Speech Recognizer Quality Assessment for Linguistic Engineering (SQALE)

Lori F. Lamel

LIMSI-CNRS
91403 Orsay, France
{lamel}@limsi.fr

ABSTRACT

The aim of the LRE-SQALE project (Speech recognizer Quality Assessment for Linguistic Engineering) is to experiment with establishing an evaluation paradigm in Europe for the assessment of large-vocabulary, continuous speech recognition systems in a multilingual environment. This 18 month project is will define and carry out the assessment experiments, paving the way for future projects with a larger scope and wider participation of European sites. The SQALE Consortium consists of a coordinator, the Institute for Human Factors at TNO, and three laboratories (CUED, LIMSI-CNRS, PHILIPS) who will evaluate their recognition systems using commonly agreed upon protocols, with the evaluation organized by the coordinating laboratory. Multiple sites will test their algorithms on the same database, so as to compare the merits of different methods, and each site will evaluate on at least two languages, so as to compare the relative difficulties of the languages, and the degree of independency of the algorithm to a given language.

INTRODUCTION

The objective of the European Community LRE (Linguistic Research and Engineering) Speech Recognizer Quality Assessment for Linguistic Engineering project (SQALE) is to experiment with installing in Europe an evaluation paradigm for the assessment of large vocabulary, continuous speech recognition systems in a multilingual environment, to identify the problems encountered, and improvements needed. The SQALE project started at the end of December 1993, and will have a duration of 18 months, so as to get a quick evaluation of the feasibility of the installation of such an assessment infrastructure in Europe. In order to do so, a small consortium was formed with four partners, so as to be able to efficiently define and carry out the assessment experiments. The merits of periodic assessment are becoming more widely apparent, and as such SQALE hopes to pave the way for future projects with a larger scope and wider participation of European sites.

The SQALE Consortium is coordinated by the Human Factors Research Institute (TNO-TM, former Institute for Perception) which belongs to the Netherlands organization for applied scientific research (TNO) with extensive experience in speech recognizer assessment. The role of the coordinator is to organize the assessment experiments which will be carried out by the remaining three partners of the Consortium. These three laboratories are from three different countrie s each with their own language: Cambridge University Engineering Department (CUED) in Great Britain, the Laboratory for Mechanics and Engineering Sciences of the National Center for Scientific Research (LIMSI-CNRS) in France, and the Man-Machine-Interface group from Philips Research Laboratories (PHILIPS Aachen) in Germany. Each of the testing sites will evaluate their own recognition system using the commonly agreed upon protocol, with the evaluation organized by the coordinating laboratory. All three test sites have participated in at least two evaluations organized by the US ARPA (Advanced Research Projects Agency) Speech and Natural Language programme, demonstrating their willing compliance with the test protocols, and to exchange information, and their capability to conduct large scale testing. All three testing sites will evaluate their algorithms on the same database, so as to compare the merits of different methods, and each site will evaluate on at least one otherlanguage, so as to compare the relative difficulties of the languages, and the degree of independency of the algorithm to a given language.

THE ASSESSMENT PARADIGM

The ARPA Speech and Natural Language programme which started in 1984 is based firmly on an 'assessment' paradigm. This paradigm involves the sharing of speech and text data for training and testing the recognition systems according to common test protocols, and comparing results and methodologies in order to improve speech recognition technology. This approach has resulted in the creation of large speech and text corpora which have been distributed among participants for benchmark tests, such as the well known and widely used Resource Management Corpus[9] and the Wall Street Journal Corpus[7]. The systems are assessed on a common basis in order to both develop and test speech recognition/understanding systems. This paradigm is also applicable to natural language analysis, or to other areas such as character recognition or computer vision.

For speech processing applications, the benchmark tests are organized by the National Institute of Standards and Technology (NIST) who distributes the test material and receives the results which are then scored and analyzed. Initially only ARPA contractors participated in the tests, which were later opened to include other laboratories outside the ARPA community. This approach has been shown to be an effective method of raising the scientific and technical standards of the participants, and to lead to fruitful exchange of scientific information. While the approach taken in this project will be largely based on the assessment activities used in the U.S. ARPA community, it is an open and challenging issue as to how to adapt this paradigm to the inherent multilingualism of Europe.

PROJECT OVERVIEW

As stated earlier, the goal of the SQALE project is to experiment with applying the ARPA-style assessment of speech recognition technology in a multilingual environment. The systems to be evaluated are state-of-the-art, research speech recognition systems[1] that are able to recognize large vocabulary continuous speech, so as to improve technology and assessment methodology.

The assessment experiments in this project will make use of common speech and text data for training and testing the recognition systems. Since these assessment experiments will be carried out in a multilingual framework, a primary concern is how to define comparable conditions for the different languages. The consortium will define an assessment methodology, based on the members' combined knowledge of evaluation methods and measures (taking into account the experience gained within ARPA and SAM), and of building recognition systems. The definition will consider the following points: specification of training material and training protocols, selection of a vocabulary list and of a common language model, selection of test materials, format of the recognizer output for scoring, scoring procedures (string matching), performance metrics (recognition/understanding), definition of reference answers, tabulation of official results, and statistical significance of results.

The two independent research questions addressed by this project are:

1. What are the merits of different recognition algorithms applied to the same data?
2. What are the relative difficulties in speech recognition across languages?

To do so, it will be necessary to define the conditions for the evaluation so as to be as equivalent as possible across the languages. The recognition task will be large vocabulary (at least 20,000 words), speaker-independent, continuous speech recognition. A baseline condition for comparison of systems is to use comparable amounts of acoustic and language model training data and the same vocabulary size for each language.

Each site will evaluate their system for American English and at least one other language. By having multiple sites testing their algorithms on the same database, it will be possible to compare the different methods for the same data. By testing the same algorithm for two databases in two different languages, it will be possible to determine the relative difficulties of the two languages, and the degree of independence of the algorithm to a given language.

Corpora

Each testing laboratory is responsible for providing data in their own language for use within the project. These data include spoken corpora with sufficient data for training speaker-independent recognition systems, corpora for training language models, as well as training and recognition lexicons. At the time of this writing the decision has been made to use the following training corpora, each containing at least 10 hours of speech material recording under high quality conditions. The recordings were all made digitally at 16kHz using a and have a signal to noise ratio of at least 35 decibels. The corpora are all orthographically transcribed. These corpora are all either already or are expected to soon be publicly available. Each site will also provide a common language model and lexicon to be used for evaluation.[2]

American English: A portion of the ARPA Wall Street Journal Corpus[7] will be used within this project. The standard WSJ0 SI-84 speaker-independent training data include 7240 sentences from 84 speakers. The language model training texts consist of 37 million words which were standardized[7]. This corpus is available through the LDC.

British English: For British English, the British English Wall Street Journal (WSJCAM0) corpus[13] will be used. This corpus is a British English version of the American English WSJ0, and uses the same prompting texts. 7000 spoken sentences from 90 speakers will be available for training. The Wall Street Journal text material will be used to provide language model training data. This corpus will soon be available through the LDC.

French: For French, the subcorpus BREF80 a part of the BREF-*Le Monde* corpus[8], produced at LIMSI will be used. The corpus containing about 5500 sentences from 80 speakers is about the same size as WSJ0. Language model training texts will come from 37M words of newspaper text from *Le Monde*. This corpus will soon be available from LIMSI.

German: For the German acoustic training data will be a portion of the PHONDAT corpus[11] containing about 2000 sentences from 100 speakers. The language model training data will come from about 36M words of the newspaper *Frankfurter Rundschau*. This corpus is available through the University of Munich.

The development data will consist of 10 sentences read by each of 20 speakers. The BREF and WSJCAM0 corpora both contain sufficient development data that will be used within SQALE for the dryrun test. Development data for German will be recorded by TNO, under the comparable conditions used for recording the training data. While the BREF corpus also contains evaluation sufficient evaluation test data for the evaluation test, data from a few speakers will need to be recorded to complement the available test data from WSJ0 (and WSJCAM0. Evaluation test data will be recorded for German. TNO will be responsible for assuring the quality of the test data.

From the available (or to be recorded test data), the test sentences will be selected taking into account the following criteria:

- The percent of words which are out-of-vocabulary (OOV). It will be attempted to have the same OOV rate, about 1.5-2.0%, for all languages.
- The test set perplexity should be about the same for all test sets.
- The test data will be selected to cover a wide range of words and contexts, and to achieve good phonetic balance.

[1] While a description of the partner's recognition systems is beyond the scope of this paper, such descriptions may be found in [10, 2, 3, 12].

[2] All sites will be required to use a common language model and a common vocabulary list. A common lexicon will be provided, but it's use is not imposed. Each site is free to modify the lexical transcriptions as they so desire.

- An attempt will be made to select sentences containing between 10 and 30 words.
- The test speakers are to be balance for gender.

Assessment Experiments

The assessment experiments will be organized by the coordinator, who will verify the quality of the test data prior to distribution, and will have the final decision as to the reference answer. This procedure is similar to the organization of the benchmark tests by NIST (National Institute of Standards and Technology) for the ARPA test paradigm. The assessment protocol, including the use of the training and test material, the reporting of each sites experimental results, and the (statistical) evaluation, will be exercised in a dry-run evaluation. The dry run will provide important feedback on the assessment protocols and statistical evaluation of the reported results. After the dry run the assessment guidelines will be reviewed and modified as appropriate before the final official evaluation. The proposed cycle of evaluation, followed by development and refinement of assessment methodologies and guidelines, is expected to lead to improvements in speech recognition technology and in assessment methodology. Human benchmark tests will be performed on the same data.

Time Scale

The project started in December 1993. The first 6 months of the project were devoted to defining the assessment protocols and to preparation of the acoustic and language modeling training corpora. These corpora have been delivered to the coordinator, who will distribute them shortly to all of the partners. In October 1994 any needed development and evaluation data will be recorded by the coordinator at the appropriate partner's site, with the development data to be distributed by the end of November. The dry run evaluation is planned for January 1995, with the evaluation test to follow about months later.

INTERNATIONAL COLLABORATION

While the aim of SQALE is to carry out a practical assessment experiment in a multilingual context, it clearly has relationships with other ongoing LRE and international projects. The closest links are with other assessment activities, primarily ARPA, EAGLES and COCOSDA. ARPA has the most extensive experience world-wide in conducting coordinated evaluation tests of state-of-the-art, large vocabulary, continuous speech recognition research systems. Having all participated in the November 1993 ARPA benchmark test, the SQALE partners maintain close contact with the ARPA community and can draw on common experience for this project. All partners will also participate in the upcoming November 1994 ARPA benchmark test.

The LRE Expert Advisory Group on Language Engineering Standards (EAGLES) Spoken Language Working Group is working to coordinate and define standards for corpora and assessment methodologies. The standards defined in EAGLES, many of which are extensions for the ESPRIT SAM and SAM-A projects, will be taken into consideration in defining the SQALE assessment activities and the practical experience gained in carrying out the SQALE assessment experiments will provide valuable input to this group. The aims of the SQALE and SAM projects are quite different, however. Where SQALE aims to evaluate speech recognition technology by evaluating research systems, the objective of SAM-A was to assess commercial and near commercial systems in relation to a task and rate them.

The importance of assessment activities is the subject of substantial international attention. The Coordinating Committee on Speech Databases and speech I/O systems Assessment (COCOSDA) was founded with major support coming from the European Speech Communication Association (ESCA) as a result of meetings at Noordwijkerhout (ESCA ETRW, 1989), Kobe (ICSLP, 1990), and Chiavari (Eurospeech, 1991) with the aim of providing international cooperation for the development of corpora and assessment methods. As speech products start to reach the market place, these issues will become even more important. More recent meetings have been held in Banff (ICSLP 1990), Berlin (Eurospeech, 1993) and Yokohama (ICSLP, 1994). At these last two meetings a presentation of the SQALE project was made with an ensuing discussion on the issues involved in multilingual assessment methodologies.

SQALE also has relationships with projects concerned with corpora development and dissemination. Due to funding and timing constraints, SQALE will use existing corpora for training, and will only record test data when needed. The LRE RELATOR project aims to catalog existing linguistic resources, and to define an infrastructure for the reusability of such existing resources. The LRE project EUROCOCOSDA aims to coordinate participation in COCOSDA at the European level, which has actions in Corpora, Synthesis and Recognition. Any corpora created within the framework of the SQALE project will be made available to the European Community via appropriate dissemination infrastructures.

CONTRIBUTION TO LRE PROGRAM OBJECTIVES

In the LRE'1992 Call for Proposal, the theme 2 of the Sub-Area 1 ("Research Aimed at the Improvement of the Scientific Basis of Linguistic Technologies") was on the "Assessment and Evaluation of NLP Systems." The workprogram made specific mention in the Project Profile of the ARPA program and evaluation paradigm: "DARPA's assessment-led program methodology has proven successful in stimulating scientific interchange with respect to shared tasks and data, thus contributing to a better understanding of the problems and to cooperative efforts aimed at solving them."

This project focuses on a practical experiment on the feasability of adapting the ARPA assessment paradigm in the European context, taking into account the cross-language aspects. By carrying out an assessment cycle, it will identify problems related to this adaptation in an international framework. A mechanism will be initiated in which it will be possible to enlarge the number of participants and number of languages, as well as to increase the difficulty of the tasks assessed.

CONCLUDING REMARKS

The results of this project will be of key importance for the development of future speech and natural language systems in Europe and will serve as a guideline for future projects or a future European infrastructure concerned with assessment of technology. The SQALE project is the first European (or worldwide) attempt to adapt the ARPA "Assessment paradigm" to a multilingual context, and will serve as a baseline from which more advanced research tools and metrics can be developed. The project reinforces relationships among European laboratories by providing a common, collaborative framework for testing systems and sharing data. Assessment of state-of-the-art speech research systems is crucial for development of technology, which is in turn crucial for the eventual development of applications and future commercial products.

Future extensions of this work will include comparisons to human benchmarks, and tests based on spontaneous speech and speech recorded in more realistic conditions.

Some concrete results of the project will be guidelines of the assessment protocols and recommendations for improvements, the availability of multilingual corpora (text, speech, language models), the results of evaluations and their analysis. Dissemination of the project results will be through public presentations at related conferences and workshops, as well as through written publications. The SQALE project has also been presented at the ARPA Workshop on Spoken Language Technology in March 1994, and at the COCOSDA workshop in September 1994. A final workshop organized by the coordinator will be open to researchers representing major European projects (LRE and ESPRIT speech projects), as well as representatives from major international projects (ARPA, VERBMOBIL (funded by the German government)), and representatives from the European Commission.

SQALE Partners:

- TNO-TM Human Factors Research Institute (Coordinator, Netherlands)
- LIMSI-CNRS (France)
- PHILIPS-Aachen (Germany)
- CUED (England)

Project Duration: 18 months

References

1. Steeneken, H.J.M. and Lamel, L.F. "SQALE: Speech Recognizer Quality Assessment for Linguistic Engineering," *Proceedings* ARPA *Workshop on Spoken Language Technology*, Plainsboro, New Jersey, March 1994.
2. Gauvain, J.L., Lamel, L.F., Adda, G., and Adda-Decker, M., "The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA Wall Street Journal Task," *Proceedings IEEE ICASSP-94*, Adelaide, Australia, April 1994.
3. Gauvain, J.L., Lamel, L.F., Adda, G., and Adda-Decker, M., "Continuous Speech Dictation in French," *Proceedings ICSLP-94,* Yokohama, Japan, September 1994.
4. Pallett, D.S., and Fiscus, J.G. "Resource Management Corpus - Continuous Speech Recognition - September 1992 Test Set Benchmark Test Results," *Proceedings Final review of the D*ARPA *ANNT Speech Program,* Palo Alto, California, September 1992.
5. Pallett, D.S., Fiscus, J.G., Fisher, W.M., and Garofolo, J.S. "Benchmark Tests for the DARPA Spoken Language Program," *Proceedings* ARPA *Human Language Technology Workshop*, Plainsboro, New Jersey, March 1993.
6. Pallett, D.S., Fiscus, J.G., Fisher, W.M., Garofolo, J.S., Lund, B.A. and Pryzbocki, M.A., "1993 Benchmark Tests for the ARPA Spoken Language Program," *Proceedings* ARPA *Human Language Technology Workshop*, Plainsboro, New Jersey, March 1994.
7. Paul, D.B. and Baker, J.M., "The Design for the Wall Street Journal-based CSR Corpus," *Proceedings ICSLP-92*, Banff, Canada, October 1992.
8. Lamel L.F., Gauvain, J.L. and Eskénazi, M., "BREF, a Large Vocabulary Spoken Corpus for French," *Proceedings Eurospeech-91*, Genoa, Italy, September, 1991.
9. Price P., Fisher, W.M., Bernstein, J., and Pallett, D.S., "The DARPA 1000-word Resource Management Database for Continuous Speech Recognition," *Proceedings IEEE ICASSP-88*, New York, 1988.
10. Woodland, P.C., Odell, J.J., Valtchev, V. & Young S.J., "Large Vocabulary Continuous Speech Recognition Using HTK," *ProceedingsICASSP-94* , Adelaide, Australia, April 1994, Vol. 2, pp. 125-128
11. TIllmann, H., *Proceedings of the 1992 COCOSDA Workshop*, Banff, Canada.
12. Aubert, X., Dugast, C., Ney, H., and Steinbiss, V., " Large Vocabulary Continuous Speech Recognition of Wall Street Journal Corpus", *Proceedings IEEE ICASSP-94*, Adelaide, Australia, April 1994.
13. Fransen, J., Pye, D., Robinson, A.J., Woodland, P. and Young, S.J., "WSJCAM0 Corpus and Recording Description," CD-rom documentation for the LRE SQALE project, CUED/F-INFENG/TR.192 ,(1994).