The ELRA Newsletter

March 1997

Vol.2 n.1

Table of contents

ELRA Catalogue, March 1997, Release 1.5 & Questionnaire about your needs enclosed

Signed articles represent the views of their authors and do not necessarily reflect the position of the Editors, or the official policy of the ELRA Board/ELDA staff.

Editor in Chief:
Khalid Choukri
Editor:
Deborah Fry
Layout:
Martine Garnier-Rizet

Contributors:
Marc Blasband
Jean-François Bonastre
Dafydd Gibbon
Paul Heisterkamp
Louis C.W. Pols
Herman J.M. Steeneken
Herbert S. Tropf
Steve Young

ISSN: 1026-8200

ELRA/ELDA
CEO: Khalid Choukri
Assistant: Rébecca Jaffrain

87, avenue d'Italie
75013 Paris
Phone: (33) 1 45 86 53 00
Fax: (33) 1 45 86 44 88
E-mail:
elra@calvanet.calvacom.fr
WWW:
http://www.icp.grenet.fr/ELRA/home.html

Dear ELRA Members,

This third and last issue of the ELRA Newsletter devoted to the different Colleges focuses on the dynamic area of Speech. In addition to providing overviews of a number of ongoing European projects and activities such as EAGLES, SpeechDat and ARISE, it looks at the state of the art in speech recognition, speaker identification and verification, and evaluation paradigms. It also contains a brief description of ESCA's activities. The next few issues of the Newsletter will concentrate on the language engineering work currently being performed in individual countries throughout Europe. As always, we would be grateful for any articles, input and suggestions which our members and other readers would like to provide.

Even more important, however, is your input on the market requirements survey enclosed with this issue of the Newsletter. This is your chance to influence our activities and priorities when it comes to marketing current and future resources. The survey, which is based on a plan drawn up over the past few months, is especially important given our decision to recruit a professional Sales and Marketing Manager (see the job advertisement on p. 15). Another "first" in this Newsletter is the appearance of advertisements from other organisations. In future, we shall be offering this service on demand, either on a reciprocal basis (in the case of other association newsletters and industry publications), or against payment. For further details, please contact the ELDA office.

ELRA/ELDA has also been liaising with a number of organisations (including COCOSDA, ELSNET, and several other associations and enterprises), with the aim of helping to co-ordinate international initiatives. Equally, negotiations with additional service providers have continued, and a number of new resources have been acquired (e.g. phonetically transcribed dictionaries and the renowned Metal lexica from GMS). For further details, see the enclosed catalogue and the last page of this issue. The full updated resource list (including the price list) is available on the Web and from the ELDA office. A large mailshot in several languages has been sent to existing and potential resource holders in written and terminology fields, in order to obtain both new resources and new members.

In order to encourage PhD students to carry out work on language resources, ELRA has also decided to institute special prizes. These will be granted in co-operation with the major LE conferences at which they are to be awarded, and will consist in part of a subsidy for attendance.

Our validation subcontracts (lexicon and corpora) are now at an advanced stage of completion, and we hope to disseminate the draft validation manuals before summer.

Last but by no means least, we are happy to inform you that the European Commission has decided to extend our contract under the LE Programme for a third year (October 1997-September 1998).

With best wishes,

Antonio Zampolli, President

Khalid Choukri, CEO

ELRA Board Profiles

Lou Boves

After obtaining his MA in Phonetics and Signal Processing (cum laude) from Nijmegen University in 1973, Louis Boves joined the staff of the then newly formed Phonetics Laboratory of Nijmegen University, where he did research on and taught acoustic phonetics and the physiology of speech production, with emphasis on the acoustics and physiology of the voice source. In 1984, he obtained his PhD from Nijmegen University (cum laude) for a dissertation on "The phonetic basis of perceptual ratings of running speech". In 1992 he was appointed KPN (Royal Dutch PTT) professor in Speech Technology and its Applications.

From the early 1980s onwards, the focus of his research and teaching started shifting away from basic speech science and towards speech and language technology. From 1985 to 1988, he was the Dutch co-ordinator of the ESPRIT project "Linguistic Analysis of European Languages", while from 1989 to 1992 he was responsible for the workpackage on text to speech in the ESPRIT project, POLYGLOT. He also was one of the members of the management board of POLYGLOT. Since March 1988, Louis Boves has been a consultant with the Speech and Language research group in what was then known as PTT Research, and is now as KPN Research, the R&D company of Royal Dutch PTT. In this capacity he was involved in the ESPRIT project SAM. He has been a member of the Board of the Dutch Speech Processing Expertise Centre, SPEX, since its foundation in 1989, and has been chairman of SPEX's board from 1994. SPEX's mission is to create spoken language resources and to make them available to the academic and industrial R&D community in the Netherlands.

Louis Boves has held positions on the boards of several scientific organisations in the Netherlands, and is currently Scientific Director of a large research programme on Language and Speech Technology which is funded by the Dutch Research Council, NWO.

Christian Galinski

Born in 1944 in Germany, Christian Galinski read oriental studies and communication studies at the University of Bonn in 1967-1971, before spending two years in Japan for further studies and research on the history of education (1971-1973). In 1975 he settled as a scientific and technical translator in Vienna, Austria, where he registered as a court translator for the Japanese language.

After founding and managing a private language and consultancy enterprise specialising in supporting business with Far Eastern countries from 1977-1979, Christian Galinski joined the International Information Centre for Terminology (Infoterm) in 1979, where he was responsible among other things for terminology standardisation, terminology planning (especially in developing countries), computer-assisted terminology work and terminography, and knowledge transfer in general. He also had special responsibility for reorganising Infoterm, and for establishing a publication and public relations programme.

In 1980, Christian Galinski started to implement the International Network for Terminology (TermNet) under the terms of a contract with UNESCO, and became its first Executive Secretary when it was founded as an international association at the end of 1988. He also served as Vice-President of TermNet from 1992-1993, and was elected President in 1996.

At the beginning of 1986, he succeeded Prof. Felber as Director of Infoterm and as Secretary of Technical Committee ISO/TC 37 "Terminology (principles and co-ordination)" of the International Organization for Standardization (ISO). Under his management, Infoterm was associated with the Department of Public Information of the United Nations, as well as becoming the International Collaborating Center for Terminology of the World Health Organization, the International Thesaurus Information Center of the International Federation of Information and Documentation (FID), the ISO Information Center for Terminology Standardization within the framework of ISONET, and a consultant to many IIGOs and INGOs. Since 1986, he has assisted in the foundation of several institutions and organisations in the field of terminology and is one of the driving forces behind activities relating to the multilingual information society, both in Europe and at international level.

In the second half of 1996, he was appointed Director of Infoterm, which was founded in August 1996 as an independent international association.

Large-Vocabulary Speech Recognition

Steve Young

Considerable progress has been made in speech recognition technology over the last few years and nowhere has this progress been more evident than in the area of large-vocabulary recognition (LVR). Current laboratory systems are capable of transcribing continuous speech from any speaker with average word error rates of between 5% and 10%. If speaker adaptation is allowed, then after 2 or 3 minutes of speech the error rate will drop well below 5% for most speakers.

This article will briefly review the architecture of a modern LVR system, describe the state of the art, and conclude with a discussion of current research issues.

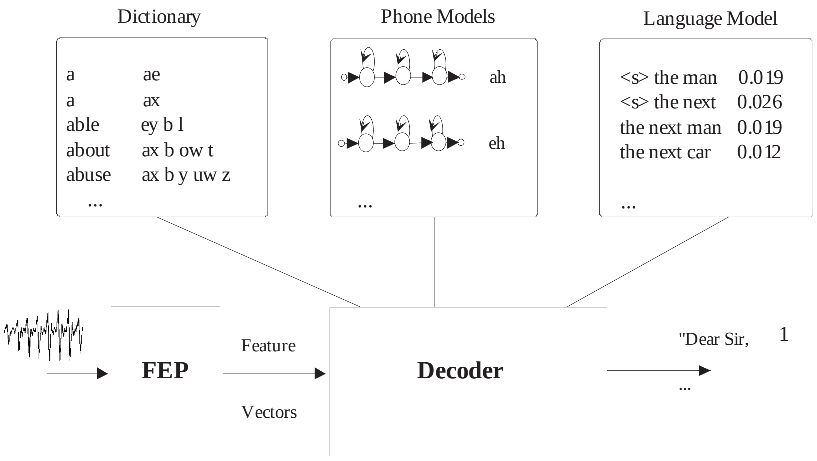

Figure 1 : **Architecture of a Large Vocabulary Speech Recognition System**

Architecture

Figure 1 illustrates the main components of a LVR system. The unknown speech wave form is converted by a front-end signal processor (FEP) into a sequence of acoustic feature vectors. Each of these vectors is a compact representation of the short-time speech spectrum covering a period of typically 10 msecs. In most systems either MFCC or PLP features are used.

Recognition is based on the principles of statistical pattern recognition and involves finding the sequence of words which is most likely to have given rise to the observed sequence of feature vectors. The probability of any given sentence is determined by decomposing it into a sequence of words and then decomposing each word into a sequence of basic sounds (phones) using a dictionary. The probability of an observed vector sequence given a specific phone is determined using a hidden Markov model, the probability of any sequence of phones given a specific word is determined from a pronouncing dictionary and, finally, the probability of any sequence of words is given by a statistical language model. Finding the most likely sequence of words for a given acoustic input is a search problem and its solution is the job of the decoder.

The language model

The probability of a given word sequence is computed using an N-gram model where N is typically 2, 3 or 4.

The probability of some word w_k in an utterance given the preceding words $W_1^{k-1} = w_1 \ldots w_{k-1}$ is approximated by $P(w_k|W_1^{k-1}) = P(w_k|W_{k-n+1}^{k-1})$ (see the language model box in Figure 1). N-grams simultaneously encode syntax, semantics and pragmatics and they concentrate on local dependencies. This makes them very effective for languages like English, in which word order is important and the strongest contextual effects tend to come from near neighbours. Furthermore, N-gram probability distributions can be computed directly from text data and hence there is no requirement to have explicit linguistic rules such as a formal grammar of the language. The major difficulties encountered when using N-grams arise from the extreme data sparsity inherent in any practical training corpus. The robust estimation of trigram probabilities thus requires considerable care. However, the problems are soluble and good performance can be obtained. N-grams do have obvious deficiencies resulting from their inability to exploit long-range constraints such as subject-verb agreement. Nevertheless, no significantly better models have been found to date.

The acoustic models

The purpose of the acoustic models is to provide a method of calculating the likelihood of any vector sequence given a specific phone. The most common approach to acoustic modelling is to use hidden Markov models (HMMs).

A HMM phone model has a number of states (typically 3) connected by arcs and a simple left-right topology (see the phone model box in Figure 1). A HMM is a finite state machine which changes state once every time unit; each time t that a state j is entered, an acoustic speech vector y_t is generated with output probability density $b_j(y_t)$. Furthermore, the transition from state i to state j is also probabilistic and is governed by the discrete probability a_{ij}. The joint probability $P(Y,X|M)$ of a vector sequence Y and state sequence X given some model M is calculated simply as the product of the transition probabilities and the output probabilities. In practice, of course, only the observation sequence Y is known and the underlying state sequence X is hidden. This is why this model is called a hidden Markov model. However, the required probability $P(Y|M)$ is easily found by summing over all possible state sequences.

The choice of output probability function in a HMM is crucial since it must model all of the intrinsic spectral variability in real speech, both within and across speakers. Early HMM systems used discrete output probability functions in conjunction with a vector quantiser. Modern systems, however, use more accurate parametric continuous density output distributions which model the acoustic vectors directly, the commonest choice of distribution being the multivariate mixture Gaussian.

HMM phone models are trained on examples of real speech. As with language models, the main problems arise from data sparsity. In real speech, contextual effects cause large variations in the way that different sounds are produced. Hence, to achieve good phonetic discrimination, different HMMs have to be trained for each different context. The simplest and most common approach is to use triphones, in which every phone has a distinct HMM model for every unique pair of left and right and right neighbours. This results in a very large number of models to train. In modern systems, data sparsity is tackled by tying components which are sufficiently similar to be able to share parameters without loss of discrimination. One of the most effective approaches is to use phonetic decision trees to tie states together.

Decoding

Decoder design is an area in which there are significant differences between systems. There are two main approaches: depth-first and breadth-first. In depth-first designs, the most promising hypothesis is pursued until the end of the speech is reached. Typically this is done using a stack-decoder. In breadth-first designs, all hypotheses are pursued in parallel. Breadth-first decoding exploits Bellman's optimality principle and is often referred to as Viterbi decoding. LVR systems are complex, and pruning of the search space is essential. In time-synchronous systems, this typically uses a process called beam search.

Research in decoder design over the last few years has sought to incorporate more complex models (e.g. cross-word triphones and long-span language models) whilst maintaining efficiency. Most schemes use multiple passes over the data and use lattices of word hypotheses to communicate information between passes. This approach allows the use of very complex models to be deferred to the later passes where the search space is much reduced. Also, a recent trend has been to incorporate adaptation into the recognition process so that the hypotheses generated in early passes are used to adapt the HMMs to give better accuracy in the later passes.

State of the art

The major benchmarks for assessing the performance of LVR systems are the US Advanced Research Project Agency (ARPA) CSR Evaluations. These began in their current form in 1989, when the task was the 1,000-word Resource Management task. In 1992, the focus switched to large-vocabulary recognition based on the Wall Street Journal (WSJ) corpus. Initially the evaluation material was from a single text source and it was filtered to lie within a specific vocabulary. Later, this filtering was removed so that by the time of the 1994 evaluation, the test material was based on a range of North American Business (NAB) news sources and the vocabulary was unlimited. In 1995, the NAB-based dictation task was extended to include speech recorded with a variety of far-field microphones in a noisy machine room. At the same time, a new focus on transcribing broadcast news (BN) started. The BN task is now the primary focus of the LVR research community. It encompasses many new challenges including handling a range of acoustic and environmental conditions (including background music and interfering speech), rapidly changing speakers and topics, non-native speakers and a mixture of both planned and spontaneous speech.

Year	LM	Training	Recognition Task	Best System	%WER
1992	5k-2g	SI84	WSJ Clean Dictation	SRI	16.6
1993	5k-2g	SI84	WSJ Clean Dictation	CU-HTK	8.7
1993	5k-3g	SI284	WSJ Clean Dictation	CU-HTK	4.9
1993	20k-3g	SI284	WSJ Clean Dictation	LIMSI	11.8
1994	20k-3g	SI284	NAB Clean Dictation	CU-HTK	10.5
1994	65k-4g	SI284	NAB Clean Dictation	CU-HTK	7.2
1995	65k-4g	SI284	NAB Clean Dictation	CU-HTK	6.6
1995	65k-4g	SI284	NAB "Noisy" Dictation	CU-HTK	13.5
1995	64k-3g	SI284+MP	MarketPlace	IBM	27.0
1996	65k-3g	BN	Broadcast News	LIMSI	27.1

Table 1: **A Summary of ARPA CSR Evaluations since 1992.**

The LM column indicates vocabulary size and N-gram order. The training column shows the acoustic training data used: SI84 = 14Hr WSJ0 database; SI284 = 66Hr WSJ0+WSJ1 databases; MP = 4Hr Marketplace broadcasts; BN = 50Hr News broadcasts. SRI = SRI International (Menlo Park, CA), LIMSI = LIMSI-CNRS group (Paris), CU-HTK = CUED HTK Group (Cambridge), and IBM = TJ Watson Research Group (Yorktown Heights, NY). Other sites which have regularly taken part include BBN (Boston), BU (Boston), CMU (Pittsburgh), CUED Connectionist group (Cambridge), Dragon (Boston) and Philips (Aachen).

Each of the tasks and evaluations above had very specific constraints, and these changed from year to year. Thus, a simple comparison of performance from year to year is difficult to make. With this proviso in mind, Table 1 summarises the tasks and the error rates of the best system over the last few years. All tests require speaker-independent continuous speech recognition.

As can be seen, performance on clean speech dictation improved steadily over the years from initially 20% word errors down to around 7%. However, performance on the more recent “real world” tasks is markedly worse, and this represents the main challenge over the next few years.

Current Research Activity

Current research activity in the field of large vocabulary speech recognition is focused on increasing robustness to variations in speaker, speaking style, background noise and varying channel conditions. Most current approaches adopt a generic strategy to dealing with these problems such as the use of MLLR. However, substantial progress will probably need a more detailed approach. Adaptation is also becoming important in the area of language modelling, where the need to track rapidly changing topics requires something more than the traditional static trigram language model. Finally, speech recogniser performance degrades rapidly with spontaneous speech. The reasons for this are still unclear but probably stem from a number of factors including poor articulation, increased co-articulation, highly variable speaking rate, and various types of disfluency such as hesitations, false starts and corrections. Good solutions to these problems may require substantial extensions to existing recognition architectures.

Conclusions

This article has reviewed the architecture of a speaker-independent continuous-speech large-vocabulary recognition system, briefly described the state of the art and reviewed current research issues. Whilst it is clear that much more needs to be done before robust, general-purpose LVR is ubiquitous, the technology is nevertheless on the threshold of usefulness for many practical applications.

For a more detailed review and a full bibliography see Young, SJ. *Large Vocabulary Continuous Speech Recognition: a Review.* Proc. IEEE Workshop on ASR, Snowbird, Utah, Dec. 1995 (also in IEEE Signal Processing, Vol. 13, No 5, pp 45-57,1996)

Steve Young
Cambridge University Engineering Department
Trumpington Street
Cambridge, CB2 1PZ
Email: sjy@eng.cam.ac.uk

Development and Assessment of Large-vocabulary Speech Recognition Systems and the Relevance of High-quality Databases

Herman J.M. Steeneken

Improving the performance of state-of-the-art large-vocabulary speech recognition systems for application-oriented conditions and the porting from a primary language (very often American English) to other languages are the major topics in present-day research in this field. Comparing systems and obtaining benchmarks are milestones in the process of quantifying the effect which improvements or different algorithms have on these systems.

The yearly DARPA assessment is an example of an effort in which different systems developed at different sites are compared under similar test conditions. In principle, the evaluation paradigm is based on shared training and test data. The test data selection and scoring is performed by an independent body. The general procedure is for all partners in the test to receive the data for development prior to the test period in order to obtain a system focused on the specific test conditions. At a certain point, all partners are given test data in order to perform a dry run. This allows for a final check of the system with representative test data and also some tuning, if required. For the final test, all partners receive the test data at the same time and are asked to deliver the recognition results within a certain period (by e-mail). The scoring is then performed by the independent body (which is NIST for the DARPA tests). The results are discussed in a workshop in which all partners and the co-ordinator participate. In general, these tests are performed for American English speech recordings (e.g. read speech from The Wall Street Journal, a North American business paper and recently also from radio broadcast recordings). In some contrasting conditions, speech of different quality is used, such as that obtained from several different microphones, telephone-quality speech, or spontaneous speech.

A similar experiment was performed in Europe, but focused instead on the multilingual European situation . In the SQALE project, which was sponsored by the European Union, three laboratories (Cambridge University, England; LIMSI, France; and Philips, Germany) made available their systems for implementation of a second or third language. In a similar way to that used with the DARPA paradigm, co-ordination and scoring were performed by an independent partner (TNO-HFRI, The Netherlands).

Specific training material is required in order to implement a new language. In general, a language-specific phone set, training material for a language model (often based on newspaper texts of 40 million words and a pronunciation dictionary) and speech-training material read by more than one hundred speakers (male/female) are necessary for the development of a large vocabulary recogniser. It is obvious that the production of the speech database in particular is costly, and that accurate specification (of the recording conditions, speaker selection and text selection) is important and specific to particular domains. Because it is widely available, the Wall Street Journal database (66 hours of speech, 284 speakers) is a starting-point for many studies. In the multilingual European situation, no full coverage for all languages is available. Four languages were involved in the SQALE project: American English (as a benchmark), British English, French and German. In the case of British English and French, two robust databases were available (WSJ-Cam and BREF); for German, a somewhat different type of database could be used (Phondat). Easy access to these databases (ELRA, LDC) is extremely important for enabling sharing.

Apart from the data specific to training, "untouched" speech material is required for the development test and final test. In general, 20 speakers (10 males and 10 females) and a minimum of ten sentences per speaker are used. If the recording conditions for this speech are significantly different from the conditions under which the training material was collected, a low(er) level of performance can be expected. Also, no indication can be obtained as to which uncontrolled variable may be responsible for the deterioration. In the SQALE project, some of the test material was specifically recorded as no untouched material was available. Some variables included in the test material were carefully controlled; for example, the "out of vocabulary" word rate (OOV) and the perplexity were carefully balanced across speakers, sentences and languages in order to obtain a fair comparison. Perplexity was distributed in three levels (high, medium and low), independently of sentence length and speaker. The OOV rate was 1.5%. However, the OOV words were not distributed randomly, but rather corresponded to the perplexity level. This allows for an analysis of correlations between these two parameters.

Efforts are currently being made to make the SQALE test material available to ELRA, thereby allowing other groups to assess their systems and to compare their performance with the benchmark set by the SQALE consortium.

Rather than just performing a benchmark test, databases can be constructed in such a way that diagnostic information on the type of errors made by the recogniser can be obtained. For example, a method for the assessment of word recognisers has been developed, allowing confusion at phone level to be quantified in a manner similar to tests for the measurement of subjective intelligibility (Steeneken and van Velden, 1989). This method was studied by the former SAM project, and the required database was recorded for many languages and made available by Eurom1.

In the SQALE project, all partners agreed to include a replica of text utterances by the same speaker in order to obtain data on inter- and intra-speaker variability (Steeneken and van Leeuwen, 1995; Van Leeuwen, 1997). Last but not least, a clear benchmark might be human performance. Again, in the SQALE project human benchmark tests were added to the original project set-up (van Leeuwen et al., 1995). For example, the average recogniser performance during the test exhibited a 12.6% (sd. 1.9) error rate. The human performance on the same data showed an error rate of 2.6 % (sd. 0.9), while non-native listeners produced an error rate of 7.4 % (sd. 1.7). There is thus some work to be done to match human performance.

Development and international comparison of assessment methods were carried out in the past by the SAM consortium, and the procedures described by SAM were adopted by the EAGLES project. ISO has now established a new working group (ISO/TC159/SC5/WG3) that will also look into the assessment of speech tools in terms of the human-machine relationship.

I hope that it is clear from this short résumé that robust assessment and high quality databases are essential to progress in the development of speech technology and language technology systems. Some standardisation of assessment methods in relation to human factors, as well as robust database construction are necessary if progress is to be made in comparable system specifications.

The efforts made by ELRA can only be successful if we all support them, while keeping in mind that commercial interests may sometimes interfere. Nevertheless, distribution to non-commercial members of ELRA and LDC should not be restricted.

Dr. Herman J.M. Steeneken

TNO Human Factors Research Institute
Kampweg 5
P.O.Box 23
3769 ZG Soesterberg
The Netherlands

Tel.: +31 346 356 269
Fax: +31 346 356 269
Email: steeneken@tm.tno.nl

For information on the SQALE database please contact:

Dr. David van Leeuwen.
Email: van.leeuwen@tm.tno.nl

Telephone Enquiries Using Speech Recognition

Marc Blasband

Travel schedule information is essential for users of public transport because they must adjust their behaviour to the latter's fixed schedules. Different media and technologies are currently being used or are under development in order to reach millions of travellers as efficiently as possible. Telephone enquiry plays a crucial role in this process, as the medium is well known by the public and its infrastructure is well established.

The public transport companies have installed a number of telephone enquiry systems that are either operated by experienced human operators or use push-button technologies. Every year, more than 200 million calls are made to railway enquiry centres throughout Europe. Operator costs limit the number of calls that can be handled and make all attempts to provide more advanced and complex services impossible.

The ARISE project aims to make the automatic handling of a significant volume of telephone calls possible by using automatic speech recognition for train schedules. The project also identifies options for handling a wider range of enquiries in the future.

ARISE starts with existing prototypes, the overall quality of which is impressive but not yet sufficient for commercial usage. ARISE uses at least one prototype per language environment (one in Dutch, one in Italian and two in French), and will improve on them. The prototypes automatically provide train schedule information by telephone for the major train stations in France and Italy, and for all stations in the Netherlands. In order to access the database, the user must specify the departure and arrival stations, and the date and time of travel. The user is free to specify either the departure or arrival time. The system recognises information provided as naturally spoken utterances.

The systems built within the ARISE project are primarily targeted at consumers, i.e. incidental users, and not at regular users such as travel agents. Therefore, the systems must be able to cope with the many ways people speak and the various questions they have. They must accept continuous speech as opposed to isolated words, and must also recognise when a traveller wishes to have additional information about a train connection, e.g. earlier or later departure times. For the caller, the process must be as natural as possible. In other words, these systems must elicit enough caller satisfaction for the traveller to decide to use the systems again, thus satisfying the financial and commercial goals of the project.

The different prototypes built throughout the duration of the project are used to measure caller satisfaction with the different systems, to decide on improvements and to test the effects of these improvements.

The project organises periodic validation to indicate the service improvements necessary in order to develop an automated system that will ultimately:

- involve minimal training for the caller;
- elicit a positive level of caller satisfaction;
- achieve a global quality appreciation that is comparable to the human-operated systems.

To reach these goals, ARISE researches topics in a number of different areas:

- usability topics and system issues will improve the position of the technology within the services provided to the caller;
- callers' perception of the quality of the speech recognition systems can be increased by improving the dialogue between them and the automatic systems;
- cheaper, more efficient and more robust language recognition will clearly contribute to this higher appreciation by reducing the number of cases in which dialogue is necessary to recover from recognition errors.

ARISE is supported by the European Union, and is a follow-up to the MAIS and RAILTEL projects, which were also supported by the European Union. Three railway companies (FS, NS-OVR and SNCF), three technology providers (CSELT, LIMSI and Philips), three universities (IRIT, KUN and RWTH) and four system integrators (LTV, KPN, Saritel and VECSYS) collaborate in ARISE. The Language Engineering sector of the European Union not only subsidises the project but also facilitates co-operation among the parties and participates actively in determining strategy.

More information may be obtained from:
Marc Blasband
Nederlandse Spoorwegen
Moreelsepark 1
Postbus 2025
3500 HA Utrecht
The Netherlands
Tel.: +30-235 57 45
Fax: +30 235 63 27

The EAGLES Spoken Language Working Group and the *Handbook of Standards and Resources for Spoken Language Systems*

Dafydd Gibbon

This brief overview is intended to provide information about the goals, authorship, content and dissemination procedure for the EAGLES Spoken Language Working Group's (SLWG) Handbook, updating and expanding the information on the SLWG's activities given in the March 1996 issue of this Newsletter.

The EAGLES Spoken Language Working Group

The EAGLES Spoken Language Group was established in 1993 within the EAGLES project (which was sponsored by the CEU's DG XIII), under the chairmanship of Roger Moore (DRA, UK), and hosted by Richard Winski (Vocalis Ltd., UK). Dafydd Gibbon (University of Bielefeld, Germany) was appointed technical editor. In consultation with a representative Core Group which includes Giuseppe Castagneri, Jean-Marc Dolmazon, Norman Fraser, John McNaught, Louis Pols and Hans Tillmann, a significant policy decision was made at an early stage to aim to produce a fully-fledged high quality *Handbook* of Spoken Language Standards and Resources over and above the basic technical report originally envisaged. At a later stage in the project, the additional goal of producing a hypertext version of the *Handbook* was adopted.

The rationale behind this decision was that, despite the additional non-financed effort involved on the part of authors and editors, a carefully produced and professionally disseminated handbook would be of more value to the spoken language community in the medium and long term than a more informal document, and that a flexible dissemination policy simultaneously using paper and hypertext electronic media would provide maximum added value to the individual reader, both at the laboratory workbench and in other working environments.

Goals of the *Handbook*

The main goal of the *Handbook* is to collect and catalogue information on spoken language resources and de facto standard procedures, and so provide an essential reference work for speech technology development. Speech technology has emerged during the past year as a market factor, primarily with dictation software and speech synthesis devices, and one main group within the potential readership will include research workers and system developers in this field, including service companies who adapt existing systems to new domains and scenarios. Other potential readers include corporate end-users who need to specify, procure, or integrate system components, and who require guidance on system specification and assessment, as well as newcomers to the field, including graduate students and workers in other countries who require access to well-documented common practice in Europe.

Authorship and content of the *Handbook*

For the specific areas to be covered, authors were recruited from major European development laboratories, both in industry and in academia. The main technical authors are Frédéric Bimbot, Lou Boves, Gérard Chollet, Khalid Choukri, Els den Os, Christoph Draxler, Norman Fraser, Dafydd Gibbon, Peter Howell, Lars Knohl, Volker Kraft, Hermann Ney, Renee van Bezooijen and David van Leeuwen. Preliminary versions of the *Handbook* have been widely distributed within the world-wide spoken language community, and feedback from experts in the subfields covered by the *Handbook* has been incorporated in a number of revision cycles.

This procedure, which was time-consuming but well worth while, was considered necessary in order to emphasise the consensus-oriented character of the *Handbook*. As a result, each chapter is not the personal view of its main technical author, but has been thoroughly discussed in the community.

The *Handbook* covers four main areas: spoken language system and corpus design, spoken language characterisation, spoken language system assessment, and spoken language reference materials, as well as providing a user guide, glossary, comprehensive bibliography and index. The section on system and corpus design includes chapters on system design, corpus design, corpus collection and corpus representation. The second group of chapters covers spoken language lexica, language models as part of the characterisation of corpora and as system components, and the physical characterisation and description of corpora. The assessment-oriented chapters cover basic experimental methodology, and the assessment of speaker verification, speech recognition, speech synthesis and interactive dialogue systems. Each chapter refers not only to current methods, but provides explicit recommendations on good practice and, where relevant, information on available tools. Finally, a collection of reference materials is included, mainly provided by representatives of other projects within the Language Engineering and ESPRIT domains, but partly specially written; these materials cover specifications of standard formats and practice, from the IPA and SAMPA standard transcription and labelling practice to corpus recording and archiving standards.

Dissemination of the *Handbook*

A number of academic publishers were unwilling to take the risk of publishing both a paper version of the *Handbook* and an electronic version on the Web (including one publisher who had previously done just this, but had apparently not been happy with the result). However, the Mouton de Gruyter department of Walter de Gruyter Publishers, Berlin, agreed to offer the following innovative dissemination structure:

(1) Library Handbook edition (including hypertext version on CD-ROM).

(2) Four paperback parts:

I. Spoken language system and corpus design,

II. Spoken language characterisation,

III. Spoken language system assessment,

IV. Spoken language reference materials.

(3) Free Web access to hypertext version with user registration.

In the interests of reducing the price, a no-royalties policy was agreed. In addition, negotiations with ELRA have been initiated with the aim of supplying the library edition to ELRA members at a significant discount.

Although the careful consensus-building procedure and the professional publishing requirement have meant a longer production process than was originally anticipated, the Spoken Language Working Group is confident that the results will justify both the effort and time spent. Editing of the *Handbook* has been completed and it is now in production.

Further information is available on the EAGLES Spoken Language Working Group's Web page,

http://coral.lili.uni-bielefeld.de/~gibbon/EAGLES

and on the central Web page of the EAGLES project,

http://www.ilc.pi.cnr.it/EAGLES96/home.html

Queries should be directed to Dafydd Gibbon at:
Universität Bielefeld
Fakultät für Linguistik und
Literaturwissenschaft
P100 131
Roau C6-138
33501 Bielefeld Germany
Tel.: +49 521 106 35 10
Fax: +49 521 106 60 08
Email: gibbon@spectrum.uni-bielefeld.de

SpeechDat: European Speech Databases for Creation of Voice-driven Teleservices

Herbert S. Tropf

At present, the market for automatic voice-driven teleservices, such as telephone banking, access to public administration information, travel information, voice mail services and directory assistance, is increasing extremely rapidly. Advances in security procedures will shortly open up a vast new telebusiness market through secure support for financial transactions. This growth will be further extended by the rapid growth expected in the mobile telecommunications market. There are two main reasons for this situation. Firstly, computer technology and speech processing have advanced far enough to be economically practical for a range of applications. Secondly, consumers are making everincreasing demands on the services they receive, such as lower costs, customised services and twenty-four hour availability. Besides this, speech remains the most natural medium for communication.

Many European companies are active in the field of creating voice-driven teleservices and delivering the necessary speech technology. Language-specific spoken language resources, i.e. speech databases, lexicons and related tools, are the basis for implementing speech processing technology, i.e. speech recognition and speaker verification. The EU-funded project SpeechDat will lay the foundations for European companies to be competitive when starting with a multilingual environment.

SpeechDat started in March 1996 and has a planned duration of 24 months. The consortium consists of 20 European partners mainly active in the telecoms sector, and the co-ordinator is Siemens AG. The project is partially funded by the Language Engineering section of the Telematics Applications Programme. The EU-funded preparatory action SpeechDat(M) prepared the ground for the current SpeechDat project to produce large-scale high quality resources.

Objectives

The main objective of SpeechDat is to produce speech databases for the development of voice-driven teleservices realising a large coverage of languages and applications. These include the 11 official European languages with their dialectal regions, some major language variants and minority languages, Norwegian and the Eastern European language Slovenian.

The SpeechDat databases will provide a realistic basis for the training and assessment of both isolated and continuous-speech utterances, employing whole-word or sub-word approaches. The specifications of the databases for the fixed telephone network, the mobile telephone network and speaker verification have been developed jointly, and are essentially the same for each language, in order to facilitate dissemination and use. (These specifications and other relevant information can be found on the official SpeechDat Web site:

http://www.phonetik.uni-muenchen.de/SpeechDat.html)

Current and future spoken language resources

Many speech databases have been created in the past independently from SpeechDat. Most often these have been produced by a single organisation, for a specific application and in a proprietary format, and have not been made more widely available. More recently, the Linguistic Data Consortium (LDC) was set up in the USA; this also co-ordinates the production and distribution of speech resources.

Currently, many automated teleservices can only operate using single words which are spoken in isolation in response to a prompt. In the future, systems will become more reliable, functional and user-friendly. Continuous speech recognition (where no pause between words is required) will become prevalent The system will be able to extract important commands, words and phrases and to discard the rest, making it more robust. The user will be able to interrupt the system during an announcement where they typically have to wait for the system to finish at present. Users will not be tied to a rigid dialogue or system of menus, but will be able to speak in a natural, spontaneous manner and provide or ask for information in any order.

The resources being produced by SpeechDat are designed to meet the requirements of developing applications with these advanced characteristics.

Table 1 gives a complete overview of the speech databases which were either completed in SpeechDat(M) or are currently being created in SpeechDat, while Table 2 lists the owners and producers of SpeechDat and SpeechDat(M) databases.

Table 1: ***SpeechDat databases for telephone applications (status: April 1997)***

	Language (variant)	Type of speech DB	No. speakers / No. calls per speaker	Status
1	Danish	fixed telephone	1000/1	complete
2	Danish	fixed telephone	4000/1	in progress
3	Dutch	mobile telephone	250/4	in progress
4	Dutch (Flemish)	fixed telephone	1000/1	in progress
5	English (British)	fixed telephone	1000/1	complete
6	English (British)	fixed telephone	4000/1	in progress
7	English (British)	mobile telephone	1000/1	in progress
8	English (British)	speaker verification	120/20	in progress
9	Finnish	fixed telephone	4000/1	in progress
10	French	fixed telephone	1000/1	complete
11	French	fixed telephone	5000/1	in progress
12	French	speaker verification	120/20	in progress
13	French (Belgian)	fixed telephone	1000/1	in progress
14	French (Luxembourgish)	fixed telephone	500/1	in progress
15	French (Swiss)	fixed telephone	1000/1	complete
16	French (Swiss)	fixed telephone	2000/1	in progress
17	French (Swiss)	speaker verification	20/50	in progress
18	German	fixed telephone	1000/1	complete
19	German	fixed telephone	4000/1	in progress
20	German	mobile telephone	1000/1	in progress
21	German (Luxembourgish)	fixed telephone	500/1	in progress
22	German (Swiss)	fixed telephone	1000/1	in progress
23	Greek	fixed telephone	5000/1	in progress
24	Italian	fixed telephone	1000/1	complete
25	Italian	fixed telephone	3000/1	in progress
26	Italian	mobile telephone	250/4	in progress
27	Norwegian	fixed telephone	1000/1	in progress
28	Portuguese	fixed telephone	1000/1	complete
29	Portuguese	fixed telephone	4000/1	in progress
30	Slovenian	fixed telephone	1000/1	in progress
31	Spanish	fixed telephone	1000/1	complete
32	Spanish	fixed telephone	4000/1	in progress
33	Swedish	fixed telephone	5000/1	in progress
34	Swedish	mobile telephone	1000/1	in progress
35	Swedish (Finnish)	fixed telephone	1000/1	in progress
36	Welsh	fixed telephone	2000/1	in progress

Table 2: ***Owners and producers of SpeechDat and SpeechDat(M) databases***

1.	Aalborg University
2.	British Telecommunications plc
3.	Centro Studi e Laboratori Telecommunicazioni S.p.A.
4.	GEC Marconi Secure Systems Ltd.
5.	GEC-Marconi Material Technology Ltd.
6.	GPT Ltd.
7.	IDIAP
8.	INESC
9.	Knowledge S.A.
10.	Kungl Tekniska Hogskolan
11.	Lernout & Hauspie Speech Products N.V.
12.	MATRA Communication
13.	Philips GmbH
14.	Portugal Telecom S.A.
15.	Siemens AG
16.	SPEX
17.	Swiss Telecom PTT
18.	Tampere University. of Technology
19.	Tele Danmark
20.	Telenor AS
21.	Universitat Politecnica de Catalunya
22.	University of Maribor
23.	University of Munich
24.	University of Patras
25.	Vocalis Ltd.

Speaker recruitment, recording and transcription

In SpeechDat, the telephone callers are recruited via market companies or internally within institutions. Each speaker is provided with a prompt sheet. The prompts consist of questions to be answered. Approximately 45 utterances are elicited per speaker; this corresponds to roughly 10 minutes recording time. An instruction sheet accompanies the prompt sheet, and acoustic prompting and instructions guide the caller through the recording process.

Speakers to be recruited for a SpeechDat database have to fulfil several requirements related to speaker-specific characteristics (e.g. sex, age), regional/dialectal factors and environment specific characteristics, i.e. environment of location of call, type of handset and type of network.

All the recordings will be performed over the telephone network through activation of a telephone server. Since the telephone interfaces are different from country to country no common recording tool will be used by all partners. Nevertheless, all recordings will be conducted via ISDN to simulate centralised services equipment. Data will be recorded on hard disk prior to the creation of CDROMs. The speech data files description and representation is unified across the languages. The SAM format is used as the standard in SpeechDat. Annotation of the speech files is made at the orthographic level. In addition, each database has a lexicon containing the standard orthographic words used in the annotations and corresponding canonical phonetic transcription using SAMPA notation.

Validation

Any database produced in the SpeechDat project must meet a set of minimum quality requirements in order to be approved by the consortium. The validation is carried out by SPEX according to criteria based on those set by the predecessor project, SpeechDat(M). Put briefly, validation examines if the correct items were recorded; if they are provided in sufficient quantities; if the speech files are properly annotated; if the database is well documented; if the concomitant lexicon is complete and well formatted; and if the database in general adheres to the predefined format specifications with regard to the directory structure and file names.

Availability

The resources produced by SpeechDat will be the property of the consortium partners, but special arrangements have been made for a wider dissemination. Each owner of a SpeechDat database is obliged to offer it non-exclusively to ELRA on terms to be agreed between the owner and ELRA, and will make it available to third parties after the end of the project at fair and reasonable commercial fees.

Herbert S. Tropf
Siemens AG, Munich
Otto-Hahn-Ring 6
81739 Munich
Germany
Tel.: +49 89 636 44 195
Fax: +49 89 636 49 802
Email: Herbert.Tropf@zfe.siemens.de

Speaker and Language Characterisation

Jean-Francois Bonastre

In this article, Jean-Francois Bonastre discusses trends and applications in automatic speaker recognition (ASR) and outlines some organisations and institutions that are currently active in the development and analysis of the technology.

Introduction

The primary goal of spoken communication is the exchange of meaning between individuals. However, spoken messages also carry additional information, such as the geographic origin of the speakers, their psychological condition, information about some of their physical features, and other elements related to their identity. Of course, these elements are not as accurately encoded as the linguistic content of the message, but they represent a significant component in the process of spoken communication.

Speaker and language characterisation is dedicated to the study of these extra-linguistic factors which are present in the speech signal.

Automatic speaker recognition

Automatic speaker recognition (ASR) is a subtopic of speaker characterisation. It aims to retrieve automatically from a speech signal the identity of the speaker who produced it, or at least to assign the speaker to a class from a set of predefined categories.

For instance, speakers can be classified according to their sex, age group, voice pathology, emotional condition, etc. These tasks are typical classification problems, hence the term "speaker classification", which is more general than speaker recognition.

A few definitions

Classical tasks in ASR can be divided into two types: speaker identification and speaker verification. Identification tasks consist of finding the speaker among a set of known speakers whose voice has the closest resemblance to a given speech signal. Verification tasks aim to test the hypothesis that a particular speaker is the actual speaker who uttered a given speech signal, or to reject the speech token as belonging to an impostor.

Within the framework of speaker identification, it is essential to specify whether the speech signal under examination can have been produced by an impostor (open-set identification) or if it can be assumed that the actual speaker is part of the set of known speakers (closed-set identification). In practice, open-set identification is a much more realistic situation, but the task is more complex.

The performance of a system is also very dependent on the degree to which the speaker is willing to take part in the recognition process. For some applications, the speaker can be considered as co-operative, e.g. if he or she utters a particular sequence of words requested by the system. But in other contexts, and in particular in forensic ones, the speaker’s goal can be to objectively evade recognition.

Robustness to the various noise and channel distortions, which is usually dependent on the context of the application, is another factor which has a major impact on system performance and is currently one of the most challenging research topics around.

Objectives

The main motivation for ASR studies is common to most areas in the domain of speech processing: the knowledge and modelling of speaker-specific characteristics within a speech signal is an essential step in the separation of its linguistic and non-linguistic content. Each factor can then be treated separately (speech and speaker recognition), adapted to each other (speaker adaptation), or recombined differently (voice modifications in speech synthesis).

From the application point of view, the authentication of the user for use in financial transactions conducted via telecommunication networks is currently a major commercial concern. ASR is a very attractive solution to this problem, as it does not require any particular equipment on the user's side besides a simple telephone.

In the forensic domain, voice recordings are being used with increasing frequency as elements of proof in legal cases. In spite of the reservation, or even reluctance, of the scientific community, many "voice experts" throughout the world are engaged in analysing such recordings. In fact, the forensic field is very eager to develop scientific methods for such tasks, not with the unrealistic hope of validating "proofs" automatically, but with the goal of aiding the orientation of an investigation in its early stages. A number of more or less automatic methods are currently used in the forensic domain, but it is particularly difficult to evaluate these, given the level of secrecy and confidentiality surrounding them.

Main difficulties

Despite the number of potential applications of ASR, the development of industrial products based on these techniques has been hindered by the difficulties of the task.

The variability of the speech signal is a very difficult factor to handle, and is complicated further by the fact that intra-speaker variability can be of the same order of magnitude as inter-speaker variability. Moreover, voice characteristics evolve with time, owing to modifications to the speakers' physiology and psychological state. Additionally, the variability of the transmission channel generally induces a mismatch between the reference and the test material, which can bias the decision towards the recognition of the channel characteristics rather than the identification of the speaker specificity.

Last but not least, the assessment of ASR techniques is intrinsically difficult: large specific databases must be collected following a protocol that allows for proper representativity. However, it is extremely difficult to simulate impostor accesses in a realistic manner: for most evaluations, impostor trials are generated by using the speech of a given speaker against all registered speakers. This procedure does not take into account the "intention" or the motivation of the real impostors.

State-of-the-art

With contemporaneous studio recordings of several seconds with collaborative speakers, it is possible to achieve very small error rates in closed-set identification, even with populations of several hundred speakers. However, as soon as one of the favourable factors deteriorates, performance invariably drops. In particular, the bandwidth limitation, together with the non-linear channel distortion caused by telephone lines, is a very significant source of degradation. Additional adverse factors, such as the possible presence of surrounding noise and the temporal drift of the speaker's voice do not, of course, have a positive effect on the performance. However, speaker identification is not the most interesting task from the application point of view.

The Equal Error Rate (EER) is the conventional laboratory performance measure used to assess speaker verification systems. EER is computed when the system is tuned so as to equalise the proportion of rejected clients (false rejections or FR) and of accepted impostors (false acceptances or FA), as EER = FA = FR. The verification performance in general and the EER in particular do not depend on the size of the client population. In practice, with realistic telephone data recorded in several sessions in non-controlled environments, an EER in the order of magnitude of 1% can be achieved.

For forensic applications, the main problem concerning evaluation is the difficulty of defining an appropriate assessment methodology and the unavailability of a representative database for assessment purposes.

Language characterisation and recognition

Language characterisation and recognition is a second area related to the classification of speech signals according to extra-linguistic features.

According to linguists, more than 5,000 languages are currently spoken around the world. Human beings have a large variety of information at hand for distinguishing one language from the others.

Although scientists have long been interested in the problem of characterising and classifying languages, it is only recently, thanks to the progress of computers, that performing this task automatically has become feasible.

The first studies in automatic language identification (ALI) started in the 1970s. During the past two decades, the progress in speech processing has had a direct impact on those in ALI and contributed to the development of several such systems.

From a very fundamental point of view, knowing the specificity of languages and finding a common and contrastive representation of them can be viewed as a major step in understanding the mechanisms of language acquisition, which is one of the keys to the cognitive sciences.

More pragmatically, the globalisation of economic exchanges throughout the world creates a growing need for multilingual telecommunication services. In this context, an ALI system plays an essential role in channelling the user to the proper language-dependent service, without requiring from him any complicated process of language selection. The example of the 911 emergency service in America is often quoted as an excellent application of ALI, allowing any call to be routed to an operator speaking the language of the current user.

Most techniques proposed for ALI are based on an acoustic-phonetic decoding of the speech signal searching for the most likely sequence of phonetic or subword-like units, followed by a linguistic post-processing stage. Typically, 90% correct identification of one language among ten can be obtained in this way. However, these methods require considerable resources, in terms of computer, data and human time, during the training phase.

Current research aims to use additional sources of information, such as prosody or a priori phonological knowledge.

Current activities

In this section, we will list several working groups or projects, the activities of which are partly or totally dedicated to speaker or language characterisation.

GT1

The GT1, a working group of the GDR-PRC CHM (a research group on man-machine communication funded by the French institution, CNRS) dedicated to speaker and language characterisation, comprises several academic, industrial and forensic laboratories. The goal of this working group is to stimulate research in the field on several common topics of interest :

- Definition and design of assessment methodology and databases for speaker and language characterisation;
- Deontological reflections concerning forensic applications to ASR;
- Studies on dynamic speaker characteristics;
- Co-ordination of activities on the phonetic typology of languages;
- Definition of the concept of impostor in speaker verification and design of a corresponding database;
- Organisation of specialised seminars and workshops.

For more information, please contact:
J.F. Bonastre
Email: bonastre@univ-avignon.fr
http://lia.univ-avignon.fr/GT1

4.2 CAVE

The CAVE project (CAller VErification) is a LE-Telematics project, the goal of which is the development of two speaker verification demonstrators over the telephone network: one for telecommunication applications, the other for banking applications. The partners are: PTT-Telecom (NL), KUN (NL), KTH (S), ENST (F), UBILAB (CH), IDIAP (CH), VOCALIS (GB), TELIA (S), and the Swiss-PTT (CH). This two-year project started in December 1995.

COST-250 and PolyCOST

Within the context of the COST programme (European Co-operation in the field of Scientific and Technical Research), the European Union has set up a specific action (COST-250) dedicated to speaker recognition over the telephone network. Fifteen European countries regularly exchange the results of their work twice a year during meetings and workshops that are open to the international community (see http://www.fub.it/cost250/). In order to develop and assess speaker verification systems, the PolyCOST database was recorded and annotated. This database is distributed by ELRA.

M2VTS: multimodal speech

The ACTS programme (Advanced Communication Technologies and Services) contributes to the funding of the M2VTS project (Multimodal Verification for Teleservices and Security Applications), the goal of which is to carry out experiments on multimodal approaches for identity verification. The project gathers industrial partners, academic institutions and technology users from six European countries. They work on face recognition, synchronisation of speech signal and lip movement, and speaker verification. An audio-video database has been recorded and annotated as part of the project. This database is distributed by ELRA.

VERIVOX

The VERIVOX project is a European project designed to reduce the false rejection rate in speaker verification systems in order to improve user acceptability of such systems. The approach adopted is based on phonetics, phonology and articulatory models.

NSA-NIST

The NSA (National Security Agency) organises annual evaluation campaigns for speaker verification systems. For the 1997 campaign, two subsets of the Switchboard database are being used as development data. A third set of speech data comprising 500 speakers (50% male, 50% female) will be distributed later for the actual evaluation. Tests are carried out on several training conditions (single or multi-session, one or several transmission channels) and various test durations. Each system is assessed both at the operating point, corresponding to 10 % false rejections, and according to a cost function weighting the false acceptance and false rejection rates.

For more information, please contact:
Alvin F. Martin
Email: alvin@jaguar.ncsl.nist.gov

EAGLES

EAGLES is a European initiative designed to produce a handbook containing recommendations on assessment methodology, and covering most of the domains in automatic speech processing. Chapter 11 focuses on speaker recognition systems.

SpeechDat

The Telematics programme of the European Union funds a consortium of industrial and academic partners which is in charge of collecting and annotating telephone speech recordings in 22 European languages and dialects. These recordings follow the "Polyphone" protocol, as defined at the global level by COCOSDA (the COordination COmmittee on Speech Databases and Assessment). For each language, 5,000 people have participated in the recordings. These data are very relevant to the study of inter-speaker variability. Additional recordings are being made in order to characterise intra-speaker variability. In this context, between 20 and 120 people are asked to record from 20 to 50 Polyphone-like sessions. These combined data sets will be a very solid basis for speaker verification assessment, with the second set being used as client data, and the first as impostor data. The resources developed by SpeechDat will be distributed by ELRA (a number of them are already available).

Workshop on speaker recognition and its applications in the commercial and forensic fields

In order to discuss the progress made in the past few years and to evaluate the potential of the current techniques from a practical point of view, the GT1 is organising an international workshop in Avignon in April 1998. This workshop will also be an opportunity to examine the specific aspects of the forensic applications of speaker recognition, and the appropriateness of current approaches to the needs in this field.

Acknowledgements

This article was written with the participation of F. Bimbot, G. Chollet, G. Gravier and F. Pellegrino.

Speech Dialogue and Corpora

Paul Heisterkamp

Human-human dialogues are extremely complex and very flexible interactive activities that involve nearly every aspect of human speech understanding and production, as well as social interaction skills. No two dialogues that intuitively seem to achieve the same exchange between partners are identical on all the levels involved; they are multi-layered. The speakers' contributions to these dialogues convey information on all communication levels. It is therefore little wonder that there is as yet no generally agreed theory of dialogue as such, nor even agreement as to what the basic units of dialogue are.

In speech dialogue, the governing factor is uncertainty. Speakers interrupt each other, they do not say exactly what they mean, they use differing speaking styles, etc. All of these difficulties are overcome by humans through exchange, leading finally (in most cases) to the establishment of a mutual belief that understanding has taken place. These uncertainties are difficult for technical systems to handle.

Still, in the course of the last few years, several man-machine speech dialogue systems of varying degrees of naturalness have been implemented, and some of them have been successfully made available to the general public. This has been possible by reducing complexity by limiting the vocabulary, linguistic means, task domain, etc. Very simple dialogues use single-word recognition and guide the user strictly through a series of predefined dialogue steps. Systems that use connected-word recognition (such as the "Linguatronic" hands-free dialling system now available in Mercedes Benz cars) allow more flexibility on the part of users, in that they can handle freely chosen groupings of a set of words, although the latter is still rather limited. Dialogue systems designed to handle spontaneous, fluent speech and mixed user/system initiative require more complex dialogue management architectures in order to establish a human-like exchange. Current approaches include "form filling" systems, partly self-organising dialogue and dialogue management as a special case of rational behaviour.

As all of these approaches attempt – although in a limited domain – to achieve dialogue behaviour that does not constrain the user more than a human agent would, their common starting point is the observation of dialogues for the required task. These can be human-human dialogues, but, because the way humans interact with each other can differ significantly from how they interact with computers, so-called Wizard-of-Oz (Woz) settings are used to record those dialogues which will be handled by the system to be implemented. In these experiments, a human agent plays the role of the computer, talking to test persons who are led to believe they are interacting with a machine. The dialogues gathered in this way are then transcribed and annotated. Currently, this is done according to the standards and purpose which the implementer of the envisaged system has in mind, i.e. on a more or less isolated and specialised basis. This approach is time consuming and expensive.

The growth of the speech industry and the need to deliver more and more applications (and, therefore, dialogues) and to meet tighter and tighter deadlines means that this state of affairs is increasingly unsatisfactory. Furthermore, other uses of speech understanding systems are being realised which need to monitor human-human dialogues even though they normally have no influence or control over the latters' structure (one example is the German VERBMOBIL system for face-to-face speech translation). For this reason, efforts have recently been made to reach at least some common coding standard for dialogue transcription which would allow the use of task-independent corpora for modelling a variety of dialogue types.

In view of the "absence of an agreed theory and an agreed use" (Norman Fraser), this is not an easy task, especially as researchers from all over the world are concerned. In addition, there is no agreement as to the relevant units or their boundaries. Thus what one can hope for is some kind of analogy to the ToBi standard for prosodic annotation, which gives some basic information on the general outlines without specifying too much detail. One of the main objectives then would be to ensure that the coding scheme allows access to corpora in order to extract those dialogues that may serve as relevant examples for a given dialogue task. Hopefully, another possibility would be to have an independent frame of reference against which to evaluate the appropriateness of some dialogue systems' behaviour, an aim also pursued by the newly launched Esprit Fourth Framework Project, DISC – Spoken Language systems and Components – Best practice and evaluation (Co-ordinating partner Roskilde University, Centre for Cognitive Science, Niels Ole Bernsen (nob@cog.ruc.dk)).

Paul Heisterkamp
Daimler-Benz AG
Research Center ULM
Wilhelm Rurge Str. 11
D-89081 Ulm
Germany
Tel.: +49 731 505 21 42
Fax: +49 731 505 41 05

The European Speech Communication Association, ESCA

Louis C.W. Pols

The European Speech Communication Association ESCA was established in 1988 with a great deal of moral support from the European Union, although the association does not rely at all upon actual grants from the EU. The first president was Joseph Mariani. The association is a democratic and self-supporting interest group with paying members (fees since 1997: 45 ECU for full members and 15 ECU for student members). ESCA is most visible through its workshops and Eurospeech conferences, and through its publications. By last year, 17 ESCA Tutorial and Research Workshops (ETRWs) had been organised on a wide variety of topics and in many different European cities; there was also one workshop in the United States (on Speech Synthesis). In 1997, four ETRWs will take place:

- Robust speech recognition for unknown communication channels, Pont-à-Mousson, France, 17-18 April
- Larynx 97, Marseilles, France, 18-20 June
- Intonation: Theory, models and applications, Athens, Greece, 18-20 Sept.
- Audio-visual speech processing, Rhodes, Greece, 27-28 Sept.

The latter two are satellite events of the biennial Eurospeech conference in Rhodes, Greece, 22-25 Sept. Eurospeech '99 will take place in Budapest, Hungary.

More details about these events, as well as a wealth of other information, can be found on the ESCA web-site: (http://ophale.icp.grenet.fr/esca/esca.html) or can be obtained from the ESCA secretariat (esca@icp.grenet.fr). Together with the International Conferences on Spoken Language Processing ICSLP (held every even year), the Eurospeech conferences held every odd year can be considered as the major speech events in the world. Well over 1,000 abstracts for this year's Eurospeech '97 have been submitted for evaluation.

ESCA distributes a newsletter called NESCA to its members (Editor: rubiuo@hal.ugr.es). The international journal *Speech Communication* also has strong links with ESCA (Editor: sorin@lannion.cnet.fr). In addition, an electronic student journal "web-sls") has been available since last year (http://web-sls.essex.ac.uk/web-sls; this is supported by ESCA, ELSNET and EACL.

From its limited resources, ESCA provides grants to members of the speech community to enable them to participate in various speech events.

To date, over 500 members have realised that ESCA membership is value for money, at the very least because members are entitled to various discounts. Members can be found in both speech and language communities, in Europe and elsewhere; they come both from industry and academia. All readers of this ELRA Newsletter are kindly invited to join ESCA, as its links with ELRA are also excellent.

Prof. Dr. Ir. Louis C.W. Pols
Institute of Phonetic Sciences / IFOTT,
University of Amsterdam
Herengracht 338,
NL-1016 CG Amsterdam,
The Netherlands
Tel.: +31 20 525 2183
Fax: +31 20 525 2197
Email: pols@fon.let.uva.nl

Further Reading

Corpus-based Methods in Language and Speech Processing

Steve Young and Gerrit Bloothooft (Eds.)

ELSNET has published its first book, which provides an in-depth introduction to corpus-based methods. A group of summer school lecturers has written chapters describing statistical modelling techniques for language and speech, the use of hidden Markov models in continuous speech recognition, the development of dialogue systems, part-of-speech tagging and partial parsing, data-oriented parsing and N-gram language modelling.

The book attempts to give a clear overview of the main technologies used in language and speech processing, along with sufficient mathematics to understand the underlying principles. It provides newcomers with a solid introduction to the field and offers existing practitioners a concise review of the principal technologies used in state-of-the-art language and speech processing systems.

Contents include:

- Corpus-based statistical methods in speech and language processing (Hermann Ney),
- Hidden Markov models in speech and language processing (Kate Knill and Steve Young),
- Spoken language dialogue systems (Egidio Giachin and Scott McGlashan),
- Part-of-speech tagging and partial parsing (Steve Abney),
- Data-oriented language processing (Rens Bod and Remko Scha),
- Statistical language modelling using leaving-one-out (Hermann Ney, Sven Martin and Frank Wessel).

A bibliography is also provided.

ELSNET wishes to help this excellent book (hard cover, 234 pages, ISBN 0-7923-4463-4) achieve a wide distribution and is therefore offering it the special price of HFL 85 (c. 39 ECU, or 45 USD, including 6% VAT). Corpus-Based Methods in Language and Speech Processing was published as part of the Kluwer series entitled "Text, Speech and Language" (series editors: Nancy Ide and Jean Véronis).

To order, please complete the Web form to be found at URL: http://www.elsnet.org/publications/elsnetBook.html for your order to be processed automatically.

You can also send a message to the ELSNET secretariat:
Utrecht Institute of Linguistics OTS,
Trans 10, 3512 JK, Utrecht
The Netherlands
Tel.: +31 30 253 6039
Fax: +31 30 253 6000
Email: elsnet@let.ruu.nl
http://www.elsnet.org

ELRA, the European Language Resources Association, has an immediate vacancy for a

Sales & Marketing Manager (m/f)

for ELDA, its Paris-based distribution agency. ELRA, a non-profit association registered in Luxembourg, was established in 1995 and receives financial support from the European Commission and national governments to promote the development and exploitation of Language Resources - monolingual and multilingual lexica, text corpora, speech databases and terminology - in Europe. Enjoying strong backing from the language engineering industry, ELRA's operations are conducted by the CEO and his team at ELDA.The role of the new Sales & Marketing Manager will be to develop and implement sales/marketing strategies for the various categories of language resources, actively sell language resources, monitor market requirements and play a driving role in the development of ELRA's commercial activities. This position will suit an experienced self-starter used to negotiating with top management in industry and research. Terms and conditions of employment are subject to negotiation, but will be commensurate with the responsibilities of the post and will include performance-based incentives. ELRA will pay relocation expenses for the selected candidate. This is initially a one-year appointment with a strong possibility of a further two years or permanent employment.

Qualifications:

- Excellent track record in sales and marketing.
- Very good communication and negotiation skills.
- National of, or resident's/work permits for an EU member state.
- Excellent command of at least two European languages including English.
- Commercial experience in language engineering or related fields, e.g. software localisation. A history of marketing language resources, software or other forms of intellectual property is desirable but not essential.
- Willingness to relocate to Paris.

Please apply in writing*, including a full CV with salary history, to:

ELRA Distribution Agency (ELDA)
Dr. Khalid Choukri, CEO
87, Avenue d'Italie
F-75013 Paris, France
Fax +33 1 45 86 44 88; e-mail elra@calvanet.calvacom.fr

For more information on ELRA, see: http://www.icp.grenet.fr/ELRA/home.html (English) or http://www.icp.grenet.fr/ELRA/fr/home.html (French)

** Initial applications by e-mail will be accepted with follow-up by post/fax*

New Resources

ELRA-S0035 Phonolex (BAS/DFKI)

PHONOLEX consists of a simple list of word forms (666,237 inflected words) with a set of features e.g. orthography (German 'Umlauts' in LaTeX format, capital nouns, old German spelling rules), linguistic information (nouns, verbs, etc.), pronunciation and a list of empirical pronunciations.

Language:	German
Format:	ASCII
Mark-up:	extended SAM-PA (PhonDat-Verbmobil)

ELRA-S0037 Speri-Data AG Technical dictionaries

All dictionaries contain phonetic transcription with related phoneme lists. The following dictionaries are available (the label basic dictionary refers to the above ELRA-S0036):

Domain	Entries
Banking French	10,200
Banking German	10,200
Banking Italian	10,200
Banking Spanish	10,200
Radiology German	42,000 (including basic dictionary)
Radiology English	16,000
Medical German	130,000 (including basic dictionary)
Jurisprudence German	31,000
Jurisprudence German	55,000 (including basic dictionary)
Insurance German & English	37,000

A peculiarity of medical dictionaries in German speaking countries has to be taken into consideration: doctors in Germany, Austria and Switzerland may not use the original technical terms in Latin but the Latin word in a spelled manner or a German technical term (see examples below). Medical dictionaries therefore have to contain three different terms or have to be printed out in three different editions.

Technical term	Technical term	Technical term
in Latin	in German spelling	in German
Appendicitis	Appendizitis	Blinddarmentzündung
Eccema	Eczema	Ekzem
Diarrhoe	Diarrhö or Diarrhöe	Durchfall, Durchfluss
Carbunculus	Karbunkel	Geschwür

ELRA-L0021 Dictionary of French verbs - CORA

This dictionary contains 25,610 verbs with usage domains, level of language (familiar, popular, literary, Quebec and Swiss terms, etc.), conjugation, auxiliary, verbal adjectives in -able, -ant or -é, encoded syntactical constructions (subject, direct & indirect object, adverb), sample phrases, synonyms, operators enabling semantic-syntactic classification, encoding of derived forms in -age, -ment, -tion, -oir, -ure, deverbal nouns, base words from which verbs can be derived, a scale of usage ranging from 1 to 6, like those used by commercial dictionaries (basic vocabulary, extended, specialised, etc.).

Codes enable automatic production of conjugation forms, derived nouns and adjectives and, if necessary, the production of potential forms.

ELRA-S0036 Speri-Data AG Basic dictionaries (colloquial language)

These dictionaries contain a daily-life vocabulary. They include phonetic transcription with related phoneme lists. The following languages are available:

Language	Entries
Danish	8,000
Dutch	12,000
English (UK)	8,000
Finnish	10,000
French	19,000
German	13,000
Italian	23,000
Norwegian	8,000
Portuguese	9,000
Spanish	13,000
Swedish	10,000

ELRA-L0022 Dictionary of words - CORA

This dictionary is composed of 126,844 words, with usage domains, grammatical category, gender, number, uncountable, collective, adjectival, nominal, verbal, adverbial derived forms according to the type of words.

ELRA-L0023 Dictionary of affixes - CORA

4,286 suffixes and prefixes, plus information on their verbal, nominal or adjectival bases or on the verbal basis of greco-latin items. This dictionary does not include the suffixes contained in the dictionary of French verbs (ELRA-L0021) and words (ELRA-L0022) such as -age, -ment, -if, -oir.

ELRA-L0024 Dictionary of verb phrases - CORA

Dictionary of 3,480 entries based on the model of the dictionary of French verbs (ELRA-L0021).

ELRA-L0025 Dictionary of invariable forms and phrases - CORA

Dictionary of 4,783 entries based on the model of the dictionary of words (ELRA-L0022).

ELRA-L0026 Dictionary of exclamatory stereotyped phrases - CORA

Dictionary of 1,901 entries based on the model of the dictionary of invariable forms and phrases (ELRA-L0025).

ELRA-L0027 Dictionary of French local authorities - CORA

38,965 entries in lower cases with accents, controlled on the guide Michelin, without localities; A link can be made to the dictionary of words (ELRA-L0022) which contains inhabitants’ names and their correspondence with town names.

ELRA-L0028 Dictionary of noun phrases and plural-only words - CORA

2,138 compound names and 1,397 entries of plural-only words.