The Development of a Text-To-Speech System for Tshivenda Language

Tsumbedzo Mukange*, MJD Manamela
Department of Computer Science
University of Limpopo (Turfloop Campus),
Private Bag X1106,
Sovenga, 0727
Tel: +27 15 2682751, Fax: +27 15 2683183
Email: tsumbedzo.88@gmail.com

Abstract- This paper introduces the development process of a Tshivenda text-to-speech (TTS) baseline synthesis system as part of a speech technology research project for major African languages used in the Limpopo Province. It highlights the importance of TTS in human-computer interfacing, the brief background of TTS systems and the description of the concatenative approach to the development of the system. The baseline TTS Synthesis system will find direct application as assistive technology for visually impaired and illiterate end users of the computer technology.

1. Introduction

Speech is one of the most fundamental and natural forms of communication among people. As it is a primary mode of communication among human beings, it is reasonable for a human being to expect the use of speech in communicating with modern computational devices.

A TTS synthesis is the automatic generation of a speech signal by a computer. It can be viewed as a process of developing a computer system that converts written text into spoken words. The main goal of speech technology is to have machines that can speak, read, understand, or even carry out dialogs like human beings.

Most languages, especially in the Asian and European sides, have already developed baseline text-to-speech synthesis systems. In South Africa there are TTS systems in few African languages and having eleven official languages makes it imperative to address the great need to cover all official languages in TTS systems production.

The development of a TTS speech synthesis for Tshivenda will enable Tshivenda language speakers to access e-service related online electronic information easily. Currently there is no Tshivenda TTS system and through this research project we want to close the gap in the shortage of TTS systems for indigenous languages of South Africa. We intend to develop a general-purpose Tshivenda TTS synthesis system based on the concatenative synthesis approach using diphones in the Festival Toolkit [3]. The system will be able to read out any Tshivenda typed or stored text in the most intelligible and natural manner.

As one of the South African official languages, Tshivenda is spoken in Vhembe District of the Limpopo province. Tshivenda is also used in other parts of South Africa such as Mpumalanga and Gauteng, and in neighbouring counties such as Zimbabwe. 2.3% of South Africa population use Tshivenda as their home language according to the 2001 census.

2. Application of TTS Systems

The TTS synthesis system in general can be applied in telephony, data entry verification, reading and communication support for visually impaired. Deaf people may communicate with people who don't understand sign language and spelling and pronunciation teaching for several languages.

3. Text-to-speech system

The TTS systems convert written text to synthetic speech and are based on the human speech production process. The basic structure of a TTS system consists of three main components; text analysis, linguistic analysis and the speech synthesis

Most existing TTS systems were developed using one of the following synthesis approaches: concatenative synthesis, formant synthesis and articulatory synthesis.

- Concatenative synthesis concatenates pre-recorded speech units to produce natural sounding speech. It suffers distortion from discontinuities in concatenation points and it requires larger memory for implementation. The collection of training speech data and the labelling of speaker-dependant speech samples is time-consuming.
- Formant Synthesis uses acoustic models and varies parameters such as fundamental frequency, voicing and noise levels over time in order to create artificial speech.
- Articulatory synthesis imitates the human speech production system. It is a difficult method to implement and the computational work is also considerably higher than other methods.

Most TTS researchers and system developers adopt concatenative synthesis and formant synthesis. Concatenative synthesis is the most widely used method that produces utterances that are both more natural sounding understandable than formant synthesis [6].

4. Festival TTS Synthesis System

The Festival TTS synthesis system is a concatenative TTS system developed at the University of Edinburg [1]. It offers

a free, language independent, run-time speech synthesis engine for various platforms under various Application Program Interfaces (APIs). Festival offers a general framework for building speech synthesis system. It has a capability of dealing with symbols that have internal structures that requires special processing, e.g., money, time, etc. It deals with those using token-to-word rules. To prove its capability of producing high quality speech on local languages, it has been shown by IsiZulu TTS and Northern Sotho TTS systems [2-5].

5. Development Phase

We intend using a concatenative synthesis approach that**,** concatenates audio samples of natural speech form pre-recorded voice. This approach is likely to give natural sounding and intelligible synthesis speech [4]. Unit size is the most important aspect to consider when using this approach. Speech unit can be words, syllables, demi-syllables, phonemes and diphones. We are going to adopt diphones concatenative synthesis approach. Diphones are the most appropriate units to use for sample-based text-to-speech synthesis, since they reduce the amount of distortion at the concatenation points [2].

5.1. Data collection and labelling

In order to design this TTS system, the first thing we needed to construct is a diphones database. Pre-recorded speech data is going to be used and a Tshivenda speaking person, who may or may not be a voice artist is required to produce such data.
In concatenative speech synthesis, labeling is used to find out where a sequence of known spoken phones is in the signal. Festival Toolkit has the capability of allowing us to label.

5.2. Text Analysis

This phase is responsible for converting non-textual content into text and it consists of word identification, text normalization and tokenization. During text analysis process, words are identified from the text, abbreviations, numbers, time, date and acronyms are transformed into their corresponding full pronunciations and also homographs are resolved [4-7].

5.3. Linguistic analysis

This phase is responsible for finding the corresponding pronunciations of words and assigning prosodic features to the phonemic string to be spoken. It is the stage that depends on the structure of the language. Letter-to-sound rules and lexicon are used to guide the construction of a proper utterance [7].

5.4. Speech synthesis

Speech is generated during this phase, by using the phonetic and prosodic descriptions and it is guided by the synthesis approach adopted. Pitch and duration of phones are tuned in this process for natural sounding utterance [4].

6. Evaluation method

Mean Opinion Score (MOS) is mostly used for evaluating the quality of speech coding algorithms and synthesized speech. For the evaluation of the quality synthetic voice, we'll carry out formal MOS evaluation. MOS is obtained by asking several people to rate the system from bad to excellent using a scale range from 1(bad) to 5 (excellent). The TTS system will be evaluated for naturalness and intelligibility.

7. Conclusion

In this paper we described the TTS synthesis system to be developed, synthesis approach and evaluation methods to be used to ensure the quality of the system and some applications of TTS systems. The Festival TTS synthesis system was highlighted as the toolkit to be used for experimentations. Major aspects of TTS synthesis system has being described. The Tshivenda TTS synthesis system forms an essential component in the human language technology terrain using all official and indigenous languages of South Africa.

8. References

[1] A. W. Black, P Taylor and R Caley, "The Festival speech synthesis system", System Documentation, Edition 1.4 for Festival Version 1.4.0, January 1999. [Online]. Available: http://www.cstr.ed.ac.uk/projects/festival, 1999.

[2] J.A. Louw, M.Davel and E. Barbard. (2005) A general-purpose IsiZulu Speech Synthesizer. *South African Journal of African Language. 2,* pp1-6.

[3] L. Mohasi and D Mashao, "Improving Fluency in a Sesotho Text-to-Speech Hybrid System", MSc (Eng) Thesis, University of Cape Town, 2006.

[4] S T. Dutoit, "A Short Introduction to Text-to-Speech Synthesis". [Online]. Available: http://tcts.fpms.ac.be/synthesis/introtts-old.html

[5] S.T. Phihlela, "Text-to-Speech synthesis in Northern Sotho", unpublished MSc Thesis, University of Limpopo- Turfloop Campus, 2005.

[6] J W. Barkhoda, B. ZahirAzami, A. Bahrampour and O. Shahryari, "A comparison between Allophones, Syllable and Diphones Based TTS system for Kurdish Language", Kurdish Academy of Language. [Online]. Available: http://www.kurdishacademy.org/?q=node/730

[7] X. Huang, A Acero and H. Hon, "Spoken Language Processing", in *A Guide to Theory, Algorithm, and System Development*, Jane Bonnell, Ed. New Jersey: Prentice Hall PTR, 2001.

Tsumbedzo Mukange received a B.Sc degree from the University of Venda in 2006 and Honours degree from the University of Limpopo in 2009. He is studying towards an MSc at the same institution. His research interest is speech Technology.