Vol 8. No. 2 June, 2015

African Journal of Computing & ICT

www.ajocict.net

Comparison of Concatenative Text-To-Speech Synthesis Approaches for Ẹdo Language

S.T. Apeh[1] & C.K. Nwaekwu[2]
[1,2]Department of Computer Engineering
Faculty of Engineering
University of Benin
Benin City, Nigeria
[1]E-mail: apeh@uniben.edu, [2]enkaycee@yahoo.com
[1]Phone: +234- 803-495-6812

ABSTRACT

In text-to-speech (TTS) conversion systems, concatenative technique involves concatenation of natural speech units and putting them together to form a word or sentence. In this paper, a comparative study of two concatenative approaches is presented. The two approaches are Word Level unit selection concatenation and Phoneme concatenation. Word concatenation involves selecting a whole word as speech unit, whereas phoneme concatenation involves use of phones and syllables as unit of speech. This study was carried out to compare the effectiveness or otherwise of the two methods in TTS systems for Ẹdo language, with selected sample texts. The Word Level achieved 82% on the scale rating while Phoneme approach scored 64%. This means that Word concatenation shows to be of better quality compared to phoneme concatenation in terms of naturalness.

Keywords: Text-To-Speech, Concatenative synthesis, Ẹdo language

African Journal of Computing & ICT Reference Format:
S.T. Apeh & C.K. Nwaekwu (2015) Comparison of Concatenative Text-To-Speech Synthesis approaches for Ẹdo Language. Afr J. of Comp & ICTs. Vol 8, No. 2. Pp 198-202. -

1. INTRODUCTION

TTS synthesis is a means of converting text to audible speech with the aid of automated system. A TTS system is a computer-based system that reads out text aloud automatically by transferring linguistic information stored as data or text into speech [4]. There is growing trend in research studies all over the world on TTS synthesis. Two main techniques have always been employed. These are *synthesis by rule* and *concatenative synthesis.* [5] While the rule-based approach exploits the expert knowledge of speech scientists in speech production and perception by putting the human expert in the design, the concatenative approach employs recordings of a human speaker, inherently putting more emphasis on data [5].

1.1. Concatenative synthesis

This technique uses short segments of recorded speech that are cut from recordings and stored in an inventory (voice database) either as uncoded waveforms or encoded by a suitable speech coding method. Concatenative synthesis has become very popular in recent years due to its improved sensitivity to unit context [6]. Many TTS systems are developed based on this corpus-based synthesis technique and it has become very popular for its high quality and natural speech output.

Concatenative TTS synthesis technique can be categorised into two, which according to [7] are:

1. Word concatenation – involves recording of all required words, thereafter concatenating the separate words to form a sentence. This technique, however, is applicable where a limited vocabulary is required e.g. announcement of arrivals in train stations or telecommunication customer care lines.
2. Phoneme concatenation – involves generating phonemes of the given language naturally occurring texts, and concatenating them to form the desired word or sentence.

1.1.1 Issues with Concatenative Synthesis

Though the concatenative synthesis has proven to be the most used approach due to its high quality output, it however has some significant drawbacks.

1. Multi-hour recording – it requires long period of hours and even weeks or days to perform the voice recording for concatenative synthesis
2. Large database – the voice database is always very large, even with compression techniques. This results in higher cost.
3. Narrow-domain application – this is an obvious setback in this technique. It can only be used in limited domain such as telecommunication services, audio books, etc.

4. Joining problem – how to successfully join sections of waveforms, such that the joins cannot be heard so that the final speech sounds smooth continues always has a cost. It may result in click being heard because of the effect of the jump.

1.2. Unit Selection Synthesis

The Unit Selection Synthesis (USS) has become the dominant concatenative synthesis technique in TTS today [8]. USS deals with the issue of how to manage large numbers of units.
In unit selection synthesis, appropriate sub-word units are selected from multiple examples in a database of natural speech. It has been shown to produce high quality natural sounding speech [9].

Arguably, the USS is the most data-driven technique as little or no processing is performed on the data, rather it is simply analysed, cut up and recombined in different sequences. In general, unit selection databases are acquired by having a single speaker read a specially prepared script, which should be normal text materials. The speech is recorded and stored in an inventory.

1.3 Creating USS Database

In terms of the actual recording, there are two main issues: the choice of speaker and the recording conditions. Speaker choice is perhaps the most vital, and all commercial TTS operations spend considerable time choosing a good speaker, who must be able to read the script accurately, without straining his or her voice, and keep the voice at a regular style throughout the exercise. Besides, it makes sense to pick a speaker who has a good voice. Though, there may be no objective means of selecting a good speaker, but a good idea is to play recordings of potential speakers to a listening target group and asking them of their preferred voice. A good unit selection system will sound very similar to the speaker it is based on. As far as recording conditions are concerned, it is highly desirable that there be as high quality as possible as background noise, reverberations and other effects will all be heard in the synthesized speech.

1.4 TTS quality requirements

Understandability, flexibility, naturalness, and pleasantness are the quality requirements of an advanced TTS system [10]. The concatenative speech synthesis has been the most successful technique in meeting all these requirements. Naturalness shows how well the quality of the system is, compared to that of the human voice, while understandability shows how well the synthesized message is understood by the listener after the first time of listening. Pleasantness is a measure of how pleasant the synthesized voice is to listen to, which is an indication on whether or not a listener will be willing to use the system again. While phoneme concatenation has high degree of flexibility in implementation, it lacks understandability, naturalness and pleasantness. On the other hand, the USS has the capability of producing TTS system with high degree of naturalness, understandability and pleasantness.

The main drawback of the USS is that it is employed in narrow domain applications as it is easily faced with out-of-vocabulary words.

2. COMPONENTS OF CONCATENATIVE SYSTEM

Concatenative synthesis can be broken down into three main components, namely:

i. Front-end
ii. Intermediate processor, and
iii. Back-end

These components are represented in Figure 1.

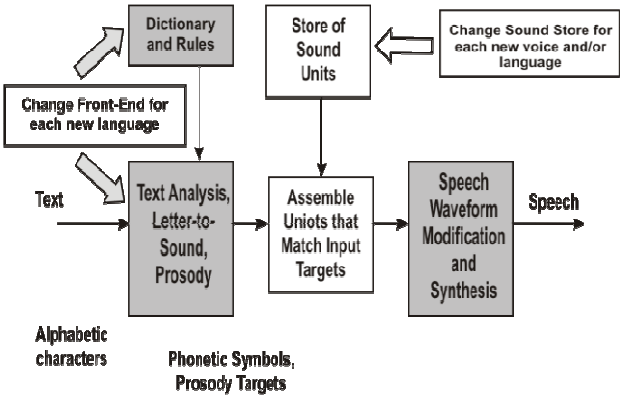

Figure 1: Components of concatenative synthesis

a. The front-end converts a given input text string into a string of phonetic symbols and prosody targets, by employing a set of rules and/or a pronounciation dictionary [5]. It is the part of the system closer to the text input.
b. The intermediate processor (centre block) assembles the units according to list of targets set by the front-end. These units are selected from a store (top centre block) that holds the inventory of available sound units. Lying in-between the front-end and back-end, the intermediate processor performs phonetic analysis, such as homograph disambiguation and grapheme-to-phoneme conversion
c. The back-end is the part of the system that is closer to the speech output. This module controls the voice rendering that corresponds to the input text.

2.1. Methodology

This study was carried out to compare the effect or otherwise of word level unit selection concatenation and phoneme concatenation methods for native Ẹdo language. Selected Ẹdo text, *Osanobua* (which means ('God') was used in the study.

a. GUI implementation – this was designed and implemented using xhtml scripts generated in Adobe Dreamweaver application environment. The application is webpage editor.

b. Speech recording – the recording was carried out using Audacity™ audio recorder, though in a non-professional setting. The speech samples were stored as *.wav* files.
c. Database creation – the speech database was created with MySQL, a popular web database application.
d. Web server – WAMP™ 5.0 was used to test the sample. The server side application is PHP.
e. Evaluation – the evaluation of the comparison was based on the subjective mean opinion score (MOS) and Degradable MOS (DMOS) rating. Ten (10) listening subjects were randomly selected for the listening group.

2.2. Implementation and evaluation

The implementation flowchart is shown in Figure 2.

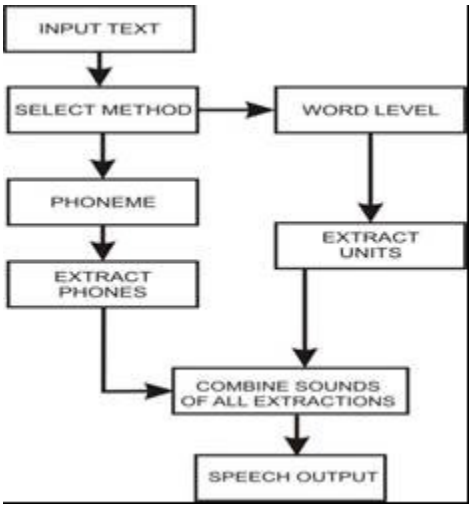

Figure 2: Concatenation flowchart

3. RESULT PRESENTATION

Table 1 is the presentation of the subjective scores by 10 listening subjects selected for the evaluation. The last row is the average of all the scores while the last column is the arithmetic mean of the individual scores for the two approaches.

Table 1: Subject rating by subjects

Subject	Word Level	Phone	Mean
1	4	3	3.5
2	5	3	4.0
3	4	2	3.0
4	4	4	4.0
5	5	4	4.5
6	3	2	2.5
7	4	4	4.0
8	3	3	3.0
9	5	3	4.0
10	4	4	4.0
Average	**4.1**	**3.2**	**3.65**

The MOS and DMOS scales are expressed as in Tables 2 and 3 respectively.

Table 2: Mean opinion score (MOS)

MOS Rating	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Table 3: DMOS categories

MOS scale	Subjective rating/response
1.00 – 3.00	Unacceptable
3.00 – 5.00	Acceptable

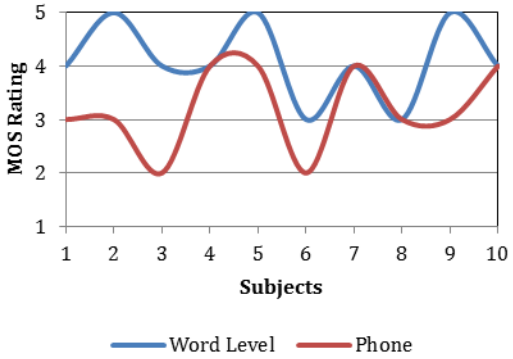

Figure 3: Plot of MOS Rating for Naturalness

Figure 3 presents the plot of the subjective scores.

4. DISCUSSION

From the plot of Fig.3, it is observed that all the values for Word Level rating lie between 3 and 5 on the MOS scale. Also, the ratings for Phoneme lie between 2 and 4, with only two values lying below 3. This shows that both methods are quite acceptable, but the Word Level unit selection ranks higher. This means that Word Level produces higher output than the Phoneme approach for Edo language.

From Table 1, the Word Level average score of 4.1 represents 82% while the Phone approach score is 64%. This also confirms that on the overall scale, the two approaches produce acceptable quality of speech output.

5. CONCLUSION

In this paper, comparison of the methods of unit concatenative synthesis – *Word Level and Phoneme* – for TTS in Ẹdo language using the text *Osanobua was examined, using subjective rating*. Results of the evaluation tests showed that even though any of the two approaches can give an acceptable speech quality in terms of naturalness, the Word Level approach produces higher quality speech compared to Phoneme.

This result shall guide the choice of approach for an ongoing work on TTS system for Edo Language.

REFERENCES

[1] Ogieva, U. (2012). Edo the Largest Ethnic Group in Nigeria (Part 1). Accessed on 27th July, 2015 at <ihuanedo.ning.com/group/wazobiaisalienotnigeria/ forum/topics/edo-the-largest-ethnic-group-in-nigeria-part-1>

[2] NBS (2012). Annual Abstract of Statistics, 2012. National Bureau of Statistics, Federal Republic of Nigeria, p. 33. Available online at <nigerianstat.gov. ng/pdfuploads/annual_abstract_2012.pdf>

[3] Ejiofor, C. (2012). Top 10 Most Popular Languages in Nigeria. Retrieved on 17th June, 2015 from <https://www.naij.com/383776-top-10-most-popular - languages-in-nigeria.html>

[4] Patra, T.K; Patra, B; Mohapatra, P. (2012). Text to Speech Conversion with Phonematic Concatenation. Int. J. of Electronics Comm. and Computer Tech. (IJECCT), Vol. 2 Iss. 5.

[5] Schroeter, J. (2005). Text-to-Speech (TTS) Synthesis. Electrical Engineering Handbook, 3rd edition. Publication of AT&T Laboratories, Chapter 16. Accessed on 26th June, 2015 at <www2.research.att. com/~ttsweb/tts/papers/2005_EEHandbook/tts.pdf>

[6] Sasirekha, E., Chandra, E. (2012). Text to Speech: A Simple Tutorial. International Journal of Soft Computing and Engineering (IJSCE). Vol. 2, Iss. 1.

[7] Ngugi, K., Okelo-Odongo, W., Wagacha, P.W. (2005). Swahili Text-to-Speech System. African Journal of Science and Technology, Vol. 6, No. 1, pp. 80-89.

[8] Taylor, P. (2007). Text-to-Speech Synthesis. University of Canbridge. p. 485-530.

[9] Saikachi, Y. (2003). Building a Unit Selection Voice for Festival. University of Edinburgh.

[10] Rousseau, F. (2005). Design of an Advanced Text-to-Speech System for Afrikaans. University of Cape Town.