SYNTHESIS OF DIALECT SPEECH FOR AN UNDER-RESOURCED LANGUAGE.

Ranta Langa*, Jonas Manamela and Nalson Gasela
Department of Computer Science
University of Limpopo, Turfloop Campus, Private Bag X1109, Sovenga, 0727
Tel: +27 15 268 3261, Fax: +27 15 268 3487
email: {ranta.langa, jonas.manamela, nalson.gasela}@ul.ac.za

Abstract- As part of the broader human technology (HTL) national research agenda, speech synthesis systems are developed to automatically convert text to speech. In this research paper, we propose to enhance an existing prototype general-purpose text-to-speech (TTS) synthesizer trained on a speech database of standard Northern Sotho sentences read by a young adult male voice. We want to enhance the prototype TTS synthesis system by adapting it to Lobedu and Tlokwa dialects of Northern Sotho using a Hidden Markov Model-based speech synthesizer (HTS). The plan is to use HTS-based parametric approaches to TTS synthesis system design to train a dialectal speech synthesizer which can later be adapted to the selected dialects of Northern Sotho. The resulting TTS synthesis systems will find regional applicability in the support of localized voice-enabled software applications that may target specific language empowerment rare and dwindling skills as numeracy in mother tongue education.

I. INTRODUCTION

The Human Language Technology (HLT) is actively involved with human-computer interface research projects. One such aspect of Spoken Language Processing (SLP) is speech synthesis (or commonly referred to as "Text-to-Speech" (TTS) synthesis). The Unit selection synthesis method is currently a popular speech synthesis technique and has shown to synthesize high quality speech. Although it is very hard to surpass the quality of the best examples of unit selection, it does have a limitation in that the synthesized speech will strongly resemble the style of the speech recorded in the database. As we require speech which is more varied in voice characteristics, and speaking styles, we need to record larger and larger databases with these variations to achieve the synthesis we desire without degrading the speech quality [1]. However, recording such a large database is very difficult and costly [2]. We need a more feasible approach to speech synthesis with diverse speaker's voices and styles. The Hidden Markov Model (HMM)-based speech synthesis is a statistical parametric model that extracts speech parameters from the speech database, trains them and produces the sound equivalent to the input text. This method has the advantage of being able to synthesize speech with various speaker characteristics, speaking styles, and still maintain some acceptable degree of naturalness. This method, however, makes it very easy to adapt new speakers and has very little memory requirements. HTS is the toolkit that is used to develop HMM-based synthesis systems.

Figure 1 shows a typical HMM-based speech synthesis system. It consists of training and synthesis parts. The training part is similar to that used in speech recognition systems. The main difference is that, both spectrum and excitation parameters are extracted from a speech database and modeled by context dependent HMMs (phonetic, linguistic, and prosodic contexts are taken into account). As a result, the system models spectrum, excitation, and durations in a unified HMM framework [4].

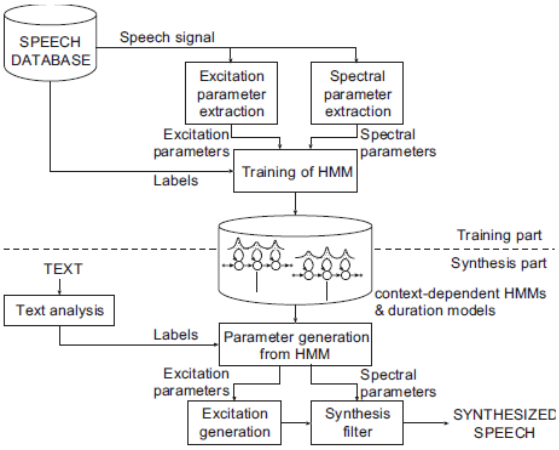

Figure 1: A typical HMM-based speech synthesis system (adapted from[3])

The synthesis part does the inverse operation of speech recognition. Firstly, a given text to be synthesized is converted to a context-dependent label sequence, and then an utterance HMM is constructed by concatenating the context-dependent HMMs according to the label sequence. Secondly, state durations of the utterance HMM are determined based on the state duration probability density function (PDFs). Thirdly, the speech parameter generation algorithm generates the sequence of spectral and excitation parameters that maximize their output probabilities. Finally, a speech waveform is synthesized directly from the generated spectral and excitation parameters using the corresponding speech synthesis filter. This system offers the ability to model different styles without requiring the recording of very large databases. The most attractive part of this system is that its voice characteristics, or speaking styles, can easily be modified by transforming HMM parameters using various techniques.

In the research project of Phihlela et al. [5], the goal was to use Festival as a toolkit to develop a general-purpose Northern Sotho speech synthesis system. The intentions of the present project are to gain knowledge into prosodic differences in major dialects of Northern Sotho and attempt

to model voices of these Northern Sotho dialects using HMM-based speech synthesis system (HTS). The Northern Sotho has various dialects; this research project has chosen only two dialects, which are Lobedu and Tlokwa, because these are the most prevalent in Greater Polokwane District of the Limpopo Province.

A basic cause of dialectal differentiation is linguistic change. A given set of varieties must meet certain minimum linguistic conditions, such as sociological conditions, where the speakers feel that they belong to that speech community. Dialectologists have confirmed this dichotomy by defining dialect as 'a variety of language, spoken in one part of a country (regional dialect) or by people belonging to a particular social class (social dialect), which is different in some words, grammar and pronunciation from other forms of the same language [6].

II. PHONOLOGICAL VARIATIONS

Phonology involves the way language sounds and the way words are pronounced in a particular language. Phonology is that discipline of linguistics, which limits itself to the manner in which human speech sounds function when sound patterns are created. Where there are variations in vocabulary and grammar, these are normally accompanied by phonological variations. They usually tend to replace the following Lobedu phonemes and morphemes with the standard Northern Sotho phonemes and morphemes, as in the following examples:

A. Phoneme example

The use of the standard ejected lateral *tl [tl']* for the Lobedu voiced interdental *d* as the following examples:

Sepedi	**Lobedu**	**English**
-tlotša	*dotša*	'smear'
-ntlo	*ndo*	'house'
-tlala	*dala*	'hunger'

B. Morpheme example

The class prefix *se –* is *khe* in *Lobedu*.

Sepedi	**Lobedu**	**English**
selepe	*khelebe*	'axe'
sefepi	*khefepi*	'whip'
selemo	*khelemo*	'summer

C. Lexical items

The following are some of the lexical items, which distinguish Tlokwa from other dialects. Most of these examples differ mostly in form but sometimes in form and meaning from their counterparts in the standard affricative *kg* as used by Batlokwa.

For example:

Sepedi	**Tlokwa**	**English**
kgoši	*khosi*	'chief'
kgoro	*khoro*	'court'
kgofa	*khofa*	'tick'

The use of glottal *h i*nstead of standard *g* as in the following examples:

Sepedi	**Tlokwa**	**English**
-gola	*hola*	'grow'
-gapa	*hapa*	'herd'
-gana	*hana*	'refuse'

III. RESEARCH DESIGN

An HMM-based speech synthesis method has been selected for use in this research project. Its flexibility in the ease of adaptability to speaker's voice characteristics and speaking styles, less memory requirements for the runtime engine and less speech data required for training the system influenced its choice. The focus of this research will be to produce better natural sounding speech for the TTS to be developed using the Hidden Markov Model approach. The HMM-based speech synthesis system (HTS) toolkit will be used for experimentation.

IV. CONCLUSION

A natural sounding and intelligent speech synthesizer which can generate Northern Sotho dialects of limited vocabulary is proposed. The HMM-based speech synthesis as statistical parametric model will be used. The result will be TTS synthesis systems of selected dialects of Northern Sotho.

V. ACKNOWLEDGEMENTS

This research project is facilitated by the University of Limpopo Center of Excellence for Speech Technology together with Telkom SA Limited and National Foundation Research as the main sponsors.

VI. REFERENCES

[1] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli, "A corpus-based approach to <AHEM/> expressive speech synthesis," in *Proc. ISCA SSW5*, 2004, pp 79-84.

[2] A.W. Black, "Unit selection and emotional speech," in *Proc of Eurospeech*, 2003, pp. 1649–1652.

[3] T. Masuku, "HMM-base speech synthesis and its applications", PhD thesis, Tokyo Institute of Technology, November 2002

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-basedspeech synthesis," in *Proc. Eurospeech*, 1999, pp. 2347–2350.

[5] S.T. Phihlela, HJ Oosthuizen, and MJD Manamela, "Text-to-Speech Synthesis in Northern Sotho", unpublished Masters dissertation, University of Limpopo, December 2005

[6] L. Allen, and M.D. Linn, Dialect and Language Variation., Academic Press., Harcourt Brace Kovanorich Publishers, London Montreal, 1986

Ranta Langa received his Bachelor of Science degree and his Bachelor of Science (Honours) degree in computer Science in 1999 and 2000 respectively from the University of Limpopo. He is presently studying towards his Master of Science degree at the same institution. His research interests include Speech Technology, Human Language Technology and Computer Auditing.