International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 1, Issue 1, pp: (14-17), Month: April - June 2014, Available at: www.paperpublications.org

Segmentation Words for Speech Synthesis in Persian Language Based On Silence

Sohrab Hojjatkhah, Ali Jowharpour

Islamic Azad university, Dehdasht branch

Abstract: **In speech synthesis in text to speech systems, the words usually break to different parts and use from recorded sound of each part for play words. This paper use silent in word's pronunciation for better quality of speech. Most algorithms divide words to syllable and some of them divide words to phoneme, but This paper benefit from silent in intonation and divide words at silent region and then set equivalent sound of each parts whereupon joining the parts is trusty and speech quality being more smooth . this paper concern Persian language but extendable to another language. This method has been tested with MOS test and intelligibility, naturalness and fluidity are better.**

Keywords: **TTS, SBS, Sillable, Diphone.**

1- INTRODUCTION

The correct pronunciation of unknown or novel words is one of the biggest challenges for text-to-speech (TTS) systems. Generally speaking, correct analysis and pronunciation can only be guaranteed if a direct match exists between an input string and a corresponding entry in a pronunciation dictionary or morphologically annotated lexicon. However, any well-formed text input to a general-purpose TTS system in any language is extremely likely to contain words that are not explicitly listed in the lexicon. lexicon entries of every language is unbounded; Thus, in unlimited vocabulary scenarios we are not facing a memory or storage problem but the requirement for the TTS system to correctly analyze unseen orthographic strings. The Persian language is notorious for its extensive use of compounds. What makes this a challenge for linguistic analysis is the fact that compounding is extraordinarily productive. Linguistic analysis has to provide a mechanism to appropriately decompose compounds and, more generally, to handle unknown words. Most of the existing commercial speech synthesis systems can be classified as either formant synthesizers [1,2,3] or concatenation synthesizers [4,5]. Formant synthesizers, which are usually controlled by rules, have the advantage of having small footprints at the expense of the quality and naturalness of the synthesized speech [6]. On the other hand, concatenative speech synthesis, using large speech databases, has become popular due to its ability to produce high quality natural speech output [7]. The large footprints of these systems do not present a practical problem for applications where the synthesis engine runs on a server with enough computational power and sufficient storage [7]. In Persian language exist short vowel and long vowel but both kinds assume same in TTS systems. This paper isolate short & long vowels. This property is used to get components of word that restrict by silent in word's pronounciation. This method can be concatenative rule.

2- RELATED WORK

2.1 SYLLABLE BASED TTS SYSTEM

Syllable is one of the method, which is used for developing a text-to-speech system. Various languages have different patterns for syllable. In most of these languages, there are many patterns for syllable and therefore, the number of syllables is

large; so usually syllable is not used in all purpose TTS systems. For example, there are more than 15000 syllables in English [6]. Creating a database for this number of units is a very difficult and time-consuming task.

In some languages, the number of syllable patterns is limited, so the number of syllables is small, and creating a database for them is reasonable; therefore this unit can be used in all-purpose TTS systems. For example, Indian language has CV, CCV, VC, and CVC syllable patterns, and the total number of syllables in this language is 10000[1].Syllable is used in some Persian TTS systems, too. This language has only CV, CVC, and CVCC patterns for its syllables and so, its syllable number is limited to 4000 [10].

2.2 DIPHONE BASED TTS SYSTEM

Nowadays diphone is the most popular unit in synthesis systems. Diphones include a transition part from the first unit to the next unit, and so, have a more desirable quality rather than other units. Also, in some modern systems, a combination of this unit and other methods such as unit selection are used. Persian has 23 phonemes, so as an upper hand estimate, it has about 900 diphones.

2.3 Allophone based TTS system

Naturalness within the allophone, especially vowels; much fewer units need be prerecorded than with other choices Good articulation between most allophones is difficult to achieve, making individual words unintelligible; stop consonantal allophones are difficult to isolate for prerecording.

Hypothetically, allophones appear to be the perfect unit for concatenation. A few hundred of them will serve as the basic building blocks for all utterances. The allophonic database is a set of allophone wav files, each file being named accounting the allophone itself and its phonetic context. According to the input text, proper allophones from database have been chosen and concatenated to obtain the primary output.[9]

3- SILENCE BASED SEGMENTATION (SBS)

This paper uses silence to word segmentation and replace properly sound for each segment. Because if silent in speech be according to silent in play sound by computer then speech is more natural. For example look at the below word "مناظره" (Monâzere). If this word become segmentary by syllable will be "Mo_na_ze_re" that for speech needed four piece recorded sound. Thus concatenatenation of these parts decrease naturalness of speech. So must to try number of parts become minimum. When it is said "Monâzere" ,intonation is "Monâ",silent, "zere" then in TTS systems it is better that segmentation to do like to natural intonation.

Here explain that how do segmentation. The syllable structure variety of Farsi is limited to only three types: *CV*, *CVC* and *CVCC*. Among all of syllables, which can be constructed from 23 consonants and 6 vowels of Farsi in the above three syllable types (structures), only about 4000 syllables are used in Farsi words and the others are not used [10]. In this paper vowels divide to kind, short vowel and long vowel. Short vowels don't show in written text but long vowels are written in text. Short vowel show with "V" and long vowel with "W". the boundaries between parts are four types "WCV", "WVW","CCV" and "CCW" that word is isolated after first letter in types. Finding these types help to see silent in the word. If the verb "Monâzere" is given to below function (iso) it will gives "*CVCW*CVCV" string that "WCV" is position of word's division. Therefore "CVCWCVCV" divide to two substring "CVCW" and "CVCV" or "Montazer" become "CVCCVCVC" that divide to "CVC" and "CVCV". Each unit isolated to another by silent in speech. See iso function

```
function iso(s:string):string;

var
```

```
ss:string;

kk:byte;

begin

 ss:='';

 for kk:=1 to length(s) do

 begin

   if pos(s[kk],c)<>0 then ss:=ss+'C'; {if kk's letter in s is consonant then add "C" to result string}

   if pos(s[kk],v)<>0 then ss:=ss+'V'; {if kk's letter in s is short vowel then add "V" to result string}

   if pos(s[kk],Ww)<>0 then ss:=ss+'W'; {if kk's letter in s is long vowel then add "W" to result string}

 end;

 iso:=ss;

end  ;
```

after segmentation use to database that contain of all Persian parts of verbs. The number of units in database is more than syllable rule but its utilization as the basic unit of concatenation will reduce the necessary smoothing process.
figure 1 shows segmentation by silence.

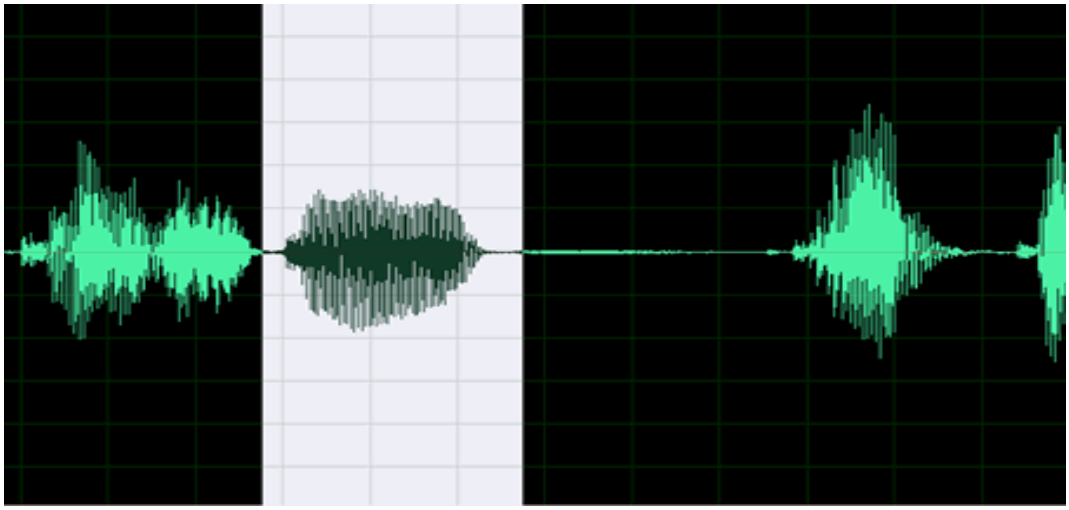

Fig 1: position of signal cutting

In SBS method number of parts that must is stored in signal dictionary increase, but in concatenate phase operation is less.

4. EVALUATION SBS METHOD

In this section compares SBS and syllables and diphones based segmentation. The evaluation do by MOS method on five person (3 men, 2 women). The persons hear produced word by different methods and determine score between 1-5 that 1 is bad and 5 is excellent. The results are shown in table 1.

Table 1. results of MOS test

Audience	Intelligibility	Naturalness	Fluidity
M1	5	3	4
M2	4	4	3
M3	5	3	4
W1	5	3	4
W2	5	4	4
Mean	4.8	3.6	3.8

Table2. most test on triphone model and dicision tree [11]

Applied method	Intelligibility	Naturalness	Fluidity
Decision tree	4.2	4.4	4.1
Triphone model	3.8	3.9	3.5

The result is acceptable. In table 2 is seen that syllable based method has been concatenated by PSOLA algorithm and intelligibility of SBS is properly.

REFERENCES

[1]. W.Barkhoda, B.ZahirAzami, O-K.Shahryari1. A comparison between allophone, syllable, and diphone based TTS systems for Kurdish language.

[2]. T. Styger and E. Keller, *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges Formant synthesis*, In Keller E. (ed.), 109-128, Chichester: John Wiley, 1994.

[3]. D. H. Klatt, "Software for a Cascade/Parallel Formant Synthesizer," *Journal of the Acoustical Society of America*, Vol 67, 971-995,1980.

[4] W. Hamza, *Arabic Speech Synthesis Using Large Speech Database*, PhD. thesis, Cairo University, Electronics and Communications Engineering Department, 2000.

[5] R. E. Donovan , *Trainable Speech Synthesis*, PhD. thesis, Cambridge University, Engineering Department, 1996.

[6] S. Lemmetty, *Review of Speech Synthesis Technology*, M.Sc Thesis, Helsinki University of Technology, Department of Electrical and Communications Engineering, 1999.

[7] A. Youssef , et al, "An Arabic TTS System Based on the IBM Trainable Speech Synthesizer," *Le traitement automatique de l'arabe*, JEP–TALN 2004, Fès, 2004.

[8] H. R. Abutalebi and M. Bijankhan, "Implementation of a Text-to- Speech System for Farsi Language," *Sixth International Conference on Spoken Language Processing (ISCA)*, 2000.

[9] A. Sharifova .A COMPARISON BETWEEN ALLOPHONE, SYLLABLE, AND DIPHONE BASED TTS SYSTEMS FOR AZERBAIJAN LANGUAGE Cybernetic Institute of Azerbaijan National Academy of Sciences 29, F.Agayev str., AZ1141, Baku, Azerbaijan

[10].Y. Samareh. *Phonetics of Farsi Language*, Iran, Tehran, Academic Press Center, 1995. (in Farsi)

[11]. Mm.Homayoonpour , musavi.sm, produce speech syntheses parameter in Persian by HMM and decision tree.2004.