An Overview of the REC Project - Speech Recognition Applied to Telecommunications

Isabel Trancoso, Diamantino Caseiro, Rui Amaral, Frederico Rodrigues, A. Serralheiro
INESC-Lisboa
R. Alves Redol, 9
P-1000-029 Lisboa
Isabel.Trancoso@inesc.pt

Fernando Perdigão, Eduardo Sá Marta
IT-Coimbra
Univ. Coimbra, Pole II
P-3030-290 Coimbra
fp@co.it.pt

Carlos Espain, Vítor Pera, Luís Moreira
FEUP
R. dos Bragas
4050-123 Porto
espain@fe.up.pt

Abstract

This paper is intended as an overview of the REC project. Hence, its starts with a description of the original goals and the reasons which led us to restructure the proposed workplan. A short summary of the main results in each task is then included, followed by a section on which were, in the opinion of the consortium members, the strongest and weakest points of this *umbrella* project. The paper ends with a brief section on future work.

1. Goals

The main goal of REC project was to gather expertise at a national level in the area of speech recognition, particularly in what concerns its applications to the telecommunications domain.

Although research on speech recognition for European Portuguese has achieved an international level through the participation in worldwide conferences and European projects, it is far from being able to compare with the one invested in languages of greater technological impact. Given the wide gap that must be conquered to achieve comparable levels and the relative small dimension of the few national research teams, it is of the utmost importance to join efforts in a common project, sharing corpora, software tools, methodologies, prototypes, etc. This was the main motivation for joining together 3 research teams associated with the Universities of Lisbon (INESC), Coimbra (IT) and Oporto (FEUP).

In the original proposal, 7 priority topics were identified. However, the recommendations of the evaluators of the proposal, together with the fact that almost two years went by since the proposal was submitted in May 1995 until the effective beginning of the project in April 1997, have led us to reformulate the tasks and the corresponding effort, in order to guarantee their actuality. After restructuring, the list of tasks is as indicated below, together with the institution responsible for each task:

- T1 - Recognition of spelled names (INESC)
- T2 - Word and topic spotting (IT)
- T3 - Robust recognition of digits and natural numbers (INESC)
- T4 - Large vocabulary speech recognition (INESC)
- T5 - Speaker recognition (FEUP)
- T6 - Automatic spoken language identification (INESC)

The duration of each task was the duration of the project (3 years) except for task T6 which ended after 2 years.

2. Main results

This section includes a necessarily brief summary of the main results achieved in each of the 6 tasks.

2.1. Task 1 - Recognition of spelled names

In this task we have tried to combine two different motivations: on one hand the study of auditorily-formulated features for place discrimination in stop consonants and on the other hand the development of robust systems for recognizing letters in spelled names, a common fall-back possibility when name recognition fails in directory assistance tasks. It is well known that the most difficult letters to recognize belong to the so-called *e-set* among which the stop consonants are particularly relevant.

The first study was carried out by the IT team [Marta 1999]. The set of features was modeled using fuzzy logic and applied to large databases of monosyllabs and spelled letters in several languages in full-band and low-pass conditions. The rationale behind this is that any valid model of human communication should replicate the human listeners "feats" of very good (albeit not perfect) discrimination of stop place even from speakers of different languages, and of "graceful degradation" when faced with markedly low-pass filtered sounds (e.g. via the telephone network).

The full set of features is too extensive to describe here, but one illustrative example is that of sequence cells, which have recently been studied in primates' higher auditory structures. Ascending sequence cells react to stimuli composed of a lower-frequency event followed by a higher-frequency event, and provide intrinsic robustness against changes in the frequency-response curve (including low-pass conditions) since the sequence survives as long as neither of its two components is completely obliterated. Ascending sequence cells were found to provide a primary feature for the discrimination of labial consonants against confusable categories. Just this one feature is able to yield correct discrimination of labial against dental consonants upwards of 94% for both full-band and low-pass telephone-like letter sounds with no adaptation measures. This result was obtained, in speaker-independent mode, in the well-known ISOLET database, for B and D letters. The low-pass result is close to human listener performance in this discrimination and exactly the same model yields similar results in the discrimination of the Portuguese letters P and T. Thus, in some difficult sub-tasks, human capabilities (not shown by conventional automatic recognizers) are being approximated.

The INESC work in this task used a subset of the SpeechDat corpus [Trancoso 1998] containing: one spontaneous item corresponding to the spelled forename of the speaker, and two read items corresponding to spelled city names (500 different cities) and words (set of 4000 different words, mostly proper names). Following the consortium rules for train and test set definition, an overall ratio of 80% / 20% between the two sets respectively, was achieved, maintaining the same proportion of age, gender and region distribution in each of them.

The feature extraction stage used conventional MFCC (14 cepstra + 14 delta-cepstra + energy + delta-energy), computed over 25 ms Hamming windows, updated every 10 ms. Some preliminary experiments using RASTA-PLP parameters were also performed.

As in all the other tasks in which the INESC team was involved, the acoustic modeling stage was based on continuous density Hidden Markov Models (CDHMM). The model topology was left-right, but the number of states varied according to the entities modeled (e.g.: 7 for the letter M pronounced as [Em@] and 5 for the same letter pronounced as [me]). Some words that were used to specify diacritics such as "acento" and "circunflexo" were naturally modeled using a larger number of states.

Using an ergodic model that allows every letter in every possible position, we have achieved a correction rate of 59%, on the training set. Using a pronunciation lexicon with all the 2676 different names in the training set, we have obtained 95%. This corresponds to an average perplexity of 1.7. When the pronunciation lexicon included the most frequent 12,000 proper names in Portuguese, the rate decreases to 92%.

The fact that there is not a single standardized way of pronouncing letter names in Portuguese significantly contributed to the difficulty of this task.

2.2. Task 2 - Word and topic spotting

After some initial work on keyword spotting that started even before the beginning of the project, and in which state-of-the-art filler models were adopted, the major effort in this task was redirected towards topic detection in spoken documents and was carried out by the INESC team.

The approach that was implemented at INESC is still undergoing preliminary tests. The basic idea is to develop a topic detection system for text and apply it later to the output of a continuous speech recognizer. As a train and test corpus, a subset of the BD-PÚBLICO newspaper text corpus was adopted. It includes 20,000 stories, spanning a 6-month period from September 95 to February 95. The stories are topic labelled using very broad categories, and each story contains between 100 and 2000 words.

The training phase is based on a two-stage unsupervised clustering algorithm, using nearest-neighbor search and the Kullback-Leibler distance measure. Topic models were created using the CMU-Cambridge Statistical Language Modeling Toolkit and are based on smoothed unigram statistics. Topic decoding is based on HMMs. We expect to have results for inclusion in the final paper.

2.3. Task 3 - Robust recognition of digits and natural numbers

The work carried out in this task was two-folded: one of the teams (IT) established as a major goal the improvement of noise robustness through the use of auditory models. The other (INESC) adopted more conventional acoustic feature extraction techniques, concentrating instead on modeling issues.

One important aspect of a speech recognition system is the analysis front-end. The dominant analysis technique to obtain speech characteristics for recognition purposes is MFCC analysis. This technique has shown reasonable properties when speech is corrupt by additive noise and is appropriate for normalization against convolutional noise. However, there are perceptual cues present in the speech signals that are not exhibited by this features. In order to investigate the advantages of incorporating the properties of the peripheral auditory system in the features vectors, we carried out a study on auditory models. Firstly we developed a physical cochlear model intended to represent the human cochlea [Perdigão 1998a]. Then a study of several auditory models was conducted in order to investigate their differences. A computationally efficient auditory model was developed incorporating a noise suppression mechanism [Perdigão 1999]. Finally, experiments with isolated digits was conducted in order to compare different analysis methods.

All the proposed auditory models, despite some differences, represent the main characteristics of inner-hair cells and auditory-nerve fibers: half-wave rectification, dynamic range compression and adaptation. We believe adaptation is one of the most important property of the auditory functioning and it is useful for recognition. Adaptation consists on a decrease of fiber response for a sustained excitation. This means that when a spectral component of a signal begins, the firing rate is very high before attaining a steady and much lower rate. This mechanism is important for the detection of plosive sounds or the onset/offset of vowel sounds or even formant trajectories.

An auditory model was developed using a gamma-tone filter-bank and the Martens-Immerseel adaptation model. Adaptation is simulated using the square root of the short-term energy at the output of each channel of the filter-bank. The first recognition results we obtained with auditory models showed they tend to be very sensitive to noise. Noise not only increases the means of firing-rate based features but also adapts the responses due to speech. In order to overcome this problem, a suppression mechanism was included in the proposed model. The recognition tests with additive and convolutional noise show that, in fact, noise degrades the recognition performance in this model and tested front-ends [Perdigão 1998b] but the noise suppression mechanism turned out to be very effective. The proposed auditory model is not the best in all tests, however it globally presents a good and consistent performance. As a conclusion, auditory models seems to present some properties useful to speech recognition but are also sensitive to noise. Further research must continue in order to understand how speech is encoded in the auditory system and how humans attains so much robustness in speech recognition.

The work carried out at INESC in this task used a subset of the SpeechDat corpus: the isolated digits, the sequences of 10 different digits and the natural numbers. This subset was then divided into train and test sets following the above mentioned criterion.

The feature extraction stage used conventional MFCC. Cepstral mean subtraction was adopted to achieve some channel and speaker normalization. Acoustic modeling was based on gender dependent CDHMMs. The models topology was basically left-right but varied according to the type of data modeled. For digits, word models with no skips were adopted, with a different number of states (3 to 8) depending on the average digit duration. For natural numbers, 3-state context-independent phone models were adopted, still keeping the word models for the digits.

The problem of modeling extraneous events was carefully investigated in this task. The orthographic annotation included marks of speaker noises, stationary noises, intermittent noises and filled pauses. Best results were achieved using a unique filler model topology with 9 different filler and silent models in parallel and a backward arc.

The isolated digit models were first trained using the Baum-Welch algorithm, and the number of mixtures per state was incrementally increased up to 3. The isolated digit models were then used as bootstrap models for training connected digit models with 5 Gaussian mixtures per state. The subword models were initialized from models trained from the directory assistance task. Best results were achieved with 5 mixtures.

The performance of the system was evaluated using Viterbi decoding. For isolated digits, it reached 99.4% accuracy and for connected digits it reached 98.0% or 96.1%, depending on the grammar (known length vs. unknown length). For natural numbers, using a grammar ranging from zero to hundreds of millions, the accuracy reached 98.4% [Rodrigues 1999].

An application was developed to evaluate on-line the system performance and acceptability. The user can phone the system and choose (using isolated digit recognition) whether he intends to test the connected digits recognition system or the natural numbers one. The recognized string is synthesized using either the SVIT directory service application (for connected digits) or the DIXI Text-to-Speech synthesizer (for natural numbers).

Further progress in this area can certainly be achieved through the inclusion of further training material, for instance: digit strings with consecutive repeated digits (not present in the adopted subcorpus). The natural number task can also be adapted to other applications such as recognizing money amounts. But apart from these training data issues, we think that the use of context-dependent phone models could be potentially advantageous, as well as the adoption of discriminative training techniques.

2.4. Task 4 - Large vocabulary speech recognition

Two teams were involved in this task. The FEUP team has not been involved in telecom-oriented speech recognition, because of the unavailability of appropriate spoken corpora which could be shared with this team. Instead, the team focused their work on the design of a semi-continuous HMM recognizer which could serve as a baseline system for comparing single-stream and multi-stream approaches. Latest results with the single-stream approach have yielded an acceptable accuracy: 6.7% word error rate on the February 89 test set for the Resource Management (RM) database.

The basic idea of their multiple-stream approach is to extract multiple information streams from the speech signal and to investigate different ways of recombining model scores at certain points in the recognition process. Two main approaches are planned for the near future: one is to recombine the information contained in word-graphs [Ney 1994] produced by separate recognizers; a second one is to process the input streams independently up to a pre-defined sub-word level, and to recombine the model scores at that level [Bourlard 1996].

The INESC team could take advantage of the availability of the SpeechDat corpus to concentrate their efforts in the recognition of items which are particularly important for directory assistance applications. Two Master Theses were devoted to this theme in the beginning of the project, for toponyms and person names. The work described here, however, is much more recent and intends to be more general, in the sense that instead of developing task-specific models, the aim was to train general purpose acoustic models which could later be tuned to specific tasks. The SpeechDat subset used for training included only phonetically rich words and sentences from 80% of the total set of speakers (close to 70 hours). The test was done with the remaining 20%, and with different items: spontaneous forenames and city names, read city names and read forename-surname pairs.

The feature extraction stage was the same as described for digit recognition and gender dependent monophone models were initially built for 39 phones, using 3-state left-right CDHMMs, with no skips. Silence and filler models used forward and backward skips. After training monophone models, word internal tied state triphone models were trained, using tree based clustering. a total of 13k triphone models were built, with 8498 shared states.

On a development set using a lexicon of 2356 phonetically rich words, the recognizer achieved an accuracy of 90.7% with 6 Gaussian mixtures per state. On the different test sets, the scores varied: 84.8% (spontaneous names - closed set of 640), 91.2% (spontaneous city names - open set of 500), 92.6% (read city names -closed set of 500) and 89.1% (forename-surname pairs - closed set of

150). The results are promising but further work is needed to improve the general purpose models, create task-specific ones and fine tune the recognition system.

2.5. Task 5 - Speaker recognition

The work done in this task involved only the FEUP team. During the first year of the project, three basic speaker recognition systems were developed, based on DTW, neural networks and HMMs. These text-dependent systems did not provide encouraging scores, specially due to the reduced training material.

During the second year, the efforts were concentrated on developing a cepstral based text independent VQ system using sentence codebooks [Moreira 1999]. The system was designed by producing first a codebook from each utterance and then quantising this codebook into the claimer's codebook. This way, silence is reduced to a few points of the sentence codebook, whatever its duration and/or location might be. This eliminates the need of endpoint detection, and reduces the effects of noise.

In the third year, several techniques for combining the results of different 64 points codebooks based speaker recognizers using different features sets were experimented, using TIDIGITS and a small locally recorded database in Portuguese. Multiple experiments were made by varying the amount of data used for training and the length of the test sentences. Best performance was achieved with codebooks generated from 25.2 s of average training time using cepstral and delta cepstral coefficients as the features set. One recognizer used the normalized error quantisation of the sentences frames into the claimer's codebook as a response and the other used the quantisation error of the sentence codebook into the claimer's codebook. Using both the product and the average between the two responses as the final decision criterion an EER of 0.6% with testing utterances of 3.2 s was obtained for the TIDIGITS, with a speaker dependent threshold. A total of 5940 impostor's sentences and 1210 claimer's sentences were used involving 110 speakers. A prototype over the telephone was built and is currently under test. The application runs in DELPHI using a ProLine/2V interfacing board.

2.6. Task 6 - Automatic spoken language identification

The responsibility of this task belong to the INESC team which established as a major goal the development of an automatic spoken language identification system easily extendable to new languages. The best systems reported in the literature make heavy use of linguistic data, using multiple large vocabulary continuous speech recognizers, one for each language. Due to the difficulty of adding a new language, those systems are generally limited to a very small set of languages.

Relatively good results can be obtained by exploiting the phonotactic properties of the languages. A particularly successful approach is parallel language dependent phoneme recognition followed by language modeling, in which the utterance is decoded by multiple language specific phoneme recognizers and the phonotactics of the resulting sequences are matched to language models. This approach does not need textual data and is able to achieve identification rates in excess of 80%, using 10-second utterances in 6 languages. Its biggest drawback is the requirement of labelled speech for a large subset of the languages used. As it is based on multiple language-specific phoneme recognizers, it requires labelled speech to train those recognizers.

Instead of using multiple language specific phoneme recognizers, it is possible to obtain the same level of identification using only one language independent phoneme recognizer. By performing multiple recognitions of the input utterance with the same models, and constraining each recognition by a different phone-bigram grammar (obtained from manually labelled transcriptions), it is possible to obtain multiple phoneme streams from the same utterance. Those streams can then be fed to stream- specific language models. This approach, called double-bigram decoding [Navrátil 1997], requires less data than the previous one, but, nevertheless, it still requires labelled speech in each language to model the language independent subword units, and it also requires textual data and pronunciation dictionaries to create the bigram grammars.

In the approach adopted by the INESC [Caseiro 1998], a single language specific phone recognizer is used (in our case the Portuguese one). The global architecture is similar to double-bigram decoding, but, by using only one language specific phone recognizer and a language model bootstrapped from the sequences recognized from the train utterances, it requires no more than speech data to be extended to new languages.

We studied the problem of language identification in the context of the 6 European languages present in the first phase of the SpeechDat collection (English, French, German, Italian, Portuguese and Spanish). Using only the phonetically rich sentences of the 6 languages, our preliminary system has achieved an identification rate of about 80% on 5-second utterances. The identification of some languages was clearly impaired due to proximity with other languages, as was the case with Spanish and Italian.

The baseline system was improved by the introduction of a classifier based on Multi-Layer Perceptrons (MLP) which replaced the decision module of the previous system that did not take into account the high redundancy in the output streams of the models for the same language. The extended system reached an average identification rate of 83.4%. The influence of sentence duration in language identification was also studied. As expected, the performance increased with duration, reaching 86.1% for utterances of 7s. Beyond 8-second duration, our results were not significant, due to the very limited number of such long sentences.

3. Internal evaluation

Being involved for the first time in such an *umbrella* type of project, we thought it was worth trying to identify the most positive and negative non-technical aspects of the project. The most positive aspects were, in our opinion, the coordination of efforts among teams, the dissemination of results (close to two dozen papers), the formation of highly qualified post-graduate students (close to a dozen "doutoramento" and "mestrado" theses, four of which are still in progress), and the intense cooperation with other projects. As the most negative aspects we have identified the difficulties we faced in sharing linguistic resources (due to unforeseen reasons, which were not in the scope of the project), and in recruiting new human resources.

The high point of the REC project was the final REC Workshop, which took place on May 8th, organized at INESC, in Lisbon, by the three teams. The workshop was open and free of charge to all participants, having gathered close to 50 attendees from industry and academia from all over Portugal. The morning session included an overview of the project and 3 keynote presentations from distinguished researchers (Mazin Rahim, Borge Lindberg and Lori Lamel), from which we obtained very valuable (and positive) feedback. The afternoon session included 8 presentations from the team members and 1 from a companion project. A final panel session was centered around *the role of research centers in terms of ASR for the telecom environment*. The role of developing new systems is definitely taken over by the big telecom operators and companies, and the time it takes to port an application to a new language is getting increasingly shorter. What should then be the role of the small research centers such as the ones in this consortium? Consultants? Providers of linguistic resources? Validation centers? Probably all of the above, but primarily, such research centers should have the role of seeking new challenges and finding new methods to deal with them.

The public website of the project contains information about papers, theses, demos, companion projects and the final workshop[1].

4. Future work

In spite of the amount of effort which has been devoted to speech recognition in the telecom environment in a worldwide scale, this cannot be considered at all a solved problem. Recent progress has been mainly achieved at the cost of larger amounts of data for training and larger computational power. Yet, the scientific community keeps setting increasingly difficult goals which tend to approximate automatic recognition to human recognition.

A major challenge in ASR is to cope with increasingly larger vocabularies (500k?). Another one is robustness to the wide variety of speaking styles, adverse environments and transmission channels. In this context, the problem of modeling pronunciation variation in fluent speech is particularly relevant, and the same can be said about non-native pronunciation; another major robustness issue

[1] http://www.speech.inesc.pt/rec/rec_en.html

is the quality degradation in mobile network recognition, a problem which we could not deal with yet, due to the fact that we do not have this type of spoken corpus. An application area where all these issues are particularly relevant is broadcast news recognition.

Nowadays, a significant part of the research effort in this area is centered in the development of robust spoken dialogue systems. We think that it is of the utmost importance to invest as soon as possible in this truly multi-disciplinary topic, combining the efforts for the development of high quality speech synthesis with those for speech recognition and understanding, natural language generation, dialogue design, etc.. A major challenge in this area is to find better ways of combining syntactic and semantic knowledge sources (e.g. topic-based language models and parsers).

The robustness of a spoken dialogue system can also be potentially improved at the cost of integrating prosodic information, a relatively new field of research in terms of speech recognition and understanding. For instance, prosodic features could be combined with information available to the recognizer to help detect recognition errors or attempts from the user to correct these recognition errors.

The need to invest in the development of multimedia multilingual systems is also unquestionable. The integration of several modalities is not nowadays a problem for applications in the telephone channel environment, but it will become undoubtedly more relevant in the near future. The access via the telephone channel to remote websites in one language using queries expressed in another language is one of the applications that makes it urgent to invest in the integration of machine translation techniques with spoken dialogue systems.

For European Portuguese, the research topics that were listed may seem too ambitious given the current state of the art and the relatively scarce human resources in this area. However, we think that it is important to define a strategic plan for speech recognition research in our language that encompasses the listed topics among its research goals. Although the examples that we mentioned do not exhaust by all means the list of research topics that should be included in such a strategic plan, the quoted examples are in our opinion some of the most pressing ones.

Acknowledgements

Our thanks go to FCT for having funded this project and to all the colleagues and students that collaborated with us. Special thanks are due to the teams of Prof. Francisco Vaz (Univ. Aveiro) and Prof. João Neto (INESC) for having honoured us with their participation in several technical meetings.

References

Bourlard H., and Dupont S. (1996), A new ASR approach based on independent processing and recombination of partial frequency bands. *Proc. ICSLP'96*, Philadelphia, U.S.A.

Caseiro D., and. Trancoso I. (1998). Spoken language identification using the SpeechDat corpus. *Proc. ICSLP'98*, Sydney, Australia.

Marta E, and Sá L. (1999), Auditory features for human communication of stop consonants under full-band and low-pass conditions. *Proc. EUROSPEECH'99*, Budapest, Hungary.

Moreira F., and Espain C. (1999), Text-independent speaker verification using string codebooks. *Proc. COST 250 MCM*, Porto, Portugal.

Navrátil J., and. Zühlke W. (1997), Double-bigram decoding in phonotactic language identification. *Proc. ICASSP'97*, Munich, Germany.

Neto J., Martins C., Meinedo H., and Almeida L. (1997), The design of a large vocabulary speech corpus for the Portuguese. *Proc. EUROSPEECH'97*, Rhodes, Greece.

Ney H. and, Aubert X. (1994), A word graph algorithm for large vocabulary continuous speech recognition. *Proc. ICSLP'94*, Yokohama, Japan.

Perdigão F., and Sá L. (1998a), Modelo Computacional da Cóclea Humana. *Proc. Acústica'98* - Congresso Ibérico de Acústica, Lisbon, Portugal.

Perdigão F. and Sá L. (1998b), Auditory Models as Front-Ends for Speech Recognition. *Proc. NATO ASI on Computational Hearing*, Il Ciocco, Italy.

Perdigão F., and Sá L. (1999), A Noise Suppression Technique using an Auditory Model. *Proc. CONFTELE'99*, Sesimbra, Portugal.

Rodrigues F., and Trancoso I. (1999), Digit Recognition using the SPEECHDAT Corpus. *Proc. CONFTELE'99*, Sesimbra, Portugal.

Trancoso I, and Oliveira L. (1998), SpeechDat Portuguese database for the fixed telephone network, *Final report*.