Development of Prototype Text-to-Speech Systems for Northern Sotho

Oosthuizen, H. J.; Phihlela, S. T.; Manamela, M. J. D.

Department of Computer Science
University of Limpopo, Turfloop, South Africa
oosthuizenhj@ul.ac.za

10.21437/Interspeech.2006-394

ABSTRACT

Two text-to-speech synthesis systems were developed for one of the eleven official languages of South Africa, *viz.* Northern Sotho. A diphone synthesis system, based on extraction of diphones from nonsense words, was constructed. A cluster unit selection synthesis system, based on recordings of sentences containing a selection of most common words in Northern Sotho, was also built. The Festival speech synthesis system was used for both systems. Both of these systems performed well in a subjective evaluation.
Index Terms: text-to-speech systems, diphone, unit selection.

1. INTRODUCTION

Prior to 1994, two official languages were recognized in the Republic of South Africa, *viz.* English and Afrikaans. A multitude of indigenous languages other than Afrikaans are spoken by the population. With the advent of the democratic government in 1994, another nine indigenous languages were recognized as official. Northern Sotho is one of these, and is the mother tongue of mainly black people living in the Northern part of South Africa, i.e. in the Limpopo province.

According to the census data of 2001, 4 208 980 people has Northern Sotho as their first home language out of a population of 44 819 778 in South Africa – roughly 9.4% [1]. Many of the rural speakers of this language, especially the older generation, are fairly illiterate with respect to English, and also technologically disabled with respect to computers. With the current drive to bridge the digital divide by South Africa's government, as well as their implementation of so-called e-government, these people are excluded. By constructing text-to-speech systems, as well as speech recognition systems for the indigenous languages, these people can be empowered fully in the above regard.

In what follows, the development of two prototype text-to-speech systems for Northern Sotho, based on the Festival Unit Selection Speech Synthesizer [2], will be described.

2. PHONE SET OF NORTHERN SOTHO

Although quite a number of dialects of Northern Sotho exist, the main dialect, Sepedi, is regarded as standard [3]. This is the dialect used for the system described here.

2.1 Phone set definition

The phone set for Northern Sotho [4] comprises the five vowels:

[a, e, i, o, u]

As well as 38 consonants:

[b, bj, d, f, fs, fš, g, h, hl, j, k, kg, kh, l, m, my, n, ng, ny, p, ph, ps, pš, psh, pšh, r, s, š, t, th, tl, tlh, ts, tš, tsh, tšh, w, y]

An additional phone, "pau", representing silence, was included with the above.

Since the letter x does not appear in Northern Sotho, this has been used in the development of the system to depict the š, because of the difficulty in using the latter symbol from the keyboard. It is pronounced like the *sh* in the English word *sh*ell.

3. DIPHONE SYNTHESIS SYSTEM

In setting up the basic directories and template files needed in the diphone synthesis system for Festival, all the phones in Northern Sotho were defined, adding feature values for each, like vowel length, vowel frontness, consonant type and places of articulation. This is needed to show how each phone is to be pronounced.

3.1 Diphone schema and generating code

As indicated in 2.1 above, Northern Sotho has 43 phones, and thus the number of phone-to-phone transitions is the square, i.e. 1849. This, however, could be reduced to 464 diphones, since most of the diphones do not occur in the language. The pruning mainly occurred for the consonant-consonant pairs, for example, "psh-r" and "t-kg" is not found in Northern Sotho. It was also necessary to include diphones where the first part is silence, as well as the case where the second part consists of silence. The classes of carriers used to construct nonsense words containing the diphones, were consonant-vowel-consonant, consonant-consonant, consonant-vowel, vowel-consonant, vowel-vowel, vowel-silence, silence-vowel, consonant-silence and silence-consonant. Examples of such nonsense words are

"pau t a t a t a pau" for the diphones "t-a" and "a-t"
"pau t a t a y a pau" for "a-y"

"pau t a m a t a pau" for "m-a"

It was also necessary to include consonant clusters to distinguish the phone-phone transitions like "ng-w". Note that "ng-w" when used in the string "Lebogang Waleng", will be pronounced differently from when it is used in the word "ngwana". Commands are available in Festival to create a complete diphone schema and code for generating nonsense words. The code contains the skeletons of the carrier words that encompass the diphones. During synthesis, the diphones are extracted from the middle of the carrier word.

3.2 Recording

After the suitable prompts were generated – these are necessary for guidance during recording, as well as for the necessary labeling thereof – the nonsense words were recorded. Since this is a prototype, a professional speaker was not employed, but one of us, Phihlela, did the recording. A Sennheiser PC130 noise-canceling microphone was used, and the recording was done on a laptop running off its battery. The recording was done as 16-bit PCM, the Festival standard [5]. Since it was not possible to do the recording in one sitting, the recordings were done on successive days at the same time to ensure consistency.

3.3 Labeling and Alignment

Automatic labeling of the recorded nonsense words was attempted. This did not, however, yield satisfactory results. Hand labeling was necessary. Wavesurfer 1.8.3 (an open source speech manipulation program developed at the Centre for Speech Technology at KTH in Stockholm, Sweden [6]) was used for transcription. The alignments of the labels were found to be crucial for the correct functioning of the system.

3.4 Pitch mark extraction

To effect a proper join of two segments, it is necessary to join the two segments at the same position of the glottal cycle for both segments. The position of each instance of glottal closure, also referred to as the pitch pulse, should thus be marked [7]. An electroglottograph signal was not available to ascertain the correct places for the pitch marks in the recordings. These were extracted from the recorded data using commands available in the Festival system, adjusting the parameters carefully to obtain the required results. The extracted pitch marks were always shifted to the nearest peak in the waveform after extraction.

Since the open distribution of Festival use the residual excited Linear-Predictive Coding (LPC), these coefficients and LPC residual files were subsequently created for each nonsense word. These are also referred to as the pitch-synchronous LPC coefficients.

3.5 Lexicon and Letter to Sound Rules

Since it is a mammoth task to include all the words with their pronunciation in the lexicon, letter-to-sound rules is required to cater for out-of-vocabulary rules. The system will look up the pronunciation of a word in the lexicon; if it is not found, it will invoke the letter-to-sound rules for the pronunciation. The entries in the lexicon comprises of the word, part of speech, and the pronunciation.

Scrutiny of the phones in Northern Sotho reveals the fact that some phones are made up of two or more phones. For example, "tsh" is made up of the phones "t", "s" and "h" – for correct pronunciation, these composite phones are defined first in the letter-to-sound rules.

3.6 Tokenization

Since written text may comprise of many characters such as acronyms, abbreviations, numbers, etc., any text-to-speech system should be able to take care of this. A function was thus added to take care of these tokens. A call to this function is executed when a token is encountered in the text, which will then return a word, or word list, representing this token. As an example, if the token *R110.25* is encountered, the word list *one hundred and ten Rand and twenty five cents* would be returned (the example is in English for clarity).

3.7 Prosody

The default prosodic models were used in the system. The prediction of the phrase breaks is done through a CART tree applied to each word to determine whether a break needs to be added. A break will be inserted where any of the punctuation marks

[' / " ? . , : ;]

are encountered in the text.

Prediction of phone duration is done through a CART tree, which predicts zscores for phones, zscores being the number of standard deviations from the mean. This duration model uses the predicted zscores with information stored in a relevant file to determine the absolute durations for the phones.

4. CLUSTER UNIT SELECTION SYNTHESIS SYSTEM

Another way in which to build a text-to-speech synthesizer is to use a database of real speech, and use a clustering method to extract the speech segments needed. It is conjectured that this method should produce better, more natural sounding speech than the diphone synthesis system [8].

4.1 Brief background

This description is based on [8]. This technique basically uses a database of general speech. Each phone type is then clustered into groups of acoustically similar units. This is based on phonetic context, prosodic features, stressing, word position and accents. The acoustic distance between clusters is needed. This may typically be based on some kind of cepstrum and fundamental frequency, or pitch synchronous analyses.

CART trees are then constructed so that at synthesis time, the correct cluster of candidate phones may be selected without incurring unnecessary complex and expensive calculations. It

is necessary to decide on the unit type that will be used, be it demi-syllable, diphone, or phone.

Black and Lenzo [8] estimates that a rough guide for constructing a reasonable system is in the vicinity of 460 phonetically balanced sentences. This method is still open to some experimentation for choosing the representations giving the best results.

4.2 Recording and labeling the database

The Festival speech synthesis system contains the necessary tools for the unit cluster selection system. A prompt file was set up containing 400 sentences in Northern Sotho representing the most common words of the language. These sentences were then recorded. After recording, every sentence was examined for audibility and cleanliness of the voice. The latter refers to absence of unwanted noise, both environmental as well as microphone induced.

A script is available in the Festival speech synthesis system for automatic labeling of the database. It was found, however, that it failed to give reasonable labels. Hand labeling was then resorted to. This was a tedious and slow task – listening to each word in a sentence and then to each phone in a word and adjusting the labels accordingly. Only about ten sentences could be correctly labeled in a day.

4.3 Extracting pitch marks and building clusters

The same technique as in the diphone system was used for the extraction of pitch marks. Because it was assumed that the speaker was not speaking with a constant prosodic style when recording the database, the parameters in the default script was not modified. Pitch synchronous, Mel-frequency cepstral coefficients were subsequently determined.

The process of building the clusters for each unit type is automatic – the necessary procedures are available in the Festival speech synthesis system. This process took quite a while to complete, the time depending on the number of instances of each type.

After the clusters were built, the system could be used to synthesize words and / or sentences, based on the cluster unit selection system.

5. EVALUATION OF THE TWO SYSTEMS

5.1 Evaluation procedure

Subjective listening tests were adopted to evaluate the two TTS systems. The evaluation was primarily focused on intelligibility, the naturalness of the speech output and the front-end text processing. For evaluation, the Mean Opinion Score (MOS) [9] were utilized, using the usual 5-scale system.

The subjects were required to evaluate the systems by answering the following questions:

i. How difficult is it to understand the synthesized message the first time you listen to it?
(1. Cannot grasp the message; 2. Vast amount of effort required; 3. Fair amount of effort required; 4. Little effort required; 5. Quite easy to understand)

ii. How well does the system handle non-standard words like abbreviations, numbers, etc.?
(1. Horrible; 2. Poor; 3. Acceptable; 4. Good; 5. Excellent)

iii. How natural does the speech output sound?
(1. completely unnatural; 2. A little bit unnatural; 3. Acceptable; 4. Natural enough to listen to; 5. Sounds like human voice)

iv. What is your overall impression of each system?
(1. Horrible; 2. Poor; 3. Acceptable; 4. Good; 5. Excellent)

In addition the subjects were asked to comment on the performance of the two systems, indicating what they perceive as being necessary to improve them.

To prevent the subjects from being biased to one or the other system, the order in which the two systems were invoked to synthesize the test sentences, were randomized – keeping careful track of the order in which they were utilized.

Eleven different Northern Sotho sentences were used in the evaluation. Some of the sentences contained non-standard words like money strings and numbers representing ranges, etc.

The subject pool consisted of 14 people, 5 females and 9 males, aged between 21 and 30, all of them being undergraduate students at the University of Limpopo. Although all of them understood Northern Sotho, they were not all Northern Sotho mother tongue speakers. Two subjects spoke Setswana, one Xitsonga and one Selobedu, the latter being a dialect of Northern Sotho.

5.2 Evaluation results

The results for each question were calculated in terms of the number of responses per score, and the averages calculated to determine the Mean Opinion Score for each system. The results are contained in table 1.

Question	Diphone Synthesis System	Cluster Unit Selection Synthesis System
Understandability	3.9	3.8
Token-to-words processing	4.2	3.5
Naturalness	3.5	4.0
Overall impression	4.0	3.9

Table 1: Experimental results of the evaluation

It is clear from the table that there were very slight differences in the rating of the two systems. The subjects did, however, prefer the cluster unit selection synthesis system.

The results obtained in this study differ from those found for two similar systems for Afrikaans, reported on in [10]. In this case, the diphone synthesis system was rated very poorly.

6. CONCLUSION

The results obtained in the construction of the two prototype text-to-speech synthesis systems for Northern Sotho indicates that it is quite feasible to build a full-fledged system that will perform well for this indigenous South African language. Some work needs to be done on the prosodic models for the diphone based system to overcome the monotony in the speech, and thus increase the naturalness. This system did well in the token-to-words processing. The speech data for the cluster unit system needs to be extended and a professional speaker used for the recording.

7. ACKNOWLEDGEMENTS

This work was done in the Telkom Centre of Excellence for Speech Technology at the University of Limpopo, South Africa. This Centre is sponsored by Telkom, Marpless, HP and the THRIP programme of the NRF. The authors wish to thank the sponsors for their generosity in this regard.

8. REFERENCES

[1] P. Lehohla. ``Stats in brief'', p:17-19, 2004.
[2] A. Black, P. Tayler and R.Caley. “The Festival speech synthesis system”, http://www.cstr.ed.ac.uk/projects/festival.html, 1998.
[3] Pan South African Language Board (PanSALB), http://www.pansalb.org.za
[4] C.J. Esterhuyse and P.S. Groenewald. “Die geskiedenis van die Sepediskryfwyse”, Journal for Language Teaching, Vol 33, No. 4, December 1999.
[5] A. W. Black, K. A. Lenzo. Building Synthetic Voices, “Recording under Unix”, http://www.festvox.org/bsv/bsv-recording-sect.html
[6] “KTH Speech, Music and Hearing”, http://www.speech.kth.se/wavesurfer
[7] J. Holmes & W. Holmes. “Speech synthesis and Recognition”, 2nd Ed, Taylor & Francis, New York, 2001.
[8] A.W. Black and K.A. Lenzo. Building Synthetic Voices, “Chapter 12. Unit selection databases”, http://www.festvox.org/bsv/bsv-unitsel-ch.html
[9] M. Viswanathan and M. Viswanathan. “Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale”, Computer Speech & Language, Vol 19(1), pp. 55-85.
[10] F. Rosseau and J. Mashao. “A hybrid Text-to-speech Approach for an Advanced Afrikaans System”, Proceedings of PRASA 2005, 23-25 November 2005, Langebaan, South Africa, 2005.
[11] S.T. Phihlela. “Text to speech synthesis in Northern Sotho”, unpublished M.Sc. dissertation, University of Limpopo, South Africa, 2006.