Enabling IT Usage through the Creation of a High Quality Hindi Text-to-Speech System

Kalika Bali, N. Sridhar Krishna, Sameer Badasker, KSR Anjaneyulu
HP Laboratories India

HPL-2007-44
March 26, 2007*

text-to-speech,
local language

Emerging markets have a significant business potential for Information and Communication Technology (ICT) products provided we address the problems that are characteristic of these economies. Providing interfaces that support local language interaction is one of the major barriers that can be resolved by using appropriate technology. On the other hand, these markets are witnessing a boom in the telecom sector with unprecedented growth in the use of telephone, both landline and mobile. In this context, speech output is a natural way of dispensing information and providing access to ICT services and products.

In this paper, we describe a high quality Text-to-Speech (TTS) in Hindi created by HP Labs India. HP Labs India Hindi TTS was created with a modular and scalable framework developed using Festival (an Open Source TTS System) and language processing tools developed in the lab. The TTS engine can be used for voice-based services running on Open Call Media Platform (OCMP), applications running on iPAQ and along with digital libraries such as DSpace to provide a more natural interface for Hindi-speakers. This could allow HP products to address new application requirements using local language speech synthesis and thereby enable HP to address this market more effectively.

* Internal Accession Date Only Approved for External Publication

Enabling IT Usage through the Creation of a High Quality Hindi Text-to-Speech System

Kalika Bali, N. Sridhar Krishna, Sameer Badaskar, KSR Anjaneyulu

HP Labs India

anji@hp.com

Abstract

Emerging markets have a significant business potential for Information and Communication Technology (ICT) products provided we address the problems that are characteristic of these economies. Providing interfaces that support local language interaction is one of the major barriers that can be resolved by using appropriate technology. On the other hand, these markets are witnessing a boom in the telecom sector with unprecedented growth in the use of telephone, both landline and mobile. In this context, speech output is a natural way of dispensing information and providing access to ICT services and products.

In this paper, we describe a high quality Text-to-Speech (TTS) in Hindi created by HP Labs India. HP Labs India Hindi TTS was created with a modular and scalable framework developed using Festival (an Open Source TTS System) and language processing tools developed in the lab. The TTS engine can be used for voice-based services running on Open Call Media Platform (OCMP), applications running on iPAQ and along with digital libraries such as DSpace to provide a more natural interface for Hindi-speakers. This could allow HP products to address new application requirements using local language speech synthesis, and thereby enable HP to address this market more effectively.

Introduction

The role of Information and Communications technology (ICT) in the growth of emerging economies is crucial but is constrained by problems particular to these economies. As HP expands its market in these economies it is confronted by multilingualism coupled with limited use of English, low computer and literacy levels and the need for simpler interfaces in local languages. Some of these barriers can be resolved by providing ICT based services and products in local languages with voice enabled interfaces. These can be provided by HP directly or by HP partners with significant strengths and expertise in these areas.

In emerging markets like India and China, nearly 90% of the population does not use English as a means of communication. Given the small number of English speakers in these markets ICT penetration can only be achieved by supporting local language in a user-friendly way. Speech Synthesis that enables ease of access and use, become important in such a situation.

Further, emerging markets like India are witnessing a telecom boom with an estimated 190 million phones (landlines and mobiles) by 2008. The rapid increase in the use of mobile technology (several million phones a month) has contributed to this. In comparison, PC penetration is growing at a negligible pace. The easy availability and use of telephones is an ideal opportunity to provide voice-based services on the phone.

Hindi is the official language of India and is spoken by 180 million Indians as a first language and 300 million as a second language (as per 1991 census). It is in fact the lingua franca of India and due to the high popularity of Hindi cinema, it serves as a link between the numerous language groups in India. If HP or its partners were able to provide a Hindi TTS as a part of certain products and services in India, the value addition would make HP products more attractive.

A local language TTS system can help overcome the barriers to accessing ICT for the non-English speaking population. Voice Portals, web-page readers, speech enabled navigators and browsers, readers for the visually challenged can all help achieve this. Also, a lightweight local language TTS engine on HP PCs and handhelds like iPAQ would give HP an edge over competitors in India, by allowing it to provide the capability to have multimedia applications that use speech synthesis in a significant way.

Creating a Hindi Text-to-Speech System

A Hindi TTS was developed at HP Labs India on a generic TTS framework that it created based on Festival [8]. This involved extending the Natural Language Processing module of Festival by providing tools that were more appropriate for handling non-European languages like Hindi. It also incorporated methodology and tools for creating TTS speech databases, from best practices for choosing a voice talent to tools for automatic segmentation and annotation of the database. The TTS framework is illustrated in Figure 1.

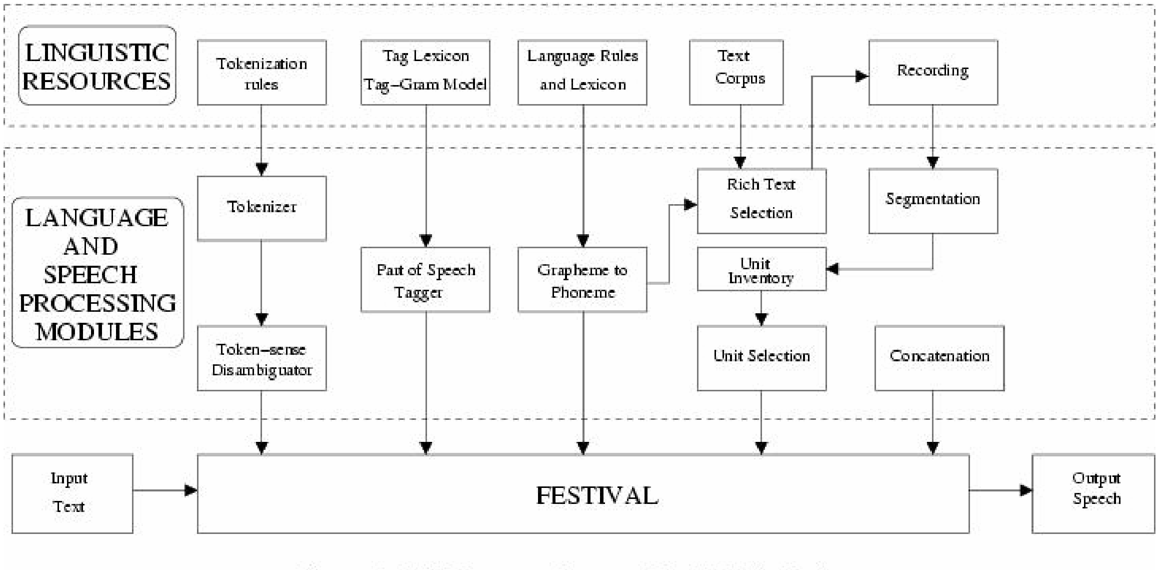

Figure 1: TTS Framework created by HP Labs India

While the focus was on creating a high quality Hindi TTS, efforts were made to make the tools as generic and language independent as possible. To enable this, the project worked towards:

- A toolset extending the NLP module of Festival by tools for Grapheme to Phoneme (G2P) conversion, Part of Speech (POS) tagging, Text Normalization, creation of lexicons and other morphological analysis. These are a part of the framework.
- Tools and processes for creation of the speech database including a text selection algorithm for phonetically balanced texts, and a tool for automatic speech segmentation and annotation.

This framework can also be used to create TTS systems for other new languages.

Language Processing Modules for a TTS System

Grapheme-to-Phoneme (G2P) conversion plays a crucial role in Text to Speech synthesis, as each character of the script has to be mapped to the correct phonetic representation. The phonetic transliteration should correctly depict the nature of the speech sound represented in terms of linguistic as well as acoustic features. Intonation and durational features should be appropriately marked. Building a G2P for each language is time consuming and resource intensive. Hence, a language independent G2P framework has been developed. The G2P system consists of a language independent rule processing engine and a database for language specific linguistic information. The G2P has been customized for Hindi for this project.

Hindi, like most Indian languages, uses a syllabic script that is largely phonetic in nature. Thus, in most cases, there is a one-to-one mapping between graphemes and the corresponding phones. Special ligatures are used to denote nasalization and homo-organic nasals. The Hindi phoneset consists of 86 phones due to the fact that it has a large set of distinct phonemes with a relatively small number of allophones. For example, the stop system in Hindi shows a four way distinction, viz., voiced vs. unvoiced, aspirated vs. unaspirated, for five places of articulation. There is also a phonological distinction between retroflex and alveolar, aspirated and non-aspirated taps. Also, each of the vowels has a phonologically distinct nasalized counterpart. Hindi has a number of borrowed phonemes of Persian-Arabic origin as well, like fricatives /z/ and /f/.

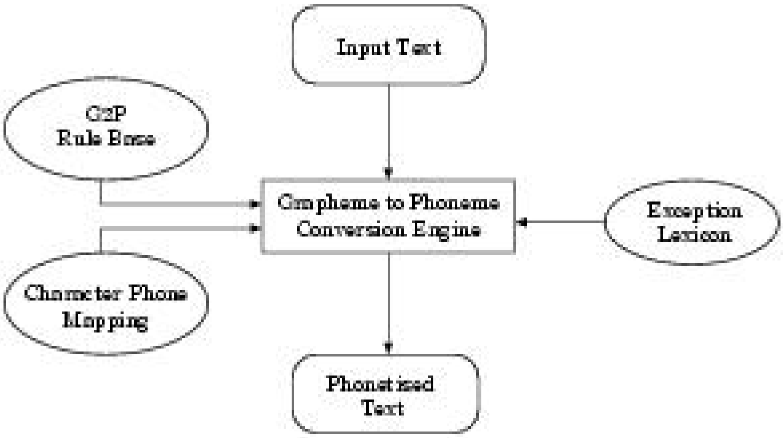

Fig. 2 Architecture of HP Labs Grapheme to Phoneme Converter

A generic G2P conversion system has been developed at HP Labs India. The system consists of a rule processing engine which is language independent. Language specific information is fed into the system in the form of lexicon, rules and mapping. The architecture of the G2P is shown in Figure 2. This G2P framework, has then been customized for Hindi. The default character to phone mapping is defined in the mapping file. The format of the mapping is given below.

Character Type Class Phoneme

Character The orthographic representation of the character.

Type Type of the character, e.g. C (for Consonant), V (for Vowel) etc.

Class The class to which the character belongs. These class labels can be effectively used to write a rule representing a broad set of characters.

Phoneme The default phonetic representation of the character.

Specific contexts are matched using rules. The system triggers the rule that best fits the current context. The rule format is shown below:

$$\alpha_1\ \alpha_2\ \ldots\ \alpha_m\ \{\beta_1\ \beta_2\ \ldots\ \beta_n\}$$

α_i Class label of the ith character as defined in the character phone mapping. Together these αs represent the context that is being matched.

β_j jth action specification node. Each such node has the form:

Action_Type:Pos:Phoneme_Str

Action_Type This field specifies the type of this action node. Possible values are K (Keep), R (Replace), I (Insert) and A (Append).

Pos The index of the character which is covered by the context of the rule (1≤Pos≤m)

Phoneme_Str Phoneme string used by this action node.

The G2P first looks for each input word in the exception lexicon. For the Hindi G2P, this lexicon is the phonetic compound word lexicon which is generated using the algorithm described [1]. If the word is present in the lexicon, the phonetic transcription of the input word is taken from the lexicon itself. Otherwise, the G2P applies the rules on this word and produces the phonetic transcription.

Though Hindi phonology makes it easier to devise letter-to-sound rules, there are certain exceptions to this. Some of these exceptions can be handled by straight-forward rules and some cannot. Schwa deletion is one such problem encountered in Hindi G2P conversion. This problem arises with the phonetization of words containing the schwa vowel (ə). Depending on certain constructs, the schwa vowel is deleted in some cases and retained in others. The orthography, however, provides minimal cues in recognizing these constructs. Further details of how schwa deletion is dealt with in HP Labs Hindi G2P can be found in [3]

A Text Normalization module has also been developed to normalize real text and convert non-standard words like dates, numerals, URLs, etc into standard words or tokens [2]. In the first stage, tokenization and classification of non-standard words is carried out followed by a second level of sense disambiguation. Tokenization and classification is carried out using a lexical analyzer that is derived from the definitions of various tokens in the form of regular expressions. Token sense disambiguation is carried out using decision lists and decision trees. Token-to-word rules are then applied for each token type and each format within a type.

Part of speech (POS) tagging provides important cues for intonation and phrasing modules in a TTS, and thus, a POS tagger has also been developed to be used for prosodic modeling. The tagger consists of two components: a lexicon of words with corresponding tags, and a trigram model of tag production. A Viterbi search is conducted using the two components to find the most probable sequence of tags given the sequence of words. A set of 6000 tagged and validated sentences was used as the training set. The POS tagger currently gives about 80% accuracy and can handle out of vocabulary words.

Creating the Speech Database for the TTS System

In a concatenative TTS system, where output speech consists of concatenated real or coded speech segments, a large database of phonetically balanced and prosodically varied speech is required. The quality and accurate annotation of the speech database is critical as it is one of the most important factors in determining the quality of the synthesized speech. Both of these require a large amount of text from which such databases can be recorded.

An Optimal Text Selection system has been developed as a part of this module. The idea is to select a minimal set of sentences that cover the phonetic variations of a language. These selected sentences that cover all the speech sounds of the language in all possible contexts are then recorded to create the speech database from which units are selected and concatenated to give the appropriate speech output. Working with an optimal set of phonetically balanced sentences cuts down the time in segmenting speech and also reduces the search space of the unit selection algorithm. The selected sentences were recorded at Jawaharlal Nehru University using best practices in recording speech databases for TTS.

One of the most important tasks in building speech databases is the annotation of speech data with its contents (labeling) and the time alignment between labeling and speech (segmentation). Phonetic segmentation and labeling are highly desirable and useful for TTS as this information is used for classifying the speech units that help to select and concatenate the right units in terms of linguistic and acoustic features. The most precise way to annotate speech data is manually by linguistic experts. However, manual phonetic labeling and segmentation are very costly and require significant time and effort. Even well trained, experienced phonetic labelers, using efficient speech display and editing tools require about 200 times the duration of the recorded speech to segment and align speech utterances. To reduce this effort considerably and aid the phonetic labelers, an automatic segmentation tool was developed at HP Labs India, adapting a Hidden Markov Model (HMM) based phonetic recognizer to the task of automatic phonetic segmentation. The entire process of speech corpora creation for the Hindi TTS is discussed in detail in [4]

Conclusion

An unrestricted domain Hindi TTS has been created using the framework. Methods are also in place for converting the Festival voices to Flite [9] (supports multi-threading), and for porting the Flite voice to an embedded platform. We have done a comparison of this TTS with the only commercially available Hindi TTS from Prologix and have found our TTS to be distinctly superior. The quality is extremely important since the acceptability of any system or product that uses the TTS would be critically dependent on this.

Festival is a very large system with a footprint of over 60 MB and is not multi-threaded. This implies that not only are the memory and space requirements very high for running Festival, it can only handle a single query at a time. For any real world deployment, a TTS engine would need to handle several queries at a time. Flite is a small, fast run-time synthesis engine and is primarily designed for small embedded machines and/or large servers. We have been able to convert Festival voices to Flite, and also in porting Flite on to embedded platforms. We have successfully ported a version of the Hindi TTS on to HP iPAQ. This would have applicability for multimedia applications on the iPAQ or for less sophisticated users who may need to use the iPAQ for specific applications. The TTS could also be used when the form factor of the iPAQ makes it more convenient to use speech synthesis as a part of the user interface.

Substantial innovation was required for building the Hindi TTS grounds up. As no databases existed for Hindi, a significant challenge lay in the creation of a high quality speech database for Hindi. Also, though an open source framework was used there was substantial work required in terms of the tools we have described above for the selection of phonetically rich text from a text corpus, the language processing tools, the automatic segmentation tool; all of which were necessary in order to create the final TTS.

In addition to the TTS, we also created with support from HP GDIC, an MRCP stack that the TTS can integrate into. This makes it much easier to integrate the TTS into any system that supports the MRCP protocol without additional integration effort. The stack was used in the Nilgiri voice service [5] (to interface it with OCMP) and is also being leveraged for the integration of the TTS into the DSpace digital library software.

HP Telecom products like OCMP, that support speech engines for Text-to-Speech Synthesis (TTS) and Automatic Speech Recognition (ASR), have a significant potential in emerging markets by leveraging growth in the telecom sector. This can be done by effectively addressing the barriers and exploiting new opportunities that emerge in these contexts. For example, the rise in mobile telephone usage has resulted in the widespread use of SMS (Short Message Service). SMS is not only used for personal messages but it is possible to book tickets, order food, make bank inquiries, or read news through SMS. An interesting phenomenon is the use of SMS by the entertainment industry, mainly the broadcasting channels, for conducting opinion polls, participating in interactive programmes, and TV contests. However, SMS is currently available only between mobile phones and primarily used in English. Also, a text message is not always accessible to a person. A useful application of local language TTS would be conversion of SMS to a Hindi Voice that would allow the message to be relayed either to a person or a voice-mail box as a voice message in Hindi. Such a service would be useful for anybody who wants to send someone without a mobile a text message that is spoken out.

Voice-based services that employ either a combination of ASR and TTS, or DTMF with TTS for an IVR system can prove extremely effective for providing access to information and services in local languages. Though the telecom industry is growing at a rapid speed in these markets, the local telecom companies do not have the capability to provide voice-based services to their customers in their own languages. Localization of technology by Speech technology companies is extremely limited. As a result, any interactive service that is deployed in a local language is constrained to simple interactions based on recorded speech, cut and joined as required. This fails to provide the customer with a natural, more personalized interactive experience in their own language. Though there is a very significant need for effective customer support – all enterprises are also trying to reduce the costs for handling this. Providing local language voice portals is one way to do this cost effectively.

Another application is to provide access to "content in context" through speech-enabled eBooks. This integration of visual and speech user interfaces to provide smart content could find a number of uses not only in the field of education but any field that uses digital content in a significant way. For example, integration of a local language TTS system to a digital library can help the user create and access local language audio content corresponding to the digital text documents. This could then be streamed to the users either synchronized to the text or separately (for those with visual impairment), when the users are not able to access the document in the textual form.

The Hindi TTS was used in the Nilgiri railway enquiry service, a proof-of-concept voice-based service in local languages using HP OCMP that has been developed in the lab. HP Labs Hindi TTS has also been integrated with DSpace [7] (the open source digital library software developed by HP and MIT) in the Educentre project in the lab. This would enable users of the digital library to access Hindi content in audio form – thus providing a value addition for those who opt for HP's digital library solutions [6].

References

1. AG Ramakrishanan, K. Bali, PT Talukdar, and SR Deepa. (2004) **Automatic Generation of Compound Word Lexicon and its Importance in Text to Speech Synthesis**. Poster presented at Language Resources and Evaluation Conference (LREC), Portugal, May 2004.
2. K. Panchapagesan, Partha Pratim Talukdar, N. Sridhar Krishna, Kalika Bali, A. G. Ramakrishna (2004). **Hindi Text Normalization.** Fifth International Conference on Knowledge Based Computer Systems, Dec. 2004, Hyderabad, India.
3. Kalika Bali, Partha Pratim Talukdar, N. Sridhar Krishna, A.G. Ramakrishna (2004). **Tools for the Development of a Hindi Speech Synthesis System.** Paper presented at 5th ISCA Speech Synthesis Workshop. Pittsburgh, USA.
4. Kalika Bali, Satinder Pal Singh, RNV Sitaram, N Sridhar Krishna, PT Talukdar, and S. Manocha. (2004) **Optimal Creation of Speech Databases for Indian Language Speech Technology**. Paper presented at International Conference on Speech and Language Technology/O-COCOSDA-2004. Noida, India.
5. Sitaram, RNV, Anjaneyulu, KSR, Prasad GVD, Pinto, Joel, and Bali Kalika. **Local Language Voice Services – Enabling OCMP for Emerging Markets**. Paper presented at Techcon 2004, Orlando, Florida. 2004
6. Srinivasan Ramani, Raphael Baecher, Kalika Bali, Vivek Singh, Jagadeesan Srivathsan. (2005) **Educenter: A system to support digital libraries to acquire and handle multi-media content.** HP TechCon India 2005, Bangalore, India, November 2005.
7. https://dspace.mit.edu/
8. http://www.cstr.ed.ac.uk/projects/festival/
9. http://www.speech.cs.cmu.edu/flite/