SQALE: Speech Recognizer Quality Assessment for Linguistic Engineering

Herman J.M. Steeneken

Human Factors Research Institute TNO
Soesterberg
The Netherlands

Lori Lamel

LIMSI-CNRS
Orsay
France

PROJECT GOAL

The objective of the 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.

INTRODUCTION

The LRE SQALE project started at the end of December 1993, and will have a duration of 18 months, in order to get a quick evaluation of the feasibility of the installation of such an assessment infrastructure in Europe. In order to efficiently define and carry out experiments, the SQALE consortium is made up of only four partners. However, it is hoped that SQALE will pave the way for future projects having a larger scope and a 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 performed by the remaining three members of the Consortium. These three laboratories, from three different countries with their own language are: 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 Speech and Natural Language programme. The ARPA programme, 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 as reported in these proceedings. 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.

PROJECT DESCRIPTION

As in the ARPA program, the assessment paradigm 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.

Two independent research questions addressed by this project are:

1. what are the merits of different recognition algorithms applied to the same data; and
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. A baseline condition for comparison of systems using the same acoustic training data, the same vocabulary and the same language model will be defined for each language. The task will be large vocabulary (20,000 words), speaker-independent, continuous speech recognition.

Each site will evaluate their system for 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 following decisions have been made regarding training corpora. These corpora are all either already or expected to soon be publicly available.

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

French: For French, the corpus BREF80, 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 40M words of newspaper text from *Le Monde*.

German: For the German acoustic training data the consortium will use a portion of the PHONDAT corpus. The language model training data will come from about 40M words of the newspaper *Frankfurter Rundschau*.

The BREF and WSJCAM0 corpora both contain development data that will be used within SQALE. Development data for German and evaluation test data for all three languages will be recorded by TNO, under the same conditions used for recording the training data.

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.

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, CSR research systems. Having 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. The LRE Expert Advisory Group on Language Engineering Standards (EAGLES) Spoken Language Working Group is working to coordinate and define standards for data, corpora, and assessment methodologies. The standards defined in EAGLES 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 importance of assessment activities is the subject of great 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.

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 will be a first attempt to adapt the ARPA “Assessment paradigm” to the multilingual European context, and will serve as a baseline from which more advanced research tools and metrics can be developed. The project will also reinforce relationships among European laboratories by providing a common, collaborative framework for testing systems and sharing data. It will be a first step towards a European project dedicated to assessment of state-of-the-art speech research systems, which can be considered a formally coordinated "technology-driven" approach.

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

Dissemination of the project results will be through public presentations at related conferences and workshops, as well as through written publications. 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