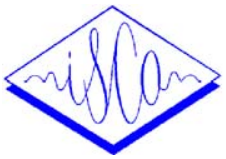

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

ON FIELD EXPERIMENTS OF CONTINUOUS DIGIT RECOGNITION OVER THE TELEPHONE NETWORK

D. Falavigna, R. Gretter
IRST – Istituto per la Ricerca Scientifica e Tecnologica
38050 Pantè di Povo, Trento, Italy.
e-mail:falavi@irst.itc.it gretter@irst.itc.it

Abstract

In this paper a continuous digit recognizer over the telephone network in real time will be described. The activity has allowed the realization of a system, installed in some Italian telephone exchanges, for providing semi-automatic collect call services. Data collection has also been performed, and a field database was built. Either a continuous digit recognition task and a confirmation task, requiring rejection, have been defined. Recognition results are presented.

10.21437/Eurospeech.1997-318

INTRODUCTION

The activity reported in this paper led to the realization of a system, installed in some Italian telephone exchanges. It provides two semi-automatic collect call services, called "Italy Direct" and "170". These systems require the recognition of digit sequences, as well as of yes/no. In the last case rejection of unforeseen sentences must be used to assure sufficient robustness with respect to user inexperience.

To train and test the system some telephone speech databases, later described, have been used. In particular a database, called FIELD, acquired during the system usage, will be introduced and discussed.

Finally, results will be presented, both for the digit recognition task and for the confirmation task.

SYSTEM DESCRIPTION

"Italy Direct" and "170" services allow an Italian user (user A) to communicate with another Italian user (user B), without interacting with a foreign operator. User A is in a foreign country and user B is in Italy (service "Italy Direct"), or vice-versa (service "170"). Briefly, user A calls an Italian operator, introduces himself, and gives him the telephone number of user B. Then the operator calls user B and asks if he accepts a collect call from user A. If yes, user A and B are connected. The automatic system, called POA, can handle both phases, requiring the presence of the operator only when the speech recognizer fails. The POA (see figure 1) consists of the following basic modules:

- a supervisor which controls the interaction flow;
- an interface with the public exchange UT100;
- a speech recognizer for digit sequences and confirmations.

Other features are: DTMF recognition, recording and play capabilities, possibility of operator intervention in case of recognition failure.

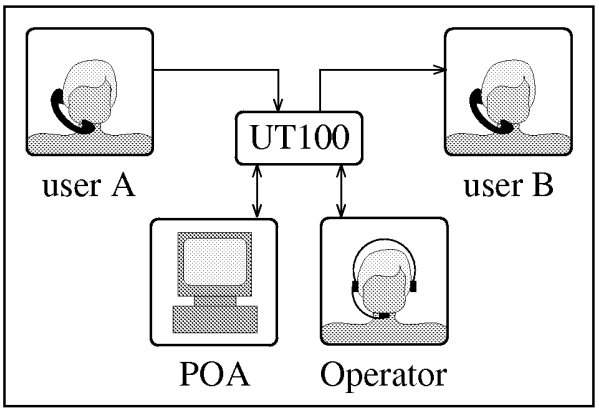

Figure 1: *Services "Italy Direct" and "170": UT100 is the public exchange, POA the automatic system. In case of failure a human operator intervenes.*

A typical call is first passed to a POA, which interacts (possibly using DTMF) with user A. After some minor informations, the telephone number of user B is requested and digit recognition is performed. Then a confirmation is asked. If user A does not confirm the digit sequence after a maximum number of trials (usually 2 or 3), the call is passed to an operator, otherwise the POA calls user B, which has either to accept or to refuse the incoming call.

The speech recognition module is responsible for executing some microroutines, that basically allow to get some elementary information. Examples are GetUserNumber(), which has to prompt a message, get and recognize the user sentence, ask for confirmation, and possibly repeat the whole process for a programmable number of trials. It returns either a user-confirmed digit sequence or a FAIL. The user can speak only after the vocal prompt, i.e. there is no barge in capability. A start-end point detector with a dynamic threshold [2] is used to detect the speech signal, that can be also stored for further analysis. At present, a total of 104 POAs have been installed near Rome and Milan, and the whole system is under test by Telecom. System design and development, apart speech recognition, have been carried out by AT-System and Italtel.

SPEECH DATABASES

The speech databases involved in the training / evaluation of the system are the following:

- CLEAN, a band filtered version (between 300 Hz and 3700 Hz) of IRST databases, APASCI and SPK, acquired in an acoustically isolated room. APASCI [1] consists of 5,215 phonetically rich sentences (194 speakers). SPK was collected for speaker recognition purposes, and contains about 30,000 digits (both isolated and sequences) uttered by 107 speakers. All this material is labeled and segmented in words and phonemes.
- PHONE, formed by 5,210 sentences (280 speakers) recorded through the public telephone network [3] by means of an automatic system that calls a previously advised speaker. Sentences include confirmations, digit sequences and phonetically rich sentences. This database has been manually checked and transcribed. Due to the inexperience of the speakers involved in the collection with speech recognition applications, several spontaneous speech phenomena (hesitations, breaths, false starts, etc.) are present in the recordings. Particular care has been taken for the transcription of these phenomena. Each file in the database has been classified into one of the following classes:
  - *higQ*, sentences without spontaneous speech phenomena;
  - *medQ*, sentences with some weak spontaneous speech phenomena (breaths, noises, hesitations, isolated laughs, etc.);
  - *lowQ*, sentences with strong spontaneous speech phenomena (false starts, speech and laughs together, etc.).

  For what concerns confirmations, it has been observed that, even if the system explicitly required to utter only "yes" or "no", a large percentage of the answers differs from them. In particular, the database contains the following distribution of *yes/no* answers:
  - 64.1% clean *yes/no*;
  - 19.2% *yes/no* with some weak spontaneous speech phenomena;
  - 2.7% other expressions clearly meaning *yes/no*;
  - 3.7% *yes/no* followed by other words (motivations, comments, etc.);
  - 10.3% expressions without a clear *yes/no* meaning.
- SIRVA, connected digit training material, provided by CSELT. It contains 4,372 digit sequences having length 2-4, uttered by 1,096 speakers covering all Italian regions.
- FIELD, speech material collected by the system in the public exchange during the service operation. These data have been divided according to their meaning (digits, yes/no of user A, yes/no of user B). At present, more than 6,000 files have been collected and checked, each one containing a single sentence; more than 2,000, mainly from users A, have been labeled as *garbage*. In the next section a description of these data will be given.

FIELD DATABASE

Data collected so far have been divided into three groups, according to the month in which they were recorded (December 1996, January and February 1997). The number of files collected in February is larger than those collected in the previous months. Each file has been manually checked, transcribed and assigned to a predefined class, according to its content. Several files have been classified as *garbage*, typically those containing silence (the user either speaks during the vocal prompt, or simply remains silent), or those due to children that make calls just for fun (note that the call by user A is free). In general, these data can be considered hard to handle. Since many calls come from public boxes, background voices and noises are frequently present. Sometimes the user himself comments to other persons what is going on.

	Dec	Jan	Feb	tot
digits	133	135	899	1167
	(54.1%)	(65.9%)	(60.0%)	(59.8%)
tens	27	19	74	120
	(11.0%)	(9.3%)	(4.9%)	(6.2%)
oov	21	5	88	114
	(8.5%)	(2.4%)	(5.9%)	(5.8%)
garbage	65	46	438	549
	(26.4%)	(22.4%)	(29.2%)	(28.2%)
tot	246	205	1499	1950

Table 1: FIELD: statistics on the collected digit sequences.

Apart from *garbage* data, digit sequences have been divided into the following classes: *digits* (valid digit sequences), *tens*, (sequences containing also tens, instead of digits alone), *oov* (sequences containing out-of-vocabulary words, false starts, etc.). Table 1 reports some statistics. Note that only 60% of all the files are valid digit sequences. We have estimated that the children are responsible of about one fourth of the *garbage* files. It is worth noting that the percentage of sequences including tens is decreasing from about 10% (Dec, Jan) to 5% (Feb). This may be due to the fact that users have learned to interact with the system correctly. Out-of-vocabulary words include digit sequences in different languages (English and Spanish), injuries ("0 7 ... 0 5 oh Cristo 'sto computer"), comments to another person ("0 3 ... 5 4 ora lo ripete ..." - "... now it repeat it ..."), explanations ("eh devo telefonare in Francia ma non so il prefisso di Parigi" - "I have to make a call to France but I don't know the prefix of Paris").

Tables 2 and 3 show statistics for the confirmation task. Each file has been assigned to one of the following four classes:

- *yn*: only "sì" and "no", possibly with weak spontaneous speech phenomena;
- *richyn*: expressions clearly meaning yes/no ("sì,

	Dec	Jan	Feb	tot
yn	291	308	564	1163
	(46.5%)	(47.1%)	(43.4%)	(45.1%)
richyn	23	10	29	62
	(3.7%)	(1.5%)	(2.2%)	(2.4%)
garbage	299	322	656	1277
	(47.8%)	(49.2%)	(50.5%)	(49.5%)
doubtful	13	14	51	78
	(2.1%)	(2.1%)	(3.9%)	(3.0%)
tot	626	654	1300	2580

Table 2: FIELD: statistics on the confirmation task of user A.

	Dec	Jan	Feb	tot
yn	101	139	941	1181
	(44.3%)	(66.8%)	(70.4%)	(66.6%)
richyn	72	17	109	198
	(31.6%)	(8.2%)	(8.2%)	(11.2%)
garbage	38	39	176	253
	(16.7%)	(18.7%)	(13.2%)	(14.3%)
doubtful	17	13	111	141
	(7.5%)	(6.2%)	(8.3%)	(8.0%)
tot	228	208	1337	1773

Table 3: FIELD: statistics on the confirmation task of user B.

sì", "sì accetto" - "yes, I do accept", "no, non è corretto" - "no, it is not correct");

- *garbage*: expressions without a clear yes/no meaning (children playing, background, "pronto" - "hallo", "no g'ho mia capio" - "I didn't understand", etc.);
- *doubtful*: typically yes/no expressions followed by other words ("sì, accetto la telefonata dalla Gran Bretagna" - "yes, I do accept the call from Great Britain", "sì pronto pronto - "yes hallo hallo", "sì sì Alfonso").

This subdivision reflects what we think a reasonable system should do. At present, in the system we explicitly model the most frequent expressions, i.e. the most frequent *richyn* sentences, in addition to yes/no. The distinction between *garbage* and *doubtful* allows to better understand rejection performance. In fact, if the system does not reject a *garbage*, this is certainly an error; but this is questionable for *doubtful* sentences. In these cases an "intelligent system" (from the user's perspective) should understand the correct meaning, while a good rejection system should ask to repeat.

Confirmations by users A are quite different from confirmations by users B. In fact, since users A have to confirm the telephone number previously uttered, they easily realize they are talking to a machine. Therefore, their confirmations are often solely yes/no. On the contrary, users B sometimes do not realize they are talking to a machine, especially because they are excited to receive a call from a relative or a friend in another country, and do not pay attention to the messages. Hence, their confirmations contain a large number of out-of-vocabulary words. This is clearly shown in tables 2 and 3 where the *richyn* total percentage is 2.4% for users A and 11.2% for users B. Note that, for user B, the *richyn* percentage drops from 31.6% (Dec) to 8.2% (Jan, Feb). This was due to a change in the final words of the vocal prompt: in December it was "... dica sì se accetta, no se rifiuta" - "... say yes if you accept, no if you refuse", then changed to "... per accettare deve dire sì, per rifiutare deve dire no" "... to accept you must say yes, to refuse you must say no".

SPEECH RECOGNIZER

The recognizer utilizes a set of phonetic units represented by continuous density Hidden Markov Models (HMMs). The acoustic features used are LPC Cepstral coefficients and log-energy, with the corresponding first and second order time derivatives. RASTA filtering is applied for channel equalization. In order to model any word (for confirmation expressions) without losing accuracy on digit strings, we decided to duplicate the phones corresponding to digits. Table 4 shows the training sets of the various databases.

CLEAN	5,215 phonetic sentences + $\sim$ 30,000 digits
PHONE	1,473 phonetic sentences + 1,184 digit sequences + 423 confirmations
SIRVA	2,194 digit sequences
FIELD	507 digit sequences + 697 confirmations

Table 4: Total number of sentences used for training.

A previous work on the PHONE database, reported in [3], showed the importance of explicitly modeling some weak spontaneous speech phenomena. A large difference in Word Accuracy (WA) was observed when recognition was performed on the *higQ* (97.21% WA) or on the *medQ* (85.33% WA) digit sequences, which contain those phenomena. To overcome this problem we introduced three new models: @eh, @br and @ns, representing hesitations, breaths and noises of various types, respectively. The HMMs of these new units were trained using the corresponding labeled occurrences in the PHONE database. Introducing them in the recognition network had no tangible effects on the *higQ* test sentences, but performance on the *medQ* part raised to 95.81%. The global effect (on a mixed test set which contains both *higQ* and *medQ* sentences) was a performance improvement from 94.75% WA to 96.95% WA.

The training procedure we adopt considers one word at a time, and performs Baum-Welch on the corresponding phone sequence. This means that a preliminary segmentation in words is needed, which was available both for CLEAN and PHONE databases. In order to obtain a reliable word segmentation on SIRVA and FIELD, we first performed a recognition (which also produces a segmentation in words) on each digit sequence. Then we retained only the digit

strings correctly recognized, following the reasonable assumption that, if the recognition is correct, the segmentation is sufficiently reliable[1]. In this way a percentage of the training data is not used, but the segmentation is reliable without hand-checking.

Following [4], the HMM parameters are first initialized using CLEAN (hmm1), then PHONE is used to retrain models and to introduce the units @eh, @br and @ns (hmm2). Further retraining has been done using PHONE + SIRVA + FIELD (hmm3).

During recognition, a network which allows any combination of digits, background and the special units @eh, @br and @ns is used. For confirmation, a network representing both *yn* and all the expressions in *richyn* is used, in parallel with a rejection network composed by a subset of the phones. In additions, some other expressions explicitly represent garbage ("pronto" - "hallo", "chi parla" - "who is speaking", etc.). Different weights are applied to different paths of the network; their value either favours or penalizes garbages.

RESULTS

In table 5 recognition results are reported for two different test sets. The first one, PH+SI, consists of digit strings of PHONE and SIRVA (1465 sequences containing 5051 digits), the second one is composed by correct digit sequences (*digits*) of FIELD (488 sequences containing 4624 digits, with an average of about 9.5 digit per sequence). Performance on the FIELD test is also given in terms of Sentence Accuracy (SA). Furthermore, the various acoustic models, namely hmm1, hmm2 and hmm3, are considered separately. As expected, performance improves by increasing the training material. The worst performance on FIELD is due, as previously observed, to the low quality of its signals. For this reason a spectral subtraction procedure [3] has been applied in order to increase the signal to noise ratio. However, as also observed in previous experiments on the PHONE database, no improvements have been obtained.

	PH+SI	FIELD	
	WA	WA	SA
hmm1	92.91%	93.51%	69.1% (337/488)
hmm2	95.51%	95.48%	76.2% (372/488)
hmm3	96.50%	95.96%	77.9% (380/488)

Table 5: Recognition results for digit sequences, using HMMs trained on increasing material.

As discussed before, for the confirmation task we did not consider the *doubtful* data, but only *garbage*, *yn* and *richyn*. Two test sets were defined, one concerning users A, the other concerning users B. Their sizes are reported in table 6. Evaluation was done by considering two parameters: the percentage of garbages detected (number of garbages correctly detected over total number of garbages) and the percentage of yes/no detected (number of yes/no correctly detected over total number of yes/no). The trade-off between yes/no detection and garbage detection can be controlled by means of weights in the network used by the recognizer. In this way a curve, shown in figure 2, can be drawn for users A and B.

	Users A	Users B
yn + richyn	907	1223
garbage	955	214
total signals	1862	1437

Table 6: Test set for the confirmation task.

Users A present the worst performance; this is mainly due to fact that their signals are often affected by a high background noise, as many calls come from public boxes.

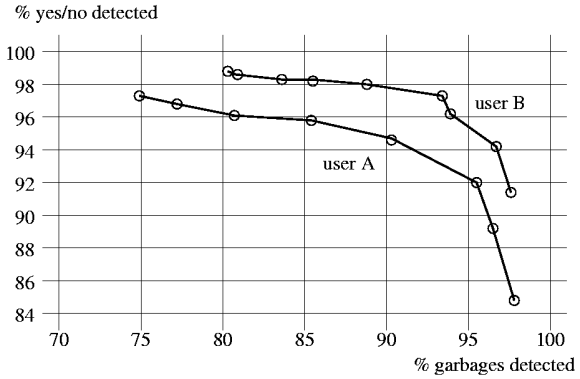

Figure 2: *Rejection results for the confirmation task. Users A and B are considered separately.*

ACKNOWLEDGMENTS

We wish to thank G. Podda for having labeled part of the field data. We also thank CSELT for having provided the SIRVA database.

References

[1] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo. Speaker Independent Continuous Speech Recognition Using an Acoustic-Phonetic Italian Corpus. In *Proceedings of ICSLP*, Yokohama, September 1994.

[2] G. Carli and R. Gretter. A Start-End Point Detection Algorithm for a Real-Time Acoustic Front-End Based on DSP32C VME Board. In *ICSPAT*, Boston, November 1992.

[3] D. Falavigna and R. Gretter. Evaluation of Digit Recognition Over the Telephone Network. In *Proceedings of ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a'-Mousson, France, April 1997.

[4] M. Weintraub and L. Neumeyer. Constructing Telephone Acoustic Models from a High-Quality Speech Corpus. In *Proceedings of ICASSP*, pages 85–88, Adelaide, Australia, 1994.

[1] We retain also strings having only word insertions at the beginning or at the end, labeling these insertions as @ns.