SIRVA - A LARGE SPEECH DATABASE COLLECTED ON THE ITALIAN TELEPHONE NETWORK

G. Castagneri, G. Di Fabbrizio, A. Massone, M. Oreglia

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

ABSTRACT

speech database was collected over the Italian Public witched Telephone Network from more than 2000 lephone customers.
n automatic workstation has been developed in order to erform the collection; it was connected to two speech cognizers that have been tested on-line during the tabase collection.
esults and solutions to different problems faced during the llection are reported.

eywords: Speech database, recogniser training, cognizer test.

1. INTRODUCTION

The collection of a large speech database is not a ere technical problem but requires different kind of nowledge and it is an occasion to find pragmatic solutions practical and theoretical problems.

Furthermore, telephone speech databases involve culiar problems due to the characteristics of the audio annel and to the lack of control over the speaker haviour. For this reason the recording workstation should carefully designed according to the network aracteristics and the specific task. Better performances of e workstation in inferring what is happening from the put analysis, will decrease the final data discard rate.

In the second half of the 1992 a country-wide speech tabase of nearly 3000 calls has been collected over the alian Public Switched Telephone Network for training and sting speech recognition systems.

The speech corpus was composed of 68 isolated ords and four strings of connected digits. An exhaustive port of the experiences gathered during this collection is ported in the following.

2. COLLECTION SET-UP

Speech samples were collected by the automatic TElephone Speech COllection System TESCOS, composed of three PC workstations connected via LAN. Two workstations were assigned to the speech collection and equipped with:

- a Telephone Interface Board Dialogic D/41D,
- an isolated words recogniser VR/40,
- a high quality 16 bit acquisition board OROS AU21,
- an Ethernet board
- a four channel programmable attenuator for level adaptation.

The third workstation performed remote maintenance and monitoring.

The system was able to answer the call, to play messages and prompts (PCM 16 bits), to recognise DTMF digit sequence for caller identification and to collect speech at 8000 Hz of sampling rate. The signal was low-pass filtered with a telephone bandwidth filter at 3.4 kHz. The gain of the acquisition channel was held constant at a value that maximised the input dynamic.

TESCOS controlled the acquisition of speech tokens by a speech recogniser which automatically detected speaker errors such as:

- early response (speech over the initial beep);
- too low level of the speech signal:
- saturation: the signal exceeds the DAC dynamic,
- too high level of background noise.
- no speech signal.

In these istances the workstation re-prompted the speaker with appropriate messages for immediate repetition. A local relational database (dBase IV compatible) was updated run-time in order to document system activities.

Two different telephone paths were used in order to maximise the similarity of the collected material with the possible application environment. Half database has been stored through Toll Free Telephone path; the second half has been collected by workstations connected to an intercontinental central office.

The workstation could also be connected to a speech recognizer. Two devices have been tested on-line during the database collection.

3. DATA BASE DESCRIPTION

Target Speaker Distribution

Italian dialects are distributed across Italy according to geographical boundaries which approximately correspond to administrative ones. For the purpose of this speech database representative samples of speech have been stored from twenty-six regional areas. They include the twenty Italian regions plus six important cities (i.e.Bari, Milano, Napoli, Palermo, Roma, Torino) that have been added as representatives of different dialect varieties.

The target of collection was to reach 2000 complete calls; a sample of 3000 speakers was chosen in proportion of the telephone density per number of inhabitants, as reported from the Italian Telephonic Company (SIP) in 1989 (fig. 1)

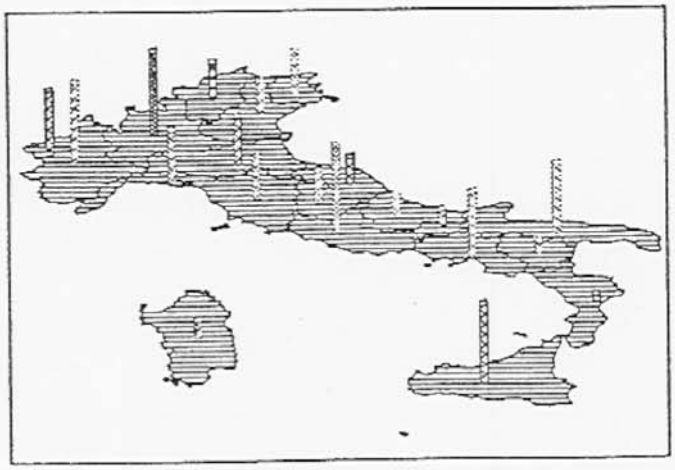


Fig. 1 - Subject density for geographical areas

Speakers from each dialect were proportionally distributed according to sex, age, telephone handset and calling sites.
Four age classes have been selected:

- from 14 to 20 years (20 %)
- from 21 to 40 years (35 %)
- from 41 to 60 years (35 %)
- over 60 years (10 %)

Four calling sites have been selected:

- home (45%)
- office (30%)
- telephone booth (15%)
- mobile telephone (10%)

Half of the speakers were male, half female.

A marketing firm selected the subjects according to the above specified criteria. Each caller was personally contacted and instructed by a trained interviewer, who collected speaker private information (town where the call was dialled from, birth place, sex...).

Obtained Distribution

Post-recording data analysis highlighted that the obtained caller distribution fits reasonably well with the expected target (average deviation rate of 6%). The main differences were due to the impossibility of recording calls from two regions during the last phase of the collection.

Callers could learn to interact with the system by a dummy code in order to reduce errors due to man-machine interaction; these calls were not considered in the final database even if they were completed.

The database archive reported all errors detected during the call; only those completed and containing less than 5 errors on each token have been accepted.

The average number of call received per day was 27; fifteen of them were correctly completed and accepted. The maximum number of call per day was 110 (60 accepted). Fig. 2 shows average call distribution over the twenty-four hours.

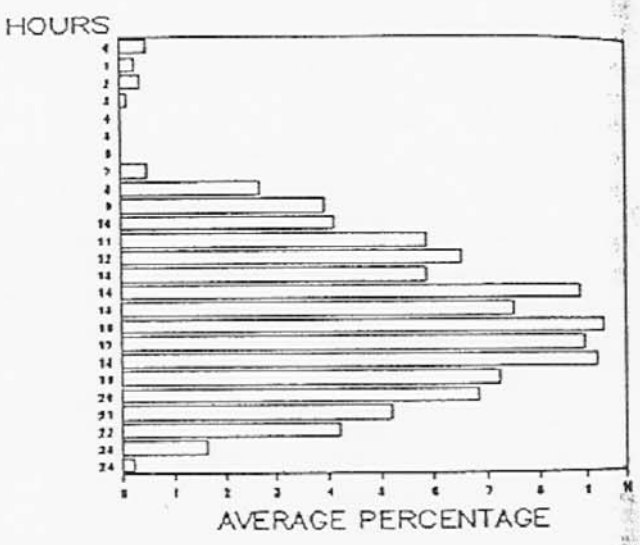


Fig. 2 - Call distribution per day

Speech Database Corpus

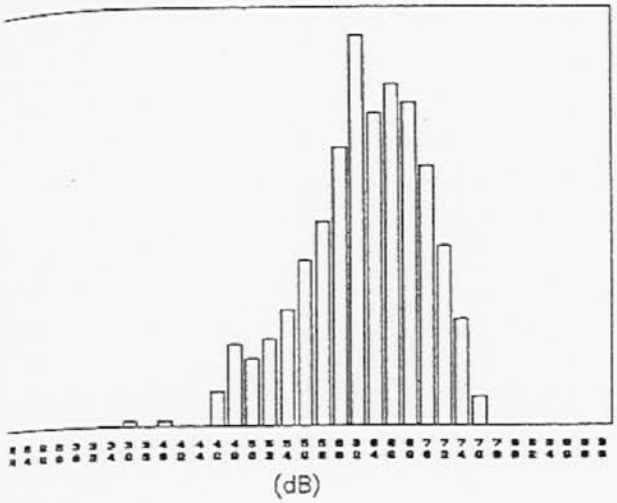
The speech corpus was composed of 68 isolated words and four strings of connected digits, organised in 6 lists :

list 1 : ten digits, *, # and PAUSE;
list 2 : 14 command words ;
list 3 : 15 Alternate Billing and banking commands;
list 4 : 4 strings of two, three and four different connected digits for each speaker;
list 5 - 6 : 26 typical Italian words used for spelling.

Each list has been uttered word by word, after a beep. Speakers answered to some additional questions pronouncing name, last name, year of birth, identification code (6 digits), area code, telephone number.

ech Level

The speech level has been computed for every speech :n. The same distribution has been obtained from data rded by the Toll Free telephone connection and from : stored through the intercontinental central office.



3 - Signal Level Distribution

Speech-level measurements has been computed rding to [1]; this method was used for specifying the mic range of the signal. Speech level distribution is vn in fig. 3

5. MAN-MACHINE INTERFACE

:edure

Subjects could call the system at any time and were led by an automatic procedure during interaction.
i subject received these informations:
an instruction sheet with the exact task description,
the list of words to be uttered,
a pocket tone dialler. This device generated DTMF tones and caller could enter its identification code allowing automatic identification.

After a short introduction, the system asked some stions and the speaker was instructed to answer them r the beep. The caller then read the word lists; TESCOS utored the item sequence and, when an error occurred, it mpted the speaker with the word he was supposed to nounce.

Speakers were instructed to read the word list before test, in order to reduce pronunciation errors and to ease speech naturalness.

An encouraging percentage (95.6 %) of lists of word out any procedure errors, highlighted the relative ness of the task.

Problems and solutions.

After the first week of recording the correctness of the collected material has been verified. The main problems were due to errors in pronouncing the connected sequences of digits, and to the percentage of clipped words.

The first problem has been solved by adding specific instructions to the voice prompt with explicit examples.

As to the signal saturation, 10% of the collected material was corrupted because the signal of at least one word of the list exceeded the A/D board dynamic; in the actual fact, the possible range of speech level over the Italian Telephone Network can be more than 20 dB and the gain of the audio channel had been calibrated in order to avoid recording of too low signals.
Possible solutions to this problem were:

1. to accept this discard rate,
2. to attenuate the signal channel reducing the overall signal to noise ratio,
3. to control the saturation for each word.

The latter solution has been adopted by computing in real time the maximum value of the samples of each recorded word and reprompting the speaker by asking to decrease the speech level if any saturation occurred. The saturation rate dropped to zero.

This solution has been chosen because the saturation of the DAC is an experimental artefact depending on the available board and it is not an intrinsic characteristic of the signal. On the other hand, the controlled solution is coherent with the normal behaviour of a real recogniser that prompts speakers with appropriate error messages when receiving speech signals out of its range.

6. DISCUSSION AND CONCLUSIONS

This work presented an isolated word speech database collected over the Italian telephone network. The collection procedure has been very carefully designed and allowed to record a very high percentage of completed call.

All collected material can be used both for training and for testing of isolated word speech recognisers.

The percentage of corrected recognised word obtained from two recognisers is avaliable for this database. This percentage has been obtained by two systems connected to the recording workstation during the whole collection. The two recognisers received the same signal that has been stored. These data will be used to check the laboratory testing enviroment. This material will allows to verify if tests performed in laboratory provide the same results obtained in the field test. This aligning capability results of great importance in order to develop testing enviroments as close as possible to thr real telephone network situations.

REFERENCES

[1] *Danielsen S., Velden J.G van:* "Speech Level Meter" n: "User Guide to Input Assessment" Esprit SAM document SAM-UCL-G005