LLSTI isiZulu TTS Project Report

M.Davel and E.Barnard

November 2004

1 Introduction

This document provides an overview of the development of an isiZulu text-to-speech (TTS) system at the CSIR, South Africa in 2003/4. The development consisted of two phases. Initially, a fairly primitive synthesizer was developed using a limited set of sentences and early versions of components such as the letter-to-sound converter were used. This phase was primarily aimed at familiarization with the Festival toolkit, and the exploration of various language-specific issues; it is described in [Davel04]. During the second phase, a more sophisticated system was developed, and its usability was evaluated with speakers from a variety of social backgrounds.

The current report details work done during the second phase, and conclusions reached.

2 Main activities

2.1 Support for user-friendly development within Festival

A number of tasks related to general Festival development and porting to Flite were undertaken – for example, guidelines on pitch marking in Festival were developed, the MultiSyn synthesizer was integrated with Flite, and intonation models were developed. These activities are described in [Louw04summary].

2.2 Task domain and text corpus collection

At project initiation, there was not a clear indication of the choice of pilot. Various options were considered, including:

- As part of the CSIR Disability Project: TTS for assistive devices.
- As part of the CSIR E-gov project: TTS for telephony-based access to government info.
- As part of the Digital Doorway project: TTS included for the sole purpose of testing user acceptability.

None of these projects was at a stage where a specific task domain or vocabulary was confirmed. To harmonize with activities of other LLSTI project participants, a “weather-related” task domain was therefore chosen – this is attractive since the dynamic nature of weather-based information demonstrates the ability of TTS (over the use of voice recordings.) The system developed is nevertheless a general-purpose text-to-speech system, with performance optimised for the specific task domain.

A previously prepared isiZulu text corpus was not available to project participants. In order to collect such a corpus, information was drawn from different sources. A general domain text corpus was collected from the Internet. (Current status: 30 000 words.) This corpus mainly consists of government-oriented documents, relating to domains such as health, tourism and governance. The documents were processed, and official approval was obtained where copyright restrictions were a concern. The corpus was validated by an isiZulu text validator,

A small weather-specific text corpus was developed based on:

- TAF (Terminal Airport Forecast) information: translated in isiZulu
- Televised weather reports: recorded and transcribed
- Manually developed texts

Prior to the release of the Hyderabad Optimal Text Selection Tool, a small tool was developed to choose phonetically balanced sentences. That has since been replaced. Scripts were developed to generate the OTS required format from general text. A subset of 70 sentences selected for v0.1, and 153 sentences for v0.2. After the completion of the v0.2 voice, an additional 27 sentences were added to compensate for missing diphones and frequently occurring English loan words. (See section 2.5)

2.3 Phone set definition and letter-to-sound rules

An initial phone set was defined, originally based on the phone set defined in standard texts on isiZulu. During the development of the v0.1 system, this was adapted to the current version of the phone set. Phone set characteristics were captured in the format required by Festvox.

The orthography of isiZulu is fairly phonetic, but not entirely so. Systematic letter-to-sound rules were not available and were developed as part of this project, using a dictionary creation tool developed in parallel to this project. The dictionary creation system was developed to allow a speaker fluent in a target language to develop a pronunciation dictionary when phonetic expertise is not available. Along with the pronunciation dictionary, a related set of grapheme-to-phoneme (G2P) rules is created automatically. The system applies a bootstrapping approach that attempts to simplify and minimise the human intervention required during the process. A word list and phoneme set for the target language are required as inputs to the system, after which the target language speaker is guided through the dictionary creation process. More information with regard to this tool is available, and if required, this tool can be made accessible to other project participants.

The letter-to-sound rules were used in three ways:

- Scripts were created to phonetisize plain text for pre-processing during text selection and automatic alignment of audio files.
- Scripts were created to generate Festival-format lexicons from the rule set and a given word list (used during voice building).
- The rules were converted to Festival-format for general domain synthesis.

2.4 Speaker selection and voice recordings

The variety and distribution of isiZulu dialects is fairly complex. Specifically, a “neutral isiZulu” does not seem to exist. The pure isiZulu spoken in kwaZulu-Natal is experienced as overly formal by Gauteng isiZulu-speakers, while, vice versa, Gauteng isiZulu can be experienced as a “slang” version. These are two extreme examples of a complicated landscape of dialects.

For the v0.1 recordings, a Gauteng-accented female isiZulu speaker was used. For the v0.2 recordings, a kwaZulu-Natal-accented male isiZulu speaker was used. Both voices are very pleasant, distinctive and clear.

Prior to recording, speakers were not asked to keep their intonation stable and various intonation patterns are included in the database. Speakers were asked to keep their prosody stable, but both speakers had great difficulty controlling volume and speed, especially at the end of a sentence. In normal spoken isiZulu, the last segment of a sentence is typically not distinct, faster and lower in volume.

Both sets of recordings were made using a head-mounted microphone and a normal PC. The first set was made in an office-environment, and the second set of recordings was collected in a specially prepared quiet room.

2.5 Speech segmentation and voice development

As in the first phase of the project, the built-in Festival voice was used to perform initial alignments of the isiZulu voice. Scripts were generated to replace the true isiZulu phones with their closest English counterparts, which in several instances required one isiZulu phone to map to a sequence of two or even three English phones. Prompts were loaded by Festival as phone strings, and the alignments provided a baseline for subsequent hand alignment. Two labellers were trained to produce alignments according to a set of conventions that were developed for this project; these labellers worked on different subsets of the data, since an automated method was used to verify the alignments.

The automated method [Barnard04] computed the average spectrum of each phone (by averaging over all aligned segments labelled as that phone), and then flagged those outliers that had the highest Mahalanobis distance from the mean spectrum. (The pooled variance across all samples was used in the distance calculation.) The flagged segments were manually checked – and corrected if necessary.

The voice development process is described in detail in [Louw04build].

Subsequent to v0.2 voice development, the diphone coverage of the speech corpus was re-evaluated. Changes in the phone set, letter-to-sound rule set and the difference between target and realised pronunciations in the speech database, resulted in a number of missing diphones.

The v0.2 voice was evaluated without adding additional diphones, in order to obtain realistic feedback on the quality of the TTS system for general synthesis purposes. Two schemes for dealing with missing diphones were tested:

- Back-off phones were defined. A missing diphone triggers the next best diphone candidate based on a set of replacement candidates defined.

- Replacing a diphone with two monophones. This was tested manually and the newly created ‘pseudo-word’ added to the corpus. When closely matched, the effect was almost indiscernible. When not closely matched, the effect was clearly audible, but still better than backing off to a weak candidate diphone.

After evaluation, an additional set of 13 sentences were defined in order to cover all diphones occurring with a frequency of more than 3 in the general text corpus.

IsiZulu text often contains English words (such as numbers and dates). In future we would like to create both an English and isiZulu voice using the same voice artists, in order to best deal with this phenomenon. For the time being, we've added an additional 17 sentences containing frequently used numbers and dates as a very limited inventory of English phones.

2.6 Morphological Decomposition

There is much linguistic information to indicate that morphological decomposition (MD) is crucial for tasks such as part-of-speech (POS) determination and tone assignment. Significant effort was therefore devoted to develop an appropriate understanding of MD, and also a computational approach to automate this task. In this regard, we were initially assisted by Prof S Bosch and Prof L Pretorius from the University of South Africa. These researchers have been active in this area for several years, and have made substantial progress in developing MD for isiZulu [Pretorius02], using the finite-state tools from Xerox [Beesly03]. However, they were not able to continue their collaboration on the LLSTI project, and alternative avenues were subsequently explored. It turns out that the morphological structure of isiZulu, although complex, has relatively localized alternations. It is therefore possible to define a much simpler notation than the full two-level formalism to describe this morphology.

Using such a simplified specification language makes it much easier to develop the MD rules for a particular language, and members of our project team have developed preliminary rules for all the word classes except the verbs. However, this was not integrated into the final TTS system for several reasons:

- After analysis of the salient intonation patterns produced by an isiZulu speaker producing continuous sentences, it became clear that tone production in natural speech is much less regular than in a language such as Ibibio [Gibbon02]. In particular, pronounced tone is only produced on certain "marked" words (either to provide emphasis, or to distinguish between otherwise homophonous words). Accurate prediction of the words to mark in this way is beyond our current capabilities, and the consensus of both isiZulu speakers and system developers was that treating all words as "unmarked" would produce more understandable speech. (Unfortunately, this analysis was performed after the main recordings for the system had been made, and the speaker could therefore not be requested to produce explicitly monotonous utterances. Investigation of this approach is interesting material for further work – see below.)

- Analysis of the same recordings, as well as discussions with isiZulu linguists and other isiZulu speakers, indicated that the placement of short pauses between each pair of (conjunctively written) words would sound acceptably normal – thus, POS determination was not necessary for the purposes of chunking for synthesis.

- The availability of the main LLSTI tool for MD was delayed somewhat, and could therefore not be synchronized with our development.

A standalone version of our MD analyzer has nevertheless been completed, and its refinement and integration with the TTS system will undoubtedly be an important part of our future work, as we strive to perform more refined chunking – and eventually for more refined intonation.

2.7 Intonation and duration modelling

IsiZulu is considered to be a tone language (although there is some debate on the meaning of this classification when applied to the Nguni family of languages [Roux00]). We therefore expected that accurate production of appropriate pitch contours would be an important element of an understandable system. However, as described above, our measurements of natural whole-sentence recordings indicated that the pitch levels were much less regular than expected – nominally high tones would frequently be produced on the same pitch level as surrounding nominally low tones, and vice versa. This lead us to implement all tones as "unmarked".

Although the resulting system is therefore "tone deaf", this fact does not particularly bother isiZulu speakers – for example, one listener simply commented that "this speaker comes from a different region". We therefore believe that we can improve the quality of synthesis substantially by explicitly aiming for monotone recordings. Of course, the eventual aim is to produce "natural" tones, but that would require the development of an algorithm to determine to predict which words are to be marked (and, of course, to predict the appropriate intonation of those words). These tasks are currently under investigation.

Regarding duration, both linguistic knowledge and informal measurements point to the lengthening of the penultimate syllable of each word as the most salient effect. We have therefore not implemented an explicit duration model, but weigh the syllable position heavily in the calculation of the target costs during synthesis. Again, listeners find this to be an acceptable compromise.

2.8 Development of appropriate target-cost function

To select appropriate units during synthesis, a concatenative synthesizer combines a target cost and a join cost. The target cost is the sum of a user-definable set of weighted components, each of which adds a penalty cost if some feature of the candidate diphone does not match the target, or if some default penalty feature is set in a candidate (which can be used to penalise candidates with poor labelling or bad pitch marking). We have developed a number of components, which are appropriate for isiZulu and related languages.

The weightings for the current target costs have been derived empirically to provide a baseline acceptable performance, but these can easily be changed to values based on statistical training or perceptual evaluation, should data be available.

The costs that were found to have a significant influence on the quality of the unit selection were:

1. Stress patterns match:

   The stress of the parent syllable of the candidate unit is compared to the stress of the parent syllable of the target unit. A mismatch adds a penalty of weight 10.

2. Word syllable positions:

   Two syllable positions were found to have a particular influence on speech quality: the final syllable in a word and the penultimate syllable. Thus, a candidate unit's syllable position is compared to the target. If a mismatch occurs, a penalty of weight 8 is added.

3. Number of syllables in word:

   It was found that if a candidate unit is extracted from a word where the number of syllables differs significantly from that of the target word then a perceptual mismatch occurs. A weight of 3 is added if the number of syllables in the target word and candidate word differ by more than a factor of 1.5.

4. Left and right contexts:

   The left and right side contexts of the candidate and target units are compared. If a mismatch occurs a weight factor of 3 is added to the unit's target cost.

3 Pilot evaluation

A pilot evaluation was designed in order to determine the usability of the current version of the isiZulu TTS system. In particular, the LLSTI initiative is aimed at providing information to users from all walks of life, and we wanted to determine whether the TTS would be understandable to users with limited literacy and limited exposure to such technology.

The voice was found to be quite understandable by the majority of evaluators – both those with high levels of literacy and those with limited or no literacy. Since much of the information was outside the weather domain, which was the focus of this development, it is safe to consider this as a domain-independent result.

Clear room for improvement remains, both in terms of the characteristics of the TTS system, and in our evaluation thereof. The design of comprehension tests for users of limited literacy will require particular attention. Interestingly, the lack of attention to prosodic information did not attract any specific comments from the evaluators, suggesting that this may not be as important for understandable isiZulu TTS as had been believed.

The evaluation process is described in detail in [Davel04eval]

4 Other Activities

4.1 System Familiarisation

Time was invested in building competence in the use of Festival, Festvox and related speech manipulation tools. Two isiZulu speakers with no prior experience in speech processing or Text-to-Speech technology are currently gaining TTS-related skills in this process (with two additional isiZulu speakers being drawn in during Phase 2.)

4.2 Mysore workshop

Participation in Mysore workshop (August 2003)

4.3 Hosting of LLSTI Partners in SA

Hosting of LLSTI Partners in SA (February 2004)

4.4 Lisbon workshop

Participation in Lisbon workshop (February 2004)

5 Conclusion

A first-generation isiZulu synthesizer has been developed and evaluated within the Festival framework. Much remains to be done in order to improve the naturalness of the system, and some improvements in intelligibility are also envisaged. It is nevertheless clear that the current version is already quite usable. Plans are therefore being made to deploy it in various contexts – for example, as part of a Web reader for visually impaired users, and in a telephone-based information service where up-to-date information can be provided in textual or electronic format and spoken back with TTS.

Although progress has been made towards simplifying the use of Festival and Festvox, more work along these lines would also be of much value.

6 References

[Davel04][1] M. Davel and E. Barnard, “LLSTI progress report”, March 2004.

[Barnard04] E. Barnard and M Davel, “Automatic error detection in alignments for speech synthesis”, to be submitted.

[Roux00] J.C. Roux,” “Xhosa: A tone or pitch-accent language?” South African Journal of Linguistics, Supplement 36, 33- 50. 2000.

[Pretorius02] L..Pretorius and S. Bosch, “Finite-State Computational Morphology -Treatment of the Zulu Noun”, South African Computer Journal, Vo. 28 pp. 30 – 38, 2002.

[Beesley03] K.R Beesley and L. Karttunen, “Finite State Morphology”, Center for the Study of Language and Information, 2003.

[Gibbon02] D. Gibbon, “Typology of African Prosodic Systems.” Bielefeld: Bielefeld Occasional Papers in Typology 1. Ed. by Ulrike Gut & Dafydd Gibbon, 2002.

[Davel04eval][1] M. Davel, “LLSTI isiZulu TTS Evaluation Report“, Oct 2004.

[Louw04build] [1] J. A. Louw, “Building MultiSyn Voices”, Sept 2004.

[Louw04summary][1] J. A. Louw, “Tasks performed for the LLSTI Project”, Sept 2004.

[1] See www.llsti.org

Appendix A: Final Deliverables

The final project deliverables are listed below:

Documentation	Project report (this document)	isizulu_tts_final.doc
	Evaluation report	isizulu_tts_eval.doc
	Summary of tool-related tasks performed for the LLSTI project	tasks_summary.pdf
Technical documentation	Building INTSINT intonation models in the Festival Speech Synthesis System	INTSINT.pdf
	Building MultiSyn voices	multisyn_voices.pdf
	A framework for bootstrapping morphological decomposition	morph_analysis.pdf
	Flite experiences and recommendations	flite.pdf
	Adding new modules to Festival	adding_festival_modules.pdf
	A short guide to Pitch-marking in the Festival Speech Synthesis System and recommendations for improvements	pitch.pdf
Zulu voice v.1	Speech database: recordings	
	Speech database: annotation & segmentation	
	Phone set, letter-to-sound rules and isiZulu-specific cost optimisation	
	Integrated TTS system	
Textual resources	General isiZulu text	zulu_general.txt
	Weather-related isiZulu text	zulu_weather.txt
	Phonetically balanced sentences selected	zulu_selected.txt
	Diphone counts and other data used during Optimal Text Selection process	ots_data
Software	An incomplete MD module, in Festival format	EST_morph_incomplete.tar.gz
	INTSINT Festival module	INTSINT_Festival_v1.0.tar.gz
	INTSINT_Festvox_v1.0.tar.gz	INTSINT Festvox module
	INTSINT updates	INTSINT_v1.0_updates.tar.gz
	INTSINT standalone module	intsint_tools_v1.0.tar.gz
	INTSINT (MOMEL) standalone module	momel_tools_v1.0.tar.gz
	LLSTI MultiSyn Festival contribution	festival_multisyn.tar.gz
	LLSTI MultiSyn Festvox contribution	festvox_multisyn.tar.gz
	LLSTI MultiSyn Flite contribution	llsti_flite.tar.gz