Simplified Way of Producing Sinhala Concatenative Text To Speech for Embedded Systems

Dammika Prasad Wijethunga[1], Asiri Nanayakkara[1] and Janaka Wijayakulasooriya[2]
Institute of Fundamental Studies, Hanthana Road, Kandy
[2]*Department of Electrical & Electronic Engineering. University of Peradeniya*
email: asiri@ifs.ac.lk

ABSTRACT

A new method for simplifying the production of Sinhala Concatenative Text to Speech system for embedded systems was developed. The system is most suitable for low power applications as well as environments with low computational resources. The new speech database consists of only 572 sound units for unlimited domain representation as compared to Sinhala diphone databases which usually contain over 1413 sound units. The quality of the new system was evaluated with six subjects and found that the intelligibility is over 80% and the naturalness was rated as "good".

1. INTRODUCTION

With Text-to-Speech (TTS) technology, synthetic voice output can be produced using textual information which serves as a more natural interface in human machine interaction. In order for TTS technology to be widely accepted, almost natural voice output quality has to be attained.

TTS technology is now engaged in a wide range of applications, covering from assistive technology and education, to telecommunications and entertainment. For desktop applications and server-based environments, adequate resources are available for developing high quality TTS technology. Therefore when developing TTS for these systems, a little attention has been paid to minimize the computer resources required for producing high quality natural voice. As a result, its use on low power embedded or portable devices has become next to impossible.

On the other hand, the ever increasing demand for enhanced consumer applications and the extensive use of portable devices such as Mobile phones or personal digital assistants (PDA) in everyday life have increased the need to efficiently adapt TTS technology in environments with limited computational Resources.

In order to address the challenge of developing a high quality TTS system for embedded devices, during last few years, several approaches have been adopted. Widely used solutions include Formant synthesis, Concatenative synthesis and HMM-based synthesis.

Formant based TTS systems are well suited for low resource environments, but Formant synthesis generate artificial, robotic sounding speech which is often not acceptable for high quality speech generation [1]. Generally, the Concatenative synthesis which is

based on the concatenation of segments of recorded speech units produces more natural sounding speech. Concatenative synthesis can be divided into few sub types. They are Unit selection synthesis, Domain specific synthesis and diphone synthesis. Unit selection synthesis can be used for general speech production while Domain specific synthesis is limited to a particular domain, like transit schedule announcements or weather reports. Although diphone synthesis is producing relatively low natural speech when compared to the Unit selection synthesis and Domain specific synthesis, it has a considerably good speech quality. Unit selection synthesis uses phones, diphones, words, phrases, and sentences in the database, while Domain specific synthesis uses only prerecorded words and phrases. Diphone synthesis on the other hand uses only the diphones (sound to sound transitions) occurring in a language. HMM based synthesis is also a commonly used synthesis method which is based on hidden Markov models.

Even though the unit selection TTS is still the dominant approach for producing high quality speech and has several good features such as it uses only a minimum digital signal processing (DSP) for producing synthesized speech (DSP often makes recorded speech sound less natural and increases the Computational load), the resources such as storage capacity needed for the unit selection TTS is quite large. Therefore the diphone Concatenative synthesis which has a small capacity database is a good alternative for constructing quality Sinhala Text to Speech system which is to be used for embedded systems.

Five years ago, based on the Festival speech synthesis framework [2], the Language Technology Research Laboratory, University of Colombo School of Computing has developed a diphone Concatenative TTS system [3]. There are 1413 Sinhala diphone units in this system. This system produces smooth fairly intelligible Sinhala speech. However, the quality of the speech is tainted by Robotic sound. Further, the implementation of high quality Sinhala diphone TTS is not a trivial task as determination of accurate diphone boundaries is difficult even using automated segmentation methods.

In this paper, we introduce a modified diphone Concatenative synthesis method with reduced database size which is suitable for embedded systems while keeping the quality of synthesized speech with minimal Robotic sound. The detailed database construction method and evaluation results will be presented. In particular, emphasis is given to three main topics. Firstly, in section 2 Database structure of the traditional diphone database is briefly reviewed and a new Sinhala Sound unit database structure is proposed. Section 3 provides details of the construction of the proposed database. In section 4, Sound units concatenation method described and in section 5, both evaluation method and results are presented. Finally, a summary and some conclusive remarks are given in section 6.

2. SINHALA SIMPLIFIED CONCATENATIVE SYNTHESIS DATABASE STRUCTURE

Spoken Sinhala has 14 vowel and 26 consonant phonemes. Therefore there are 40^2=1600 hypothetically possible diphone combinations. Since not all phoneme-

phoneme pairs occur in the speech, the spoken Sinhala diphone inventory has only 1413 diphones [4].

In order to implement a Sinhala diphone TTS we have to record all 1413 diphones and it's not a trivial task. Even though there are some automated segmentation methods, they are not completely accurate in finding diphone boundaries and hence automatic diphone segmentation is needed to be hand corrected. Therefore we introduce a database with different sound units which can be easily constructed with segmentation of sound units. The new database not only consists of reduced number of sound units but also will be useful in producing quality synthesized Sinhala speech.

Any Sinhala utterance is made of combination of vowels, consonants and silence periods. We describe the new method with the aid of a word with two letters "ගද" and "ද" as shown in figure 1. This word can clearly be separated into five Phonemes as Consonant(C), Vowel (V), Silence(S), Consonant(C), and Vowel (V). In this case, we can practically observe by hearing that there is a sound distribution of the Consonant is embedded in the subsequent vowel. However, it can be proved by using methods such Linear Predictive Coding that the end part of the vowel is nearly free from this contamination.

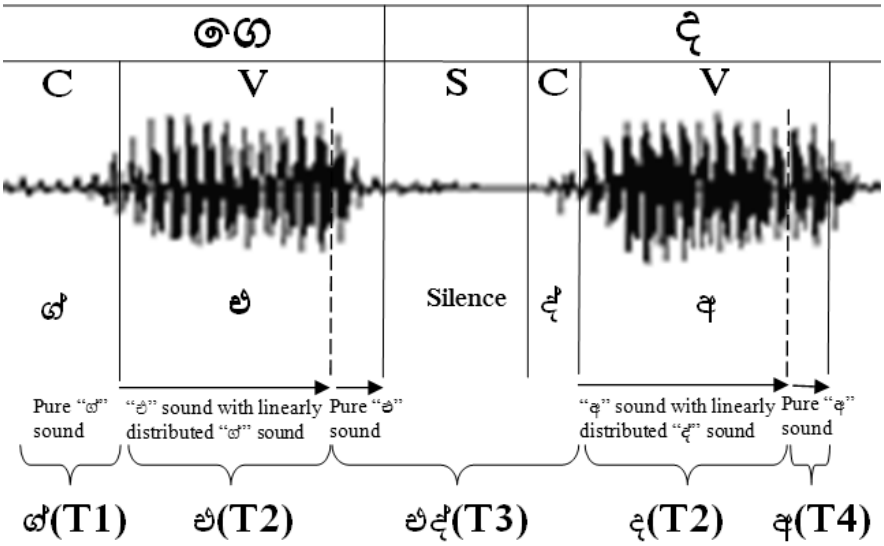

Figure 1: New invented Sinhala sound unit structure for concatenative synthesis

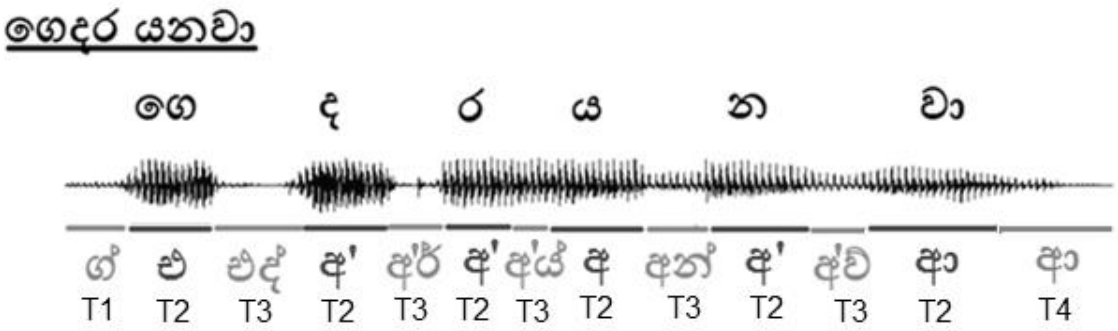

Figure 2: Sinhala sound unit structure in a short Sinhala sentence.

As a result, as shown in figure 1 those five Phonemes described earlier can be constructed with four types of sound units. With these four sound units any Sinhala utterance can be constructed. Four sound units can be generalized as follows.

Type 1 sound unit (T1): Pure consonant sound units
Type 2 sound unit (T2): A vowel unit, but it is distributed with a consonant sound.
Type 3 sound unit (T3): A unit that consists with a pure vowel sound, silence and a consonant sound. But some units consist of a pure vowel and a consonant sound (without a silence period).
Type 4 sound unit (T4): Last part of pure vowel sound units.

Whole sound database consist of these four types of sound units and the distribution of the sound units can be properly explained from the second example given in Figure 2.

Even though there are four types of sound units in this speech database, it is not necessary to record all four types. Because T1 sound unit can be created using T3 sound units. Short Time energy calculation was used for this extraction process. Final database consists of 572 sound units (as compared to 1413 Sinhala diphone database) as given below.

Number of sound units in T2 = 21 X 13 + 13 = 286 (except ඟ, ඩ, ඳ, ඹ sounds)
Number of sound units in T3 = 21 X 13 = 273
Number of sound units in T4 = 13 (except ඍ)
Total Units = 13 + 286 +273 = 572

ඟ, ඩ, ඳ, ඹ, and ඍ related sounds were created with combining the excising units. Therefore those units were also not recorded.

3. CONSTRUCTION OF THE DATABASE

A male speaker was chosen for pronunciation of the words. He was advised to maintain a constant pitch, volume, and fairly constant speech rate during the recording. However recording sessions were limited to one hour per day. There is a significant voice changing effect during the day for the same speaker and therefore the recording sessions were scheduled at the same time period of the day (9am to 10am).

Speech signals were recorded by a normal condenser microphone and a laptop with free digital audio editing software Audacity 1.2.6. File format is Microsoft PCM file format (.wav) with sampling rate of 44,100 Hz and 16 bit A/D conversion. All segmentations were manually done in the Audacity 1.2.6 software environment. No automated method was used for this process and only visual observation of the wave forms and hearing of the sound were used. When concatenating the sound units, we decided to not to use any Digital signal processing techniques in order to reduce the computational load in embedded systems. Therefore, to prevent hearing the perceptible clicks in the joining points, the edges of the segmented units were kept at a very low or zero amplitude.

Recording of T2 sound units: In order to construct a single target sound unit, six different words containing the same sound unit were recorded. These six words may include meaningful words as well as nonsense words. The target sound unit is segmented from the middle syllable of a word, minimizing the articulatory effects at the start and the end of the word.

The best sound unit was selected as follows. From each recorded word, the target sound unit was extracted. Then it was inserted into the other five words replacing the same sound unit within each word. The sound unit which has the ability to produce intelligible pronunciation in all six words is selected as the best sound unit. This is illustrated in Figure 3. The same procedure was followed for constructing T3 units. The procedure for constructing T4 is also the same except the target sound unit is now located at the end of the word.

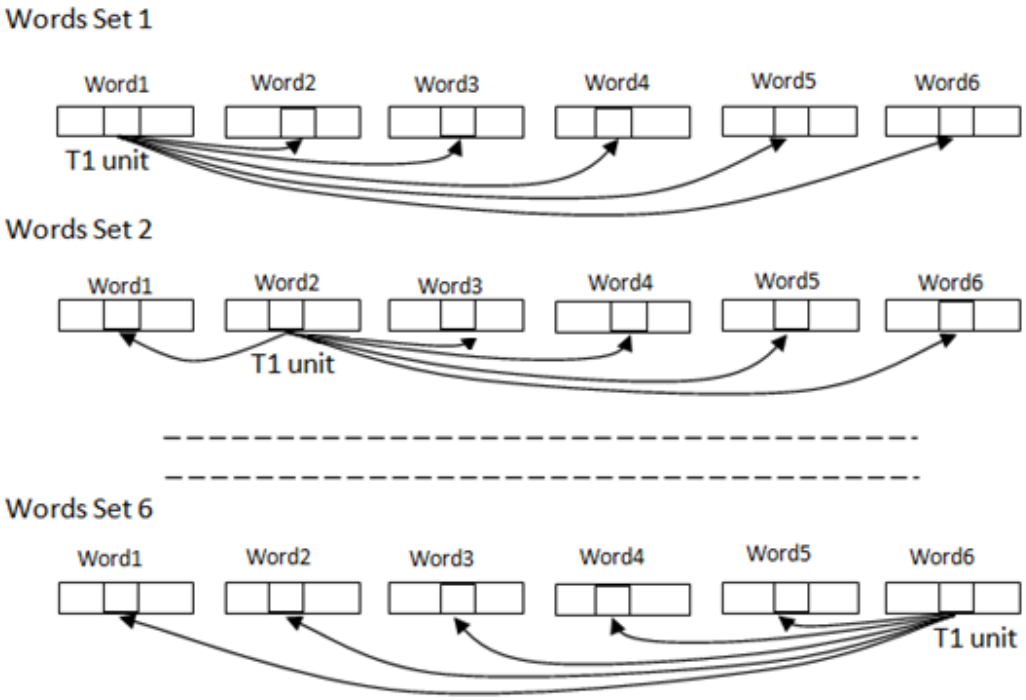

Figure 3: Procedure for selecting best target sound units

4. SOUND UNITS CONCATENATION

A MATLAB program was used for concatenation and generation of synthesized speech for evaluation purposes. As mentioned above in section 2, T1 sound unit extraction process need to be done using short time energy calculation in runtime. Since we have already added some features to speech segments when designing and constructing the database, any other DSP techniques such as pitch synchronization or windowing the edges were not used for concatenation process. Further, additional rules were also imposed for special utterances.

5. EVALUATION

The synthesis speech output was directed to a small size speaker system which is connected to the laptop. The preliminary evaluation was performed with six subjects. They were asked to judge 20 words and 7 sentences which are all collected from newspapers, internet sources such as news web sites, and literature magazines. The intelligibility of the synthesized speech was evaluated on two levels; Word level and sentence level. Subjects, participating in the test were asked to write down everything they heard. Figure4 shows the percentage of correctly understood words and sentences.

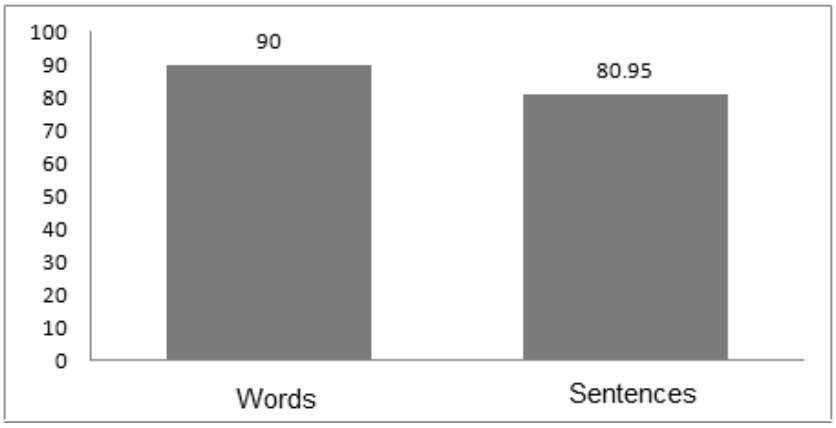

Figure 4: Intelligibility test. Percentage of correctly understood words and sentences

Next, for evaluation of the quality of the synthesized speech, a long paragraph was played and subjects were asked a question about naturalness and sound quality and asked to note how well the voice performs. The exact question for this test was:"Did you consider the synthesized voice has good sound quality?" The results are shown in Figure 5.

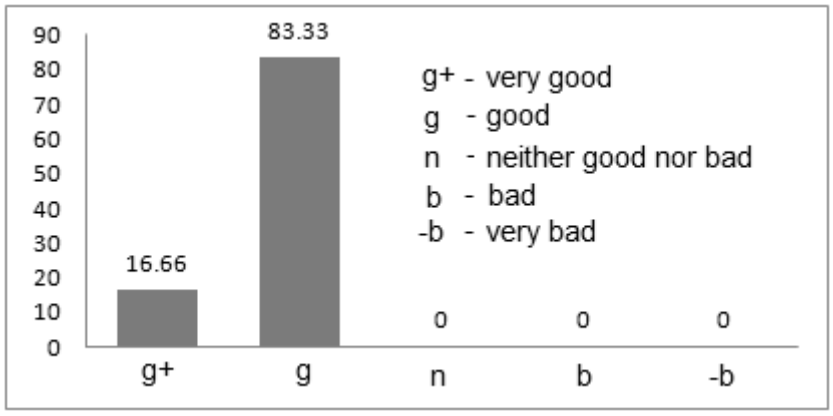

Figure 5: Results of Sound quality of the synthesized speech

Among the six subjects, around 17% considered the voice has "very good" quality while 83.33% thought the sound quality of the voice was "good".

6. CONCLUSION

The produced synthesized speech is intelligible, but utterances sometimes suffer from a lack of naturalness and fluency. Improvement of intelligibility and naturalness depends, in particularly, on proper lexical stress assignment and a more sophisticated generation of prosodic features. In this study we mainly considered the Sinhala speech database construction with certain features especially suitable for embedded systems with limited computational resources and we were able to achieve a quality and efficient TTS technology using a simplified speech database system.

REFERENCES

[1] Schnell, M., Jokisch, O., Hoffmann, R. and Kustner, M., *Text-to-speech for low-esource systems,* IEEE Workshop Multimedia Signal Processing (MMSP), St. Thomas, (2002) 259-262
[2] Black, A.W. and Lenzo, K.A., *Building Synthetic Voices,* Language Technologies Institute, Carnegie Mellon University and Cepstral LLC. Retrieved from: http://festvox.org/bsv/, (2003)
[3] Weerasinghe Ruvan, Wasala Asanka, Welgama Viraj and Gamage Kumudu, *Festival-si: A Sinhala Text-to-Speech System,* Language Technology Research Laboratory, University of Colombo School of Computing, (2007)
[4] *Proceedings of Text, Speech and Dialogue,* 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 3-7, (2007) 472-479
[5] Karunatillake, W.S., *An Introduction to Spoken Sinhala, 3rd Ed*., M.D. Gunasena & Co Ltd, (2004)