International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 5 (2017), pp. 891-898

http://www.ripublication.com

Generation of Diphone Database for Telugu

D. Nagaraju
Research Scholar, Bharatiyar University, Coimbatoor ,Tamilanadu, India.

Dr. R.J .Ramasree
Professor Dept of Computer Science, RSVP, Tirupati, AP, India.

Abstract

This paper describes the process of building Telugu diphone database. Diphone database developed an acoustic data base of Telugu diphones by studying the different possible combinations of phonemes to list all the possible diphones. The diphone is an acoustic unit which requires the combination between two phonemes. Diphones are speech units that begin in the middle of the stable state of a phone and end in the middle of the next phone. Diphones concatenation is efficient method in speech synthesis (TTS) of many languages. Consider the pronunciation of – అమ్మ- amma . It consists of phonemes అ [a], ◌ం [m], ◌ం [m], అ[a]. The diphones generated while pronouncing the above are a) – అ [-a] b) అ ◌ం [am], c) ◌ం ◌ం [mm] d) ◌ం అ [ma] e) అ- [a-]. The advantage of a diphone database is fast to build and maintain but results usually in less good output speech quality[09].

Keywords : Telugu – concatenation – diphone database –TTS - etc

I. INTRODUCTION

Several developments in TTS have produced synthesizers with high intelligibility but the sound quality is still a major problem. Voice quality directly impacts on speech synthesis system [8]. This paper presents the building of Telugu diphones database for TTS systems supported by MBROLA frame work. This work focused on Telugu language.

II RELATED STUDIES

Romanian diphone database has 1156 diphones including 30 allophones. It is designed with 684 diphones[09]. Arabic language has 28 consonants, 3 long vowels and 3 short vowels. Arabic diphone database was designed with 1190 diphones[11]. Lithuanian language has 58 phonemes and to get high quality 29 stressed phonemes are added. Two [It1&It2]Lithuanian database created with 5000 diphones[12]. Bengali diphone database has 44 phones and 1936 diphones[02]. Korean diphone database was designed with 1853 diphones out of possible 3977 diphones, these diphones are extracted from 400 sentences text corpus [03]. .

III TOOLS USED

A. PRAAT TOOL

Praat is a software package, useful to develop, analysis and reconstruct of speech signal in phonetics. It was designed by Paul Boersma and David Weenink of the University of Amsterdam[04]. It is freely available for most platforms. It includes articulator synthesis, spectrographic analysis etc, operates on various operating systems like Windows, UNIX, Linux and Mac [05]. Advantage of praat is easy interface and default options try to learn. Praat tool can perform waveforms generation, intensity contour, pitch tracks, recordings, edit recorded sound, extract sounds, get pitch, intensity, draw a plot etc [06].

B. MBROLATOR

The Mbrolator, is a software suite for MBROLA supported database. Input to the Mbrolator is 1. Diphones are in wav format files 2. diphone database file in the SEG format. Output is Mbrola supported diphone database.

Diphones follows the following three rules to generate Diphone database [10]

1. The sampling rate of diphone WAV files is 16 KHz.
2. Maximum of 10000 diphone samples.
3. For each diphone a context of 500 samples needs to be left on the left and on the right sides [07].

If the rules 1 or 2 are violated, the Mbrolator will exit and no database will be created. The rule 3 give easy to pitch extraction and which is large enough for the correct analysis.

IV STEPS TO CREATE A DIPHONE DATABASE

Creation of diphone database is mainly achieved in four steps [01], they are

A. CREATING A TEXT CORPUS:

A list of phones [P], including allophones for Telugu language is prepared. Indian language scripts are originated from the ancient Brahmi script. Akshara are depends on its composition of consonants and the vowels. Properties of Aksharas are as follows (1) An Akshara is an orthographic representation of a speech sound in an Indian language; (2) Aksharas are syllabic in nature (3) The typical forms of Akshara are V, CV, CCV and CCCV, thus have a generalized form of C*V[02]. There are 16 Vowels and 36 consonants in the Telugu language. Total phones in Telugu language is 52[13].

A list of diphones [D] is generated, general total diphones are $|D| \leq |P|^2$. Diphones are adjacent pair of phones. There are 52 phones in Telugu. Maximum possible diphones are 2704(52*52= 2704)[13]. Diphones in Telugu language are /-a/-aa/-i/-ii/-u/-uu/-r/-ru/-e/-ee/-ai/-o/-oo/-au/-am /-ah/-k/-kh/-g/-gh/-~N/-c/-ch/-j/-jh/--n/-T/-Th/-D /-Dh/-N/-t/-th/-d/-dh/-n/-p/-ph/-b/-bh/-m/-y/-r/-l/ -v/-L/-S/-Sh/-s/-h/-Ra/-kS/a-a/a-aa/a-i/a-ii/a-u/a-uu/a-r/a-ru/a-e/a-ee/a-ai/a-o/a-oo/ a-au /a-am/a-ah/a-k/a-kh/a-g…… … … ……. … ../kS-s/kS-h/kS-Ra/kS-kS/.

B. RECORDING THE CORPUS:

Selection of speaker is an important task in speech recording. In generally a speaker have clear and more consistent voice. Professional speakers are general better for synthesis than non-professional.

The recordings were made at semi-professional recording studio in two sessions. Recordings are made without interruption and maintaining constant pitch. In this study a set of 106 sentences/622 words are recorded and 1356 diphones are extracted. The signals were sampled at 16 KHz and quantified 16 bits per sample. The speech is digitally recorded and stored in a digital format.

C. SEGMENTING THE CORPUS:

Diphones are extracted either manually or automatic system from speech signal. In this design manual system is used to build Telugu diphone database. Phones are searched in first step; Diphones are identified in the next step as shown in fig 1&2. Extract the identified diphones from annotated speech signal as shown in fig 3. Text grid file saved from praat tool as shown in figure 5. Text grid file gives the time intervals of the annotated speech signal. Segmented database file was created [SEG file]. Diphone database SEG file contains Diphone Name, 1st Half phone ,2nd half phone, Diphone starting time, diphone ending time and diphone boundary time. Example SEG file as shown in fig 4

Example word :- ullipaaya

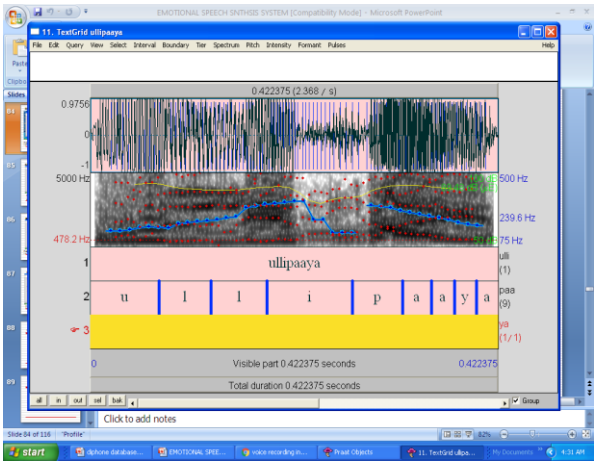

Figure 1. Phone selection

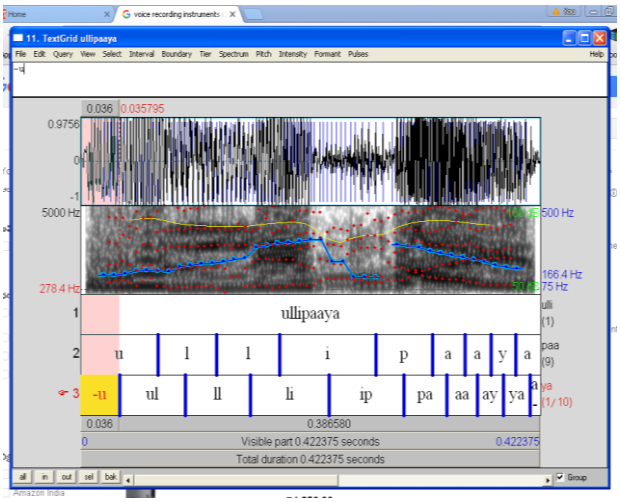

Figure 2. Diphone identification

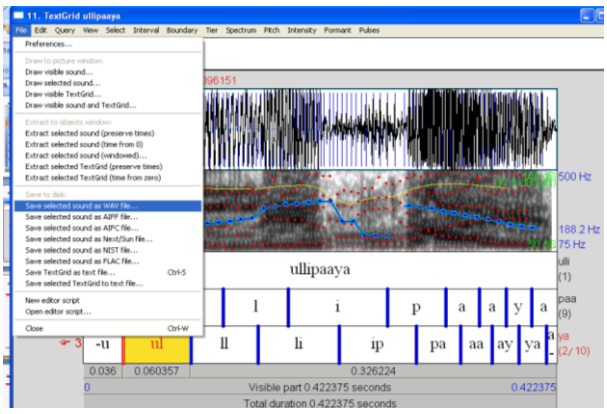

Figure 3. Diphone extraction- save selected file in wav format

Dipone Name	1st Half phone	2nd Half phone	Diphone Startng Time	Diphone Ending Time	Diphone Boundary Time
-u.wav	-	u	2000	2904	2452
ul.wav	u	l	2000	2878	2439
ll.wav	l	l	2000	2810	2405
li.wav	l	i	2000	3026	2513
ip.wav	i	p	2000	2945	2472
pa.wav	p	a	2000	2783	2391
aa.wav	a	a	2000	2918	2459
ay.wav	a	y	2000	2810	2405
ya.wav	y	a	2000	2918	2459
a-.wav	a	-	2000	3107	2553

Figure 4. Sample SEG file

sample text grid file : ullipaaya

```
File type = "ooTextFile"
Object class = "TextGrid"
    item [3]:
        class = "IntervalTier"
        name = "diphone"
        xmin = 0
        xmax = 0.424
        intervals: size = 10
        intervals [1]:
            xmin = 0
            xmax = 0.032675394533995844
            text = "-u"
        intervals [2]:
            xmin = 0.032675394533995844
            xmax = 0.087632992122896 1
            text = "ul"
        intervals [3]:
            xmin = 0.087632992122896 1
            xmax = 0.138519656557063
```

```
        text = "ll"
    intervals [4]:
        xmin = 0.138519656557063
        xmax = 0.21138936002679
        text = "li"
    intervals [5]:
        xmin = 0.21138936002679
        xmax = 0.278152663764417
        text = "ip"
    intervals [6]:
        xmin = 0.278152663764417
        xmax = 0.31479106215701713
        text = "pa"
    intervals [7]:
        xmin = 0.31479106215701713
        xmax = 0.34369468755562393
        text = "aa"
    intervals [8]:
        xmin = 0.34369468755562393
        xmax = 0.3742266862161241
        text = "ay"
    intervals [9]:
        xmin = 0.3742266862161241
        xmax = 0.4047586848766242
        text = "ya"
    intervals [10]:
        xmin = 0.4047586848766242
        xmax = 0.424
        text = "a-"
```

Figure 5. Sample Text Grid file

D. EQUALISING THE CORPUS:

The energy levels at the beginning and at the end of a segment are modified in order to eliminate amplitude mismatches – the energy of all the phones of a given phoneme is set to phones' average value [11].

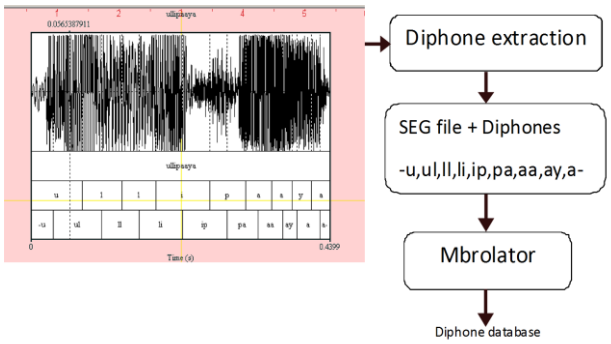

Figure 6. General architecture of generation of DDB

V. EVALUATION

The extracted diphones are then put forward to the evaluation stage in which both database files and the diphone files are evaluated by manually. The annotation diphone files allow manually comparison of the annotations with signals in the diphone files. Mbrolator generates diphone database by taking Database SEG file and all available diphones in the given language.

VI. CONCLUSION

In this paper we presented Telugu male diphone database. Quality of the diphone database is depends on the Voice quality, recording medium, type of content and manner of segmenting. Evaluation experiments showed that minimizing distortion at diphone junctions generally increased the naturalness of the output speech. However, we noticed exceptions to this rule for some particular words; other distance measures must be tested too. Unit selection method uses a costly database and requires a very large corpus of natural speech to extract the units. The run-time choice of the best instance makes this method inherently slower, but a better speech quality usually results.

REFERENCES

[01] Dutoit, T., Pagel, V., Pierret, N., Bataille, F. & van der Vrecken O. 1996. The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. In: Proceedings of ICSLP 96, vol. 3. Philadelphia, pp.1393-1396

[02] http://kathak.sourceforge.net/diphone.htm

[03] Kyuchul Yoon " A prosodic Diphone database for Korean Text to Speech synthesis system" A. Gelbukh (Ed.): CICLing 2005, LNCS 3406, pp. 425–428, 2005._c Springer-Verlag Berlin Heidelberg 2005

[04] http://web.stanford.edu/dept/linguistics/corpora/material/PRAAT_ workshop_manual v42 1 .pdf

[05] https://en.wikipedia.org/wiki/Praat

[06] http://ec-concord.ied.edu.hk/phonetics _and_ phonology /wordpress/ learning_ web site/chapter_1i ntroduction_ new .htm #1.2.1

[07] Dutoit, T. 2005. The MBROLA project. , accessed on 2010-09-19

[08] Aymen El Kadhi, Fadhila Gherri, Hamid Amiri. Building diphone database for Arabic text to speech synthesis system. ieeexplore.ieee.org/iel7/7194873/7232976/07233151.pdf

[09] diphone database development for a Romanian language TTS system by Dragos Burileanu, Adrian Neagu and Corneliu Burileanu. Printed and published by the IEE, Savoy Place, London WC2R OBL, UK.

[10] J. Bachan, "Efficient diphone database creation for MBROLA, a Multilingual Speech Synthesiser", XII International Phd workshop,OWD 2010, pp 23-26, October 2010.

[11] Aymen El Kadhi, Fadhila Gherri, Hamid Amiri "Building diphone database for Arabic text to speech synthesis system "http://ieeexplore.ieee.org/document/ 7233151/

[12] Pijus Kasparaitis "Diphone database for Lithuanian Text to Speech synthesis", INFORMATICA, 2005, Vol. 16, No. 2, 193–202 193 2005 *Institute of Mathematics and Informatics, Vilnius*

[13] Dr.RJ Ramasree and D.Nagaraju " extraction of diphones for Telugu :Issues and solutions"