FIELD TEST EVALUATIONS and OPTIMIZATION of SPEAKER INDEPENDENT SPEECH RECOGNITION for TELEPHONE APPLICATIONS

Christian GAGNOULET and Christel SORIN

CNET Dépt RCP 22301 LANNION-France

ABSTRACT

This paper presents, in a first part, the detailed results of several field evaluations of the CNET speaker independent speech recognition system in a context of 2 voice-activated servers accessible by the general French public over the telephone. The analysis of roughly 11 000 user's tokens indicates that the rejection of incorrect input is a major problem and that the gap between the recognition rates observed in real use conditions and in the most "realistic" laboratory tests remains very large.

The second part of the paper describes the current improvements of the system : better rejection procedures, enhancement of the recognition performances resulting from both the introduction of field data in the training data and the increase of the number of parameters, automatic adjustments of the HMM topology allowing to either reduce overall model complexity or improve recognition performance. Tested on long distance telephone databases (450 to 750 speakers), the current version of the CNET recognition system yields a laboratory error rate of 0.7 % on the 10 French digits and of 0.95 % on a 36 word vocabulary.

INTRODUCTION

At CNET, the speech recognition studies are specifically oriented toward the development of telecommunications applications. This implies the development of robust, speaker independent speech recognition systems but also the design and the evaluation of complete spoken dialogue systems, for which human factor studies are essential.

SYSTEM OVERVIEW and FIELD TEST EVALUATIONS

System overview

The connected speech recognition algorithm developed at CNET in 1986 [1] uses the HMM approach and has been implemented on several devices (RDP 20 and RDP 50 boards) [2]. In the implemented version of the algorithm, 6 Mel cepstral coefficients, the energy and its derivative are computed every 16 ms to obtain the input vectors. The observation probabilities are represented by gaussian functions with diagonal covariance matrices and are tied to the transitions of the Markov chains. Various kinds of modelling can be implemented : either word units or sub-word units such as phonemes, diphones or allophones. For each application, the network is fully compiled and includes initial and final silence models for each word. This system has been tested on several databases of isolated and connected words recorded over the telephone network with willing subjects (mainly long distance lines, speakers representing different regional accents) : DIGITS-1 (455 speakers, 10 French digits), NUMBERS (730 speakers, 00...99 in French), TREGOR (513 speakers, 36 French words). For each database, one half of the data was used for training, the other half for testing. Using word models, the obtained word error rates (for the first version of the system) were 2.1 % for DIGITS-1, 2.7 % for TREGOR and 9.6 % for NUMBERS.

Field test evaluations

Experimental server : MAIRIE-VOX

In 1988, an experimental, one-port voice interactive system, MAIRIEVOX [3], was built on a PC computer using the RDP 50 board (word models, 13 states/word, 3 gaussian pdfs per state). Designed to give various informations about local services around the city of Lannion (20 000 inhabitants), MAIRIEVOX is accessible by the general public over the telephone since mid-88. The input interface for the user is restricted to voice input without any keypad complementary command. A tree structure is used to access information. The complete vocabulary contains 21 words (extracted from the 36 words TREGOR data-base) but the dialogue module limits the active vocabulary to 6 words at each step.

Since that time, MAIRIEVOX has been the subject of several field trials allowing to identify its critical points and to substantially improve both the speech recognition performances and the acceptability of the service.

The first evaluations (during which the input signal was not recorded) mainly allowed to improve the ergonomy of the service. For example, it appeared extremely usefull to authorize the recognition of the speech commands during the delivery of the voice messages : this allows the regular users of the service to anticipate the commands and therefore to quickly reach the required information in the dialogue-tree. An echo-cancellation procedure (non recursive filter with a 8 ms window) was thus introduced on the speech recognition board. The dialogue strategy was also modified to take into account the necessity of recovering from the largest part of recognition errors ("confirmation" procedures with Yes/No commands in case of recognition difficulties). With these two main improvements, the acceptability of the isolated-word, menu-driven speech command server has been demonstrated (less than 10 % failure in the access to the requested informations).

During 1990, a new set of evaluations has been done : roughly 4600 voice inputs (corresponding to 340 telephone calls) were systematically recorded, listened to and labelled as *"correct inputs"* (55.5 %), *"incorrect speech inputs"* (i.e. non permitted by the dialogue) (17.8 %) and *"noise"* (26.7 %).

From the application point of view, the rejection of incorrect inputs appears therefore to be a crucial point : despite clear instructions to the caller, roughly 45 % of the inputs to MAIRIEVOX are incorrect (words outside the vocabulary or noise). The simple rejection procedure used in MAIRIEVOX (all the vocabulary words are candidates at any time even if the dialogue module filters the words which are not valid in the context, use of a simple duration-based rejection threshold) allowed to limit the false rejection error rate to roughly 10 % ("correct inputs"). For incorrect inputs, 82 % are corectly rejected but 18 % induce an error (false acceptance).

From the recognition point of view ("correct inputs" only), we observed a 12.2 % error rate (21 valid words), 36 % of which being due to bad end point detection (truncated words). On the other hand, contrary to previous observations, the need for modelling hesitations (or surrounding speech) didn't really appear to be crucial : less than 5 % of the speech inputs contain hesitations or supplementary words (the design of the dialogue seems to play an essential role in this phonemenons).

Industrial Server "Horoscope"

A commercial voice-activated server "HOROSCOPE" was operating since April 1990 over the 9 taxation areas of the French telephone network. Based on the same recognition technology as MAIRIEVOX, it involves the recognition of the 12 horoscope signs spoken in an isolated manner. The calling person had the ability to ask for a horoscope sign at any time (branching factor of 12), any number of times, by waiting for the end of a message playback, or by interrupting it. The very direct dialogue procedure prevented the use of any dialogue-driven rejection process (contrary to MAIRIEVOX).

During June 1990, 6446 tokens from 1724 calls [4] were recorded, listened to and labelled as *"correct speech inputs"* (73 %), *"incorrect speech inputs"* (speech without a vocabulary word) (15 %) and *"noise"* (12 %). Here again the rejection problem seems to be more important than the "word spotting" problem : less than 2 % of the *"correct speech"* inputs contain hesitations or supplementary words.
The lack of perfect noise/speech discrimination in the endpoint detector aggravates the problem, as already observed for MAIRIEVOX : from the 27.1 % word error rate observed on *"correct inputs"*, roughly 50 % are due to bad endpoint detection. The very low recognition score observed here results from 3 main short comings in the realisation of this industrial system : 1) only 1 gaussian/state was used in the 13 state word models, 2) only 150 speakers were used for training the models, 3) the implemented echo-cancellation procedure was a very simplified version of the procedure proposed by CNET (a 10 dB difference was observed between the 2 attenuation rates).

In conclusion, after assessing two general-public word-recognition applications in use over the French telephone network, it was found that, despite clear instructions to the caller, a considerable proportion of the input lies outside the permitted vocabulary. These extraneous inputs are either incorrect speech tokens or non-speech tokens for which the caller is not always responsible (DTMF dialing, line bursts, outside noise etc...). There is therefore an urgent need for efficient rejection procedures. Moreover, the gap between the recognition error rates observed in real use conditions and in the laboratory tests is very large : a multiplicative factor of 3-4 is observed ; it can reach 10 if the application is carelessly designed and modelized !...

NEW REJECTION PROCEDURES

Two rejection procedures [5] have been investigated and compared on the "HOROSCOPE" field database containing all the "correct" and "extraneous" tokens recorded during the "Horoscope" field trials, to which were added 1699 tokens from 151 willing subjects recorded through the telephone network. Half of the data was used for training, the other half for testing.

The first rejection procedure uses 3 sink models trained with the *"extraneous"* tokens (incorrect or noise inputs) of the training corpus and imposes thresholds on word-model scoring : the rejection threshold is applied on a *"corrected score"* which is the word HMM score minus the contribution of the silence models.

The second rejection procedure operates on the *"trace"* of the HMM (i.e. informations on the optimal Viterbi path). It involves the extraction of the HMM trace from a given input token and the classification of this trace into "acceptance" or "rejection" by a multi-layer perceptron (MLP). This rejection procedure is independent from the recognition process : it uses HMMs designed with the sole purpose of producing informative traces.

For the *"trace"* rejection procedure, the best results were obtained with a trace containing 1) the number of frames observed per gaussian, 2) the average energy coefficient and 3) the average first Mel frequency coefficient of the frames observed per gaussian, i.e. with a trace exhibiting both a duration and a signal representation.

The results of both procedures are illustrated on Figure 1 where the sum of the **SE** rate and the **FR** rate measures performance on *correct* tokens and the **FA** rate measures performance on *extraneous* tokens.

$$\textbf{SE rate} = \frac{\text{number of substitution errors}}{\text{number of correct tokens}}$$

$$\textbf{FR rate} = \frac{\text{number of false rejection}}{\text{number of correct tokens}}$$

$$\textbf{FA rate} = \frac{\text{number of false acceptance}}{\text{number of extraneous tokens}}$$

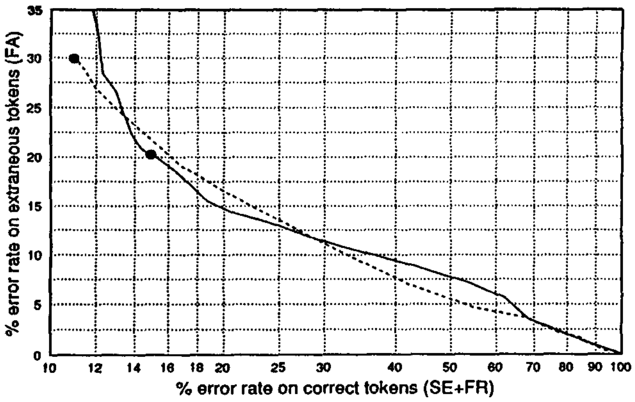

Figure 1 :
Rejection using the HMM trace *(full curve)* and rejection using sink models and a score threshold *(dashed curve)*

Although the performances of both procedures appears to be similar within the confidence interval, there is one aspect unique to the rejection by trace : its ability to reject a large proportion of the substitution errors instead of proposing them to the user (substitution rejection error rate of 66 %).

Work is currently underway to refine the trace-based procedure. Another promising direction seems to be to combine this two complementary methods.

RECOGNITION OPTIMIZATIONS

Use of field data in HMM training

The constant gap between the recognition rates observed in real use conditions and in the laboratory tests led us to investigate the introduction of field data in the training database to see if it can significantly improve the recognition performances.

In this experiment [6], 2 sets of a telephone-speech database corresponding to the 21 word vocabulary of the MAIRIEVOX server have been used :

- a "**LAB***oratory* database" corresponding to a subset of the TREGOR database : 513 willing subjects, 9797 uniformly distributed tokens,

- an "**EXP**loitation *database"* corresponding to an extension of the previously introduced field database : 1547 naive speakers (real users) produced 9536 "correct tokens", non uniformly distributed among the 21 words.

Both databases exclusively hold manually validated data, i.e. data labelled as *"perfect"* (non truncated and without hesitations or supplementary words) after listening. Each database was split into two equal parts : one for training, the other one for testing.

The training of the HMM word models was done either on the **LAB** database, on the **EXP** database or on a **MIX**ed database containing an equal proportion of laboratory and field data. The results are illustrated on table 1 (word error rate).

	LAB test database	**EXP test database**
LAB models	2.3 % [+0.4]	5.8 % [+0.7]
EXP models	9.8 % [+0.8]	4.4 % [+0.6]
MIX models	3.6 % [+0.5]	3.9 % [+0.5]

Table 1 :
Word error rate for a 21 word vocabulary (long distance telephone speech) : influence of "field" data introduced in training

It can be seen that the use of "**MIX**ed" models leads to a 30 % reduction of the recognition error rate on the field databases : the introduction of field data in the training phase does improve the field recognition performances.

Work is currently underway for achieving on-line selection of the "correct" field data to be introduced in a *"retraining"* phase of systems in exploitation.

Increasing the number of parameters

Several studies have shown the usefulness of adding time-dependent information in the HMM input vectors. Table 2 illustrates the results of various tests on the DIGITS-1 data base (455 speakers) using input vectors containing either 9 acoustic coefficients (8 MFCC and energy), 18 acoustic coefficients (the same as above plus their first derivative) [7] or 27 acoustic coefficients (second derivative added).

It is also well known that increasing the size of the models (i.e. number of states and pdf's) yields better performance, at least for isolated word recognition. Comparative results between 13 state and 30 state word models are shown in Table 2.

	9 coeff.	18 coeff.	27 coeff.
13 states	5.9 %	2.2 %	-
30 states	3.5 %	1.2 %	-
41 states*	-	1.1 %	0.7 %

Table 2 :
Word error rate on the DIGITS-1 telephone speech database (455 speakers) : influence of the number of acoustic coefficients and states

(* 775 speakers database)

The new version of the recognition algorithm implemented on the RDP 50 board (TMS 320C25) yields an error rate of **0.69 %** (41 state word models, 27 acoustic coefficients) on an expanded version of the long distance telephone DIGITS database (775 speakers) and of **0.95 %** (18 acoustic coefficients, word model size depending of the word length) for the TREGOR long distance telephone database (36 words, 513 speakers).

Automatic adjustments of the structure of HMM models

Using whole word basic units is generally a good choice for small vocabulary isolated word recognition, and increasing the size of the models usually leads to better performance. However, this also increases the computation time, due to the number of observation probabilities (gaussian functions) that must be computed for each frame. Thus, in order to use the best possible model in real time industrial devices, it was usefull to investigate the possibility of reducing the number of gaussian functions by clustering "similar" pdf's. This was done by iteratively merging the 2 gaussian pdf's inducing the smallest decrease of the total probability of the training observations, until the desired number of pdf's is reached [8]. On the 36 word TREGOR database, this procedure allowed to reduce by **40 %** the number of gaussian functions while keeping identical performances.

Using sub-word basic units leads to more compact models (since all the occurrences of a given unit share the same set of pdf's), but it is difficult to increase the a priori size of the acoustical models (they may become too long). An algorithm has thus been developped [8] around the two following basic ideas : splitting the pdf's having the highest contribution to the probability of the training data, and discarding the transitions which are scarcely used. These two operators (splitting and discarding) are applied successively, and the model is re-trained after each modification. By applying this procedure on a pseudo-diphone based model [1], the recognition error rate has been reduced from 2.5 % to 1.8 % on the 36 word TREGOR telephone database used above.

CONCLUSION

Exhaustive analysis of field trials allowed to better identify the most crucial shortcomings of the speech recognition systems developed in the laboratory and to substantially improve both the speech recognition performance and the acceptability of the resulting services. From our experience, it appears that both the rejection of incorrect inputs and the noise-speech end-point detection are among the most crucial problems.

A new rejection procedure has been presented which still requires further refinements. Introducing field data in the training database proves to be an efficient procedure for rapidly improving the performances of systems that can be re-trained during their exploitation. Recognition score improvements were obtained by increasing the number of acoustic coefficients and HMM model parameters. Finally, dynamic adjustments of the structure of Markov models allowed to either reduce the overall model complexity (a crucial point for industrial implementations) or improve the recognition performance especially for larger vocabularies where the use of sub-word basic units becomes necessary.

ACKNOWLEDGMENTS

The authors would like to gratefully acknowledge all the members of the Speech Recognition and Voice Services Groups : Jean Monné, Denis Jouvet, Michel Toularhoat, Katarina Bartkova, Dominique Dubois, Patrick Haffner, Guy Mercier, Laurent Miclet, Gaby Gargan, Jacques Réjaud, Luc Mathan, Laurent Mauuary, Dominique Morin, for their participation to the work reported here.

REFERENCES

[1] D. JOUVET, J. MONNE, D. DUBOIS (1986) : "A new network-based speaker independent connected word recognition system", Proc. IEEE/ICASSP 86, 1109-1112.

[2] J.P. TUBACH, C. GAGNOULET, J.L. GAUVAIN (1989) :"Advances in speech recognition products from France", Proc. SPEECH TECH'89.

[3] C. GAGNOULET, D. JOUVET, J. DAMAY (1991) : "MAIRIEVOX : a voice activated information system", Speech Communication, Vol 10, N° 1.

[4] L. MATHAN, D. MORIN (1991) : "Speech field databases : development and analysis", submitted to EUROSPEECH Conf., Genova, sept. 91.

[5] L. MATHAN, L. MICLET (1991) : "Rejection in an isolated word ASR system using multi-layer perceptrons and the trace of HMMs", to appear in Proc. IEEE/ICASSP 91.

[6] D. MORIN (1991) : "Influence of field data in HMM training for a voice-activated telephone server", submitted to EUROSPEECH Conf., Genova, sept. 91.

[7] D. DUBOIS (1991) : "Comparison of time-dependent acoustic features for a speaker-independent speech recognition system", submitted to EUROSPEECH Conf., Genova, sept. 91.

[8] D. JOUVET, L. MAUUARY, J. MONNE (1990) : "Automatic adjustments of the Markov models topology for speech recognition applications over the telephone", NATO/ASI Workshop, Cetraro, July 1-13 (to appear).