THE ESPRIT PROJECT POLYGLOT

Louis Boves

Dept. of Language and Speech
Nijmegen University
P.O. Box 9103
6500 HD Nijmegen, The Netherlands

ABSTRACT

The ESPRIT project POLYGLOT aims at developing multi-lingual Speech-to-Text and Text-to-Speech conversion and to integrate this technology in a number of commercially viable prototype applications. Speech-to-Text conversion is mainly concerned with very large vocabulary isolated word recognition. It uses a statistical knowledge based approach that was pioneered for Italian and is now being extended to other languages. Work on continuous speech recognition has the character of an in-depth feasibility study. For Text-to-Speech conversion a new multi-level data structure is developed that facilitates rule writing by offering a semi-graphical rule format. The recognition and synthesis technology is used to build a number of generic prototype applications that mainly address office automation.

INTRODUCTION

Polyglot is a 16.5 million ECU (i.e., approximately $23 million) project that is funded by the European Community as part of the ESPRIT-2 program. As is usual in ESPRIT, the European Community covers 50% of the total costs; the other half of the cost is paid by the partners in the Polyglot Consortium. In terms of manpower the resources amount to a total of some 133 man years. The project started in August 1989. It was approved for a duration of three years. Originally, a workplan spanning five years was submitted, so considerable cuts in the plans were necessary. An attempt will be made to obtain new ESPRIT funding for a continuation project that will probably go under the name Polyglot-2.

Polyglot builds partly on the results of a previous ESPRIT project that was titled "Linguistic Analysis of European Languages" [1]. In that predecessor project the attention was mainly focused on the acquisition of databases and statistical knowledge about the seven European languages that are being investigated in Polyglot. In alphabetical order these languages are British-English, Dutch, French, German, Greek, Italian and Spanish. The data and knowledge acquired in that project were a.o. used to build grapheme-to-phoneme and phoneme-to-grapheme conversion modules for the seven languages. Of course, the phoneme-to-grapheme conversion modules required the development of language models; for that goal Markov models based on Part-of-Speech information were developed.

Since it is necessary to have at least one partner in each of the seven language communities the Polyglot Consortium is necessarily quite large; at this moment it consists of the following partners (there have been some modifications in the past):

- Olivetti Speech and Language Laboratory, Torino, Italy, acting as Contractor, i.e., as leader of the consortium
- Bull SA, Massy, France
- Philips Hamburg Research, now located in Aachen, Germany
- Siemens AG, Neurenberg, Germany
- Centre for Speech Technology Research, Edinburgh, U.K.
- LIMSI/CNRS, Orsay, France
- I.P.O., Eindhoven, The Netherlands
- Nijmegen University, The Netherlands
- Patras University, Greece
- Ruhr University, Bochum, Germany
- Universidad Polytecnica de Madrid, Spain

The work in Polyglot is structured in two ways. First there are five Work Packages (WP), one dealing with Isolated Word Speech Recognition (IWSR), one with Continuous Speech Recognition (CSR), one with Text-to-Speech Conversion (TTS), one with Applications (APP) and one with Common Tasks (COT). Perpendicular to this structuring based on technologies there is another organizing principle, viz. the distinction between Language Dependent and Language Independent work [2]. Polyglot aims at the development of Language Independent frameworks in which Language Dependent knowledge and data can be integrated in order to build homogeneously structured multi-lingual speech systems. In this paper the five Work Packages will be the organizing principle.

Pilot Languages

In a consortium as large as Polyglot that, moreover, assembles partners from countries with widely diverging cultural and economic status and traditions it is impossible that all partners have the same level of expertise in all aspects of the work. That is reflected by the fact that some of the partners avail of high quality speech recognition and/or speech synthesis systems for their own language, whereas other partners are still in early stages of building such systems for their own language. That is not necessarily due to a lack of knowledge or expertise; it can also be the result of strategic decisions of some partner to concentrate his efforts on other topics in the past. In such a situation it is only natural that the short term goals for the languages are different. This introduces the concept of *pilot languages,* i.e., languages for which the work is ahead of the remaining languages. The experience gained in the work on the pilot languages is disseminated and used to speed up the work for the other languages.

COMMON TASKS

In ESPRIT projects are set up as collaborative enterprises. Thus explicit efforts are made to ensure that all partners cooperating in a Consortium use common, or at least compatible procedures. Ideally, they should even use common

hardware. In Polyglot the ideal of common hardware could not be reached, since most of the partners already had acquired most of the computers necessary for carrying out the research before the start of the project. The budget available for the project did not allow the purchase of completely new equipment for the project. This necessitated considerable effort in specifying standards for hardware and software in order to obtain a common platform.

One very important advantage of a collaborative project is that the costs of software development can be kept to a minimum by distributing the development tasks over the partners. Obviously, this is one aspect where the distinction between Language Independent frameworks and Language Specific data plays a major role: the Language Independent software needs to be written only once and made available to all partners. Since it is not yet feasible to produce completely system independent software it was specified that all software written as part of the project should be in 'C' and that, with few exceptions, every program should be able to compile and run on a SUN station and on an MS-DOS PC.

Another field where standardization is crucial is the recording of speech databases. Since databases for seven different languages are needed, it was not possible to do all recordings at a single site. In order to obtain compatible recordings from seven or so different sites in seven different countries, precise specifications of the recording conditions had to be developed. That process was complicated by the fact that several Work Packages had different requirements with respect to recording quality and procedures.

It has been the task of the WP Common Tasks to provide all standards and specifications. Moreover, this WP was responsible for the organization and the monitoring of the acquisition of all databases needed in the other WP's. Finally, it was responsible for the production of the software for common use.

ISOLATED WORD SPEECH RECOGNITION

The WP IWSR aims at the implementation of very large vocabulary, speaker adaptive, isolated word speech recognition for all seven languages of the consortium. In practice, an attempt is made to extend an existing system for Italian to six other languages [3]. That system was designed to offer fast speaker enrollment, easy modification of the dictionary and flexible control.

The systems run on an MS-DOS PC that uses one or two special-purpose plug-in boards. After signal processing, resulting in vectors of 20 LPC Cepstrum coefficients and two energy values for each 10 ms speech frame, each frame is given the label of the nearest phonemic template. The string of *prototypes* thus formed is then used for a fast lexical access that retrieves the 100 or so most likely word candidates. The Dynamic Programming string match used in this *preselection* phase relies on knowledge about phoneme confusions, phoneme durations and phoneme and diphone frequencies in the language. Typically, some 25 templates are used during preselection. Next, *Fine Phonetic Analysis* (FPA) is used to sort the word candidates produced during preselection and retain only the 1-5 best scoring candidates. The objective function is based on the distance from spectral prototypes (during FPA the number of prototypes is typically around 70), duration and three features derived from energy. A left-to-right beam search is used to find the optimal alignment between the speech and the phonetic representations of the words returned by preselection. Finally, a language model is used to select the best scoring word among the output of FPA. The language model combines word frequencies, a bigram model and some deterministic knowledge and the acoustic probability of each candidate in a single probabilistic score.

Equipped with just the DSP board that performs the LPC analysis the system runs in real-time with vocabularies of 20,000 words. In order to obtain real-time performance with much larger vocabularies (say between 60,000 and 100,000 words) another special purpose board, built with four different ASIC's that speed up preselection has been designed.

Speaker Enrollment

Most of the knowledge in the system is obtained from processing large amounts of speech from a large number of different speakers. Thus, only the prototypes used in preselection are speaker independent. Since the number of prototypes used during that stage is typically very small, it is an easy task to acquire personal prototypes for a new speaker. Enrollment consists of speaking some 40 carefully chosen words that are processed by automatic prototype extraction software.

Modifying the Vocabulary

Tools are provided for the maintenance of the dictionaries. When new words are added, the graphemic forms are automatically converted to phonemic forms and rules are provided for the generation of the most common pronunciation variants. Since lexical access during preselection and the scoring during FPA heavily rely on phonemic models, accurate modeling of the pronunciation is mandatory.

Flexible Control

For many applications the performance of an IWSR system can be immensely improved if the size of the vocabulary can be dynamically adapted to the state of the dialogue. In fact, very large vocabularies are only needed during free text dictation. The IWSR system developed for Italian, and in the process of adaptation for the other languages, allows on-line selection of subsets of words from the base vocabulary. Obviously, the ability to make this selection is especially important in preselection.

State-of-the-Art

Most of the work needed to implement preselection for all languages is now ready. Dictionaries comprising representations for use in preselection have been compiled for all languages. Also, prototypes for a small number of speakers in each language have been built. Preliminary tests run in January 1991 have shown that acceptable preselection results (for at least 98% of the words spoken, the correct word is in the preselection list) are obtained for all languages under consideration. Formal tests of preselection performance with lexica of 2000 words are planned for April 1991. For the pilot

languages English and Dutch much larger dictionaries are available (and will be tested).

CONTINUOUS SPEECH RECOGNITION

Unlike IWSR, where concrete applications are aimed at, the WP Continuous Speech Recognition was set up to investigate the feasibility of continuous speech recognition in situations that differ from the DARPA Resource Management task. In addition, it is proposed to carry out an in-depth investigation of the viability of alternatives for the HMM approach. Still, the DARPA RM task was chosen as a reference, in order to be able to relate the performance of the systems built in Polyglot to a generally accepted and well-understood standard.

Originally it was planned to do a large number of experiments in which integrated search should be compared with bottom-up phoneme based search. Moreover, both search strategies should be used with HMM, TDNN, and the frame labels produced by the Olivetti IWSR system described above. Finally, the approaches should be compared with respect to acoustic-phonetic decoding and word and sentence accuracy. Unfortunately, the limited resources available for this WP combined with a host of practical problems forced a drastic reduction of these plans. It is now intended to limit the investigation to TDNN and HMM in integrated search. On the other hand, more emphasis will be put on work on language models and on system integration aspects, especially with respect to the possibility of using linguistic constraints to improve phonetic decoding. Apart from the DARPA speaker dependent RM task the systems will be tested on a corpus of read newspaper text. Such a corpus, limited to a vocabulary of 5000 words, is available for British-English and German. Formal tests of the performance on the DARPA RM task will be available in August 1991. Preliminary tests with the Continuous Mixtures Densities HMM system developed by Philips Hamburg Research showed encouraging performance: using 46 monophones and 227 triphones a word error rate of 23.3% was obtained in the no-grammar condition. The triphones were selected by choosing only those that had a frequency of occurrence > 10 in the dictionary (data presented by Herman Ney during the January 1991 Review Meeting).

TEXT-TO-SPEECH CONVERSION

In Polyglot, a relatively large part of the resources is devoted to Text-to-Speech (TTS) conversion. This is because we believe that a high quality TTS system is essential for the majority of the applications in which speech technology is to provide the major user interface. Moreover, high quality TTS systems are not yet available for most languages represented in Polyglot. Last but not least, even if such systems would exist for some languages, they cannot be integrated into a single system that has an architecture that is uniform for all languages.

Advanced Features

Automatic Language Identification. The TTS system that is being developed in Polyglot will have a number of unique features. One is an automatic Language Identification Module (LIM) tIat is able to identify the language of each sentence sent to the TTS system. Since we will have a multi-lingual system with a uniform architecture for all languages the LIM will act as an intelligent switch that selects the appropriate language for each sentence. LIM is implemented as a rule-based program that uses three knowledge sources:

- a list of very frequent words for all languages
- a list of letter combinations that can or cannot occur word initially and word finally in any of the languages
- a list of letter sequences that cannot occur word internally in any of the languages

These word-level knowledge sources are not sufficient to determine the language for each word. In fact, many words can, and do, occur in more than one language. Therefore, a sentence level scoring mechanism is added that selects the most likely language for each complete sentence [4]. It has been shown that LIM in its present form performs virtually without errors. Most problem cases that were found in a test on large text corpora appeared to be due to errors of a very preliminary version of the sentence boundary detection algorithm and the inability of the LIM to recognize, e.g., addresses in foreign languages.

Syntax and Prosody. Most existing TTS systems suffer from inadequate prosody, due to the fact that syntactic processing is kept to a minimum. However, the Polyglot system will do sufficient syntactic and prosodic processing to be able to generate adequate intonation in most neutral texts. To that end it will use a medium sized lexicon (between 5,000 and 10,000 most frequent full words for each language) containing phonemic forms and word class information, a set of 'morphological' rules that guess the word classes of the words not found in the dictionary, a Markov grammar that computes the optimal ordering of the possible classes of all words and a Wild Card Parser (WPC), i.e., a deterministic parser based on a Context Free Grammar. The WPC attempts to account for the maximum number of words in an input sentence using the minimum number of major syntactic constituents. Thus, it yields a partial parse each time a complete parse of the input is not possible. Partial parses may contain words that are not part of a syntactic constituent; these unaccounted words are called Wild Cards [5]. The output of the WPC is given to a prosodic processor that implements a form of the 'Focus-Accent' theory that predicts the relation between syntax and prosody as well as the words that should carry pitch accents [6,7]. Experiments for Dutch have shown that the approach yields excellent results. Consultation with the partners working on other languages has confirmed that the same approach should work for all languages under consideration.

Multi-level Data Structure. In order to be able to take full advantage of the syntactic and prosodic information it was necessary to design a multi-level data structure for the TTS system in which information on several levels can be stored in such a way that levels can be linked with one another and each rule can access all information on all levels that are relevant [8]. In order for this quite complicated data structure to be addressable by phonetic rule writers, a rule formalism had to be designed that allows expression of rules in the form of a graphical representation of the relevant levels of the data structure. It is expected that the prosodic information will not only be helpful in generating high quality intonation contours, but that it will also enable us to improve the segmental rules,

because it offers an easy way to account for interaction between prosodic and segmental phenomena.

The Architecture

The architecture of the Polyglot TTS system is highly modular. Thanks to the flexibility of the multi-layered data structure and the access functions that come with it, it is possible for each language to use exactly those modules that are needed. It is possible to add layers to the data structure, and that can be done in such a way that only those languages that use the new layers will actually implement them. Thus it is possible for each language to choose exactly those types of information (i.e., layers) that are necessary. For instance, morphological analysis is essential for English, but it may not be necessary for Italian. If the Italian version of the TTS system does not need the morphology layer, it simply does not use it. Of course, it is then not possible for rules in the Italian system to refer to morphological data.

The Polyglot TTS system will use rule based synthesis. Segmental rules that produce highly intelligible speech are available for Italian, Spanish and Dutch. For Dutch the rule based system that is under development, partly in the framework of Polyglot, partly under the national Dutch SPIN program, has been shown to equal the best competing diphone system in intelligibility both on the level of segments and paragraphs [9]. It is believed that rule based synthesis offers better opportunities to improve speech quality than diphone systems. The rules for the other languages will be obtained by adapting existing rules for other languages. In order to support that conversion task a very flexible working environment has been built that allows the rule developer to look at parameter tracks and spectrograms of both natural and synthetic utterances, to listen to both natural and synthetic versions of an utterance, and to change rules interactively.

For all languages a prosody data base has been recorded, that consists of large numbers of sentences covering most syntactic and prosodic structures of interest as well as a number of short prose passages. All material has been read by two professional speakers, a female and a male. The speech material is transcribed on the segmental and the suprasegmental level and the resulting information is stored in a data base that allows one to access the data in many different ways. It is intended to segment and label the speech on the level of segments, so that the information can be used to develop and test duration rules. In addition, the suprasegmental transcriptions are used to derive and test rules that predict pause locations and pitch contours. Most of the linguistic and phonetic rules are implemented in such a way that they can be executed on a PC. The conversion of phoneme target values to filter parameters and the computation of the eventual speech signal are done on a DSP32C board.

APPLICATIONS

ESPRIT projects are mainly application driven. Unlike the situation in the DARPA community, one of the major evaluation criteria for ESPRIT projects is the commercial interest of the results and the commitment of the industrial partners to commercialize those results. Thus, it is only natural that a considerable amount of the resources in Polyglot are spent to the development of prototype applications. Unlike most groups working on speech technology the Polyglot Consortium is not exclusively aiming at applications that involve information access via the public telephone network. Most of the applications that are under development are intended for use in the office or in the classroom. Of course, many commercially viable applications will require the combination, if not the integration of speech recognition and text-to-speech conversion. As can be seen below, this is reflected in some of the prototypical applications in Polyglot. Also, the applications that are at this moment limited to either speech input or speech output can be easily extended to exploit a combination of both technologies.

Application Architecture

In Polyglot it was decided that the work on applications should not be limited to the development of a number of prototypes. Instead, it was felt that the development of a uniform Application Architecture that would enable application developers to integrate speech technology in an easy way was at least as important from the point of view of future exploitation of the technology. Therefore, one of the major tasks of the APP WP is the specification and implementation of so called Application Programming Interfaces (API) that will allow almost every application program to interface with the Polyglot IWSR and TTS system. The API's will be provided for MS-DOS and MS-WINDOWS; the desirability to provide API's for UNIX is still under investigation.

Application Prototypes

Dictation. Polyglot is working on a number of prototype applications. Perhaps the most ambitious one is report preparation. Although there are many domains where such a system would be useful, the demonstration prototype will be limited to medical dictation. The prototype will work in two modes, viz. interactive and free dictation. In interactive mode the system will guide the user through a predefined protocol. The system will ask a number of questions, using the TTS system, and at each point in the dialogue the user is offered a number of possible answers that he or she can speak. Each alternative will generate the appropriate passage in the eventual report. In free dictation mode the dialogue will still be the same, thus ensuring that the report will be complete under all circumstances, but now the user is offered an extra alternative after each question; if (s)he does not choose one of the fixed alternatives, but the added alternative *free text,* the system will go into dictation mode and the user can enter arbitrary text that will be copied to the report. A laboratory version of the complete system is available for a Radiology application in Italian. A fully developed prototype will follow soon. For British-English a laboratory version of the interactive mode is available for applications in Radiology and Pathology. Prototypes for the other languages are expected to be ready by the end of 1991.

Office Automation. Bull SA already markets an application named **Microname** that offers access to a data base of telephone numbers in a company. At present, access is only via a terminal. Ergonomic and marketing studies have shown that there is a need for access by voice. The speech version of **Microname** will be offered integrated with a speech-driven version of another product, called **Micropost,** that offers telephone access to Electronic Mail. The user will

be offered the possibility of accessing his or her E-Mail system via the public telephone network, scan through the messages and ask that one or more messages be read by the TTS system. Obviously, this is one of the applications where automatic identification of the language of the message is essential.

Teaching Aids. In a multi-lingual community like Europe there is an ever increasing need for language training and teaching. One of the Polyglot applications addresses this need by offering flexible computer assisted instruction in learning the meaning of words, in spoken language comprehension and in spelling proficiency. Based on a CD-ROM containing a large number of multi-lingual dictionaries the user will be able to enter any word in one of the languages and see and hear the translation in another language. Since most words will have several, if not many, translations, it will be possible to view and to hear the words in sentence contexts that help to make the correct choice. In another application the TTS system will read sentences or passages to the user in one of the languages. The user is then asked questions that test comprehension. On request, the computer can check the spelling proficiency of by asking the user to type the sentences that were spoken and compare the user input with the text sent to the TTS system.

CONCLUSION

The ESPRIT project Polyglot concentrates on the extension of existing technology to make it available in uniform architectures for large numbers of languages. It is believed that this is extremely important for the successful launching of many applications in the multi-lingual community that the EC is now, and that it probably will remain for a long time to come. In addition to extending existing technology to ever more languages, the project should make substantial contributions to to our knowledge about the structure of speech and spoken language. These contributions will be most apparent in the fields of text-to-speech conversion and continuous speech recognition. With respect to isolated word speech recognition we believe that the attempt to extend existing technology to many other, phonetically quite different languages will also advance our understanding of the technology.

REFERENCES

1. Boves, L., "A multi-lingual language model for large vocabulary speech recognition," *Proceedings from EUROSPEECH-89,* Vol. 2:168-171, Paris, France, 1989.

2. Vittorelli, V., Adda, G., Billi, R., Boves, L., Jack, M., and Vivalda, E., "Esprit POLYGLOT Project: Multi Language /Speech Technology," *Acoustics Bulletin*:17-20, October 1990.

3. Billi, R., Arman, G., Cericola, D., Mollo, M.J., Tafini, F., Varese, G., and Vittorelli, V., "A PC-based very large vocabulary isolated word speech recognition system," *Proceedings from EUROSPEECH-89*, Vol. 2:157-160, Paris, France, 1989.

4. Henrich, P., "Language identification for automatic Grapheme-to-Phoneme conversion of foreign words in a German Text-to-Speech system," *Proceedings from EUROSPEECH-89*:220-223, Paris, France, 1989.

5. Willemse, R., and Boves, L., "Context Free Wild Card Parsing in a Text-to-Speech System," *Proceedings from ICASSP-91*, Toronto, Canada, 1991.

6. Baart, J.L.G., *Focus, Syntax, and Accent Placement: towards a rule system for the derivation of pitch accent patterns in Dutch as spoken by humans and machines,* Ph.D. Dissertation, Leiden University, 1987.

7. Dirksen, A., and Terken, J., *Specification of the procedures for prosodic marker assignment*, Deliverable DEL-TT-11, Polyglot Project, 1991.

8. van Leeuwen, H., and te Lindert, R., "Speech Maker: Text-to-Speech synthesis based on a multi-level, synchronized data structure," *Proceedings from ICASSP-91*, Toronto, Canada, 1991.

9. van Bezooijen, R., and Pols, L. C. W., "Evaluation of allophone and diphone based Text-to-Speech conversion at the paragraph level," *Proceedings from the XIIth Intern. Congress of Phonetic Sciences*, Aix en Provence, France, 1991.