

AN OVERVIEW ON RECOGNISER TESTING ACTIVITIES PERFORMED IN CSELT

F. Canavesio, G. Castagneri, G. Di Fabrizio, F. Senia

*CSELT - Centro Studi e Laboratori Telecomunicazioni S.p.A.,
via G. Reiss Romoli 274, I-10148 Torino, Italy*

tel: + 39 11 228 5111 - fax: + 39 11 228 6207 - e-mail: URCF@UZ6000.CSELT.STET.IT

ABSTRACT

The goal of this paper is to describe the testing activities on telephone speech recognisers performed in CSELT during the last year. Both commercial devices and laboratory prototypes have been widely tested in laboratory following the specification developed in the ESPRIT Project SAM.

For this purpose speech databases, collected over the Italian Public Switched Telephone Network (PSTN), have been used.

Recently, a new facility developed in CSELT, allowed on-line testing of recognisers during the collection of speech database. These data have been used to verify the laboratory test setup.

Keywords: speech database, data collection, speech recogniser, laboratory test, field test.

1. INTRODUCTION

The growing availability of speech recognisers designed for the telephone network environment and the whole range of applications made available with the new audiotex services, has increased the need of systematic, reliable and repeatable tests.

Recognisers used for these applications are speaker independent, with small vocabularies (digits, alphabet, command words, ...); this class of devices is addressed in the following.

The telephone environment introduces many variability sources as:

- speakers (skill, dialect, age, motivation, ...)
- background noise sources (radio-tv, factory, car engine, traffic, ...)
- channel noise
- microphones
- network devices (coder, echo canceller, ...).

These factors can hardly be controlled during the collection of speech material, for this reason telephone databases should

be large enough to ensure a proper representation of the above mentioned factors.

The telephone network itself it is not a stable entity; on the contrary it grows and tends to be improved changing continuously; it results that a telephone speech database, representing a "still picture" of the network at a given time, can have a time-limited validity.

On the other hand, it is not trivial to reproduce exactly the situation of the collection in laboratory. Even if digitally recorded data are used, a number of discrepancies between field and laboratory setup can hardly be avoided.

The need of reducing the gap between field and laboratory tests and to enhance the correct estimate of the performance of the recognisers, has been the basis ground of most of the testing performed in CSELT in the last period. For this purpose, a new recording workstation has been developed to allow recogniser testing during the collection of SIRVA, a large telephone database. The results of these tests are compared with those obtained in the laboratory. Two recognisers, designed to work in the telephone environment, have been tested.

In order to have a complete representation of the performances of a recogniser, a definite set of test condition has been defined covering the following aspects:

- recogniser robustness toward signal level;
- type of call (local or long distance);
- word presentation mode (non pre-segmented words, pre-segmented words or full sequence of isolated words).

2. SPEECH DATABASES DESCRIPTION

The recognizers have been tested using two different speech databases: COLLECT and SIRVA. They have been collected in digital form on the Italian PSTN, on local and long distance connections, with a variety of live conditions and with speakers covering different regional accents. The vocabularies include the italian digits; this subset has been used for laboratory and field test.

COLLECT consists of more than 1000 speakers recorded in the 1987, from local and long distance calls (50% each). This characteristic is well documented in fig. 1 that shows the bimodal distribution of the speech level computed for each file of the database. This method was used for specifying the dynamic range of the signal [1].

SIRVA has been collected in 1992 from more than 2000 speakers by toll-free lines over the Italian PSTN [2]. Calls were equally distributed all over the country; the speech level distribution is shown in fig.2. Here the distribution curve fits approximately a Gaussian.

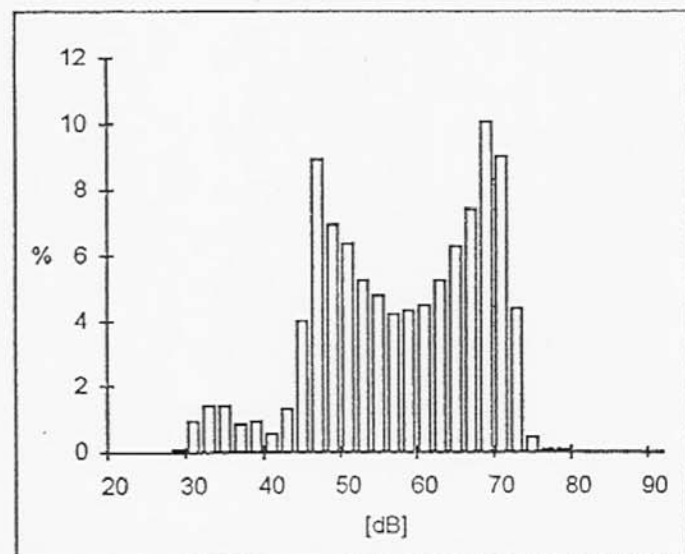


Figure 1 - COLLECT speech level distribution

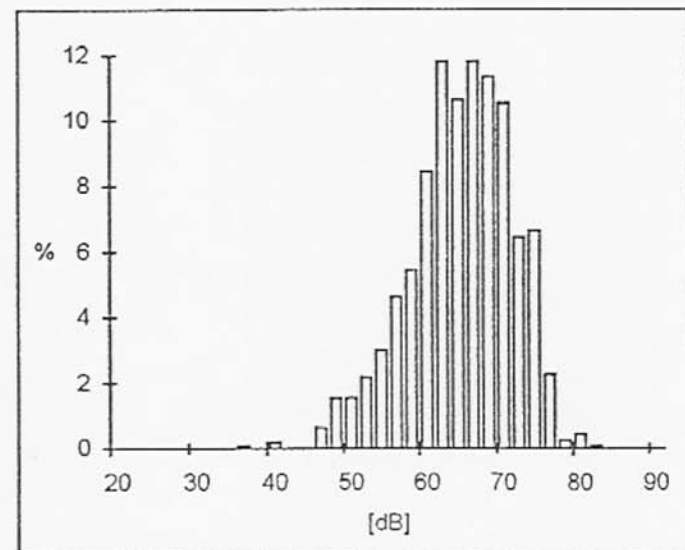


Figure 2 - SIRVA speech level distribution

Comparing figs.1 and 2, the improvement occurred on the Italian telephone network can be noted. Speech samples collected in SIRVA from long distance call present a speech level distribution with a higher mean value and a smaller variance than those recorded in COLLECT.

A peculiarity of SIRVA is the availability, for the DIGIT subset, of the result of a speech recognizer, activated

by each recorded word. That was possible as two recognizers were connected to and controlled by the recording workstation. They received as input the same audio signal recorded in the database and provided on-line test results. This characteristic of SIRVA enables the use of this database to calibrate the laboratory testing environment, by verifying the congruency of results obtained on the same speech material, both on field and in the laboratory.

3. DESCRIPTION OF THE RECOGNIZERS

Two external recognizers have been used; they have similar characteristics even if they are implemented on different hardware.

Recognizer A is a speaker independent system trained with telephone material on the Italian digit vocabulary. It is a commercial device designed to work with a widely available telephone interface. It can manage up to four telephone channels but during the test only one of them was active.

Recognizer B is a speaker independent prototype developed on a commercial general purpose floating point DSP board, hosted on a PC. It has been trained using the Italian digit of a telephone speech database. Both the systems are equipped with a programmable input attenuator.

4. TEST RATIONALE

Each recognizer has been evaluated in three different stages:

STAGE 1 - First Laboratory Test

During the first stage the recognizers have been evaluated using the test subset of the database COLLECT (500 speakers). Tests have been repeated at different attenuation levels, to study the interaction of this factor with the system performance.

Two methods have been used to test recognizers [3][4]. In the Isolated Mode Test (IMT) single words, narrow segmented, have been played to the recognizer and system answers were collected word by word. If no answer was received within a preset timeout period the system skipped to the next word. In the Continuous Mode Test (CMT) a complete file was played and responses were collected as they arrived.

STAGE 2 - On-Line Field Test

In this stage results of the recognizers connected to the recording workstation have been collected on line during the speech database acquisition. The devices were fed with speech data directly uttered by the speakers in the acquisition window.

STAGE 3 - Second Laboratory Test

During this stage the speech material collected in the previous stage, i.e., the digit subset of the SIRVA database,

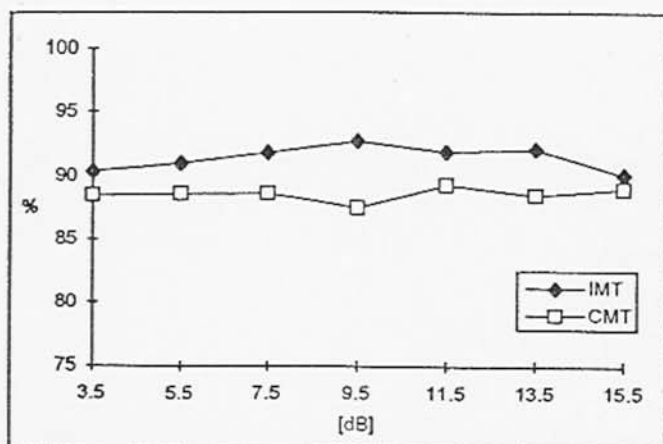


Figure 3 - Recogniser A, Stage 1, Long Distance Calls correct performances vs. attenuation level

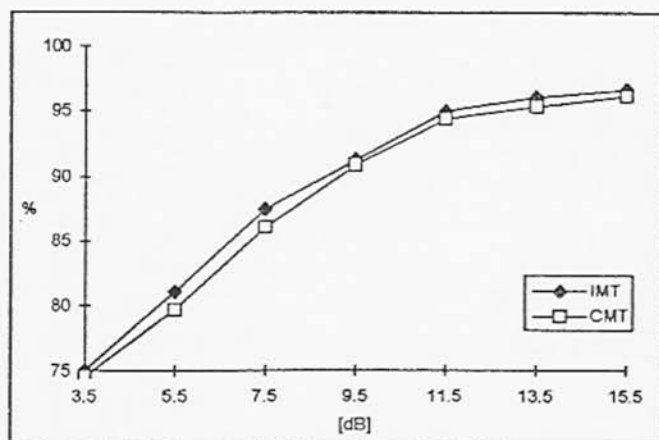


Figure 4 - Recogniser A, Stage 1, Local Calls correct performances vs. attenuation level.

has been used to test the two recognisers in the laboratory. Also, here the test has been repeated at different attenuation level. The recognisers were tested by playing the speech material recorded on digital form, using the same acquisition windows of the stage two, i.e., with non-segmented speech and data.

5. TEST

Two different test setups were adopted. In stage 1 and 3 the test workstation SESAM, developed in the SAM ESPRIT Project, was connected with the recognisers directly or by a telephone bridge, depending on the characteristics of the system under test. This setup has been used for all the tests performed with stored speech material.

The second setup was based on the workstation TESCOS developed for the collection of telephone speech databases. TESCOS is PC based and it can be connected to a speech recogniser by RS232 port or by internal BUS. Its main features include the control of the recogniser, the collection of its responses and the creation of the related response files, formatted according to the SAM file specifications [5].

Rec.	Calls	Mode	Att	Corr.	Sub.	Miss
A	Long dist.	IMT	9.5	92.7	3.1	4.2
		CMT	15.5	89.0	3.4	7.6
	Local	IMT	15.5	96.7	2.8	0.5
		CMT	15.5	96.2	3.0	0.8
B	Long dist.	IMT	7.5	97.1	2.0	1.0
		CMT	12.0	90.5	3.6	5.9
	Local	IMT	18.0	97.8	1.4	0.8
		CMT	25.5	93.8	3.0	3.3

Table I - Stage one results

During the recording of the SIRVA database the two recognisers have been connected to the recording workstation in different stages of the collection. They were activated when the speakers were uttering the list of digits. If any procedure error (speech over the initial beep, no speech signal, too low level or saturation of the speech signal, too high level of background noise, ...) was detected during the test of the recognisers A, the unused utterance was discarded and the recogniser result was not considered. That is not true for recogniser B and miss errors can be found in its performance results.

Test results

During the first stage the speech database COLLECT was used. To evaluate the influence of the speech level over the recogniser performance the test has been repeated at different attenuation level. The same test has been repeated using both the isolated and the continuous mode [6]. The best results are reported in table I and the figs. 3 - 4 show the performances vs. attenuation level, obtained using the Local Calls subset. Results achieved using the continuous mode are always lower than the ones obtained with the isolated testing mode in all the performed tests. Miss errors are generally increased but also substitutions occur more frequently with the CMT method.

In the second stage each recogniser has been tested during the recording of about 500 speakers of SIRVA. The input gain has been adjusted to optimize the recogniser performances, considering the results obtained by the COLLECT database. As Recogniser A was embedded into TESCOS and was responsible of the speech detection, no miss errors could occur in this stage; "missed word" were lost both for the acquisition and for testing and speakers were reprompted to repeat the word.

During the third stage, the recogniser has been tested with the same speech material of stage two, using the SESAM workstation. The global results of stage two and three are compared in table II.

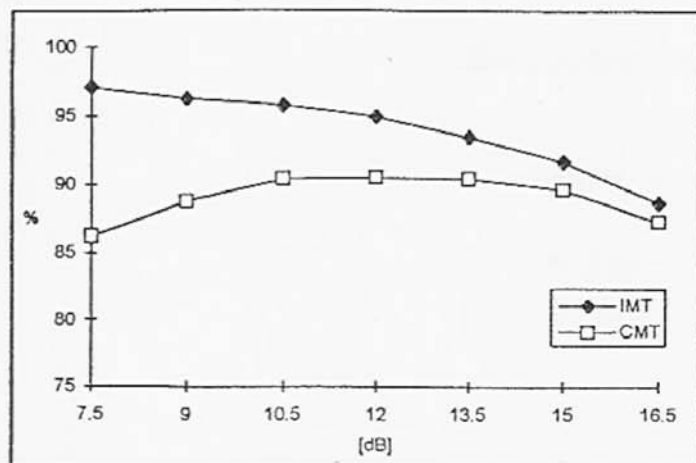


Figure 5 - Recogniser B, Stage 1, Long Distance Calls correct performances vs. attenuation level

6. DISCUSSION

Measurements performed in this work clearly show that COLLECT and SIRVA telephone speech databases are different. The differences are due to changes happened on the PSTN and the distance between local and long distance calls is reduced.

Both the recognisers are quite sensible to the speech level as highlighted in the first stage tests. This characteristic made difficult to exactly reproduce in laboratory the same situation obtained in a live test; in fact the results obtained in stage 2 and 3 are similar but not equal. The difference is smaller for recogniser B. It probably depends on the test setup. Recogniser B has a LINE input that can be directly connected to the D/A board of the testing workstation. The speech signal, in this case, should not be affected by any distortion.

Recogniser A is connected to a telephone interface board and cannot be directly linked to the test workstation. In this case the test setup is more complicated because the signal is sent to the tested system by a telephone bridge and an external attenuator. All these devices can introduce not documented signal distortion that can justify the differences in the global scores.

7. CONCLUSIONS

This work was focused on the comparison between laboratory and field tests. TESCOS workstation, allowing the on-line recogniser testing during a speech database collection, gives the unique opportunity of characterizing a laboratory

Rec.	Setup	Corr.	Subst.	Miss
A	field	96.5	3.5	-
	lab.	95.1	4.7	0.2
B	field	92.9	3.7	3.4
	lab.	93.6	3.6	2.8

Table II - Test results obtained in stages two and three

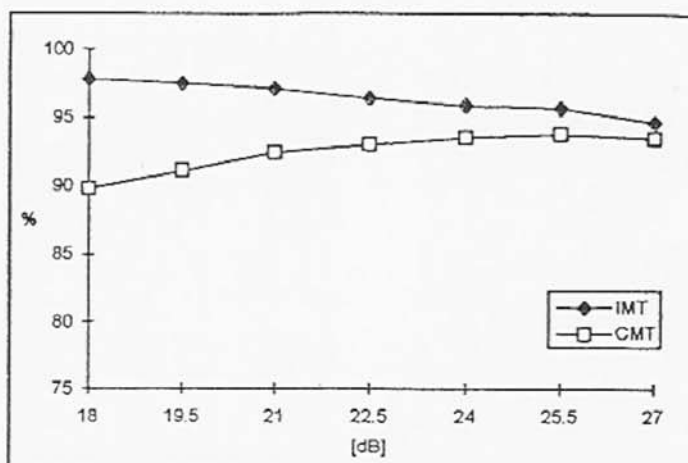


Figure 6 - Recogniser B, Stage 1, Local Calls correct performances vs. attenuation level

test environment.

The recogniser sensitivity to the signal level and to the audio channel make quite difficult to reproduce, in laboratory, the same situations occurred during the speech collection. This is particularly true when the tested device cannot be directly linked to the test workstation.

The availability of test workstation specially designed for telephone recogniser could help to solve the main problems (level adjustment, impedance adaptation, ...). A substantial improvement is possible sending directly PCM A-law signal to the recogniser. Unfortunately at the moment only a small subset of devices allows this test modality.

REFERENCES

- [1] *Danielsen S., Velden J.G van*: "Speech Level Meter" in: "User Guide to Input Assessment" Esprit SAM document SAM-UCL-G005
- [2] *G. Castagneri, G. Di Fabrizio, A. Massone, M. Oreglia*: "SIRVA - A large speech database collected on the italian telephone network" in: EURO_SPEECH '93 proceedings.
- [3] *G. Castagneri, G. Di Fabrizio, F. Senia*: "SamPac v 3.10 Documentation" in: "User Guide to Input Assessment" Esprit SAM document SAM-UCL-G005
- [4] *F. Canavesio, G. Castagneri, G. Di Fabrizio, F. Senia*: "Comparison between two methodologies of testing isolated word speech recognisers" in: proceedings ICSLP '92
- [5] *G. Castagneri, G. Di Fabrizio, M. Oreglia, A. Massone*: "TESCOS - an integrated workstation" in proceedings AVIOS '93 (in progress)
- [6] *F. Canavesio, G. Castagneri, G. Di Fabrizio, F. Senia*: "Comparison of two methods of performance assessment of commercial Speech Recognition Systems operating in telephone environment" in: CCITT contribution COM XII- -E May 1993