African Speech Technology (AST) Telephone Speech Databases: Corpus Design and Contents

Philippa H. Louw, Justus C. Roux*, and Elizabeth C. Botha***

*Research Unit for Experimental Phonology
University of Stellenbosch, South Africa
**Department of Electrical, Electronic and Computer Engineering
University of Pretoria, South Africa
phlouw@maties.sun.ac.za, jcr@maties.sun.ac.za, botha@ee.up.ac.za

Abstract

The African Speech Technology project is developing telephone speech databases for five of South Africa's eleven official languages, i.e. South African English, Afrikaans, and three African languages, Zulu, Xhosa, and Southern Sotho. These databases will be fully transcribed – orthographically and phonetically – and will be used for the training and testing of phoneme-based, speaker-independent speech recognition systems. This paper describes the design and contents of the speech corpus that is currently being collected over both mobile and fixed networks. In particular language coverage is discussed within the framework of the multilingual character of the South African population. Some language-specific differences with regards to the contents of the different databases are noted. Methods and tools applied in the acquisition of phonetic information are discussed.

1. Introduction

The *African Speech Technology* (AST) project undertaken in South Africa is the first of its kind involving the development of the indigenous languages of the country at technological level. The project aims, *inter alia*, to deliver a telephone speech application developer's toolkit that will function in five of the country's eleven official languages, i.e. in South African English, Afrikaans, and in three African languages, Zulu, Xhosa, and Southern Sotho. A prototype multilingual enquiry and booking system for the hotel industry will be developed as a first application. A detailed report on this project was presented at LREC 2000 [1] and more information is also available at http://www.ast.sun.ac.za.

This paper describes the design and contents of the speech corpus that is currently being collected over both mobile and fixed networks. Experiences and standards of colleagues working in the field of multilingual databases were closely followed [2], and hence the final speech database specification is based on the LE2-4001 SpeechDat(II) project [3]. The full AST corpus design specifications are available on the AST web site [4]. In this paper only a general description of the contents of the databases is given. In Section Two the issue of language coverage is discussed within the framework of the multilingual character of the South African population. A description of utterances is given and some language-specific differences with regards to the contents of the different databases are noted. In Section Three methods and tools applied in the acquisition of phonetic information are discussed.

2. Database Contents

The databases will be used for the training and testing of phoneme-based, speaker-independent speech recognition systems. The content of the databases was designed to support general information retrieval and transaction teleservices and the hotel booking system in particular. Speakers recruited to participate in the recording of the databases were each given a unique datasheet on which the items to be read appear in table format.

2.1. Language coverage

Many South Africans are at least bilingual or trilingual. In addition to their mother tongue, they are usually fluent in either English or Afrikaans. Although English only ranks fifth in the hierarchy of the number of mother tongue speakers across the population [5], it is the most commonly used language in business and industry. The internal speech variation in spoken South African English and Afrikaans is, however, 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 are being 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 as well as four groups of non-mother-tongue speakers were targeted within the South African population, i.e. Black, Coloured, Asian and Afrikaans speakers. The Afrikaans database includes speech produced by Afrikaans mother-tongue speakers, as well as Black and Coloured speakers. 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 are included.

The speech database specification has as its goal to provide recordings of between 300 and 400 different speakers for each sub-database. Speakers were recruited countrywide targeting 400 speakers per sub-database. For the Xhosa, Zulu and Southern Sotho databases only mother-tongue speakers were recruited; therefore these databases totaled up to 400 speakers each. Table 1 below lists the main databases and their sub-groups according to the language spoken and the speech varieties within each database.

Language Spoken	Code	No. of Speakers
1 **English (E)**		**1500-2000**
Speech varieties:		
Mother-tongue English	EE	300-400
Black English	BE	300-400
Coloured English	CE	300-400
Asian English	ASE	300-400
Afrikaans English	AE	300-400
2 **isiXhosa (X)**	XX	**300-400**
3 **Sesotho (S)**	SS	**300-400**
4 **isiZulu (Z)**	ZZ	**300-400**
5 **Afrikaans (A)**		**900-1200**
Speech varieties:		
Mother-tongue Afrikaans	AA	300-400
Black Afrikaans	BA	300-400
Coloured Afrikaans	CA	300-400

Table 1: AST Telephone Speech databases: Languages and speaker populations (min 3 300 – max 4 400)

2.2. General description of contents

The AST contents specification totals 38 to 40 utterances per phone call comprising a mixture of spontaneous and read speech. This results in call durations of approximately 7-10 minutes each. The types of utterances elicited are summarised in Table 2. All utterances / items are read speech unless marked spontaneous. The AST column indicates the number of a specific utterance type per phone call for each speaker database of 300-400 calls. The specification for each database will be the same except for a few language specific-differences.

AST (300 – 400)					Utterance/ item description
X	**E**	**A**	**S**	**Z**	
2	2	2	2	2	isolated digit items
4	4	4	4	4	digit/number strings including repetition of Data Sheet no.
1	1	1	1	1	natural number
1	1	1	1	1	money amount
1	1	1	1	1	gender (spontaneous)
1	1	1	1	1	age (spontaneous)
1	1	1	1	1	home language (spontaneous)
3	3	3	3	3	yes/no questions
1	1	1	1	1	level of highest education (spontaneous)
1	1	1	1	1	type of phone used (spontaneous)
4	4	4	4	4	dates
2	2	2	2	2	times
2	2	2	2	2	application/ domain keywords/key phrases
2	2	2	2	2	word spotting phrases
6	6	6	6	6	directory assistance names
3	3	3	3	3	spellings
0	2	2	0	0	phonetically rich words
3	3	3	3	3	phonetically rich sentences
38	40	40	38	38	TOTAL utterances

Table 2. AST Speech Database Corpus contents

2.3. Language Specific Differences

The contents specifications as well as the expected responses varied with regards to only a few item types due to language specific factors.

2.3.1. Numbers, dates and times

Most items within the corpus were dealt with in the same way for all the languages. For example, the same set of isolated digits, isolated digit sequences, number strings, telephone numbers, natural numbers and money amounts were used as read speech items for all eleven databases. However, responses by speakers of the African Languages were expected to be different from those of English and Afrikaans speakers for these items. English and Afrikaans speakers were expected to only cite the digits, numbers, amounts etc. in their mother tongue, but for the African languages, it is accepted to cite these items in the mother tongue or in English (for Xhosa and Zulu) or even in Afrikaans (for Southern Sotho speakers). In Xhosa, for instance, the item "2353" could be read as "Two thousand three hundred and fifty three" or as "Amawaku amabini namakhulu amathathu namashumi amahlanu anesithandathu" meaning "Thousands-that-are-two and hundreds-that-are-three and tens-that-are-five and three". Code-switching is also likely to appear in the spontaneous citing of dates and times e.g. a Xhosa-speaking person might cite the time as "Ixhesha ngoku ifive past ten" meaning "The time now is five past ten".

2.3.2. Read items

Read items such as dates, times, application key words, and word-spotting phrases were generated for English and then translated to the other four languages. The English set served as a guide-line for each translator to generate a similar list with examples as they would occur in the particular language, taking cultural factors into account.

2.3.3. Names of cities and towns

South African towns or cities typically have names with one or more alternative in another language. If a city name is English e.g. Cape Town, this name would be taken up in the English city and town name list that would ultimately form part of the English database. If this city has a Xhosa alternative e.g. "iKapa", this name would be taken up in the Xhosa city and town name list together with the locative version of this name (in this case "eKapa") that would ultimately form part of the Xhosa database. The same will apply to a city that has more than one alternative name, for example Johannesburg has a Zulu ("iGoli/eGoli") a Xhosa ("iRhawutini/eRhawutini") and a Southern Sotho ("Gauteng") alternative.

2.3.4. Phonetically rich sentences

Phonetically rich sentences were included for all the languages. Three phonetically rich sentences are read by each caller with the explicit intention to obtain adequate training coverage of all phonemes and a good coverage of the most frequent diphones within a language for continuous speech modelling. It is not intended to provide phonetically balanced material, where the frequencies of occurrence of the phones and contexts mirror that of typical linguistic material in that language.

The number of phonetically rich sentences differs for each database. Between 259 and 800 unique sentences per language were used as well as one particularly "rich" sentence that was constructed for each language and was included once in each data sheet.

2.3.5. Phonetically rich words

Phonetically rich words were included to ensure the occurrence of sounds in end silence contexts. The African languages, however, have a consonant-vowel (CV) structure where most words end on vowels. Subsequently phonetically rich words were only included in the English and Afrikaans databases where closed syllables do appear.

3. Methods and Tools

Reliable and quantifiable information on the frequency of occurrence of diphones in the African languages in general, and in the languages of the project are basically non-existent. Hence, a software package was used to accurately and automatically extract this information from large texts.

An existing sound pattern analysis program of the Research Unit for Experimental Phonology, called `Patana`, was further developed and a Windows-based version, `Patana2000`, is currently in use. This software package is utilised to automatically convert orthographic texts to phonetically transcribed texts using grapheme-to-phoneme rules and to analyse these texts to determine the most frequent sounds and sound combinations as well as the distribution of these elements in a particular language.

`Patana2000` can currently transcribe and analyse Xhosa, Zulu, Southern Sotho and Tswana texts. A grapheme-to-phoneme rule set was developed for each language and the user can select the appropriate rule set before phonetically transcribing a text. There are no limitations on the size of the text that the program can transcribe. Once the conversion from orthographic to phonetic text is completed, a search for any phoneme or combination of phonemes can be done using the "Query" function. The user drags one or more phoneme from the "symbols" palette to the query box, indicates in which position in a word string this sound combination should occur, e.g. at the start of the word, in the middle, anywhere etc., and clicks on the "Run Query" button. The results will show the number of words in the text, the number of words searched, the number of occurrences of the specified sound combination (in the specified position), as well as the words in which these matches were found. Figure 1 shows an example of such a search done in `Patana2000`.

As shown in Figure 2, the user may also switch to the "diphone" or "tri-phone" view, in which a list of all diphones / triphones that appear in the text is given with each diphone / triphone's corresponding occurrence frequency results.

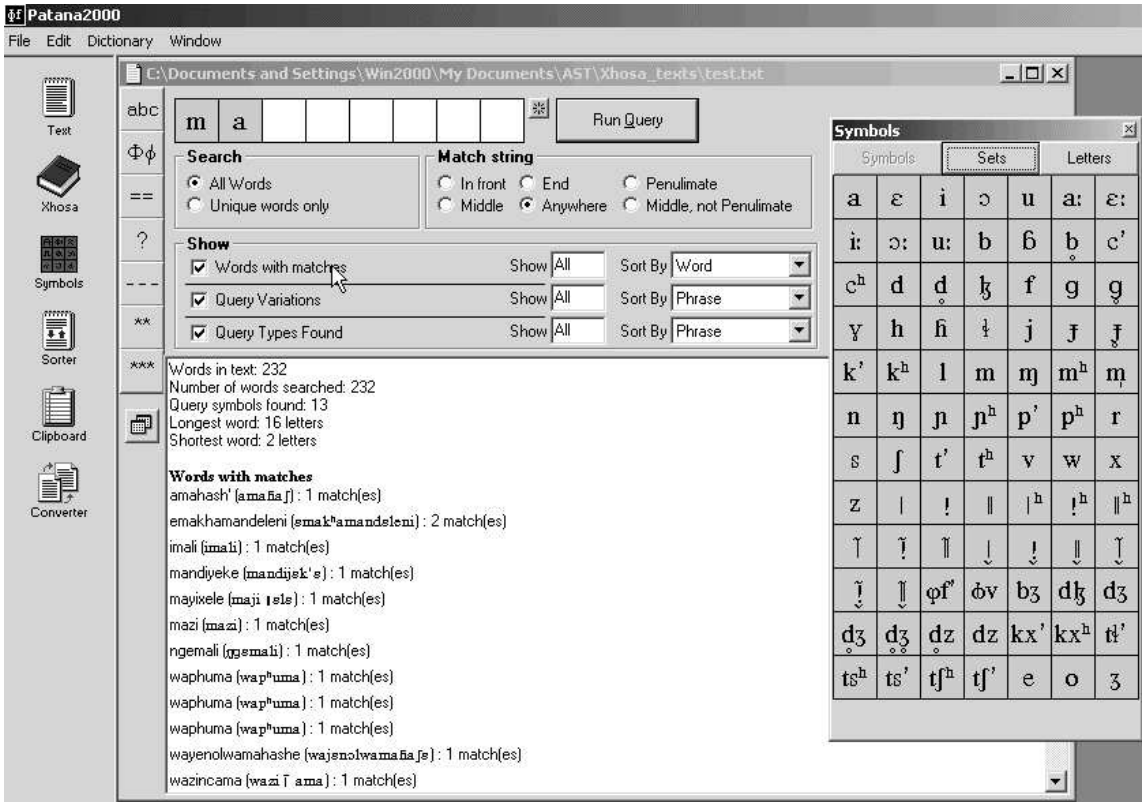

Figure 1. Doing a search on the sound combination /ma/ in `Patana2000`

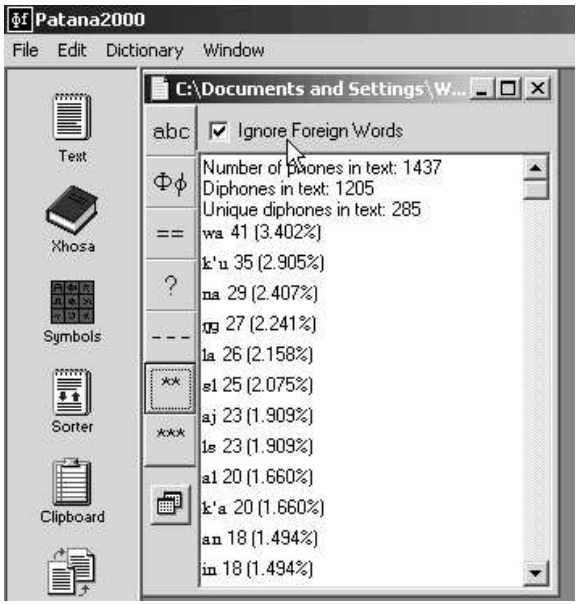

Figure 2. Results of a diphone count in `Patana2000`

3.1. Diphone frequency

To ensure that the phonetically rich sentences used in the databases cover the most frequently used diphones in a language, the following procedure was used. Relatively large texts (±55 000 words for Xhosa) were analysed to determine which diphones appear in the language. The diphones were identified and counted using `Patana2000`. Once all the diphones in the text were identified the frequency of occurrence of each diphone within the text was determined. A graph of the results was drawn and a cut-off-point was determined to identify the number of most frequent diphones that would explicitly be represented within the corpus of phonetically rich sentences. Once the list of sentences was compiled, the sentences were transcribed phonetically in `Patana2000`. All the phonemes and diphones in this list of sentences were then identified and counted. A cross-check was then done to ensure that all phonemes that appear in the language, as well as all diphones do indeed appear in the list of sentences (as determined in the previous analysis of the large text).

The list of English phonetically rich sentences was compiled somewhat differently. A subset of 800 sentences was extracted from the SX (phonetically-compact) and SI (phonetically-diverse) sentences in the TIMIT database [6]. The sentences were extracted in such a way that the representation of diphones in the extracted corpus was similar to the representation of diphones in the TIMIT SX sentences.

4. Conclusions

The design of the corpus of the African Speech Technology Telephone Speech Database was influenced by the multi-lingual character of the South African population. The project is currently developing orthographically and phonetically transcribed speech databases for five of the country's eleven official languages, taking into account the speech varieties of mother-tongue and non-mother-tongue speakers. The five languages are South African English, Afrikaans, Xhosa, Zulu and Southern Sotho. Since certain essential phonetic information on these languages (e.g. phone and diphone frequencies) was not available, software was developed to extract this from large texts. The recruiting of speakers for three of the databases (Xhosa, Afrikaans Mother-tongue English and Black English) have been completed and the in-coming phone calls are currently being screened and transcribed by trained transcribers.

Over and above the application to which these databases are to be put, new cross-linguistic phonetic evidence may come to the fore that might provide answers to the question whether the Black English variety within South African English should be regarded as a monolithic whole or as a variety comprising Xhosa English, Zulu English and Southern Sotho English.

5. Acknowledgement

This project is financially supported by the Innovation Fund (project number 21213) of the Department of Arts, Culture, Science and Technology (DACST) of the South African National Government.

6. References

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

[2] Draxler, C., Van den Heuvel, H., and Tropf, H. "SpeechDat experiences in creating large multilingual speech databases for teleservices." *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain. 1:361-366. 1998.

[3] Winski, R. Definition of corpus, scripts and standards for Fixed Networks. *SpeechDat Technical Report* SD1.1.1, 1997.

[4] Louw, P.H. et al., *Corpus Design, Speaker Recruitment Documentation and Standard for Telephone Speech Database.* AST Technical Report 05, 2001. www.ast.sun.ac.za/ast/verslae/T Reports/

[5] Du Plessis, T. South Africa: from two to eleven official languages. In Deprez, K. & Du Plessis, T. (Eds) *Multilingualism and Government.* Pretoria: Van Schaik Publishers, 95-110, 2000.

[6] The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT). http://www.ldc.upenn.edu/readme_files/timit.readme.html