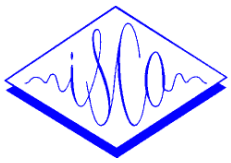


First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

BILINGUAL SPEECH INTERFACE FOR A BIDIRECTIONAL MACHINE TRANSLATION SYSTEM

Jean-Pierre Tubach(*), Raymond Descout(**), Pierre Isabelle(**)

(*) TELECOM Paris, Signal Dept (CNRS, URA 820), France and CWARC
(**) CWARC (Canadian Workplace Automation Research Center), Montreal, Canada

ABSTRACT

This paper describes the work carried out at CWARC (Canadian Workplace Automation Research Centre) (Voice Technology Group (VTG), in co-operation with the Machine Translation Group (MTG)). It investigates the use of commercially available speech technology devices for advanced human-computer interaction applications, a very important matter nowadays.

The MTG at CWARC has developed a bi-directional machine translation system, working between French and English, for meat and cattle market reports issued by Agriculture Canada. An initial speech input/output interface (IRMA) was designed for this system by the VTG, and was demonstrated successfully at the Expotec exhibition in Montreal over the summer of 1989.

The purpose of this work is to provide CRITTER with a more advanced speech input interface than that integrated in IRMA (continuous speech, multi-speaker or high-quality), using the VECSYS and XCOM MEDIA50 recognition boards.

The initial project objectives dealt only with French, but the satisfactory performance of the study and its initial results showed that it was possible and desirable for English to also be included.

The result was a bi-directional (French <--> English) translation system with speech input, a "first" to the best of our knowledge.

This project offers interesting ergonomic potential for the translators' workstation project, since users can enter either the source language text for machine-aided translation by CRITTER, or the results of their "human" translation in the target language.

I INTRODUCTION

I.1 Brief description of CRITTER

It is recognized that while machine translation of any kind of text remains a difficult objective, attainable only in the very long term, it is feasible in the context of a language and a semantic field restricted to a specific application.

That is the approach taken with CRITTER, for the weekly reports on the meat and cattle markets in the various provinces, issued by Agriculture Canada.

This software, implemented on a SUN workstation, and described in [4], is based on a transfer system. One particularly interesting feature is its reversibility between French and English.

I.2 Brief description of IRMA

Voice recognition in the first version of the IRMA system is based on the DRAGON VoiceScribe 1000 board, and must be by isolated word recognition. The vocabulary comprises 150 words in French and the same in English, and the recognition board applies no syntactical rules.

Accordingly, CRITTER can be provided only with a lattice of words, comprising at least all the homophones or quasi-homophones of a recognized word (e.g., in French, faible, faibles ; était, étaient, été). The translation system is used to remove ambiguities from this lattice, by seeking an analysis that results in a successful sentence.

All the control functions (changing languages, turning the microphone on and off, editing) are called from the SUN workstation on which CRITTER runs, and are well integrated in the system's excellent demonstration screens

Recognition errors are frequent, despite single-speaker operation, and make the system difficult to use. Some of the reasons for the errors are:
- the board used no longer represents state of the art technology;
- no syntactic rules apply to the recognition process;

I.3 Options selected for this project

We decided to use the new recognition boards, Vecsys Datavox and XCOM Media50 (described in [7]), as best we could, to provide a good demonstration of the potential of speech technology as a complement to machine translation:
- continuous dictation, rather than isolated words;
- syntactic rules for the recognition process, providing CRITTER with correctly spelled strings that can be directly assimilated;
- oral commands can be given for switching languages and operating the microphone, as well as for corrections ("speech editing").

I.4 Language definition

CRITTER is used to translate Canada Livestock and Meat Trade Reports. The first step was to conduct a study of the language specific to this application, through discussions with the Machine Translation Group, examination of Agriculture Canada reports, and trials with the CRITTER system.

A sub-set of this language was defined, of appropriate complexity for the current capabilities of speech recognition systems .

Our initial intention was to use dictation of complete sentences, and the Media50 board was first used for that purpose, as a single-speaker system.

But the language used frequently involves very long sentences, particularly in French (e.g., 27 words : "à Vancouver, les prix de gros des bouvillons de court engraissement ont gagné 3 dollars le kilo en raison d'arrivages peu abondants au début de la semaine").

Our first experiments showed that even with pauses and/or hesitations at specific places in the sentence, the dictation of such sentences as single units was highly artificial, and called

for a speaker very skilled in working with this type of system, which of course defeated our purpose. There were also technical problems linked with this mode of operation; for instance, the apparent response time at the end of a sentence was systematically increased by the maximum allowable length of a pause (at least one second).

Accordingly, we defined another dictation mode, much more satisfactory from all points of view for this application: dictation by phrases, which allows the user to dictate the major syntactic elements of a sentence as independent units: place, time, subject, verb-complement, cause (subject and verb-complement may be divided into two parts, if desired). For example, "Wholesale prices / of butcher cows // increased / three dollars per kilogram // at the end of the week // in Montreal // owing to very low supply // period" (where / indicates an optional pause, and // a mandatory pause)

Note that the system must then be informed specifically ("period" or "end of sentence") that a complete sentence has been dictated.

There is also the possibility of dictating simpler phrases without pausing (subject verb complement).

The problem of user hesitations (uh, um, ah) and pauses becomes much less important with this dictation method, since they occur mainly at the breaks we have chosen between phrases. Nevertheless, we took account of the phenomenon noted by Hauptmann and Rudincky ([3]), namely that the speaker tends to hesitate before dictating particularly precise information ("prices rose // uh three dollars per kilogram").

Syntactic constraints are weaker with the phrase approach (e.g., it is no longer "known" whether the subject was singular or plural when distinguishing between "have increased" and "has increased"). But the quality of recognition obtained at the completion of our work allows correct operation with lesser constraints (in a multi-speaker system, this applies only to French).

The syntactic refinements used in the recognition system to take account of certain specific features of French have no counterpart in English, and so disappear. Here we are speaking of liaisons (des vaches, des (z) agneaux), elided articles (de veaux, d'agneaux), and agreement between adjectives and nouns (demande faible, arrivages faibles). As for vocabulary, English words are much shorter, particularly in this field ("yearlings" for "sujets d'un an", "stockers" for "bouvillons de long engraissement"), which might at first be expected to hamper recognition, but causes no major problems thanks to the good quality of recognition.

II XCOM MEDIA50 BOARD

II.1 Introduction

This board, marketed by XCOM (Grenoble, France), among others, under the name Media50, is the fruit of work at the Centre National d'Études des Télécommunications of France Télécom by C. Gagnoulet, D. Jouvet and J. Monné. It takes up a single slot in a PC compatible computer.

The board, with its reasonable price (about $2,500) and limited hardware resources (TMS 320C25, 128K of data memory and 32K of program memory), has great potential. In fact, it is used by France Télécom in Mairievox, a telephone-based speech server, and the same algorithms are employed in the voice-operated public telephone booth, Publivox ([2]).

It applies training and recognition algorithms using Markov hidden models. Initially designed to recognize a limited number of connected words (which may be where the "50" in Media50 comes from), with a Markov model for each word, it can also be used for phonetic recognition (an elementary model for each phonetic unit), to support larger vocabularies (about 120 words) and a more detailed syntax. One very important feature is that the system can learn the voices of several speakers, for multi-speaker operation with no increase in memory required.

The development environment we used is PHIL 86, from CNET. It allows for the description of the application syntax and vocabulary, for training the resulting Markov model on a multispeaker speech database, then for recognition and assessment. More datails on the use of this software can be found in ([8]).

II.2 Markov models used

D. Jouvet, in his doctoral thesis ([5]), explored a number of Markov models for phonetic recognition. The best results were obtained for models with a large number of states, transitions and probability functions (typically 5 states and 12 transitions for vowels and consonants, and 7 states and 18 transitions for groupsof two phonemes that must sometimes be considered).

We found that their use for an application as "large" as the one discussed here unfortunately calls for more data memory than the 128K of the Media50 board.

C. Gagnoulet, who encountered this difficulty earlier (personal communication), proposed much more compact models (typically 4 states and 5 transitions for vowels, 3 states and 3 transitions for consonants, 5 states and 7 transitions for groups of 2 phonemes). We tested them and found that while the representations of vowels were satisfactory, those of consonants were unsuitable for our application.

Finally, we adopted "intermediate" models, which represent a compromise between memory requirements and recognition performance. This basically involved abandoning the shorter model (3 states) in the previous paragraph, and selecting the 4-state, 5-transition model for vowels and consonants, and the 5-state, 7-transition model for phoneme groups.

II.3 Information on the "phrase" application in French

The vocabulary comprises 120 words
48 phonetic units are used:
13 vowels
18 consonants:
15 diphones
2 additional units (euh and mmm for hesitations

The static branching factor is between 1 and 50, the average being 5.1. The dynamic branching factor is 3.0. The phrase length is between 3 an 11, the average being 8.3. The language has about 10,000 different phrases (This information is provided by VECSYS's REBUS2 utility program).

The training process was conducted with 10 speakers, 6 from Quebec and 4 from France (in fact, it would be more accurate to speak of "Montrealers" and "Parisians", since no strong regional accent was considered). This training process took about six hours using the PC/Media50 board system. Each speaker dictated 105 phrases or short sentences. Most dictated them twice, to provide evaluation data.

Other trainings were conducted for speakers from Quebec only, and for speakers from France only (see the paragraph on Evaluation).

II.4 Information on the "phrase" application in English

The vocabulary comprises 108 words.
40 phonetic units are used:
10 non-diphthongal vowels;
20 consonants;
9 diphthongal vowels and diphones;
1 additional unit (ah, for hesitations).

The phonetic transcriptions of the vocabulary are based on the "Gage Canadian Dictionary" ([1]).

The static branching factor is between 1 and 47, the average being 8.4. The dynamic branching factor is 3.2. The phrase length is between 3 and eleven words, the average being 7.4. The language has about 6,000 differnt phrases.

The training process took approximately four hours, with six Anglophone or bilingual speakers.

Each speaker dictated 100 phrases or short sentences. Most dictated them twice, to provide evaluation data. Test data were also provided by two Francophones (one from Quebec and one from France).

III VECSYS DATAVOX RECOGNITION SYSTEM

This recognition system, marketed by VECSYS (Bièvres, France), is the fruit of work by J.L. Gauvain and J.C. Gangolf, at LIMSI-CNRS in Orsay.

This is a set of two boards, for PCs and compatibles, and an external box, in charge of analog-digital and digital-analog conversion functions. The boards apply training and recognition algorithms using Dynamic Time Warping (DTW). One of the boards includes a specialized integrated circuit, called the uPCD (dynamic comparison microprocessor), the main reason for the system's excellent performance ([6]).

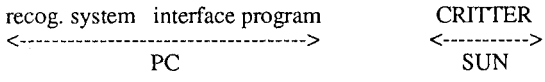
This product is oriented toward recognition in single-speaker mode, with excellent recognition quality. The software provided by VECSYS allows recognition of connected words for vocabularies of over 200 words. (Other software is being developed at LIMSI for more ambitious applications.)

In French and English, the application vocabulary is the same as in the previous section. The application language is a bit more comprehensive (e.g; numerals can be uttered as digit point digit).

From DATAVOX specifications, the product is not used up to its limits for this application, and doubling the vocabulary size could apparently be considered without any problem

IV INTERFACE PROGRAMS

A resident interface program is necessary for the connection between the recognition boards, from which it receives input, and CRITTER, to which it transmits the data after formatting it.



We had a choice of two possible designs when choosing the communications protocol between the two systems and their respective roles, which could perhaps be over-simply described as a translation system with a speech server, and a speech system with a translation server.

In fact, it would doubtless be desirable to combine the benefits of both formulas and allow both systems to take control, as applicable, for switching languages, turning the microphone on and off, and correcting errors.

In order to come up with a viable system as soon as possible, we decided to maintain exactly the same protocol between the two systems as in the first version of IRMA. Nevertheless, it is with the communications protocols developed by the STG that more sophisticated and satisfactory protocols will naturally be implemented (such protocols provide a general framework for communications between a Unix system and a PC equipped with speech devices).

V EVALUATION

The evaluation studies described below were done on the multi-speaker system implemented on the Media50 board, because the problem of the influence of the different variants of French occurs only with a multi-speaker system. But let us make first the statement that on line performance with the single speaker VECSYS system is very impressive, both for response time and recognition accuracy.

V.1 In French

To study the influence of the two variants (Quebec and France, or rather Montreal and Paris) of French pronunciation, we carried out:

- a training process involving all the 10 recorded speakers, which we will call T for total
- a training process with the six Quebec speakers only, or Q, for Quebec;
- a training process with the six speakers from France, or F, for France.

The speakers during the training process included only one woman, from Quebec.

In all cases, the initialization was done with the results of the training process for a fairly similar network, albeit more limited, done for a single speaker (French).

The results are as follows. The percentages of errors in terms of phrases (dictation units) and in terms of words (without distinguishing between substitutions, insertions and omissions) are presented below, in table form. We distinguish between the results on the training sample and the test data.

T training

Results for all speakers

	phrases	words
training	7.7	2.9
test	9.2	3.9

Q training

Results for speakers from Quebec

	phrases	words
training	6.8	2.5
test	7.2	2.7

Results for speakers from France

	phrases	words
test	13.2	6.6

F training

Results for speakers from France

	phrases	words
training	4.2	1.5
test	6.3	2.7

Results for speakers from Quebec

test	15.6	9.0

First of all, it can be seen that performances are fairly satisfactory for the T case (all speakers, for both training and recognition). Also, the results are slightly better for the two cases in which the two sets of speakers were separated.

The differences between the F and Q cases should be interpreted with care, since they probably reflect only the different numbers of speakers (6 for Q, 4 for F).

There is a noticeable deterioration in the case of cross F/Q tests; it must be remembered, however, that this is the only case in which speakers who did not take part in the corresponding training phase participated. Thus it would be more appropriate to compare these figures with the results of other speakers from Quebec for Q, and other speakers from France, for F. This was not possible in the time available. The much more subjective impression during the on-line cross-tests (at the microphone) is not as unfavourable as the above error rate, twice as high or greater, suggests.

Our conclusion, then, is an error rate of 3.9% for words, and 9.2% for sentences, for the test case with all speakers. Those rates allow acceptable system operation, with the "correction" speech editing function, (or editing on IRMA/CRITTER).

V.2 In English

The training process was conducted for six Anglophone or bilingual speakers in English (three men and three women). The tests were done with those same speakers, and separately with two non-Anglophones (one from Quebec and one from France). The results are as follows:

Results for Anglophone or bilingual speakers

	phraes	words
training	18.3	11.1
test	23.5	13.1

Results for non-Anglophone speakers

test	30.2	19.5

These results are poor, and prevent viable system operation. We offer the following explanations (in decreasing order of importance):

- the English vocabulary contains words that are very similar phonetically (calves / cattle / cows), (was / were) (increased / decreased). If those cases of confusion are removed, the error rates are close to those in French;
- in the phonetic description of words, no distinction is made between stressed and non-stressed vowels;
- the phonetic models used are not sophisticated enough to take account of certain distinctions (see the Chapter on the Media50 system);
- the initialization of the training process is done using a network of isolated words dictated by a single French, non-Anglophone speaker.

Nevertheless, the informal tests at the microphone with Anglophone speakers give a more favourable impression than suggested by the above figures.

VI CONCLUSION

One month appears to be a minimum for the initial development of a complex application on XCOM Media50with the PHIL 86 software, and knowledge of the principles of the algorithms used (Markov) is useful. Multi-speaker operation, on the other hand, has definite potential.

With VECSYS Datavox, training to use the recognition system and its development environment is not very difficult, and one week appears to be sufficient for putting together a first, non-elementary, application, even for engineers unfamiliar with speech processing algorithms. The drawback is single speaker operation, which is nevertheless acceptable in such an office automation task.

Our work has led to the design and implementation of a continuous speech input system for the CRITTER machine translation system, in French and English. Dictation is by phrases. It has clearly shown the value and feasibility of speech input at a translator's workstation.

VII OUTLOOK

To point the way for follow-up on this study, we can make the following remarks:

In the recognition boards, we used a fairly sophisticated syntax to represent a relatively flexible language, in an orthographically correct manner. But we are far from reaching the level of syntactic knowledge applied in CRITTER itself.

Thus it would be much more natural and efficient to ask a future speech recognition server only to recognize phonetic units, phonemes as a simple example; those units would then be processed by CRITTER's syntactic analyzer, designed to work with phonetics (the intermediate step of orthographic representation is certainly not essential).

This approach was not possible with the products currently available. Even when those products perform phonetic recognition tasks (Media50), they also operate at the word level, and require syntax for satisfactory operation; the formalism provided for expressing that syntax (equivalent to regular grammars) is insufficient for implementing CRITTER's knowledge.

To our knowledge, the only system on the market that approaches the idea of a phonetic recognition system is, in English, the "Phonetic Engine"(R) from SSI, intended specifically for connection to a CRITTER-style workstation. But it is far from certain that the recognition quality is sufficient for the needs of this project.

For French, there is the "machine phonétique" that is soon to result from CNET's work on the KEAL system.

We must closely monitor work that may lead to a recognition system providing a set of multi-speaker phonemes, application-independent (but language-dependent!). At the moment, it looks as though the teams using Markov modelling will reach that goal first, rather than those working on expert knowledge and artificial intelligence techniques.

By this we mean the work by CNET in Lannion (C. Gagnoulet et al), for French, and by Carnegie Mellon University in Pittsburgh (K.F. Lee et al), BBN in Cambridge, MA (Schwarz et al), and Bell Labs (S. Levinson et al) for English.

ACKNOWLEDGMENTS

Jean-Pierre Tubach wish to express his gratitude to:
CWARC management, in Montreal, and Télécom Paris management, in Paris, for making his sabbatical trip possible.
Pierre Hamel, Sylvain Faucher, Elliott Macklovitch and Michel Simard, of CWARC, for their great help in carrying out this project.
Denis Jouvet and Jean Monné, of CNET in Lannion, and Bernard Prouts, of VECSYS, for their long-distance assistance.

BIBLIOGRAPHY

[1] Gage Canadian Dictionary (Avis S.W. et al.) Gage publishing limited, Toronto, Canada

[2] Gagnoulet C., Jouvet D. "Développements récents en reconnaissance de la parole". L'Echo des recherches (CNET - ENST), No 135, 1989, pp 27-36

[3] Hauptmann G.A., Rudincky, A.I. "Talking to computers : an empirical investigation". International Journal of Man-Machine studies, vol 28, pp 583-604 (1988)

[4] Isabelle P., Dymetman M., Macklovitch E.: "CRITTER, a translation system for agricultural market reports", 12th International Conference on computational linguistics (COLING), Budapest, Hungary, August 1988; and CWARC document Co 28-1/25-1988E.

[5] Jouvet D. "Reconnaissance de mots connectés indépendamment du locuteur par des méthodes statistiques". Télécom Paris (ENST) Doctoral Dissertation, June 1988.

[6] Quenot G., Gauvain JL, Gangolf J.J., Mariani J.J. : "A dynamic programming processor for speech recognition". IEEE Transactions on Integrated Circuits. 1989

[7] Tubach J.P., Gagnoulet C., Gauvain J.L. : "Advances in speech recognition products from France", Speech Tech'89 Conference, New York, Mai 1989

[8] Tubach J.P. : "Continuous speech recognition as input for a machine translation system in French and English" CWARC Technical Report, Montreal, Canada, 1990.