# Improving Fluency in a Sesotho Text-to-Speech Hybrid System

Lehlohonolo Mohasi, Daniel Mashao
Department of Electrical Engineering, University of Cape Town,
Private Bag, Rondebosch 7700, Cape Town, South Africa
lmohasi@crg.ee.uct.ac.za, daniel@ebe.uct.ac.za

**Abstract – Most of the present text-to-speech systems produce an acceptable quality speech output. Text-to-speech systems that are based on limited domain techniques produce speech that is close to human speech; however, they lack flexibility in that they cannot be used to synthesize words not in their own vocabulary. One approach of dealing with the flexibility problem is to use hybrid systems which combine limited domain systems and open vocabulary systems. This only solves part of the problem as discontinuities between words generated by different systems become apparent in the produced speech. In this paper, we improve the hybrid system by implementing techniques that can mask the discontinuities so that the output speech is more fluent. The proposed system was evaluated by carrying out subjective listening tests. In the tests, 20 listeners evaluated the quality of the speech output based on the MOS scoring system. The results showed an improvement on fluency with an overall score of 3.7 from 3.05.**

**Index Terms:** ***Fluency, text-to-speech***

## I. INTRODUCTION

Communication is a vital form of interaction in human society, be it through speaking or writing. Human beings have also shown interest in communicating with machines or computers. In order to have effective communication with computers, the output speech needs to be as natural-sounding and fluent as possible. This will increase interest for people to use technology for communication. This is where speech technology comes in. Speech technology has various categories which deal with ways to enhance communication with humans. This paper investigates one of the categories of speech technology, text-to-speech synthesis (TTS) technology, and its role in telecommunications.

TTS plays an important role in the application of information access. An example is e-government services where people can access government information on the website over the phone. The information requested is read back to them through the use of TTS technology. This service has an advantage that a person does not need an internet connection in order to access information. TTS in telecommunications is also applicable in Teldem systems where deaf people can communicate with hearing people over long distances. This service also incorporates use of speech recognition technology.

Different techniques have been used in TTS technology research [5, 12] in order to produce TTS systems with quality speech output. TTS techniques can be classified into three categories, the most popular being concatenative synthesis. Concatenative synthesis connects segments of pre-recorded speech to form words. Even though this method sounds natural, it has some glitches audible at the concatenation points of speech units. In this paper, we propose a method for masking the glitches so as to get fluent TTS systems. Fluency is essential for highly effective communication and invokes interest in people to communicate with TTS systems.

Section II gives details of the motivation for carrying out this project in a vernacular language. In Section III, various techniques used in TTS systems are discussed. We propose a method on how to overcome the fluency problem in the current and available systems in Section IV. Section V covers the procedure followed in carrying out the proposal. We conclude by giving results obtained from the experiments and tests done, and finally draw conclusions from the results.

## II. MOTIVATION

Language barrier is one of the causes of digital divide in the technology world. Many technological designs need the user to understand English in order to be able to operate or use them. This is of major concern as not everyone will benefit from the technology due to their lack of understanding in English. It is therefore, essential that technology designers meet the needs of the disadvantaged by introducing systems that everyone will be able to use with their limited literacy skills. Thus, the aim of integrating Sesotho as one of the official languages into the text-to-speech technology is to accommodate people who speak the

language and eliminate fear of using the technology. This way, people will now be comfortable and willing to use such services regardless of their computer- or language literacy.

The procedure in this paper can be used for TTS implementation in other South African official languages. Having multilingual TTS systems will be beneficial in information access. For instance, text-to-speech can be used in telephone-based systems, say, government services, where they can request information which will be relayed back to them as it is read off the government website. This support for language diversity will enable everyone to have access to information in a convenient manner, regardless of the language one speaks.

## III. TTS TECHNIQUES

Speech can be synthesized using three different methods, namely:

*Articulatory synthesis* attempts to model a human speech production system directly. When speaking, vocal tract muscles cause articulators to move and change the shape of the vocal tract which causes different sounds. The articulatory method has a potential of producing high-quality synthesis because it tries to model the human speech organs directly. The challenge though, is that it is a difficult method to implement and the computational load is also considerably higher than other methods. It has therefore, not received much attention from speech synthesis researchers and it has not achieved the same level of success yet.

*Formant synthesis* models pole frequencies of speech signal or transfer function of vocal tract based on source-filter model [2]. Formant synthesis does not use any human speech samples at runtime. Parameters such as fundamental frequency (F0), voicing, and noise levels are varied over time to create a waveform of artificial speech. It also provides infinite number of sounds and this makes it the most flexible synthesis method.

*Concatenative synthesis* strings together different lengths of prerecorded samples derived from natural speech. Generally, concatenative synthesis gives the most natural sounding synthesized speech. However, natural variation in speech and automated techniques for segmenting the waveforms sometimes result in audible glitches in the output, detracting from naturalness.

Concatenative synthesis has some weaknesses compared to other methods.

- Distortion from discontinuities in concatenation points, which can be reduced using diphones or some special methods for smoothing signal.
- Memory requirements are usually very high, especially when long concatenation units are used, such as syllables or words.
- Data collecting and labeling of speech samples is usually time-consuming. In theory, all possible allophones should be included in the material, but trade-offs between the quality and the number of samples must be made.
- The method is limited to one speaker.

There are three main subtypes of concatenative synthesis:

*Unit selection synthesis* uses large speech databases (more than one hour of recorded speech). During database creation, each recorded utterance is segmented into some or all of the following: individual phones, syllables, morphemes, words, phrases, and sentences. The division into segments can be done using a number of techniques, like clustering, using a specially modified speech recognizer, or by hand, using visual representations such as the waveform and spectrogram [3]. An index of the units in the speech database is then created based on the segmentation and acoustic parameters like the fundamental frequency (pitch). At runtime, the desired target utterance is created by determining the best chain of candidate units from the database (unit selection). This technique gives the greatest naturalness due to the fact that it does not apply digital signal processing techniques to the recorded speech, which often makes recorded speech sound less natural. In fact, an output from the best unit selection systems is often indistinguishable from real human voices, especially in contexts for which a TTS system has been tuned. However, maximum naturalness often requires unit selection speech databases to be very large, in some systems ranging into gigabytes of recorded data and numbering into dozens of hours of recorded speech.

*Diphone synthesis* uses a minimal speech database containing sound-to-sound transitions occurring in a given language. In diphone synthesis, only one example of each diphone is contained in the speech database. Diphones are defined to extend the central point of the steady state part of the phone to the central point of the following one, so they contain transitions between adjacent phones. This means that the concatenation point will be in the steadiest region of the signal, which reduces the distortion from concatenation points. At runtime, a target prosody of a sentence is superimposed on these minimal units by means of digital signal processing techniques such as Linear Predictive Coding (LPC), PSOLA, or MBROLA [3]. The quality of the resulting speech is generally not as good as that from unit selection but more natural-sounding than the output of formant synthesizers. Diphone synthesis suffers from sonic glitches of concatenative synthesis and sounds robotic. This has led to its use in commercial applications to decline, though it is still being used in research.

*Limited domain* concatenates pre-recorded words and phrases to create complete utterances. It is used in applications where a variety of texts the system will output is limited to a particular domain, like transit schedule announcements or weather reports. This technology is very simple to implement, and has been in

commercial use for a long time. The naturalness of these systems can potentially be very high because the variety of sentence types is limited and closely matches the prosody and intonation of the original recordings. However, because these systems are limited by the words and phrases in its database, they are not general-purpose and can only synthesize combinations of words and phrases they have been pre-programmed with.

## IV. Proposal

With regard to the three TTS techniques mentioned in Section III, our proposed method uses concatenative synthesis, the most common method. The limitation of this method is that it produces glitches during synthesis, which we have an intention of getting rid of through the use of intonation modeling.

One of the most difficult problems in text-to-speech synthesis systems to date is prosodic modeling, which is the main contributor to fluent speech. Fluency is one of the criteria expected of high-quality TTS systems. Prosody generation can be done through either rule-based modeling or statistical modeling. Researchers have worked on these models [6-8, 11] and they have different opinions on the best model for a particular language [9-10].

In order to have a TTS system that meets the requirements of an advanced TTS system, Rousseau [1] built a hybrid system (using two voice systems) which produced speech that was natural-sounding, flexible, pleasant and understandable. This system, though, has a weakness of having glitches caused by discontinuities in the two voice systems as they read out text. The speech output, though natural, does not flow. This paper, therefore, proposes a method to mask these glitches and have a better system which produces more fluent speech. This will be achieved by applying intonation modeling techniques, duration and F0, on the unit selection hybrid systems. For F0 modeling, a linear-regression statistical model is used. The pitch contours of the original and generated modules will be compared for improvement. Subjective listening tests will then be done in order to assess the speech output quality and see if there is any difference after applying intonation.

## V. Implementation

In order to build a hybrid system, individual voices were first built following the procedure in [4]. The voices built were the diphone voice, the limited domain, and the open domain whose utterances were labeled using Sphinx. Even though the diphone voice was not going to be used in the hybrid system (due to its robotic speech output), its purpose was for utterance prompting during the recording phase for the other unit selection voices. The diphone voice is useful in using Sesotho phones during synthesis prompting, instead of the English phones which are related to the default diphone voice in Festival, kal_diphone.

Before building the voices, a vocabulary of 400 words was created. This database was used for both voices systems. The recordings were done by a female who is fluent in Sesotho. The recorded prompts were then automatically labeled using a full acoustic Hidden Markov Model (HMM) model called Sphinx. An utterance structure was then generated from the labeled utterances and the pitchmarks extracted. During pitchmarking, the pitchmark parameters were modified accordingly to fall within the speaker's frequency range, which were 303 Hz and 181 Hz in this case. Pitch-synchronous Mel Cepstral parameterization of the speech was generated, from which cluster units were built. After the voices had been built, they were put together to form a hybrid system through a python script written by Rousseau [1]. This hybrid system synthesizes text by first going through it and checking which words are in the dictionary (or database). Then, for synthesis, the system uses the limited domain voice for words in the dictionary and the open domain voice for out-of-vocabulary (OOV) words. The system does this until it reaches the end of input text.

Intonation modeling was applied on the open domain voice. The reason for this being that the limited domain voice sounds more natural and fluent than the open domain voice. The aim is to get these features (naturalness and fluency) for the open domain system to be as close to those of the limited domain voice as possible. In this way, the glitches heard between voice interchange during synthesis will be reduced. Prosody (intonation) was applied on the open domain system using duration and fundamental frequency (F0) modeling techniques. A new hybrid system was then built from the limited domain and open domain with prosody. The F0 contours of the two systems were compared for differences and subjective listening tests were carried out. The features tested were understandability, pleasantness, and fluency, with fluency being of most interest. Fluency provides information about the melody and flow of the speech output. This criterion is also related to the concatenation point level of the voices which result in glitches. These tests were done on long and short sentences to find out if the length of a sentence has any effect on the fluency of the speech output. 20 subjects were used and they were all fluent in Sesotho. They were familiarized with the system before they could take the tests. The reason for the listening tests was to get the subjects' views on the two systems and find out if they could hear any improvement in the sense of fluency. The evaluation of the systems was based on a (Mean Opinion Score) MOS score with rating between 1 and 5, 1 being the worst and 5 being the best. Results obtained from these are shown in Section VI below.

## VI. Results and Evaluation

Table 1 below shows the results obtained from duration and F0 modeling of the voice used in the new open domain system with prosody.

**Table 1:** Results from Duration and F0 modeling

| | **Duration Modeling** | **F0 Modeling** | | |
|---|---|---|---|---|
| | ***Duration*** | ***Syl_start*** | ***Syl_mid*** | ***Syl_end*** |
| RMSE (Hz) | 0.7235 | 25.6281 | 21.5318 | 24.0215 |
| Correlation | 0.6907 | 0.4808 | 0.5794 | 0.8975 |

On average, F0 modeling has a lower correlation than that of duration modeling.

The results below illustrate the F0 contour shapes of the two hybrid systems. System 1 is the baseline system on which no intonation has been applied. System 2 is the system with intonation modeling. The sample utterance tested is "morena makhetha ke mofu".

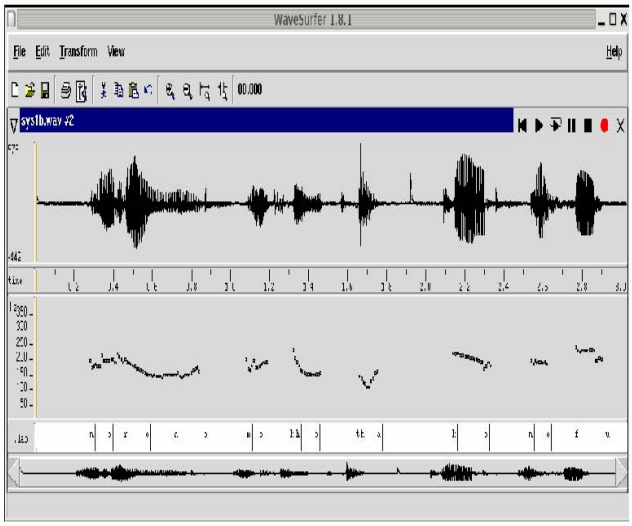


**Figure 1:** F0 contour shape for hybrid System 1.

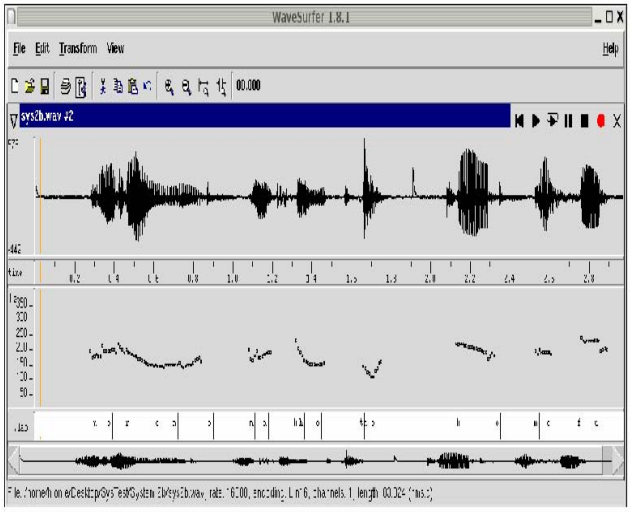


**Figure 2:** F0 contour shape for hybrid System 2

The contour shapes of the two systems look very similar, showing no major changes in F0.

The following two figures show the results obtained from the listening tests.

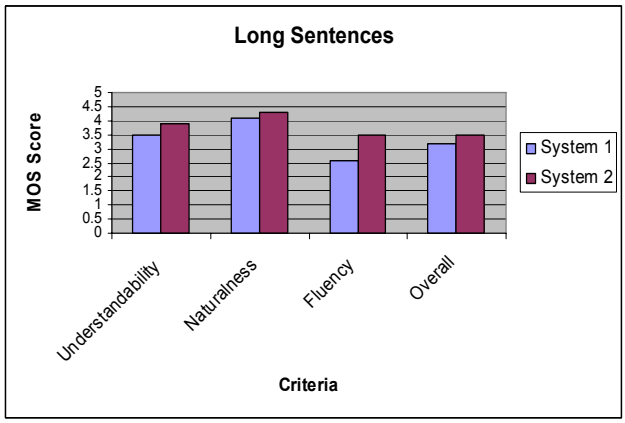


**Figure 3:** Graphical results for long sentences on the hybrid systems

For long sentences, the second hybrid system performs better for all categories, with all categories above the acceptable level of 3. System1 performs badly on fluency, with a score of 2.6.

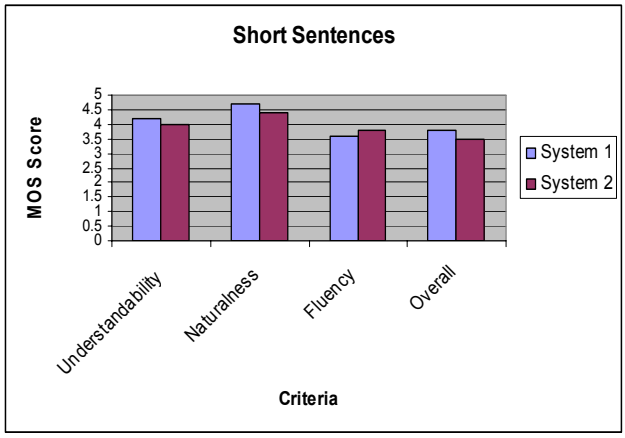


**Figure 4:** Graphical results for short sentences on the hybrid systems

For short sentences, both systems perform reasonably well, and much better than the long sentence performance. Naturalness rates the highest for both systems and fluency rates lowest. System 2 is more fluent than System 1 though overall, the listeners preferred System 1. Overall, System 2 was more preferable than System 1.

## VII. CONCLUSIONS

Based on the results obtained from subjective listening tests for long sentences, it can be concluded that intonation modeling has improved the speech output quality of the system on all categories. Fluency, in particular, has improved dramatically from a score of 2.6 to 3.5. For short sentences, improvement is

much greater and fluency has gone higher – from 3.6 to 3.8. The conclusion reached, therefore, is that the hybrid system is more fluent in synthesizing short sentences than long sentences. This improvement though, is not very outstanding in the F0 contour shapes of the two systems.

Even though fluency and naturalness of the hybrid system have improved, the overall score has not reached the desired level of 5. In order to enhance the system further, it is recommended that for future work, more studies should be done on duration analysis and modeling of Sesotho language. An investigation should also be carried out on stress characteristics in Sesotho and energy differences between phones.

## REFERENCES


[1] F. Rousseau, "Design of an Advanced TTS System for Afrikaans", MSc Thesis, University of Cape Town, 2006.

[2] S. Lemmetty, "Review of Speech Synthesis Technology", MSc Thesis, Helsinki University of Technology, 1999.

[3] "Speech Synthesis", http://www.fact-index.com/s/sp/speech_synthesis.html [Last accessed on 12 February 2006]

[4] A. Black and K. Lenzo, "Building Synthetic Voices", 2003.

[5] A. Black, et. al., "Limited Domain Synthesis", Proceedings of ICSLP, 2000.

[6] T. Saito, M. Sakamoto, "Generating F0 Contours by Statistical Manipulation of Natural F0 Shapes", *Proceedings of Eurospeech*, 2001.

[7] T. Saito, M. Sakamoto, "Applying a Hybrid Intonation Model to a Seamless Speech Synthesizer", *International Conference on Spoken Language Processing (ICSLP)*, 2002.

[8] K. Dusterhoff, A. Black, "Generating F0 Contours for Speech Synthesis using the Tilt Intonation Theory", *Proceedings of the ESCA Workshop on Intonation*, 1997.

[9] F. Tamburini, C. Caini, "An Automatic System for Detecting Prosodic Prominence in American English Continuous Speech*", International Journal of Speech Technology*, Vol. 8, pp. 33-44, 2005.

[10] E. Keller, S. Werner, "Automatic Intonation Extraction and Generation for French", *14th CALICO Annual Symposium*, 1997.

[11] C. Tseng, et. al., "Fluent Speech Prosody: Framework and Modeling", *Speech Communication*, Vol. 46, pp. 284-309, 2005.

[12] "History of Speech Synthesis, 1770-1970: Wolfgang von Kempelen's speaking machine and its successors", www.acoustics.hut.fi/~slemmett/dippa/chap2.html. [Last accessed on 17 March 2006]



**Lehlohonolo Mohasi** is currently undertaking her MSc (Eng) degree at the University of Cape Town. This is her second year of study and she specializes in Speech Technology.

**Professor Daniel Mashao** is her supervisor from the same institution. He is a member of IEEE.