The African Speech Technology Project: An Assessment

Roux, JC[*], Louw, PH[*] & Niesler, TR[**]

[*]Research Unit for Experimental Phonology
[**]Electrical and Electronic Engineering
Stellenbosch University
Stellenbosch
jcr@sun.ac.za, phlouw@sun.ac.za, trn@sun.ac.za

Abstract

This paper reflects on the recently completed African Speech Technology (AST) Project. The AST Project successfully developed eleven annotated telephone speech databases for five languages spoken in South Africa i.e. Xhosa, Southern Sotho, Zulu, English and Afrikaans. These databases were used to train and test speech recognition systems applied in a multilingual telephone-based prototype hotel booking system. An overview is given of the database design and contents. The acquisition of the data is discussed with regards to the telephony interface, as well as speaker recruitment and briefing. Particular reference is given to some of the practical implications of acquiring appropriate data in under-developed communities. Database management processes such as transcription, quality control and validation are explained. This is followed by information on the development of the prototype. Results of usability tests are discussed followed by an assessment of the Project as a whole.

Introduction

The research project entitled *Promoting the development of the official languages of South Africa through language and speech technology applications* (working title: African Speech Technology - AST) was a four year project funded by the Innovation Fund of the Department of Science and Technology (DST) of the South African national government. The Project commenced in January 2000 and was introduced to the LREC community in Athens in the same year (Roux et al., 2000). This Project was successfully completed in December 2003.

The Project was motivated by an appreciation of the need to develop the indigenous languages of South Africa at technological level in order to keep these languages abreast with developments in the ICT field, and to facilitate access to information for all citizens in a developing country. As one of its main aims, AST developed telephone speech databases for five of South Africa's eleven official languages, namely Xhosa, Southern Sotho, Zulu, South African English and Afrikaans. These databases were fully transcribed both orthographically and phonetically.

A second main aim of the Project was to develop a multilingual telephone-based hotel booking system as a prototype for demonstration purposes. This system allows speech to be both recognised and synthesised in one of three languages. A user may use the system to negotiate a reservation at a hypothetical hotel in his or her preferred language. The large and varied speech databases produced as part of the Project were instrumental especially to the successful development of the speaker-independent speech recognition module, which forms part of the overall system.

Database design and contents

Language coverage

The internal speech variation in spoken South African English and Afrikaans is considerable and, in many instances, culturally-bound. In order to make provision for these known varieties, a total of eleven databases based on the five languages was developed.

The English and Afrikaans databases are divided into five and three sub-databases respectively, based on different speech varieties used by mother-tongue and non-mother-tongue speakers. For the English database, English mother-tongue speakers (database EE) as well as four groups of non-mother-tongue speakers were targeted, namely Black, Coloured, Asian and Afrikaans speakers (databases BE, CE, IE, AE). The Afrikaans database-group included speech produced by Afrikaans mother-tongue speakers, as well as Black and Coloured speakers (databases AA, BA, CA). Within the Black speaker group, speakers having any one of Xhosa, Zulu, Southern Sotho (Sesotho), Tswana (Setswana) or Northern Sotho (Sepedi) as their mother tongue were included. For the Xhosa, Zulu and Southern Sotho databases only mother-tongue speakers were recruited (databases XX, ZZ, SS).

General description of contents

The AST contents specification totals 38 to 40 utterances per phone call comprising a mixture of spontaneous and read speech. The types of read utterances elicited include isolated digit items, natural numbers, dates, times, money amounts, application/domain specific words or phrases, and phonetically rich words and sentences. Spontaneous responses were gathered by asking the speakers to say their age, home language, date of birth and to answer yes/no questions.

The acquisition and management of speech databases

Telephony interface

An ISDN Primary Rate Interface (PRI) was required in order to digitally record the incoming calls on eleven channels simultaneously, as well as to log call information. A Dialogic D/300-SC board which interfaced directly to an ISDN PRI channel from Telkom, was used. When a speaker dialled the toll-free number, he/she would be welcomed and prompted to answer a few questions and then requested to read the items on the data sheet one by one. Each prompt was followed by a “beep” tone after which recording commenced. Each response given by the speaker was stored in a separate file. In turn, all the utterances recorded per call were stored in one folder named according to the date and time of the call.

Speaker recruitment and briefing

The AST Project involved colleagues from universities across the country recruiting speakers to partake in the recording of the databases. Between 300 and 400 speakers were recruited for each database. The aim was to recruit 50% male and 50% female speakers between the ages of 20 and 60. Fifty percent of these speakers were requested to use a land-line phone and 50% were requested to use a mobile phone when calling a toll-free number. Prior to making the phone call the speakers were briefed with information regarding the Project and how to make the phone call. Each speaker was presented with a unique data sheet containing the items to be read. On successful completion of the phone call and return of the data sheet, each speaker received a reward in the form of a Telkom telephone card.

Adhering to the above-mentioned requirements proved to be quite a challenge when recruiting speakers from different linguistic and socio-economic environments. In certain communities, for example where the XX, ZZ, SS, BE, BA, CE and CA databases were recorded, not many recruits owned mobile phones or had airtime available. In many cases the recruiter had to use his/her own mobile phone with airtime provided by the Project. As a result, the distribution of mobile phone types was less diverse than had been hoped for. Spontaneous time phrases were limited by the fact that many speakers did not wear watches. Consequently, the response recorded was often that of a speaker asking a bystander what the time was.

Although it was not difficult to recruit older speakers, these speakers could not always provide the desired responses. This was because they were not accustomed to interacting with a pre-recorded voice-based system and had difficulty reading a language other than their mother tongue. Often, these speakers responded before the tone was played. Hence many utterances were truncated. Speakers who did not have a clear understanding of the purpose of the phone call, replied in a language other than the prompt-language. Gathering speech data in multilingual environments with speakers varying in degrees of literacy, proved to be extremely taxing.

Data management

A total of 5767 phone calls were recorded over a period of twenty months. The contents of the calls varied from completely empty, to a few utterances per call, to complete calls consisting of 40 utterances. By means of a script, calls with 5 or fewer files were classified as empty and not considered further. The remainder was subjected to a screening process in which the transcription team listened to each call in order to determine the contents. Calls containing no speech or excessive noise were flagged as unusable. As the statistics in Table 1 show, 41% of the almost 6000 calls recorded were classified as empty or unusable. To some extent this data loss was anticipated by distributing 400 datasheets per database, while aiming for only 300 to 400 usable calls. Taking this into account, the number of transcribed calls actually totaled 77% of the expected 4400 calls.

11 Databases	**Recorded**	**Empty/ Unusable**	**Transcribed**
XX	632	317	315
SS	500	224	276
ZZ	399	159	240
EE	485	193	292
AE	562	218	344
BE	600	276	324
CE	425	149	276
IE	529	172	358
AA	395	122	273
BA	572	238	334
CA	668	305	363
Total	5767	2373 (41%)	3395 (59%)

Table 1. Phone call recording and transcription statistics

The 3395 usable calls were subsequently processed in the five successive stages listed below:

- Orthographic transcription
- Quality control
- Generation of deterministic phonetic transcriptions
- Manual phonetic corrections
- Internal validation

A team of approximately 70 transcribers was extensively trained to process the data. The level of experience of members of the transcription team varied from basic computer literacy and no linguistic background, to experts in the field of phonetics. Various training workshops were held throughout the course of the Project to familiarise the transcribers with the intricate transcription specifications. Depending on the level of experience of the particular group, training in the following areas were required: basic computer literacy, file management skills, orthographic transcription specifications, basic phonetics and various phonetic notations, spectrogram reading and speech signal analysis, phonetic transcription specifications and the use of AST validation tools.

The annotation of speech databases

We aimed to produce a speech database wherein the speech of each phone call was orthographically and phonetically transcribed and aligned in time with the speech signal. The data was transcribed using the open source tool *Praat*, developed in the Netherlands by Paul Boersma (www.praat.org).

Standards and specifications

The standards and specifications of the multilingual European SpeechDat Project (Draxler et al., 1998) were used as a reference when developing the orthographic and phonetic specifications for the AST Project (Louw et al., 2001). However, the specifications contain many unique features in order to accommodate language-specific issues. A phoneme set (ASTbet) that contains all phonemes and sound combinations such as affricates, diphthongs and tripthongs that occur in all five languages was compiled. The phonetic transcriptions were strictly limited to contain only these phonemes. Four different phonetic notations were utilised during the transcription process, namely the International Phonetic Alphabet (IPA), Extended Speech Assessment Methods Phonetic Alphabet (X-SAMPA), Praat and ASTbet. Praat uses so-called "backslash" sequences to display IPA symbols. Transcribers more familiar with IPA than X-SAMPA, therefore, had to learn the (sometimes very long) Praat strings. Six databases were transcribed with X-SAMPA because the transcribers of these databases were already familiar with this notation, and four were transcribed using IPA/Praat strings. Both the X-SAMPA and IPA/Praat transcriptions were eventually converted to ASTbet. ASTbet symbols correspond uniquely to IPA symbols, but contain only basic printable ASCII characters.

The transcription specifications became quite intricate, having to account for phenomena such as code switching, place names consisting of two or more parts, word fragments, mispronunciations and, in particular, cases where combinations of these phenomena occurred. It was not possible to anticipate all manners of speaking, hence the specifications were extended and became more complex as we progressed with the transcriptions. Moreover, the quality of the transcriptions were subject to human error as the transcribers made typing errors, spelling mistakes and sometimes misinterpreted what they heard.

Quality control and internal validation

Initial quality control was done by importing the transcriptions into a single document and proof-reading this for specification and spelling mistakes. Errors found in this way were corrected manually in each affected source file. Perl scripts were developed to identify and, in some cases, automatically fix errors in the data. These scripts would eventually replace the manual quality control process. The automatic validation process was initiated by running the script called `Cleanup`. This script applies corrections such as stripping redundant white spaces, inserting silence markers in untranscribed intervals, stripping invalid characters and fixing incorrect combinations of parentheses and brackets. `Cleanup` reduced the number of errors in the XX database by about 87%.

A second validation script that was used extensively is `CheckForErrors`. All allowed phonetic, orthographic and marker symbols, as well as valid combinations thereof, were listed in this script. If a transcription (either orthographic or phonetic) within a particular file did not comply with the specifications, this file would be flagged and copied to a separate directory, together with a text file reporting the exact error that occurred. Transcribers would subsequently correct these errors and re-run the script until the data was error-free.

Prototype speech recognition systems

Prototype spoken language dialogue systems operating in Afrikaans, English and Xhosa were developed during the final stage of the Project. The task of the dialogue system was to allow a user to make a hotel reservation. This was chosen for its suitability as a demonstration task and did not serve as an end in itself.

Speech recognition

Speech recognition was accomplished using an HTK-based decoder (Young et al., 1999). This hidden Markov model-based (HMM) speech recogniser performs a time-synchronous beam-search using the Token-Passing procedure (Young, 1989).

Acoustic models

A set of HMM acoustic models was trained for each of the three languages using the HTK tools. The speech was parameterised as Mel-frequency ceptral coefficients (MFCCs) and their first and second differentials. Diagonal-covariance speaker-independent cross-word triphone models with three states per model and eight Gaussian mixtures per state were trained using the phonetically-labeled training sets by embedded Baum-Welsh re-estimation and decision-tree state clustering (Woodland et al., 1994). The performance of these models, in terms of phoneme recognition accuracy, are listed in Table 2. A more detailed description of the development of these phoneme models may be found in Niesler and Louw (2004).

Language	Phoneme accuracy %	No. of phonemes
AA	67.4	72
EE	74.7	83
XX	64.3	110

Table 2 Phoneme recognition accuracy

From the table we see that the English models out-perform the Afrikaans models by a 7.3% margin, and these, in turn, outperform the Xhosa models by a 3.1% margin. Since English has the smallest number of phonemes, followed by Afrikaans and then Xhosa, this result was not unexpected.

Natural language understanding

A finite-state architecture was adopted for the natural language understanding component of the spoken dialogue system (Niesler & Roux, 2001). Meaning is associated with particular paths through a finite-state network by embedding semantic tags at appropriate points in the network's definition. The understanding process consequently consists of a parsing operation that determines whether a user utterance is contained within a given finite-state network. The semantic tags associated with a successful parse represent the result of the understanding process.

Dialogue modelling

A software toolkit (AStudio) was developed with a graphical user interface to facilitate dialogue development. A system-directed dialogue strategy making extensive use of implicit confirmation for error recovery was employed in our prototype systems. Each system consisted of 45 dialogue states.

Speech synthesis

A software interface was developed to the Festival toolbox (www.festvox.org) to achieve high quality limited-domain concatenative speech synthesis in each of the three languages.

Usability tests

Mother-tongue speakers of Afrikaans, English and Xhosa were recruited to test the respective prototype systems. In this way each system received between 78 and 88 calls. The percentage of these calls resulting in a successful booking is shown in the following table.

System	% Successful bookings
AA	83
EE	77
XX	60

Table 3 Percentage successful bookings

The Afrikaans and English systems showed the highest success rates, while the Xhosa system lagged behind by a larger margin. Careful analysis of the calls showed that calls to the Xhosa system were very frequently made from noisy environments. Furthermore, the Xhosa users were more likely to respond to the system using long sentences. Both these factors increase the difficulty of the recognition process, leading to a larger number of recognition errors and consequent deterioration in system performance.

Finally, it was observed that Afrikaans and English callers respond in a more limited variety of ways to the system than the Xhosa callers. This allows simpler understanding and recognition grammars to be used for Afrikaans and English, while more complex grammars must be constructed for Xhosa. This may again impact negatively on the latter system's performance.

Assessment

Our experience has been that the development of annotated speech databases is a non-trivial exercise. Planning is a prerequisite for a successful project, especially when negotiating such unpredictable territory as human communication. Furthermore, gathering language resources in a developing country with speakers from different linguistic backgrounds and literacy levels presents its own unique challenges to the design of data sheets and the data capturing process. We have found that, when collecting speech data on such large scale and in a relatively short time span, basing our work on established standards and ensuring that those standards comply with language specific issues, has reduced the number of potential pitfalls.

A team of about 70 people was involved in the screening, transcribing and validation of the databases. Despite receiving regular in-depth training and feedback on all aspects of the process, it remains a tedious and error-prone task. Scripting has proven to be an invaluable aid in fixing many types of errors automatically and is a powerful tool for the manipulation of data.

It was, however, well worth the effort if one considers the value of such reusable resources. We anticipate that the eventual deployment of speech-based query systems will greatly enhance the access to information for different language groups in the country and, in particular, the illiterate population. It is expected to have an impact in fields such as e-health, e-commerce, e-learning and e-government and will help to bridge the digital divide by promoting the use of local languages in modern communication systems.

Finally, the successful development of the three prototype systems has demonstrated the feasibility of spoken dialogue technology in the South African context. Indications are that dialogue design methods should be improved and refined in order for the performance of systems operating in the African languages to match that of European languages.

References

Draxler, C., Van den Heuvel, H. and Tropf, H. (1998). SpeechDat experiences in creating large multilingual speech databases for teleservices. In Proceedings of the First International Conference on Language Resources and Evaluation (pp 1:361-366). Granada, Spain: ELRA.

Louw, P.H., Roux, J.C. and Botha, E.C. (2001). African Speech Technology (AST) Telephone Speech Databases: Corpus Design and Contents. In Proceedings of the Seventh European Conference on Speech Communication and Technology (pp 2055-2058). Aalborg, Denmark: ISCA.

Niesler, T.R. and Louw, P.H. (2004). English, Xhosa and Afrikaans phoneme recognition for the African Speech Technology telephone speech databases, submitted to: The South African Computer Journal.

Niesler, T.R; and Roux, J.C. (2001). Natural language understanding in the DACST-AST dialogue system. In Proceedings of the Twelfth Annual Symposium of the Pattern Recognition Association of South Africa (PRASA) (pp 134-136). Franschhoek.

Roux, J.C., Botha, E.C., and Du Preez, J.A (2000). Developing a multilingual telephone-based information retrieval system in African languages. In Proceedings of the Second International Conference on Language Resources and Evaluation (pp 2:975-980). Athens, Greece: ELRA.

Woodland, P.C., Odell, J.J., Valtchev, V. and Young, S.J. (1994). Large vocabulary continuous speech recognition using HTK. In Proceedings of the International Conference on Acoustics Speech and Signal Processing. (pp 125-128). Adelaide: IEEE.

Young, S.J. (1989). Token passing, a simple conceptual model for connected speech recognition systems. Technical Report TR38. Cambridge University.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V. and Woodland, P. (1999). The HTK book, version 2.2. Entropic Ltd.