ISSN 2250-3153

Hindi & Telugu Text-to-Speech Synthesis (TTS) and inter-language text Conversion

Lakshmi Sahu

Raipur Institute of Technology
Raipur - 492001 (C. G.), INDIA

Abstract- In this paper, I am explaining single text-to-speech (TTS) system for Indian languages (Viz., Hindi, Telugu, Kannada etc.) to generate human voice or speech (text to a spoken waveform). In a text-to-speech system, spoken utterances are automatically produced from text. This paper present a corpus-driven text-to-speech (TTS) system based on the concatenative synthesis approach. The output generated by the proposed text-to-speech synthesis system resembles natural human voice. It accepts input in two forms: manual user entry and from file (text or MS Word document). Proposed system supports multiple way of output; direct to computer speakers, Wav file, or MP3 file. Generated output can have different accent, tone based on selected languages. The proposed text-to-speech system will be implemented in C#.Net (Windows Form Application) and runs on Windows platforms. This paper has examples for Hindi (North Indian) and Telugu (South Indian) languages' to elaborate proposed system. This It also elaborates inter-language text conversion (*not translation*). Therefore, Hindi text will be converted into Telugu text and vice-versa. The research and development of this TTS done for my M. Tech major project.

Index Terms- text-to-speech, indian language, hindi, telugu, speech synthesis, concatenation, text conversion.

I. Introduction

The function of Text-To-Speech (TTS) system is to convert the given text to a spoken waveform. This conversion involves text processing and speech generation processes. These processes have connections to linguistic theory, models of speech production, and acoustic-phonetic characterization of language. To build a voice/speech for a language text, the steps involved are as follows (elaborated in Figure 1):

- Indian Language Analysis: Preparation of phoneme & di-phoneme list used in a language. Have enumeration to represent these phones (viz. phonetics).
- Building input sound inventory to support all phoneme & di-phoneme
- Define letter to sound rules/mapping
- Text Analysis: Analysis of input text (language) and converting into phoneme enumeration.
- Getting sound file (or content) for each enumerated value and concatenating them to construct speech.
- Evaluation of resultant speech

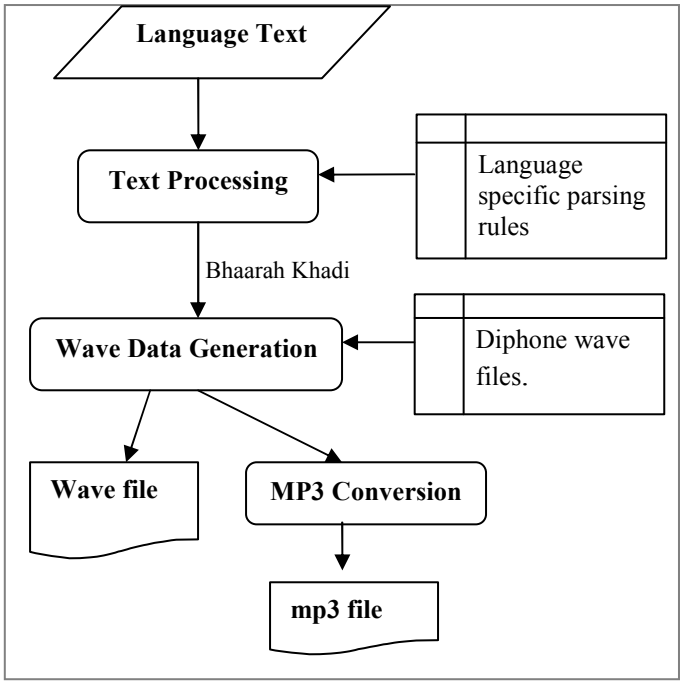


Fig. 1 Block diagram of text to wave file generation

II. Indian Language Analysis

The scripts of Indian languages have originated from the ancient Brahmi script. The basic units of writing system are characters which are orthographic representation of speech sounds. A character in Indian language scripts is close to syllable and can be typically of the following form: C, V, CV, CCV and CVC, where C is a consonant and V is a vowel. There are about 35 consonants and about 18 vowels in Indian languages.

An important feature of Indian language scripts is their phonetic nature. There is more or less one to one correspondence between what is written and what is spoken. The rules required to map the letters to sounds of Indian languages are almost straight forward. All Indian language scripts have common phonetic base.

Available character sets (windows default, UNICODE) for computers do not supports Indian languages (Hindi, Gujrati, Telugu, Kannada etc). Therefore, we use custom fonts (from different vendors, viz., Ankit (Hindi), Tikkana (Telugu)) to work with Indian languages. These fonts still use character sets like windows default or UNICODE. However, their graphical representation will be different. For example vowel V_A is represented by character "v" in Hindi and "@" in Telugu.

Table 1: Vowels in Hindi & Telugu

Alphabet	Hindi		Telugu	
V_A	अ	v	అ	@
V_AA	आ	vk	ఆ	A
V_I	इ	b	ఇ	B
V_I2	ई	bZ	ఈ	C
V_U	उ	m	ఉ	D
V_U2	ऊ	Å	ఊ	E
V_E0			ఎ	H
V_E	ए	,	ఏ	I
V_AE	ऐ	,s	ఐ	J
V_O0			ఒ	K
V_O	ओ	vks	ఓ	L
V_O2	औ	vkS	ఔ	M
V_REE	ऋ	_	ఋ	F
V_AM	अं	a	ం	=
V_AHA	अः	¿	ః	>

Table 2: Consonants in Hindi & Telugu

Alphabet	Hindi		Telugu	
C_K	क	d	క	N
C_KH	ख	[k	ఖ	O
C_G	ग	x	గ	Q
C_GH	घ	?	ఘ	R
C_ONG	ङ	³	ఙ	S
C_CH	च	p	చ	T

Table 3: Consonants in Hindi & Telugu

Alphabet	Hindi		Telugu	
C_SH	श	'k	శ	q
C_SHH	ष	"k	ష	r
C_H	ह	g	హ	v
C_TR	त्र	=		
C_GY	ज्ञ	K		
C_SHR	श्र	J		
C_KSHH	क्ष	{k	క్ష	x

English language always have vowel characters right of its associated consonant. However, in Indian languages, vowel may appear both sides (left, right) of consonant. In English, whether vowel positioned at start or mid or end of word, its appearance will not be changed (Capitalization rule is exception here.). Whereas, if vowel appear at start of word, it will have its full form. Appearance at mid or end, it will have its half form (In Hindi, we refer MAATRAA, In Telugu, GUNITALU). This will vary language to language, vowel to vowel. Table 1 shows that how vowel V_I appears in Hindi & Telugu.

Table 4: Vowel & consonant appearance in words

Hindi	Alphabet	Telugu	Alphabet
इमली	V_I,C_M, C_L,V_I2	ఇల్లు	V_I, C_L,C_L, V_U
किसान	C_K, V_I, C_S, V_AA, C_N	రాగి	C_R, V_AA, C_G, V_I

III. SOUND INVENTORY

Building a sound inventory involves making a decision on basic unit of synthesis, enumeration of phonemes, recording, labeling and finally coding the data.

A. Basic unit of synthesis

The basic unit of synthesis can be a phoneme or diphone or syllable or word or phrase or even a sentence. Theoritically, larger the basic unit, fewer will be the concatenation points during synthesis and better the quality of produced speech. I have used diphone as unit of synthesis to have optimal size of sound inventory and maintaining quality of synthesized speech.

B. Enumerating the diphone set

In Hindi, combination of a consonant & a vowel sound/phone/alphabet and placing it into a table, referred as **Baarah Khadi**. As per best of my knowledge, Hindi language has enough vowel, consonant and Baarah Khadi sounds to represent (sound) all Indian languages sounds. Therefore, if we provide a unique identifier to each tabular entry in Baarah Khadi then we refer it as language neutral sound/phone identifier (similarly, Phonetics in English language).

Table 5: Bhaarah Khadi ID and language symbols

Bhaarah Khadi	Hindi		Telugu	
BK_A	अ	v	అ	@
BK_AA	आ	vk	ఆ	A
BK_I	इ	b	ఇ	B
BK_I2	ई	bZ	ఈ	C
BK_K	क	D	క్	N+
BK_K_A	क	d	క	N{
BK_K_AA	का	dk	కా	N}
BK_K_I	कि	fd	కి	N¡
BK_K_I2	की	dh	కీ	N¤

Compare to other Indian languages, still we have few vowels and consonants that not exists in Hindi. like, V_O0, V_O0 etc. This can overcome by adding these additional alphabets into proposed system's Baarah Khadi. I have considered a phone-set which is a super set of Hindi and Telugu languages.

Thus, with a 40-phone inventory, one could collect a 16 * [1(V) + 38(C)] = 646 diphone inventory and create a synthesizer that could speak anything, given the imposition of appropriate prosody.

C. *Diphone Database Construction*

Designing, recording, and labeling a complete diphone database is a laborious and a time consuming task. The overall quality of the synthesized speech is entirely dependent on the quality of the diphone database. I have recorded voice (wave file; maintaining constant pitch, volume, and speech rate during the recording) for each phoneme & di-phone (each Baarah Khadi entries has its own sound clip file). Which helps to increase quality of output speech. All the samples were recorded at sampling rate of 48 kHz (16 bit sample; mono channel and Audio format is PCM). The recorded samples were segmented manually with WavePad sound editor. After labeling, the segmentation results were visually inspected and corrected by checking all results using WavePad sound editor.

IV. LETTER TO SOUND RULES/MAPPING

Font Characters Mapping

Mapping is required for font characters of a Indian language to represent vowel and consonant alphabet. Details of vowel appearance (before or after) with respect to consonant is also needed while parsing language text. There are also other challenges. There will possibility in different fonts that a vowel or consonant or Bhaarah Khadi may have multiple symbols or characters. Table 5 shows (partial list of Baarah Khadi) mapping of Baarah Khadi, language's alphabet and its associated Unicode characters. I am maintaining Baarah Khadi mapping into XML files to read programmatically language details.

```
<FontDefinition FontName="Ankit">
 <BK ID="BK_A" Type="V" SymbolA="v"></BK>
</FontDefinition>
<FontDefinition FontName="Tikkana">
 <BK ID="BK_A" Type="V" SymbolA="@"></BK>
</FontDefinition>
```

Conversion of One language text to another

As each language has its own font character mapping definition (alphabet & Baarah Khadi), it is each to convert one language text into other language text by using Baarah Khadi (Intermediate processing results while text to speech conversion).

V. INPUT TEXT ANALYSIS & PROCESSING

The text-to-speech conversion process can be divided into following stages:

A. *Text Entry*

If input text is provided by manual entry, then there should be form or UI to read it. I have used Windows forms for user entry.

B. *Text Analysis*

Text analysis is the task of identifying words in the text. The first task in text analysis is to make chunks out of input text - tokenizing input text. At this stage, the input text is also chunked into reasonably sized utterances. For many languages, tokens are white space separated and utterances can, to a first approximation, be separated after full stops, question marks, or exclamation points. Apart from chunking, text analysis also does text normalization. Text normalization includes "Token Identification" which is the task of identifying special symbols, numbers and "Token to Words" which convert the identified tokens to words for which there is a well defined method of pronunciation.

C. *Pronunciation*

Having properly identified the words, their pronunciation can be found by looking them up in a lexicon, or by applying letter-to sound rules to the letters in the word. For Hindi & Telugu languages pronunciation can almost completely be predicted from their orthography, pronunciation can be found using a set of Letter-to-Sound (LTS) rules (Baarah Khadi mapping).

VI. SPEECH CONSTRUCTION (WAVE OUTPUT GENERATION)

There are four basic approached to synthesizing speech, namely waveform concatenation, articulatory synthesis, formant synthesis and concatenative synthesis. One of the common approach to synthesis is Concatenative Synthesis.

A. *Concatenative Synthesis*

Concatenative synthesis uses actual short segments of recorded speech that were cut from recordings and stored in an inventory (‘‘voice database’’), either as ‘‘waveforms’’ (uncoded), or encoded by a suitable speech coding method. It involves taking real recorded/coded speech, cutting it into segments, and concatenating these segments back together during synthesis.

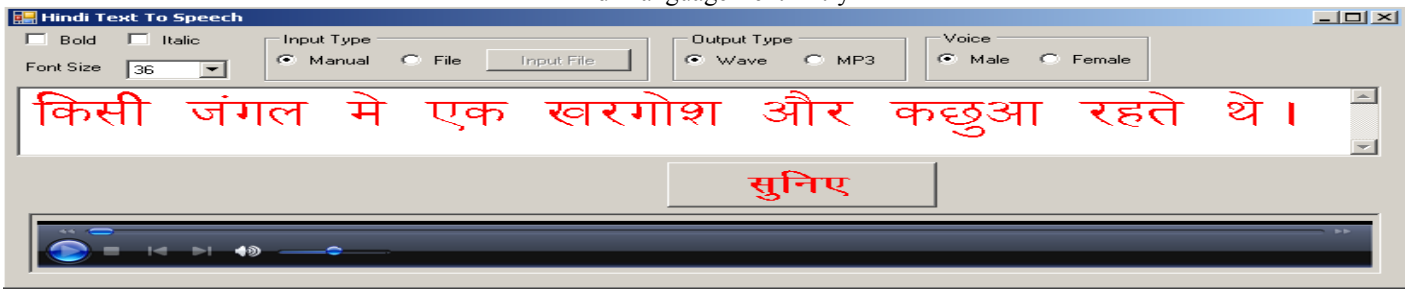
B. *Sound Generation Rules*

Sound rules vary from language to language. To synthesize a particular language, required units (diphones) from the database which doesn't contain any language specific information and these selected units were then typically altered by signal processing functions to meet the language specific target specification generated by different modules in the synthesizer. This need to be considered while converting Baarah Khadi to sound for a language. For example, when Telugu speaker speaks a word which ends with consonant (no MAATRAA; default V_A phone) then speaker will give stress to V_A sound. Sound V_A is not stresses by Hindi speakers.

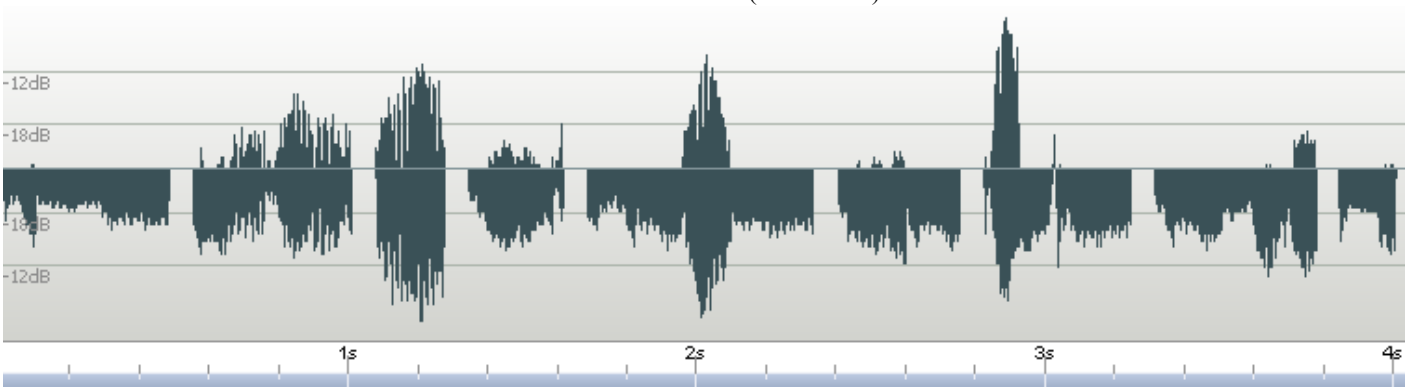
VII. EVALUATION OF RESULT

This implementation has been tested with text from Rabbit and Tortoise story for both Hindi & Telugu language.

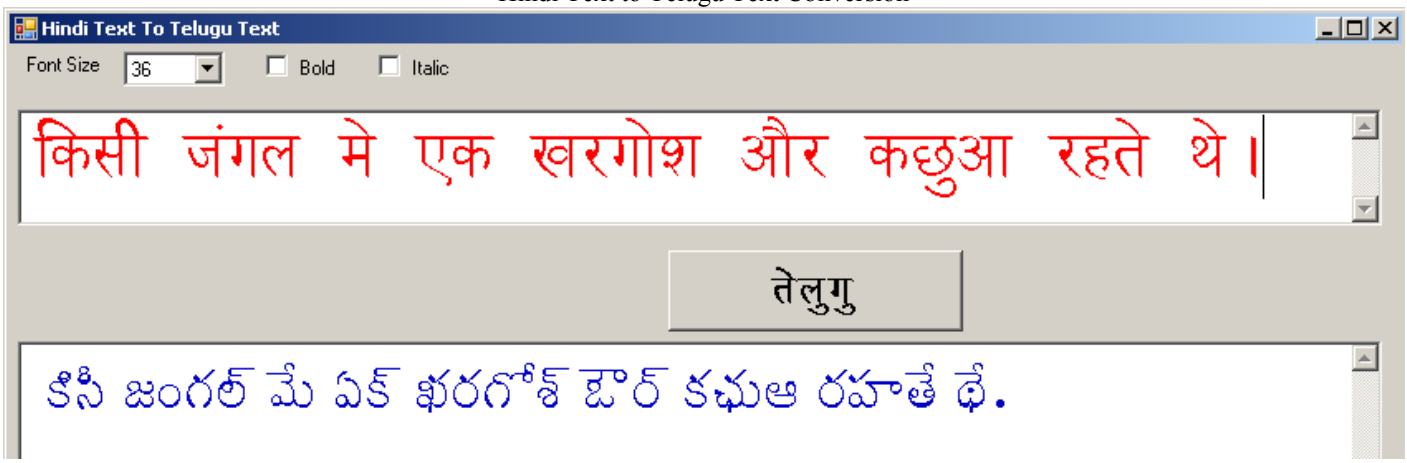
Hindi Language Text Entry



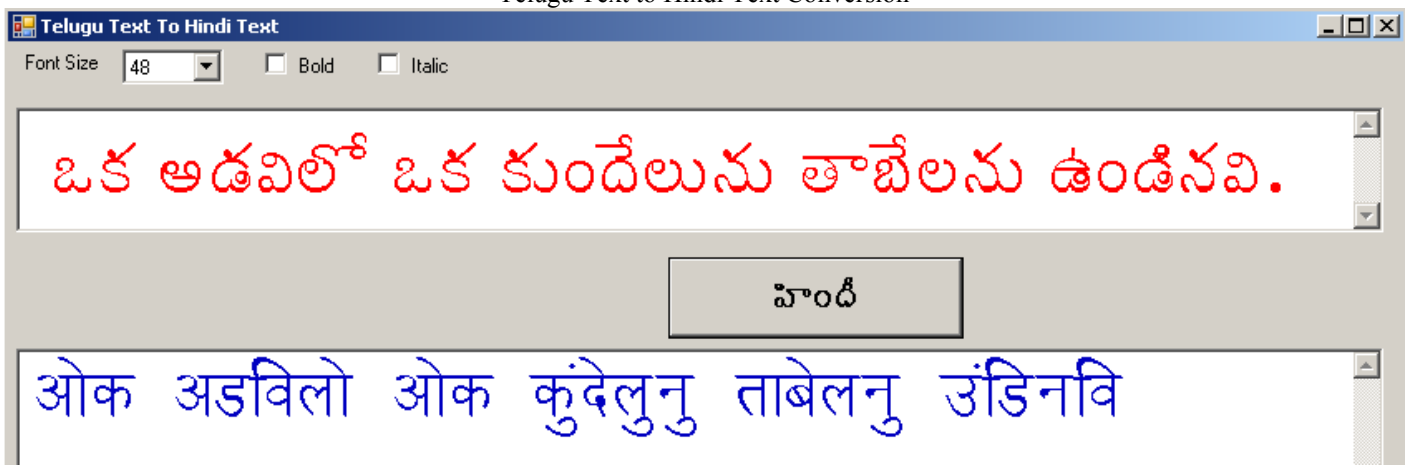
Generated wave file (wave form)



Hindi Text to Telugu Text Conversion



Telugu Text to Hindi Text Conversion



VIII. CONCLUSION

This system has limited to two voices (a male and a female). This can be enriched with multiple voices (for both male and female) with minor changes in code and reach sound inventory (recording voice samples). For now, its implementation is limited to support Hindi & Telugu. To support new language, say Kannada, it require Font Characters Mapping and little extra C# coding to plug language text parsing and pronunciation rules. This system has difficulty to support language like Tamil which words are pronounced based on context.

REFERENCES

[1] N. Sridhar Krishna, Hema A.Murthy and Timothy A. Gonsalves, ―"Text-to-Speech (TTS) in Indian Languages", Int. Conference on Natural Language Processing, ICON-2002, Mumbai, pp. 317.326, 2002.

[2] Juergen Schroeter — "Text to-Speech (TTS) Synthesis", AT&T Laboratories

[3] S.P. Kishore, Rajeev Sangal and M. Srinivas — "Building Hindi and Telugu Voices using Festvox", Language Technologies Research Center International Institute of Information Technology Hyderabad.

[4] Ruvan Weerasinghe, Asanka Wasala, Viraj Welgama and Kumudu Gamage — "Festival-si: A Sinhala Text-to-Speech System", Language Technology Research Laboratory, University of Colombo School of Computing, Colombo, Sri Lanka

[5] R. Gupta, M. V. Shastri, Rapidex Language Learning Series — "Hindi – Telugu learning course", Pustak Mahal.

AUTHORS

First Author – Author name, qualifications, associated institute (if any) and email address.

Correspondence Author – Author name, email address, alternate email address (if any), contact number.