Dhvani
Indian Language
Text To Speech System

Santhosh Thottingal
Swathanthra Malayalam Computing
http://smc.org.in

Agenda

- Text to speech system – A brief introduction
- Dhvani Introduction
- Algorithm and Architecture
- Demo –Malayalam, Hindi, Kannada,Oriya, Gujarati, Bengali, Telugu, Panjabi
- How to add a new Language support
- Discussion on the Front ends, Integration with other application

Text to Speech Systems

- Speech Synthesis: Artificial production of human speech
- Quality Measured by similarity to human voice and intelligibility
- History: There was Mechanical and Electronic attempts starting from 1000 AD- “Speaking Heads”
- Intonation and Prosody- Still a research area

TTS- Technologies

- **Concatenation Technologies**:Concatenation of segments of recorded Speech- More Natural
  - **I. Unit selection Synthesis:** Use large db of phones, syllables,words, sentences, pitch, duration, position...More Natural but big Database...Candidate selection using decision trees at runtime.
  - **II. Diphone synthesis:** Only diphones stored, Concatenation using DSP techniques. DB size depends on phonotactics of language. Eg: PSOLA, MBROLA
  - **III. Domain Specific Synthesis-** For limited number of words- Eg: Talking Clocks, Calculators

Dhvani is a Diphone Concatenation based TTS

TTS- Technologies

- Formant Synthesis
  - No database of speech
  - Artificial wave form creation using fundamental frequency, voicing, noise levels..
  - Also called rule-based synthesis
  - Eg: Speak & Spell by Texas Instruments(1970s)
- Articulation Synthesis: Based on the speech models of the human vocal tract and the articulation processes occurring there
- HMM Based Synthesis: Speech wave creation using Hidden Markov Model using Frequency Spectrum(Vocal tract), Fundamental frequency(vocal source) and duration(prosody)

Dhvani TTS

- Started as a part of Simputer Project.
- Designed By Dr. Ramesh Hariharan, IISC
- Sound database developed at IISC Bangalore
- Language Independent Design
- Based on Diphone concatenation technology
- Project was inactive for the last 5 Years.
- An attempt in India to Cover all Indian languages under a single framework
- A GPLed Project for GNU/Linux Platform

Supported Langauges

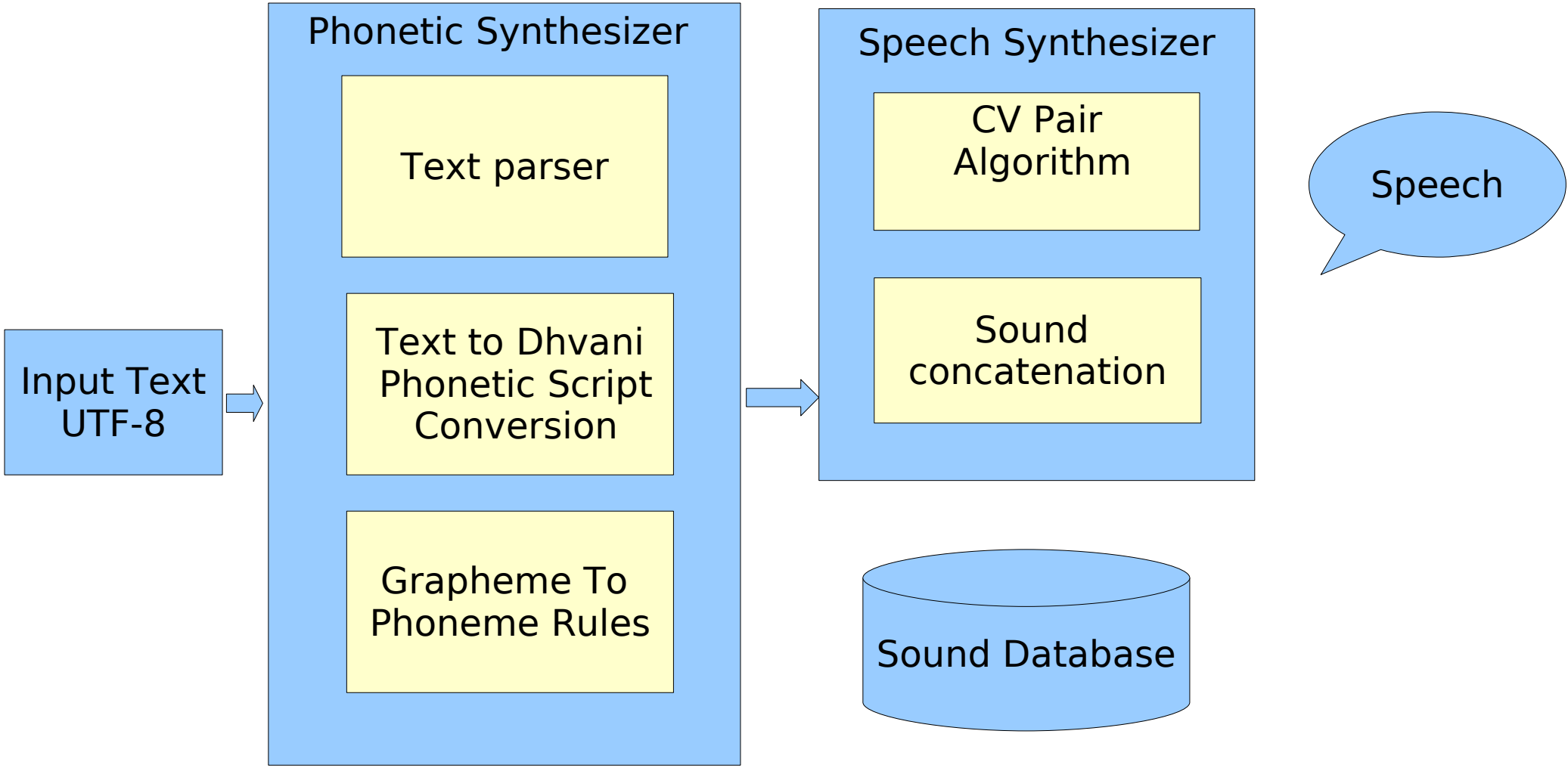
Language Modules exist for the following Languages

- Malayalam
- Hindi
- Punjabi
- Kannada
- Oriya
- Gujarati
- Bengali
- Telugu

Algorithm

- Based on the Observation that a Direct G2P mapping exits for all Indian languages in general.
- Each language requires a Unicode parser.
- A UTF to phonetic conversion system converts the Text to a phonetic script- Dhvani specific.
- Speech synthesizer takes phonetic script and concatenates the sound files to produce speech
- words are identified by space, comma, full stop, new line etc...
- Pause at each word gap, new line, paragraph

Architecture



Text to Dhvani Phonetic Script

- This makes Dhvani language independent
- Any unicode text will be converted to a common script.
- This script is the input to the Speech Synthesizer
- Examples:

  * khana (food in hindi) kh2 n2 (CV CV)
  * maun (silence in hindi) m13n (CVC)
  * kahaan (where in hindi) k1 h2an (CV CVC)
  * pratibha (talent in hindi) pHr1 t3 bh2 (HCV CV CV)
  * sankalp (resolution in hindi) s1n k1l 0p (CVC CVC 0C)
  * chandramaa (the moon in hindi) ch1n dHr1 m2 (CVC HCV CV)
  * praan (life in hindi) pHr2n (HCVC)
  * mysore (as pronounced in kannada) m10 s6 r5 (CV CV CV)
  * rashtr (nation in hindi) r2sh 0tt 0r (CVC 0C 0C)
  * aadesh (instruction in hindi) 2 d8sh (V CHC)
  * andaaz (style in urdu) 1n d2z (VC CVC)
  * ahimsa (nonviolence) 1 h3n s2 (V CVC CV)
  * vazhapazham (banana in tamil) v2 zh1 p1 zh1m (CV CV CV CVC)

Grapheme to Phoneme Conversion

- The phonetic description is syllable based.
- 8kinds of sounds are allowed (C -consonant, V -Vowel, H -Half Sound).
- V: a plain vowel
- CV: a consonant followed by a vowel
- VC: a vowel followed by a consonant
- CVC: a consonant followed by a vowel followed by a consonant
- HCV: a half consonant, followed by a CV
- HCVC: a half consonant, followed by a CVC
- 0C: a consonant alone
- G[0-9]*: a silence gap of the specified length (typical gaps

Vowels

vowels allowed are:

a as is pun
aa as in the hindi word saal (meaning year)
i as in pin
ii as in keen
u as in pull
uu as in pool
e as in met
ee as in mate
ae as in mat
ai as in height
o as in the tamil word ponni (meaning gold)
oo as in court
au as in call
ow as in cow
tamil-u : as in the tamil aanddu (meaning year)

- The phonetic description uses the numbers 1-15 instead of the pnemonics given above.

Consonants

Consonants are:
k kh g gh
ch chh j jh
t th d dh n
tt tth dd ddh nna
p f b bh m
y r l ll v sh s h
zh z

- These consonants are numbered 1..34. the phonetic description however uses the pnemonics above. Within the program and in the database nomenclature, the numbers are used.

Sound Database

- All sound files stored in the database are gsm compressed .gsm files.(GSM standard by The Communications and Operating Systems Research Group (KBS) at the Technische Universitaet Berlin)
- Recorded at 16KHz as 16bit signed linear samples.

The following sound units are stored in the database

- CV pairs: 1..33 * 2 4 6 8 9 10 12 13 14 15
- VC pairs: 2 4 6 8 9 10 12 13 14 15 * 1..34
- V: 1..14
- C: 1..34
- Halfs: ky kr kl kll kv ksh khy khr khl khv gy gr gl gv gn ghy ghr ghv ghn chy chr chv jy jv ty tr tv thy thr dy dr dv dhy dhr dhv ny nr nv tty ttr ttv ddy ddr ddv py pr pl pll fr fl by br bl bhy bhr bhl my mr vy vr vl

The total size of the database is around 1MB

Sound Concatenation

- CV files are named x.y.gsm where x is the consonant number and y is the vowel number.
- VC files are named x.y.gsm where x is the vowel number and y is the consonant number.
- V files are named x.gsm where x is the vowel number.
- Halfs files are named x.y.gsm where x,y are the two consonants involved.
- 0C files are named x.gsm where x is the consonant number.
- All files other than the 0C files have been pitch marked and the marks appear in the corresponding .marks files, one mark per byte as an unsigned char.

Sound Concatenation

- In addition to the sound files, there are four files in database/, namely cvoffsets, vcoffsets, voffsets and hoffsets, which store various attributes of the sound files.
- **cvoffsets**

CV fields:

start(start of the cv)
diphst(diphone start position: default halfway to ctov from start)
ctov(cons to vowel change position)
longvowlen(length of long vowel, currently not really used)
shortvowlen(length of short vowel)
diphend(end of diphone for long vowel, short will be obtained from long)
diphshortfactor(factor for getting short diphone from long)
halfst(place where this cv is cut to connect to previous half)

Sound Concatenation

vcoffsets

- VC fields:
  - end(end of vc)
  - diphend(diphone end position: default halfway from ctov to end)
  - vtoc(vowel to cons change position) longvowlen(length of long vowel, currently not really used)
  - shortvowlen(length of short vowel)
  - diphst(start of diphone for long vowel, short will be obtained from long)

voffsets

- V fields:
  - length (length to be played starting from 0)

hoffsets

- Halfs fields:
  - start (start of half) end (place where this half is cut and appended to the next

Language Modules

- A language Module does the parsing, grapheme to phoneme conversion
- Input is text in Unicode format.
- Output is phonetic script
- Any logic for producing it based on the language characteristics can be done in the language module
- Dhvani can detect the languages and it dispatches the text to the corresponding phonetic synthesizer
- Multiple languages in a single input text is supported

Language Modules

- Language module can handle the number reading logic
- Acronyms, Currency, other features of language can be done.
- To write a new Language module, start with one existing one and make necessary changes.

Typical use of TTS systems

- TTS can save time and money in business, when compared to studio based pre-recorded speech files
- Telephony applications- voice portals, CRM, call centers
- In-vehicle environments to read text while driving
- Hands-busy, eyes-Busy applications in industry
- Many applications if we can develop a voice recognition system and integrating it with TTS

TODO

- More Language Modules, Testing of existing Language Modules- Developers from these languages are invited!
- Integration with Desktop Environments, Text editors etc.. Already Integrated with Gedit as an External tool
- A GUI for Dhvani
- Facility to save the speech in various sound formats. Currently it saves the file in 16 bit unsigned 16KHz PCM format
- Applications that use Dhvani as a back end for various accessibility requirements

Thanks

- Developers: Dr. Ramesh Hariharan, Santhosh Thottingal
- Download: http://sourceforge.net/projects/dhvani
- Documentation/Wiki: http://fci.wikia.com/wiki/Dhvani
- License: GPL version 2 or Later