An HMM-based Text-to-Speech Synthesis System for Xitsonga

Ntsako Baloyi*, 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: {200522530, jonasm}@ul.ac.za, nbaloyi11@gmail.com

Abstract-This paper outlines the plan to build an HMM-based baseline speech synthesis system for the Xitsonga language. The system to be built should produce natural sounding synthetic speech given typed or stored text input. It should further be able to model some speaker characteristics and speaking styles. Using HMMs such a system can be built without requiring a very large speech corpus for training the system. This research project forms part of the broader speech technology project that aims to develop spoken language systems for human-machine interaction using the eleven official languages of South Africa.

Index Terms—HMM-based speech synthesis, text-to-speech

1. INTRODUCTION

This study focuses on building a general-purpose Xitsonga speech synthesis system that is flexible, intelligible and natural sounding. The system should be able to model some of the desirable speaker characteristics and speaking styles. Speech synthesis systems also referred to as text-to-speech (TTS) systems receive typed or stored text as input and produce the equivalent speech waveform as output as depicted in Fig 1.

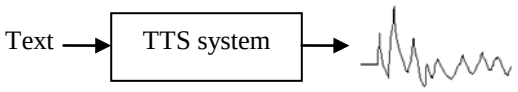

Fig1. A general outline of a TTS showing text as input, the TTS system, and speech waveform as output

The synthesis task will be carried out using an HMM-based speech synthesis (HTS) approach. This method makes it easy to accomplish our task without requiring a very large training speech corpus. The HTS approach requires much less speech data to train the system as compared to the concatenative-based synthesis approach. In addition to a smaller training corpus, HTS also requires very little memory for the synthesis engine at runtime. As a result, TTS systems based on this approach can easily be integrated into handheld devices [1].

Speech synthesis systems have over the years been developed for various languages all over the world. Most of those systems have been developed using unit selection methods which have proved to produce very natural sounding speech; though at the cost of very large speech training corpus. There is currently no TTS in Xitsonga at the University of Limpopo speech technology program, thus the system to be built will be the first of its kind for this language. The results of the 2001 census indicated that by that year Xitsonga was spoken by about 4.44% of South African citizens as their home language.

Speech synthesis systems have various areas of applications: hands and eyes-free computer interaction, email readers, translation systems, learning assistant systems for the handicapped, speaker verification systems etc. [1]

The remainder of the paper is outlined as follows: section 2 gives a brief overview of components that makeup our HMM-based speech synthesis system. Section 3 gives concluding remarks.

2. THE HMM SPEECH SYNTHESIS SYSTEM

2.1. Recording of speech database

A regular good-quality microphone will be used to record the speech corpus. The recording process will be executed in a specialized noise-free room. Speech by several speakers will be recorded and stored in the speech database, though there will be one main speaker. The recording process will occur over a number of days for all the speakers in order to capture variability in speakers' speaking characteristics. Phonetically balanced sentences in Xitsonga will be used for recording. These sentences should be from a wide variety of areas in order to make the system as much general-purpose as possible. Multiple instances of phoneme HMM utterances may be stored in the speech database in order to represent different contexts.

2.2. Training phase

Initially, spectrum (mel-cepstral coefficients) and excitation (log fundamental frequency or log F0) parameters are extracted from the speech database and their dynamic features (delta and delta-delta coefficients) are calculated [1, 2]. The extracted parameters model speaker characteristics and speaking styles and they are used to train (or model) the context-dependent phoneme HMMs [3]. Spectrum parameters are modeled by multivariate Gaussian distributions, whereas excitation parameters are modeled by multi-space probability distribution hidden Markov models (MSD-HMMs) [1].

A decision-based clustering technique which uses the Minimum Distance Length (MDL) criterion is applied separately to distributions of mel-cepstral, log F0 and state durations of the context-dependent phoneme HMMs [2, 3]. This technique ties contextual factors (i.e. phoneme identity, stress-related and locational contexts) that are almost similar. This is done because it is both impractical and impossible to prepare a speech database that can model all combinations of contextual factors. A re-estimation of the clustered context-dependent phoneme sequence will then be performed using the expectation-maximization (EM) algorithm [3]. Clustering is also used to generate excitation and spectrum parameters for newly observed vectors, i.e. observation vectors not included in the training corpus [4].

State durations are modeled by context-dependent n-dimensional Gaussian distributions which are then clustered by a decision tree. State densities capture/model the temporal structure of speech [2, 5]. Mel-cepstral coefficients, log F0 and state durations will be modeled simultaneously in a unified framework of HMM [2, 5].

2.3. Adaptation phase

A speaker adaptation technique is used to adapt a target speaker using the trained HMM from the training phase. Tikashi Masuko in [1] indicates that the adaptation technique requires only a small amount of speaker adaptation data from the target speaker. The adaptation process may be done using an individual speakers' speech data or by averaging several speakers' speech data [1]. Speaker voice characteristics, styles or even emotions can be modified/updated by transforming HMM parameters using adaptation or other methods (such as interpolation and eigenvoices) [3].

2.4. Synthesis phase

An arbitrary text to be synthesized will be input, it will then be transformed into a context-dependent phoneme label sequence. A sentence HMM should then be generated by concatenating the adapted context-dependent phoneme HMMs from the adaptation phase according to the context-dependent phoneme label sequence. State duration distributions are then used to determine state durations of the label sequence [3]. A speech parameter generation algorithm is then used to generate spectrum and excitation parameters from the context-dependent phoneme HMMs. The Mel Log Spectrum Approximation (MLSA) filter is used to synthesize a speech waveform from both spectral and excitation parameters [2, 3].

2.5. Evaluation of the HTS

A black-box evaluation method [6] will be used to evaluate the performance of the system for naturalness and intelligibility. Xitsonga mother tongue speakers will be selected ranging from less literate to professionals. These candidates will evaluate the system and rate it with respect to its observed naturalness and intelligibility separately. They will be given a range of possible options to select one that best characterizes their opinion. The mean opinion score (MOS) of all the results will be calculated at the end.

2.6. HTS Toolkit

The HMM-based synthesis system has a toolkit which is provided as a patch to HTK. For this project, the HTK 3.4.1 embedded with HTS version 2.1.1 will be used for experimentations [7].

3. CONCLUSION

The development of the first baseline Xitsonga TTS using an HMM-based speech synthesis (HTS) approach as part of the broader project of speech technology will increase the number of essential components of indigenous South African spoken language systems. The system should be highly intelligible and natural sounding; and should also model some of the desirable speaker characteristics well.

4. REFERENCES

[1] T. Masuko, "HMM-Based speech synthesis and its applications", PhD thesis, Tokyo Institute of Technology, Nov. 2002, pp. 1-84

[2] K. Tokuda, H. Zen, and A.W. Black, "An HMM-based speech synthesis system applied to English". Proc. of 2002 IEEE Workshop on Speech Synthesis, pp. 227–230, Sep. 2002

[3] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," IEICE Trans. Inf. & Syst., vol. E88-D,no. 3, pp. 503–509, Mar. 2005

[4] T. Raitio, "Hidden markov model based Finnish text-to-speech system utilizing glottal inverse filtering", Masters Thesis, 2008

[5] H. Zen, K. Oura, T. Nose, J. Yamagishi, S. Sako, T. Toda, T. Masuko, A.W. Black, and K. Tokuda, "Recent development of the HMM-based speech synthesis system (HTS)". In *Proc. 2009 Asia-Pacific Signal and Information Processing Association (APSIPA)*, Sapporo, Japan, Oct 2009

[6] X. Huang, A. Acero, and H. Hon. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall PTR, 2001

[7] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A.W. Black, and T. Nose, HMM-based Speech Synthesis System (HTS). May 2010, [Online] Available: http://hts.sp.nitech.ac.jp/

Baloyi Ntsako received his Bachelor of Science degree and his Honours degree in Computer Science in 2008 and 2009 respectively from the University of Limpopo and is presently studying towards his Master of Science degree at the same institution. His research interests include Speech Technology and Computer Security.