



COST232: SPEECH RECOGNITION OVER THE TELEPHONE LINE

Andrea Paoloni(FUB: Italy), Torbjorn Svendsen (NTH, Norway), B. Kaspar(FDZ, Germany), Denis Johnston (BT, UK),
Gunnar Hult(Telia, Sweden)

ABSTRACT

This paper reports on the various studies which have been undertaken within the COST232 project "Speech Recognition over the Telephone Line" and summarises the various conclusions which have been drawn so far. It also introduces and describes a proposed reference system and a large telephone based database.

1. INTRODUCTION

COST - the acronym for European Cooperation in the field of Scientific and Technical research - provides a framework for cooperation in basic research within Europe. COST is self funding: participants provide their own funds for projects from their own organisations. The COST umbrella simply provides a forum for them to share ideas and identify where they can work together in areas of mutual interest.

COST232 started in 1989. 12 countries currently participate, most supported by their respective telecommunications administrations and companies. The project has two main themes: Methods and algorithms (Working Group 1) and Databases (Working Group 2).

2. WORKING GROUP 1

Working group 1 is mainly concerned with the development and testing of methods and algorithms. The participants in COST 232 bring a diverse range of applied and theoretical backgrounds into the project but all share an interest in seeing speech recognition applied to telephony systems. This variety in backgrounds results in various kinds of reports being presented at our meetings and extensive, applications-oriented investigations from outside are coupled with work undertaken specifically within the COST232 program. The participants also bring various approaches to speech recognition to the project. Hidden Markov Models (HMMs) dominate but reports have also been given on artificial neural network models (multilayer perceptrons, Kohonen self organising feature maps) and phonetically and linguistically motivated knowledge based methods. At an early stage in the project it was planned that there would be separate groups for exploring each of these three approaches. However most participants showed interest in all three methods. It also turned out that applications in public telephony systems which obviously appealed to everyone were almost all based on HMMs. Presentations are now

given to the entire group of participants irrespective of the contents. It is obviously impossible to give a comprehensive presentation of the work performed by the partners in COST 232 in the present paper. A more complete overview was presented at the COST232 workshop[1], and outside presentations include [2] - [11]. However many of the presentations made in Working Group 1 address work in progress and have not been presented publicly outside the group. In addition to studies on improved algorithms for clean telephone bandwidth speech a number of studies have been presented which address the specific problems that arise when conditions typical to telephony distort the speech. Such conditions include background noise (e.g. extraneous speech, office noise, car noise), transmission line noise/attenuation and handset microphone variability.

Field trials

Extensive field trials with speech recognition in public telephony systems have been carried out by Telefonica and Jutland Telephone. Both have built services where a number of services are accessible through a single telephone number with the subsequent selection of sub-topics achieved using speech recognition. The Telefonica system has a vocabulary of ten Spanish digits plus the words "yes" and "no". It is a word-based Continuous Density Hidden Markov Model (CDHMM) system with one model per word. Training was carried out using 6000 telephone calls in a simulated environment. During training manual endpointed material is used [2]. In a field trial with Telefonica employees across Spain as participants, 546 calls containing 4549 utterances were collected. 78.3% of the words were correctly recognised. The Jutland Telephone system [3] has a vocabulary of six words (names of desired services such as Sports News, General News, ...). It uses a CDHMM with a total of 23 models for the 6 words (to cover alternative pronunciations) plus a number of models used for rejection of extraneous speech and non-vocabulary words. Training was done with real telephone speech from 128 carefully selected speakers. The field trial period ended in May 1992 and during the trial the system could be accessed by over 1 million subscribers. From May to September 1991 the system received 55000 calls with a successful transaction rate of over 87%. Both

field trials resulted in a conclusion that for applications of the above type a word spotting scheme should be introduced to achieve more satisfactory performance.

Recognition and noise

BT has conducted an experiment investigating the effects of handset and line variability on the performance of a number of commercially available word recognisers. One interesting, although unexpected, result was that the line variability had a significantly higher influence on recognition than the handset variability. At Fondazione Ugo Bordoni the robustness to background noise of some commercial speech recognizers has been investigated. This study, in which noise samples from the NOISE0 CD-ROM added to clean speech in a digit recognition task concluded that although there were differences in performance between the recognisers a 12 dB S/N ratio seemed to be the lower threshold for good operation.

Reference recognizer

Different applications, differing backgrounds and different network constraints mean that a number of diverse approaches to speech recognition are used by the COST232 partners. Also a large number of speech corpora exist in the speech community making the evaluation of the performance of a specific recognition algorithm difficult. In order to ease the comparison of speech recognition and pave the way for standardisation it was decided that a useful contribution from this project could be the definition of a reference speech recogniser.

The reference recognizer was designed based on the following basic principles:

- * It should be a baseline recognizer with no "bells and whistles"
- * It should be HMM-based
- * It should be easy to distribute between the partners and should be easily installed at each location
- * It should be easy to modify the reference recognizer to a more "state of the art" device if the users have specific needs

Based on the above, and on the fact that most partners in the project either already had or expected to obtain the Hidden Markov Toolkit (HTK) developed at Cambridge University it was decided to define the reference recogniser as a UNIX script for HTK. The basic features of the recogniser are:

- Preprocessing: 10ms frame shift using a 30ms Hamming window and a preemphasis coefficient of 0.95, 12 lifted cepstral coefficients derived from 10th order cepstral analysis, normalised log energy and their delta coefficients constitute the 26 dimensional feature vector.
- HMM specifics: 3-state context independent phone models allowing state skipping. Continuous mixture densities with 4 mixture components.

- Speech and annotation files should follow the SAM[13] format.

The first version of the reference recognizer was developed at the Norwegian Institute of Technology. It is now being tested by the other partners on a range of databases to establish its repeatability, stability and transportability.

3. WORKING GROUP 2

Working group 2 is primarily concerned with databases: both collating information about the technology and collecting speech databases.

Information gathering

One sub-goal of COST232, WG2, was to collect information concerning the status of telephony based speech recognition in Europe. This would both overview applications and collate/evaluate experiences gathered from these applications within the group. To achieve this a questionnaire was designed and distributed to different companies via the COST members. This will be available with the final report of COST-232.

Design guide-lines for telephony systems

For telephony based recognition to be both successful and universally deployed in networks, systems will have to cope with a wide range of 'real world' situations. One of the tasks of this sub-group was to identify those factors which must be taken into account when these systems are offered for public services. One assumption made was that the recognition technology deployed over the next 5-10 years will be similar to that which is available in the laboratory today. In essence this means that the technology will allow for isolated words, connected words with restricted syntax, and the limited deployment of keyword-spotting. Continuous speech is expected to be confined to specific and 'natural' domains e.g. natural numbers or amounts of money. The experiences are formulated as the following 'design rules'.

Target users: All potential user populations must be covered including children and elderly people.

Speech mannerisms: The wide range of speech mannerisms such as 'hmms' and 'aahs' exhibited by speakers must be accommodated either by incorporation into models or by using a grammar. The cost of doing this, in terms of worse substitution error rates must be taken into account.

Macro-dialogue: Vocabularies will need to be much less restrictive than they are at present - but functionality will still need to be constrained. Dialogues which use 'coded information' e.g. "say seven for help" must be avoided. Synonyms, however must be incorporated extensively. For example 'Say that again', 'What was that?', 'Sorry', 'Pardon' etc. must all be permissible and recognised as synonyms for 'repeat'. For some applications (e.g.

amounts of money), natural numbers or a minisyntax are appropriate.

Provide sufficient dialogue control: It is unlikely that 'free-form' speech input will be viable for public telephony services within the next decade. This means that the dialogue initiative and control will remain with the system within the foreseeable future. At the same time however, it is certain that the skilled ("power") users will demand sufficient flexibility to interact rapidly with systems and such users may adopt highly structured and precise sub-languages. Important considerations here are 'Barge-in' techniques (interrupting the system speech output), 'short keys' options like 'next', 'previous', 'stop', (e.g. when processing lists of information) and rapid escape paths so that users can return quickly to the root node.

Keep dialogue complexity low: The use of comprehensive audio feedback to provide guidance is not normally very satisfactory. Long messages pose serious difficulties for the occasional user of speech interactive systems who cannot normally remember what has been said and are intolerably slow for the habituated user. Unless the response prompted for is either obvious or intuitive e.g. "What day of the week do you want?" only very short menu choices can generally be supported. At present it is recommended that in any application the depth of a dialogue tree should not exceed three and the maximum number of items to remember (e.g. choices in a menu) should not exceed four at any node. Strategies for adapting to first-time/occasional/skilled/foreign users will also need to be determined.

Micro-dialogue: Validation/correction techniques appear to be very application specific and are determined by the cost penalty associated with a failure to detect as well as correct an error. The evidence so far suggests that there is no general best choice of the validation/correction method to use because it depends so critically upon the content/value of the information at each particular state of the dialogue.

Provide rejection models: The ability to reject 'invalid' words is of paramount importance in even the simplest of applications. False acceptance of an invalid utterance causes more serious problems than a straightforward substitution error creating both asynchrony in the dialogue and bewilderment in the user. As neither the user nor system may be aware of the existence of the error a complex dialogue often compounds rather than solves the problem.

Provide 'garbage-models' for keyword-spotting: This technique helps overcome many of the human factor difficulties introduced by a strict isolated word framework and has the potential to make recognition appear more robust. However the risk of error is always higher when a word is embedded in a phrase than when the same word is in quiet isolation. Keyword spotting should be treated as a 'safety net' which may help catch the occasional embedded utterance rather than a technique which can be relied upon to extract words from

unconstrained inputs at any time. As a general rule it is unwise to do anything which encourages the user to speak too freely and the importance of well designed prompts to 'discipline' users to speak isolated words cannot be overstated. To date there are no indications as to the success (or otherwise) of experiments in which different techniques like keyword spotting and connected digit recognition have been combined in a single application.

The Multi-English European Database

Although there are many speech databases available today, comparatively few contain speech spoken over telephone lines and fewer still contain speech transmitted over international telephone lines. COST provides a good environment for undertaking such data collections and we have taken the opportunity to do this within this project. Although it would have been desirable to have collected the data in all the European languages this would have made the collection too large and so we decided to consider only one language: English. As it will contain English words spoken by non-native English speakers the database will be known as the Multi-English European Database. We have set up two collection points, one at the BT laboratories in Ipswich, UK, and another at the FUB laboratory in Rome, Italy. The two collections will contain exactly the same speakers and words but be collected over different telephone routes. The word list selected was based on the TI databases [14], plus the two command 'yes' or 'no' so that comparisons between the 'clean' speech on the TI databases and the 'telephone' speech on the COST database can be compared directly. Each telephone call records the names of the speakers and their laboratories. Overall we have for each call 23 utterances, i.e. two names (speaker plus laboratory), the digits from zero to nine plus 'oh', and the following commands: erase, rubout, stop, start, help, enter, repeat, go, yes, no. The speaker populations from each laboratory consist of eight males and eight females with half the calls over 'normal' phones and the other half over cellphones (or radiophones). Some recommendations about age, weight, height, etc., are also distributed to the different laboratories in order to distribute speakers to obtain a representative spread of characteristics in the speaker population. Instructions are given both by a form and using spoken prompts. For each of the four calls each user completes a form which contains the list of the words to be pronounced. The vocabulary is divided into three lists of seven with words randomised to avoid obvious sequences like, one, two, three, etc. The database is unique in that it adopts a balanced design, has been collected over international phone lines and contains English words spoken in a very wide range of accents. This database should help us to answer a number of questions concerning telephony, accent, transmission techniques and their effect on recogniser performance. The database has no copyrights and will be made available at cost.

4.CONCLUSIONS

The COST environment provides a useful framework for undertaking cooperative research into speech recognition research. It is particularly appropriate for encouraging the exchange of ideas and news about applications and in this project has provided a focus for the setting of recognition standards and providing a unique type of speech database.

5.REFERENCES

- [1]G.Hult, T.Svendsen: "Activities in COST 232 Working Group 1: Methods and Algorithms", Proc. COST 232 Workshop, Rome, 1992
- [2]M.J.Poza, L.Villarrubia, J.A.Siles: "Influence of the Endpointing during Training and Recognition on the Performance of an Isolated Word Speech Recognition System", Voice Word Wide,pp.69-73,1991.
- [3]B.Baungaard, J.S.Nielsen: "TeleDialog - Talegenkendelse i telenettet", Internal report, Jutland Telephone, October 16, 1991 (in Danish).
- [4]K.Elenius, G.Takacs: "Acoustic-Phonetic Recognition of Continuous Speech by Artificial Neural Networks", Speech Transmission Laboratory, Quarterly Progress and Status Report, vol. 2-3, pp. 1-44, Stockholm,1990
- [5]F.Greco, G.Ravaioli: "An Experiment on Phoneme Classification Through a Time Delay Neural Network", Proceedings, Third Italian Workshop on Parallel Architectures and Neural networks, pp. 407 - 411,1990
- [6]J.Kaja: "A Neural Network Approach to Detection of Voiced Regions", Proceedings of FONETIK-90, The 4th Swedish Phonetics Conference, Umea, pp.108- 112,1990
- [7]B.Kaspar, B.Lochschmidt: "SpeechLex - Phonological Word Modelling Component of an Experimental Speech Recognition System", Proc. Eurospeech-89,vol 1,pp. 530-533,Paris,1989.
- [8]M.Blomberg: "Synthetic phoneme prototypes in a connected-word speech recognition system", Proc. ICASSP-89, vol. 1, pp. 687-690, Glasgow,1989
- [9]G.Hult: "Pulse-by-Pulse Pitch Analysis Through Zero Phase Low-Pass Filtering", Proc. Speech Science & Technology '90, pp. 134-140, Melbourne, Australia,1990
- [10]P.O.Huschy, T.Svendsen: "ANN Based Speech Recognition Using A Preprocessor for Time Compression", Proc. Eurospeech-91, Vol. 2,pp. 563- 566,Genoa,Italy,1991
- [11]A.Basu, T.Svendsen: "A Time-Frequency Neural Network for Phoneme Recognition", Proc. ICASSP'93, pp. I509-I512, Minneapolis,1993
- [12]CCITT Blue Books, VOL V. Rec G703
- [13]User Guide to ETR Tools, ESPRIT P.2589 (SAM), Final Report, UCL - G007, London 1992
- [14]Doddington G., Schalk T., Speech Recognition: Turning Theory to Practice, in IEEE Spectrum VOL. 18, No 9, September 1981