COMPARISON BETWEEN TWO METHODOLOGIES OF TESTING ISOLATED WORD SPEECH RECOGNIZERS

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

CSELT - Centro Studi E Laboratori Telecomunicazioni
Via G. Reiss Romoli 274, 10148 Torino, Italy
tel: + 39 11 228 5111 - fax: + 39 11 228 6207

ABSTRACT

The goal of this paper is to compare two different methodologies for automatic testing of isolated word speech recognisers. They are usually tested by playing single speech tokens and collecting the device response.
The second approach, allows to output speech files containing lists of words and simultaneously to collect recogniser answers. It ensures a more realistic testing of the performance of any kind of isolated word speech recogniser. The results obtained are significantly lower than the ones got with isolated word testing.
Furthermore many recognisers, designed to work in noisy environment, cannot be correctly tested in isolation as they need to be continuously fed by background noise.

I. INTRODUCTION

The laboratory testing of speech recognisers originates methodological problems and practical issues.
The need of reducing the gap between field and laboratory test to enhance the correct estimates of the real performance of the recognisers belongs to the first category.
On the other hand the methods must be completely automatized as the testing itself is very time consuming and the manual scoring is practically impossible as thousand of speech tokens are needed to obtain reliable results.

The goal of this paper is to compare two different methodologies for automatic testing of speech recognisers.
Isolated word recognisers are usually tested by playing single speech tokens and collecting the device responses.
A second approach, implemented in the SAMPAC assessment software [1] developed in the framework of the SAM Esprit Project 2589, allows to output speech files containing lists of words and **simultaneously** to collect recogniser answers.
This method allows to test recognisers that implement AGC algorithms and/or background noise level adaptation techniques. In fact tests are performed with continuous voice portions consisting of sequences of discrete words embedded in *real background noise*.
The ultimate objective is to avoid both unrealistic testing of the end-point capabilities of the device (due to the presentation of pre-segmented words) and experimental bias generated by improper stimuli presentation.
As part of the SAM project, wide tests on commercial recognisers have been performed. Their scope was primarily to test the methodology and highlight problems due to the testing procedure itself, more than to assess the performances of the devices.

II. SPEECH MATERIAL

Two kinds of databases have been used in this comparison:

- clean speech collected in controlled conditions;
- noisy speech collected on the public switched telephone network from real customers.

II.1 High Quality Speech Data

A subset of the Italian section of the European Speech Database EUROM_1 [2] has been used. This speech material has been collected in anechoic room using strictly controlled procedures.
The signal files were recorded using an high quality microphone and digitalized with a sampling frequency of 20 KHz by an OROS AU22 board mounted on PC.
The subset used in the test consists of the corpus "NUMBER" (100 different numbers extracted among the numbers between 0 and 9999).
It is divided into two parts:

- the **Many Speaker Set**, consisting of 51 speakers that recorded one repetition of the corpus; it has been used to train the recogniser;
- the **Few Speaker Set** containing six repetitions of the corpus uttered by 10 more speakers. It has been utilized as test set.

II.2 Telephone quality speech data

This speech database has been collected on the Italian Public Switched Telephone Network, both from local and long distance calls, using 1065 real customers that uttered once the Italian digits and 5 command words.
Only the Italian Digit Vocabulary has been used in this experiment.
The whole database is labelled at word level and every noise, of whatever origin, has been classified.
The database has been randomly split into two sets, the training set, containing 671 speakers and the test set, containing 394 speakers. The characteristics of the two sets are quite similar as shown by the Signal/Noise Ratio distributions presented in Fig.1 and Fig.2.
Using the segmentation of words and noises, two different test sets have been produced:

- a **clean test set** that contains only portions of speech files without any kinds of identifiable noises (excluding the telephone network background noise);
- a **noisy test set** that includes also all words and portions of files that are corrupted by any kind of acoustic events like noises due to the speaker (breathe, cough, ...) telephone interferences (dialling, clicks, ...) or environmental noises (TV set, voices, ...).

III. METHODOLOGY

The test workstation SESAM developed in the SAM Esprit Project 2589 has been used to implement and automatize the

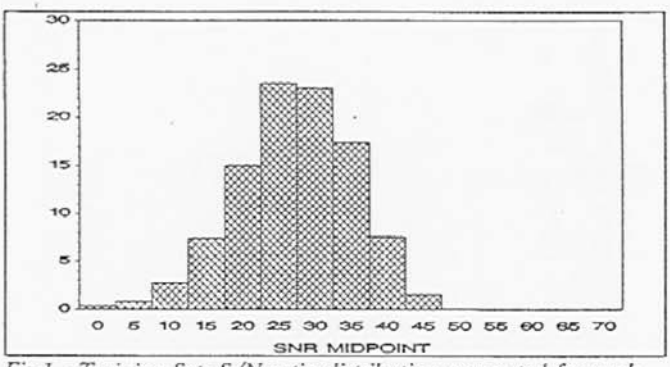

Fig.1 - Training Set: S/N ratio distribution computed for each word.

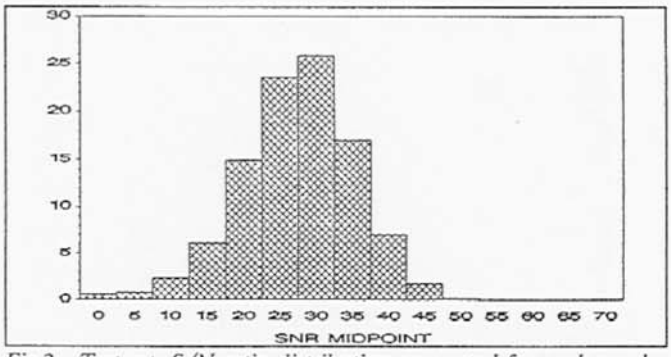

Fig.2 - Test set: S/N ratio distribution computed for each word.

methodology used in this experiment. The testing software SAMPAC, developed in CSELT, is device independent and allows to implement the same methodologies for all the automatically controls every recogniser function like configuration, training, adaptation and test.

Two approaches have been followed to test recognisers.
In the Isolated Mode single words are sent to the recogniser and the system answer is collected word by word. If no answer is received within a preset timeout period, the system skips to the next word.
The test sequence is fixed: start the recogniser, word playback, result collection and stop the recogniser. This sequence is repeated for each stimulus of the test. This method is the most obvious one, but it can generate a number of experimental artifacts that bias the test results.

In fact if we examine more closely the acoustic signal received by the recogniser, when the system is activated, a true zero signal is observed, till the workstation is able to play the appropriate word. This temporal gap depends on the speed of the mass memory device that contains speech files and can last from few millisecond to some seconds. During this period the recogniser activates all its signal analysis processes (AGC, noise adaptation, signal tracking, ...) on an *artificial* silence.
Then the recogniser receives the speech token. Depending of the closeness of labelling and of the level of the signal background noise, a number of different situations can happen, none of them reflecting the real state of the recogniser during normal operations.
It is impossible to describe exactly all the biases that the method can introduce, they are highly dependent on the recogniser implementation. The most affected capabilities are the end-point algorithm, the automatic gain control and the noise adaptation procedure.
Additionally, some databases have been recorded with DC-offset far from zero. That can originates clicks when the first sample is played, interfering with the normal functioning of the recogniser.

In the Continuous Mode a complete file is played and responses are collected as soon as they arrive. In this case, after the start, the recogniser runs continuously sending back the answers to the testing workstation which also records the exact time when it receives each response.
The signal fed to the recogniser is the same that has been digitally recorded, containing all the extra-phonetic phenomena (noises, pauses,...) that normally happen uttering a string of words. In this case the testing is more realistic and the results obtained should be more correlated to the ones got in a field test of the device.

A simple alignment algorithm is used, in this experiment, to build the right Stimulus-Answer sequence; this is the main problem of the methodology, but it can be overridden using speech files with pauses between two stimuli longer than the recogniser response time, to avoid displacement in the sequence.

IV. RECOGNISERS

Two commercial recognisers have been tested using the two different methodologies.

One recogniser can be trained on speaker independent vocabularies up to more than 100 words. It has been tested both with the Italian digits and with a 100 words vocabulary (numbers).

The second has a fixed vocabulary of ten words (Italian digits).

IV.1 Recogniser A

The recogniser A is a system designed to be plugged in a IBM PC or compatible system. It consists of a single board and relies on a template representation of speech; it can work both in speaker dependent and speaker independent modality, after appropriate training.

During the training, the recogniser creates a single composite template for each word in the recognition vocabulary by averaging, vector by vector, all training utterances for that word. In addition to the mean value, the standard deviation is also calculated for each feature.

A procedure for forced adaptive training is provided. In this case, templates are modified during the recognition process if a rejection occurs, so the recogniser can be adaptively trained during use.

IV.2 Recogniser B

It is a self-contained speaker-independent voice recognition and telephone line interface system. It is composed of a processor board in the form of a PC peripheral together with a piggy-back communication board. The processor board employs an Intel 80186 microprocessor and a Texas Instruments TMS320C25 signal processor to perform all controls and voice processing functions. The communication board provides an RS-232 data line as well a telephone line interface.
This system utilizes speaker-independent technology for recognition of speech over local and long distance telephone calls; it is provided by AGC features and noise adaptation capability.

V. EXPERIMENTAL TESTS

RECOGNISER A

V.1 Test I

The first test has been performed using the recogniser A and the high quality database, containing Italian numbers. The following paragraphs explain in details the procedure used in this test.

V.1.1 Training the recogniser The recogniser has been trained using the numbers contained in the Many Speaker set of the italian part of EUROM_1.
The training procedure consists of three steps:

1) speech material is collected through the PCM CODEC of the recogniser and stored in files;
2) recognition vocabulary is built from the stored speech material;
3) recognition vocabulary is adapted by manipulating the training material. This has been repeated three times. Each adaptation step lasts around 25 hours.

The training set size (51 talkers) is congruent with the requirements of the device user manual.

V.1.2 Recogniser Testing. The test has been performed using the Few Speaker set of EUROM_1 containing six repetitions of the corpus "NUMBERS" uttered by ten speakers not included in the training set, for a global amount of 6000 tokens.
The active vocabulary was of 100 words (the numbers).
The performance of the recogniser has been tested in 3 steps:

1) after the creation of the vocabulary
2) after the second step of adaptation
3) after the third last step of adaptation

The tests have been performed both in isolated and in continuous mode.

V.1.3 Running the tests The training, the test and the scoring control files were automatically generated by the RISE program [3]. The SAMPAC v3.10 software has been used to test the recogniser.
The scoring results have been obtained using SAM_SCOR v3.0 software [4].
Each test lasts approximately 5 hours.

	ISOLATED			CONTINUOUS		
Step	Correct %	Miss %	Subst %	Correct %	Miss %	Subst %
0	87.1	0.52	12.4	83.1	0.0	16.8
2	90.3	0.51	9.1	85.5	0.0	14.4
3	90.6	0.51	8.8	85.5	0.0	14.4

Table I - Test results - Recogniser A, Italian number database.

V.1.4 Test Results Table I shows the results obtained by the multi-step adaptation procedure. They highlight the effect of adaptation applied to the training phase.
Overall performance rises in the first steps passing from 87.1% without any adaptation to 90.6% with 3 adaptation steps for the isolated testing and from 83.1% to 85.5%, for continuous testing.
After the third step of adaptation no more improvement is registered.
Fig. 3 shows a constant difference between the Isolated Mode and the Continuous one, for each step of adaptation. The substitution errors rise while no differences are reported on the missed words. The high S/N Ratio of this database justify the low number of missed words.
The testing mode interacts probably with the accuracy of the end-point of the recogniser. In the Continuous Mode the system probably fails more in extracting the whole portion of signal containing the word, enhancing the probability of a substitution error.
The continuous test is more severe with this high quality database; the ANOVA computed on the test results confirmed the significance of the difference between the testing modes.

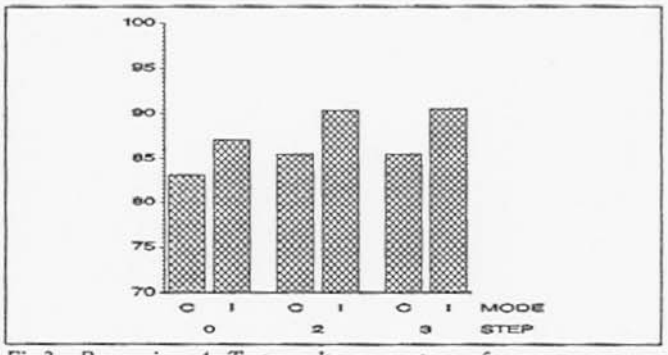

Fig.3 - Recognizer A: Test results, percentage of correct recognition.

V.2 Test II

The second test has been performed using the database of the Italian digits collected on the public telephone network.
The same training procedure was adopted, only the speech material and the number of adaptation steps (12 instead of 3) were changed.
The test has been performed using two of the subsets (long distance call, noisy and clean) above specified.

	ISOLATED			CONTINUOUS		
Database subset	Corr. %	Miss %	Subst. %	Corr. %	Miss %	Sub. %
Clean	94.4	0.0	5.6	81.3	3.7	15.0
Noisy	94.3	0.1	5.6	78.4	4.5	17.1

Table II - Test results - Recogniser A, Italian digit database (long distance).

V.2.1 Test Results The difference between the results obtained using isolated and continuous mode is dramatic (see table II). The score of 94.4 % achieved with this difficult database were quite good while the result of the continuous testing are far from acceptable.
The presentation mode have stronger effect than the presence of extra-phonetic events; in fact the difference between continuous and isolated mode is larger then 13% in term of correct recognition, while the difference between the two data set, clean and noisy is only appreciable in the continuous mode and is around 3% in term of correct recognition.
Both substitution and miss errors rise; it demonstrates that the end-point of this recogniser is substantially facilitated by the

isolated presentation of the speech signal. When stimuli are presented in isolation, but surrounded by real noise, as it happens in the continuous mode, the end-point looses both in precision of extracting the speech signal and in ability of detecting the presence of the signal.

RECOGNISER B

This recogniser was specially designed to work with the Italian telephone digits uttered in isolated mode.
It is a speaker independent system and is trained once directly from the manufacturer.
The test procedure followed was kept as close as possible to the one used with the recogniser A and described in the previous paragraph.
Tests have been performed using the four subsets (local and long distance calls, noisy and clean data) above specified. They have been repeated 5 times to counterbalance results variations.

	ISOLATED			CONTINUOUS		
Database subset	Corr %	Miss %	Subst %	Corr %	Miss %	Subst %
Long-distance clean	93.5	4.3	2.3	91.5	5.8	2.7
Local clean	92.5	4.9	2.6	91.8	5.4	2.8
Long-distance noisy	93.5	3.9	2.6	89.8	7.3	2.8
Local noisy	92.2	4.7	3.0	90.9	5.7	3.3

Table III - Test results - Recogniser B, Italian digit database.

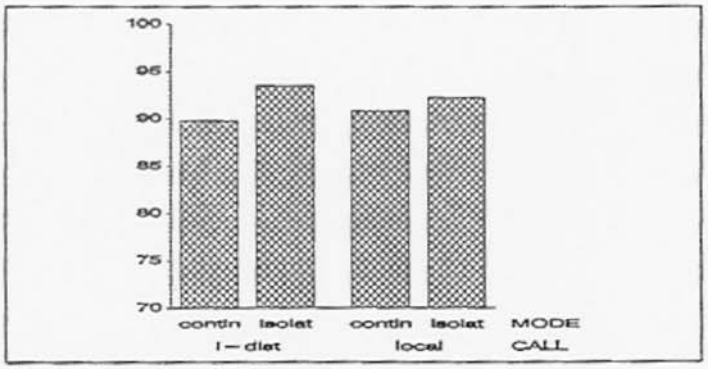

Fig.4 - Recogniser B. Test results, percentage of correct recognition.

V.3 Test Results
Table III summarizes the results obtained with recogniser B. The differences between the two methodologies are highlighted in Fig.4.

Results achieved using the Continuous Mode are lower than the ones obtained with the Isolated Mode with all the digits database subsets used. The end-point of this recogniser is not tested correctly with the Isolated Mode method; it is demonstrated by the rising of the miss errors using the Continuous Mode method.

The pre-segmented material presented in isolated mode leads to an overestimates of the device capability of detecting a voice signal in a background noise. The ANOVA performed on the test results confirmed the significance of the difference between the testing modes.

It is also important to highlight the differences between the clean and the noisy test sets.

In the Isolated Mode the same results are achieved for the two sets. A decrease of performance is registered if the noisy database is used in the continuous mode. The presence of extra phonetic events affect both the end-point and the recognition accuracy of the system.

V.4 DISCUSSION AND CONCLUSIONS
The behaviour of the recognisers is very different in the two modes.

Generally speaking the Isolated Mode seems to overestimate the recogniser performances. The main limit of the this method is that the recogniser receives end-pointed words surrounded by *artificial deep silence*. So, the end-point algorithm of the recogniser has its duty facilitate far from the normal operation condition. That is particularly true using noisy signals, but the same trend has been demonstrated with high quality signals too. In this case, even if the recogniser do not fail in detecting the word, it probably looses in end-point accuracy, generating an higher number of substitution errors.

Furthermore this method do not permit to analyze the robustness of the recogniser against noises and how these events interfere with the correct detection of speech.

The Continuous Mode allows a more realistic testing of the performance of any kind of isolated word speech recognisers. The results obtained are significantly lower than the ones got with Isolated Mode but probably closer with the scores obtained in field operations.
Moreover, Continuous Mode allows to avoid experimental artifacts like noise adaptation or automatic gain control on artificial and improper signal windows, or click and bursts due to the activation and deactivation of D/A conversion board on non zero signals.

References

[1] Castagneri G., Di Fabbrizio G., Senia F., "SAMPAC 3.10 - Documentation", Esprit Project 2589, Report SAM-CT-120, February 1992.

[2] Castagneri G., Vacchetta L., "Documentation of the Italian EUROM_1 Database", Esprit Project 2589, Report SAM-CT-122, March, 1992.

[3] G. Castagneri, F. Senia, "SAM-RISE V1.1 - User's Guide Relational Interface for Speech Evaluation", Esprit Project 2589, Report SAM-CT-105, 19 September, 1990.

[4] Lindberg B., Andersen O., Jorgensen R.K., Danielsen S.W., "SAM_SCOR V.3.0 Reference Guide", Esprit Project 2589 Report SAM-IES-062, 9 December, 1991.