IITB-Monash Research Academy

An Indian-Australian research partnership

Indian Institute of Technology Bombay | MONASH University

Synthesizing Audio for Hindi WordNet

Diptesh Kanojia[1,2,3], Preethi Jyothi[2], Pushpak Bhattacharyya[2]

[1] IITB Monash Research Academy, IIT Bombay, Powai, Mumbai, India

[2] Department of Computer Science and Engineering, IIT Bombay

[3] Faculty of Information Technology, Monash University

Introduction

Speech synthesis is the computer-generated simulation of human speech. It is used to translate written information into aural information which is more convenient for readers.

We choose to perform unit selection synthesis and build cluster units of the speech data recorded by a human voice. We use the Festival system to create a synthetic voice for Hindi. We use preexisting “voices”, use publicly available speech corpora to create a “voice” using the Festival Speech Synthesis System (Black, 1997)

Motivation

“Our goal is to enrich the semantic lexicon of Hindi WordNet by augmenting it with word audios generated automatically using a speech synthesis voice model.”

We envision that this addition to Hindi WordNet will further its use in the education domain, for students and language enthusiasts alike.

Dataset

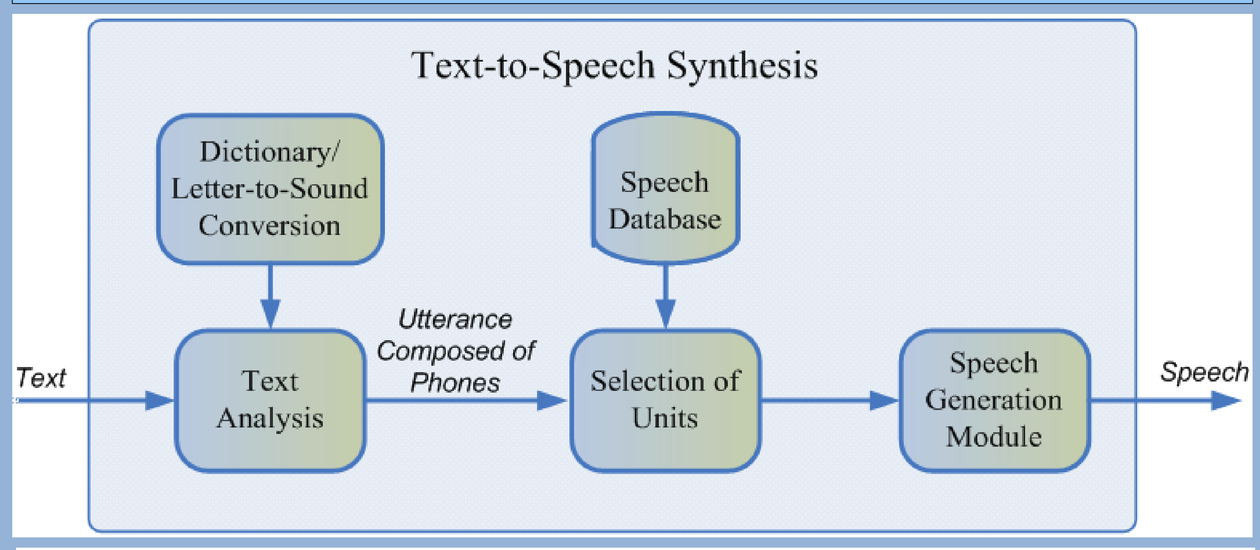
We use the Female Voice - Hindi and Female Voice – English dataset provided by the IndicTTS forum to train our system. We download 7.22 hours of Audio with English and 5.18 hours of monolingual audio. We use a total of 2318 Female Hindi sentence utterances downloaded from IndicTTS consortium, and 1378 word audios manually recorded by us to train the voice model.

Technique	Explored	Voice Models Generated	Usable for Hindi TTS
Festival+FestVox (IndicTTS Data)	***Yes***	***Many***	***Yes***
Flite Voice (Hindi - Female)	Yes	1	Yes
Flite Voice (Hindi - Male)	Yes	0	Yes
Flite Voice (Marathi - Female)	Yes	0	Yes
Flite Voice (Marathi - Male)	Yes	0	Yes
Festival (diphone)	Yes	1	Yes
Wavenet (basveeling)	Yes	1	No
DeepVoice	Yes	0	No
Merlin	No	0	No
MaryTTS	No	0	No
Tacotron	No	0	No
SampleRNN	No	0	No
Char2Voice	No	0	No

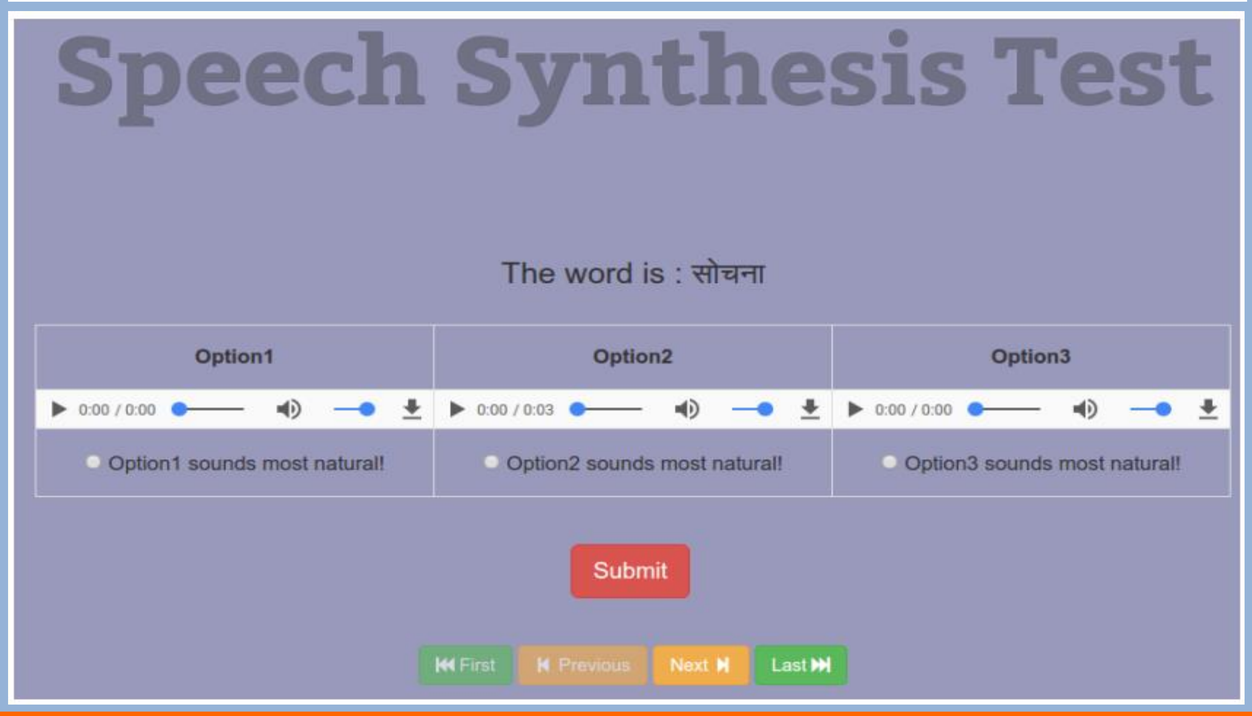
Experiments/Models

We train and create multiple models for speech synthesis of Hindi and compare the audio produces with each other by manual evaluation through human annotators.

We create a web interface and provide option for selecting the best audio (most “natural” sounding) to the annotator. The results are shown in the table below. We experiment with neural speech synthesis as well, but could not produce a model due to lack of data.



	#0	#1	#2	#1+#2	Most Liked
Model 1	79	55	99	154	**101**
Model 2	37	78	112	190	90
Model 3	72	86	58	144	51
Model 4	**55**	**117**	**107**	**224**	70



Results & Conclusion

We receive a total of 442 responses for 30 word samples. Thus, we assume that 14 people had completed the test. The results of our initial evaluation based on naturalness are as follows: (i) **The mean of our voice model win percentage is over 44%**. We beat both the other voices by an acceptable margin, (ii) **Prerecorded speech by humans was rated best somewhat less than 30%** of the times, and (iii) Grapheme based synthesized speech **scored around 26% on this scale.**

We generate **audios for 151831 words, and 40337 synset glosses/example sentences.**

e-mail: diptesh@cse.iitb.ac.in url: http://www.cse.iitb.ac.in/~diptesh