Bilingual TTS for Tamil and English

AG Ramakrishnan, Vikram LR, Abhinava, ShivaKumar HR
Medical Intelligence and Language Engineering (MILE) Laboratory,
Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012.
agrkrish, vikram.ckm, abhinav.zozo, shivahr@gmail.com

Abstract

An unlimited vocabulary text-to-speech engine has been developed, which currently handles both Tamil and Kannada Unicode text. The input text is processed by a grapheme to phoneme converter module, which uses language specific pronunciation rules to convert the text into an unambiguous phonetic representation. This text is then parsed into demisyllable like basic units. The occurrence of these basic units are searched for, from the phonetically rich spoken database, which is segmented and annotated at the phone level. An unit selection algorithm then selects the best combination of the available speech units to be concatenated to synthesize the speech, which is then converted into .wav format.

Introduction

Text to Speech (TTS) synthesis is an automated encoding process, which converts the given text in a specific language into speech. Till date, only English and some European language TTS systems have gained commercial importance due to their quality output. This paper primarily deals with developing a modular, unit selection based TTS framework for Indian languages. Bilingual Tamil and English TTS is developed for this purpose. However, this framework can be easily modified for any other language. The TTS framework developed is concatenation based, with polyphone taken as the unit of concatenation. This framework is further optimized to suit embedded applications like mobiles and PDAs.

We designed and developed corpus-based concatenative Tamil speech synthesizer in Matlab and C. A concatenation based speech synthesizer requires a rich and large speech database with varied and natural distribution of prosodic and spectral characteristics of speech sounds. The sentences to be recorded need to be selected from a text corpus. We used CIIL (Central Institute of Indian Languages, Mysore) Tamil text corpus for our research. A greedy algorithm is used to select phonetically rich sentences from this huge corpus. This resulted in 1026 sentences, which were recorded from a professional, native Tamil speaker. These sentences are segmented offline and the database is organized in such a way that it facilitates faster search.

During synthesis, from the phonetic transcription of the sentence to be synthesized, specifications of the required target units are predicted. Units are then selected from the database that best match the target specification according to a distance metric and a concatenation quality metric. These units are then concatenated to produce synthetic speech. There may be audible glitches in the output after concatenation. This could be because of either poor segmentation of the speech database or improper selection of units by the TTS frame work. In our case, we know that the segmentation is nearly error-free. Hence, post-processing is performed on the final set of units. This includes smoothing the pitch contour,

[redacted]

during concatenation, at junctions of units with unacceptable pitch discontinuity. Our experiments reveal that about 15-20% of the unit junctions require pitch smoothing. Optimal coupling technique is then used to concatenate these units at appropriate positions. This resulted in intelligible and reasonably natural synthetic speech.

Intelligibility of synthetic speech also depends on selecting the units that match the target phonetic contexts. At times, the required phonetic context may not be available in the database. In such case, we propose that similar phones that are perceptually indistinguishable may replace these phonetic contexts. The most confused pairs of Tamil phones, which can be replaced by each other in specific contexts at the time of synthesis (if they are not available in the corpus) are found. Explorative experiments to determine the applicability of incorporating these techniques resulted in high mean opinion scores for the synthesized output from the native Tamil evaluators. Hence, we consider that this possibility of replacing missing phonetic contexts can be used in practical TTS.

Finally, when any person speaks the same sentence repeatedly, the speech waveforms don't have identical characteristics. With this motivation, the final portion of my research attempts to analyze the variability of characteristics of different instances of speech, when a speaker utters the same sentence multiple times, at different times. The idea is to look at the possibility of generating a slightly different synthetic speech each time the same text is synthesized, thus trying to make the TTS sound not monotonous and more human like.

Also, we observe that incorporating prosody and pause models for Indian language TTS would further enhance the synthetic speech quality output. These are some of the potential, unexplored areas ahead, for Indian speech synthesis.

Motivation for bilingual TTS

In the present scenario, usage of English in Tamil text has become common and inevitable. If such words are omitted in TTS, the TTS would be less effective. Hence we have developed a bilingual Tamil TTS for generating Tamil and English by using the same synthesis Tamil data for both the languages. The sparsely occurring English text is converted into phonemes using a separate grapheme to phoneme converter and the corresponding phonemes are obtained from the available Tamil database for concatenation. The initial results are encouraging and we are working on some more improvements for better sounding English.

Tamil synthesis database has 5 hours of Tamil sentences recorded by a male professional Tamil speaker. The recorded database is rich in phonetic context and phonetic variations. The TTS takes Tamil Unicode input, and performs equivalent phonetic translation. In Tamil, some letters and phonetic contexts influence the phoneme of a letter, such as "*ka*" can become "*ga*" in some phonetic contexts. The rules for these phonetic changes are coded as rules in the grapheme to phoneme converter. The phonemes are then again grouped into polyphonic cluster. The clusters are searched in database to find a best choice is selected for a given context. We have developed a prosody based unit selection algorithm to further enhance the best selection for a given text.

Features/Specifications of the TTS

1. Unlimited Vocabulary : Any sentence involving any combination of native words of Tamil (or Kannada) is handled.

2. Quality : The intelligibility of the TTS is quite high, and it is also acceptably natural.
3. Text Encoding: Only text entered in Unicode will be handled. Other proprietary or public font encodings are not handled, and will not be handled, even in the future.
4. Web Demo: The TTS has been made available as a web demo. Go to the link, http://mile.ee.iisc.ernet.in:8080/tts_demo. The TTS demo page can be seen and a box, where the Tamil or Kannada text in Unicode must be submitted.
5. Test Input to the TTS: If you want to submit your own custom text input, you can do so, by typing using our open source Multilingual Indic keyboard interface: பன்மொழி வாயில் or Vishwavaangmukha), which you can download from http://code.google.com/p/indic-keyboards
6. Output Format: The TTS outputs a standard .wav file.
7. Testing: The web demo of Tamil TTS has now been tested by hundreds of people from around the world, and many GB's of synthesized .wav files have been downloaded by them.

Current Limitations

It cannot handle numerals, proper nouns and words originating outside the current language handled by the TTS, sentences needing intonations changes, such as interrogative and exclamatory sentences. There may also be some mispronunciations at times. In the case of long sentences, even the pauses (phrase breaks) may not be at the instances, where a human speaker will naturally pause for better intelligibility and clarity.

Future Enhancement

All the current limitations mentioned above are being addressed by us, and will be updated in time. The numeral handling facility will be added very soon. Some of the causes for mispronunciations have been identified, and will be corrected next. However, further enhancement of naturalness requires significant research in computational linguistics and prosody and hence, may take considerable time to reach good quality.

Conclusion

MILE Laboratory has teamed up with Bookshare.org, an International non-profit organization, to provide Tamil and Kannada digital books (copyright free or permitted by authors) online to print-disabled people (visually challenged, old people with vision disabilities and people with other disabilities that make it impossible for holding a book and turn pages of it). The MILE OCRs and TTS in the respective language (Thirukkural / Vak) will be used for this purpose, and thus the printed content can directly be heard as speech on a desktop computer or laptop.

Acknowledgment

The TTSs were developed as part of a Research Project, funded by Ministry of Social Justice and Empowerment (MSJE), Government of India. The author also thanks Karthigeyan, Jayavardhana Rama, Prathibha, Muralishankar, Lakshmish Koushik, Parthasarathy, Arun, Abhinav, Vikram, Shivakumar and others, who contributed to the TTS research and development over a period of time.