the database. The first 30 seconds are used for training, the following 30 seconds are used for testing. The obtained results, shown in table one, have a general trend that is conform to our expectation, i.e. the Equal Error Rate (EER) become smaller and smaller, when the training (or test) material become longer and longer. Female speakers show a performance score much worse than male speaker. At moment we have no justification to this fact, unless the trivial one that female user's population (FU) have a broader intra speaker variability and/or that female impostor's population (FI) have a broader inter speaker variability in relation to the male ones.

		TRAINING					
		MALE			FEMALE		
	TIME	**10*s***	**20*s***	**30*s***	**10*s***	**20*s***	**20*s***
TEST	**10*s***	15	13	12	22	20	18
	20*s*	10	9	8	18	17	16
	30*s*	9	8	7	17	16	14

Table 1: Equal Error Rate (EER) values for male and female speakers using the reference system.

5. CONCLUSION

We have introduced the Italian SIVA telephonic speech database. A description of the database, of the collected speech material and of the quality of the telephone calls is reported, as well the results obtained using a reference speaker verification system.

6. ACKNOWLEDGMENTS

Authors would like to thank CSELT Institute that contributed to the realization of the SIVA database, and that sponsored the recruit of 500 speakers. We would also thank everyone that called the automatic recording system.

7. REFERENCES

1. Bimbot, F., Chollet, G., Falcone, M., "The Assessment of Speaker Recognition System", ESPRIT P.6819 (SAM-A), first year progress, report N.9

2. Di Carlo, A., Falcone, M., Paoloni, A., "Corpus Design for Speaker Recognition Assessment", proceedings of the ESCA Workshop on 'Automatic Speaker Recognition, Identification, Verification', Martigny, 1994, pp.47-50.

3. Godfrey, J., Graff, D., Martin, A., Pallet, D., "Public Databases for Speaker Recognition and Verification", proceedings of the ESCA Workshop on 'Automatic Speaker Recognition, Identification, Verification', Martigny, 1994, pp.39-42.

4. Bimbot, F., Magrin-Chagnolleau, I., Mathan, L., "Second-order statistical measures for text-independent speaker identification", Speech Communication 17 (1995), pp.177-192.

5. Contino, U., Falcone, M., "SIVA the Muser: un Database Vocale per il Riconoscimento del Parlatore", atti del XXIV convegno Nazionale della Associazione Italiana di Acustica, Bologna, 1995, pp.145-150 (in Italian).

6. Falcone, M., Contino, U., "Acoustic Characterization of Speech Databases: an Example for the Speaker Verification", proceedings of the 'International Congress on Phonetic Science', Stockholm, 1995, pp. 290-294.