A general-purpose IsiZulu Speech Synthesiser

J.A. Louw, M. Davel, E. Barnard
Human Language Technologies Research group,
Meraka Institute / University of Pretoria
August 2005

Abstract

A general-purpose isiZulu text-to-speech (TTS) system was developed, based on the "Multisyn" unit-selection approach supported by the *Festival* TTS toolkit. The development involved a number of challenges related to the interface between speech technology and linguistics – for example, choosing an appropriate set of phonetic units, producing reliable pronunciations, and developing appropriate cost functions for selecting and joining diphone units. We show how solutions were found for each of these challenges, and describe a number of other innovations (such as automated fault detection in manual alignments) that were introduced. Initial evaluations suggest that the synthesizer is usable by a wide spectrum of isiZulu speakers.

Introduction

Text-to-speech (TTS) systems are widely used to generate spoken utterances from text. Such systems have a long history - see (DuToit, 1999) for an overview – and two broad categories of approaches have been most successful. Early synthesizers generally employed *model-based approaches,* where the speech-production process is described with a parametric model, and the parameters of the model are varied in time to produce synthesized speech. Perhaps the most successful model of this nature is based on the work of Klatt (1987), in which the parameters are related to formant frequencies and amplitudes (and corresponding excitation sources). Work on these approaches continues to attract attention, and models employing parameters related to the positions of the articulators have been used with some success in recent years.

The bulk of the attention in both the academic and commercial arenas has, however, shifted to *concatenation-based approaches.* Here, one no longer attempts to derive an explicit model of speech production – instead, speech segments are excised from a corpus of recorded speech, and spliced together to produce synthesized utterances. Concatenation is generally believed to produce utterances that are both more natural and more understandable than model-based synthesis, at the cost of increased effort in recording, preparing, storing, and searching the corpus of recorded speech. Thus, the development of a concatenative system in a new language is a significant effort.

We provide an overview of the development of a general-purpose isiZulu TTS system. The development was aimed at understanding the challenges involved in developing TTS for an Nguni language, and 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. During the second phase, a more sophisticated system was developed, and its usability was evaluated with speakers from a variety of social backgrounds.

Speech Synthesis Engine

The Festival Speech Synthesis System (Black, 1997) was used as the synthesis engine in this work. In order to obtain state-of-the-art naturalness, a new unit selection method, known as *Multisyn* (Clark, 2004) was used to select the speech units to be concatenated. For Multisyn, a large text corpus is recorded (as is also done in the *Classification and Regression Tree* (CART) unit selection method (Hunt, 1996)), but instead of using phones as a basic unit of concatenation, diphones (that is, phone-length units which extend from the temporal centre of one phone to the centre of the adjacent phone) are used. This, together with the fact that all diphones are searched for possible *candidates,* potentionally occuring as strings of adjacent diphones in the target data, produces superior synthesis quality when compared to other concatenative synthesis methods.

The database consists of multiple instances of each diphone type whereas a diphone synthesiser consists of a database with just one instance of each diphone type. Thus, no prosodic modifications, which introduce deterioration in speech quality, are performed on the selected units. It is asumed that the selected unit will contain the correct prosody, based on their context.

The total cost of selecting a candidate diphone is the sum of the target and concatenation costs. Target costs are linguistic measures of how well the candidate diphone fits the target diphone, while the concatenation cost is the acoustic cost of concatenating a candidate diphone with other possible candidates in the target string.

The advantages of the Multisyn synthesiser over the more conventional Cluster unit selection method is that the target costs can be chosen to optimize the quality of the synthesized speech. These costs generally depend on the characteristics of the language being synthesized, and are dificult to train from data, especially if the database is relatively small (fewer than 500 utterances). Another advantage is the fact that all units are candidates for selection, whereas in the Cluster unit selection the search space for possible candidates is broken up into clusters of acoustically simular units having the same target costs; thus, the selected units may not be not optimal in that case.

The disadvantages of the Multisyn synthesiser are related to its strengths: a large search space must be searched, and target costs are calculated during synthesis (whereas with the Cluster unit selection method the target costs are inherently the clusters themselves, and the clustering process significantly reduces the extent of the search space). Multisyn synthesis may therefore be slower.

Task domain and text corpus collection

A “weather-related” task domain was chosen for our initial development – 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 for this development. In order to collect such a corpus, information was drawn from different sources. A general domain text corpus of 30 000 words was collected from the Internet. This corpus consists mainly 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, and extended with a number of customized weather-specific sources (such as transcriptions of televised weather reports, and manually developed texts).

An open-source text selection tool, *Optimal Text Selection* (OTS), (developed at Hyberabad, India), was used to choose phonetically balanced sentences. Scripts were developed to generate the OTS required format from general text. A subset of 153 sentences was thus selected for recording. After the completion of an early version of the system, an additional 27 sentences were added to compensate for missing diphones and frequently occurring English loan words.

Phone set definition and grapheme-to-phoneme rules

An initial phone set was defined, originally based on the phone set defined in standard texts on isiZulu. During the development of the system, this was adapted to a more refined version to suit the domain and the selected texts. Phone set characteristics were captured in the format required by Festvox.

The orthography of isiZulu is fairly phonetic, but not entirely so. Systematic grapheme-to-phoneme rules were not available and were developed as part of this project, using a dictionary creation tool developed in parallel to this project (Davel, 2004).

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 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. The grapheme-to-phoneme 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

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 recordings, a kwaZulu-Natal-accented male isiZulu speaker was used, with a very pleasant, distinctive and clear voice. Prior to recording, the voice artist was asked to keep his prosody stable, but had great difficulty controlling volume and speed, especially at the end of a sentence. In normal spoken isiZulu, the last segment of a phrase is typically reduced significantly (that is, shorter and lower in volume). Since the speaker found it highly unnatural to modify this pattern, we accepted such recordings and took special precautions to ensure that phrase-final syllables are not used elsewhere in synthesized utterances (see below).

The recordings were made using a head-mounted microphone and a normal PC. The recordings were collected in a specially prepared quiet room.

Speech segmentation and voice development

The built-in Festival English voice was used to perform initial phonetic time 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 labelers were trained to produce alignments according to a set of conventions that were developed for this project; these labelers worked on different subsets of the data, since an automated method was used to verify the alignments.

The automated method computes the average spectrum of each phone (by averaging over all aligned segments labeled as that phone), and then flags those outliers that have 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 (Louw, 2004). Subsequent to voice development, the diphone coverage of the speech corpus was re-evaluated. Changes in the phone set, grapheme-to-phoneme rule set and the difference between target and realised pronunciations in the speech database, resulted in a number of missing diphones.

The 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 replacement diphone.

After evaluation, an additional set of 13 sentences were defined in order to cover all diphones missing from our initial set, but 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 have added an additional 17 sentences containing frequently used numbers and dates as a very limited inventory of English phones.

Intonation and duration modeling

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 (Roux, 2000)). We therefore expected that accurate production of appropriate pitch contours would be an important element of an understandable system.

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 (Gibbon, 2002). 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.

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.

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.

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.

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 linguistic 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 labeling 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 (see above) 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.

Pilot evaluation

A pilot evaluation was designed in order to determine the usability of the current version of the isiZulu TTS system. In particular, we wanted to determine whether the TTS would be understandable to users with limited literacy and limited exposure to such technology.

One part of the pilot evaluation was designed to assess the utility of our TTS system for over-the-telephone use. Information on three subjects (a weather forecast, the disease malaria, and unemployment insurance) was scripted in English, translated into isiZulu, and synthesized with our TTS. This information was embedded in an Interactive Voice Response (IVR) application, which provided initial instructions, and then prompted the evaluators to listen to each of the information clips. All prompts are in isiZulu; whereas the synthesized voice is male, a female voice was recorded for the rest of the IVR application.

A questionnaire was prepared to (a) obtain biographic information of the evaluators (e.g. their age, home language, level of education); (b) test the evaluators' understanding of the matter presented in this way; and (c) query the evaluators' subjective experience of the synthesized speech. Evaluators were asked to provide their biographic information before interacting with the IVR application; understanding was tested after each information clip had been presented, and the questions on their subjective experiences were completed after termination of the IVR application. All responses were recorded on paper – where the evaluators were not sufficiently literate to perform this task, an experimenter asked them the questions verbally, and filled out the questionnaire on their behalf.

Finally, a Web-based evaluation was also developed in order to gain access to a wider group of evaluators (although this part of the evaluation would then be limited to literate, technologically sophisticated users). The same basic protocol was followed as for the telephone-based evaluation, but users now recorded their responses directly in the Web application, and their responses were logged on the Web server.

Evaluators were canvassed in various ways, ranging from personal contacts to a company-wide e-mail solicitation. Evaluators completed the process without assistance where possible, but in cases where their level of literacy did not allow them to read the instructions and complete the questionnaire, an experimenter assisted them by reading from the questionnaire and entering their answers on the paper form. Results from the telephone-based and Web-based trials were entered into a spreadsheet, and processed to assess the performance of the TTS system.

Results

The telephony system was evaluated by a test user group consisting of twenty-three people. 40% of the test group (E1) had a low literacy level. These evaluators either indicated that they could not read or write, or had significant difficulty reading the evaluation form. Theses users were typically employed as labourers, gardeners or laundry-workers. The remaining 60% of the test group (E0) were literate, and in certain cases well educated. Occupations for this second group ranged from domestic worker and security guard to software developer and human resources practitioner. Sixteen of the telephony evaluators indicated that they spoke isiZulu as a "home language" – the remainder of the test group consisted of fluent isiZulu speakers with home languages that included Setswana, isiNdebele, SiSwati, Xitsonga and isiXhosa. A third test group (E2) evaluated the system via the Internet application.

E0 Evaluation

The following was noted during the E0 evaluation (literate evaluators):

- 57% of evaluators indicated that they found the synthesised voice "fairly easy" to "very easy" to understand. While the remainder of the test group found it "fairly hard" to "very hard", no one indicated that the system was "impossible" to understand.

- Surprisingly, 50% of the evaluators found the synthesized voice "as natural as" or only "slightly less" natural than the recorded prompts used to provide the instructions. The remainder felt that the voice was either "a fair amount" or "a lot" more unnatural with only one evaluator indicating that the synthesised voice does not "sound human at all".

- A strong correlation was observed between people's ability to answer the test questions correctly, and the "intelligibility score" they gave the system, as shown in Table 1.

Table 1: Perception of intelligibility according to number of questions answered correctly for E0 evaluators

Number of correct answers (of 3)	**Own intelligibility score (of 5)**
3	4.4
2	3.3
1	2.7

- Five of the 13 evaluators (36%) answered all three questions correctly. The average number of correct answers over all evaluators was 2.1 (of 3).

- Home language isiZulu-speakers tended to answer the evaluation questions more accurately than other isiZulu-speakers (with an average of 2.3 correct answers compared to 1.8)

E1 Evaluation

The E1 evaluation (illiterate evaluators) produced results that differed from the above in interesting ways. During observation and through after-evaluation discussions, it became clear that the test procedure was not sufficient in two ways:

- Evaluators found it difficult to differentiate between information obtained from the system, and information they knew beforehand. For example, the weather question provided a fictional weather report and on account of the information provided, asked what the weather is "expected to be today". A number of the evaluators provided the actual information, not the fictional information[1].

- Evaluators found it difficult to choose a negative answer – even where they knew the negative answer to be correct. For example, evaluator E106 answered the malaria question incorrectly, as he understood contact with human blood to be the way in which the mosquito contracted malaria – and as that is "bad", chose the other answer. In the same vein, the employment related question asked whether "only an employer" should contribute to the Unemployment Insurance Fund. Even evaluators, who in subsequent questioning indicated that they understood that both the employer and employee should pay, still chose "yes"

[1] The fact that 8 of the 9 E1 evaluators chose "sunny" rather than "cloudy" (the correct answer)

> because the employer should contribute, rather than “no” (the correct answer) since the double negative required to answer the question correctly was unnatural to them.

This demonstrates the complexity involved when using a comprehension test with low literacy evaluators: an area in which significant further research is required.

The subjective opinions of E1 evaluators were similar to those of E0 evaluators with regard to intelligibility (3.8 average compared to 3.6) and slightly lower with regard to naturalness (2.8 average compared to 3.2). While the objective evaluation (which does not compensate for the issues mentioned above) indicates a low level of intelligibility (1.4 average), the results of the subjective evaluation are very encouraging.

E2 Evaluation

The Internet evaluators (E2) produced similar trends as observed during the E0 evaluation. The Internet evaluation process has not drawn many responses, and the evaluation is being continued.

Conclusion

Initial experiments with the isiZulu test system have produced encouraging results. 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. (In fact, comprehension was best in the non-weather domains, for reasons described above.)

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

testifies of prevalent weather conditions in Gauteng during late winter.

believed. The lessons learnt will be applied in refining our experimental approach, and in the continued development of the isiZulu TTS system.

References

Barnard, E. & Davel, M. 2004. 'Automatic error detection in alignments for speech synthesis', LLSTI isiZulu TTS Project Report. (available at www.llsti.org)

Clark, R. A. J, Richmond, K. & King, S. 2004 'Festival 2: build your own general purpose unit selection speech synthesiser'. in 5th ISCA Speech Synthesis Workshop, Pittsburgh, PA:173-178.

DuToit, T. 1999 *An Introduction to Text-to-Speech Synthesis*, Berlin:Springer

Gibbon, D. 2002. 'Typology of African Prosodic Systems'. *Occasional Papers In Typology 1*. Ed. by Ulrike Gut & Dafydd Gibbon. Bielefeld: Bielefeld

Hunt, A. & Black, A. 1996. 'Unit selection in a concatenative speech synthesis system using a large speech database', in Proceedings of ICASSP, 1:373-376. Atlanta:Georgia

Klatt, D. 1987. 'Review of text-to-speech conversion for English' *J. Acous. Soc. Amer*. 82: 737-793

Louw, J.A. 2004. 'Building MultiSyn Voices', LLSTI isiZulu TTS Project Report (available at www.llsti.org).

Roux, J.C. 2000. 'Xhosa: A tone or pitch-accent language? ', *South African Journal of Linguistics*, Supplement 36: 33- 50

Taylor,P. ,Black, A & Caley, R. 1998 'The architecture of the Festival speech synthesis system',. in Proc.The Third ESCA Workshop in Speech Synthesis: 147-151.