8 Speech synthesis IN INDIAN LANGUAGES & FUTURE PERSPECTIVES

ARUN BABY, ANJU THOMAS, JEENA J PRAKASH, ANUSHA PRAKASH & HEMA A. MURTHY

I. INTRODUCTION

Building speech synthesis systems for Indian languages is difficult owing to the fact that digital resources for Indian languages are hardly available. Vocabulary independent speech synthesis requires that a given text is split at the level of the smallest sound unit, namely, phone, the waveforms or models of phones are concatenated to produce speech. The waveforms corresponding to each of the phones is obtained manually (listening and marking), when digital resources are scarce. Manually labelling of data can lead to inconsistencies as the duration of phonemes can be as short as 10ms.

The most common approach to automatic segmentation of speech is, to perform forced alignment using monophone HMM models (corresponding to the sound units in a language) that have been obtained using embedded reestimation after flat start initialisation. Parsers are available for each of the languages under consideration. These results are then used in a DNN/SGMM framework to build better acoustic models for speech synthesis/recognition. Segmentation using this approach requires large amounts of data and does not work very well for low resource languages. To address the issue of paucity of data, signal processing cues are used to restrict embedded reestimation.

There are many issues in building TTS systems for Indian languages. Although there are many languages that use the Devanagari script, there are a number of languages that use their own scripts. A common label set was first developed to generate a common framework for all Indian languages. Voice activity detection is first performed to determine the voiced regions in an utterance. Short-term energy (STE) and spectral flux (SF) are computed on intra voiced segments. STE yields syllable boundaries, while locations of significant change in spectral flux are indicative of fricatives, and nasals. STE and SF cannot be used directly to segment an utterance. Minimum phase group delay based smoothing is performed to preserve these landmarks, while at the same time reducing the local fluctuations.Boundary corrections are performed at the syllable level, wherever it is known that the syllable boundaries are correct. Embedded reestimation of monophone HMM models is then restricted to the syllable boundaries. The boundaries obtained using group delay smoothing results in a number of false alarms. HMM boundaries are used to correct these boundaries. Similarly, spectral flux is used to correct fricative boundaries. Thus, using signal processing cues and HMM reestimation in tandem, robust monophone/triphone HMM models are built. These models are then used in a DNN framework to obtain state-level frame posteriors. The boundaries are again iteratively corrected and reestimated.A major problem with HMM parameter estimation is the lack of large amounts of data.

State-tying is commonly used approach to reduce the number of HMM states for parameter estimation. State-tying is primarily based on the merging of similar sounds. As Indian languages belong to a different family of languages compared to other Western languages, a set of rules that are common across different languages was first designed. Linguists and phoneticians across the country were consulted to arrive at this.

The final waveforms are then used in a Unit selection synthesis/Statistical parametric synthesis framework to build speech synthesis systems for 13 Indian languages. Both quantitative and qualitative assessments indicate that there is a significant improvement in quality of synthesis.

II. TTS CONSORTIUM EFFORT FROM 2009-2017

The TTS consortium funded by MeitY, GoI under the guidance of TDIL, MeitY India has developed TTS systems for 13 Indian languages using this novel approach. The languages include Assamese, Bengali, Bodo, Gujarati, Hindi, Kannada, Marathi, Malayalam, Manipuri, Odia, Rajasthani, Telugu, Tamil, and 13 flavours of L2 Indian English.

The TTS consortium has created a website (www.iitm.ac.in/donlab/tts/) where a number of different resources are available. There were two phases to the consortium effort.

A. PHASE I – 2009–2011

The first phase was to develop TTS systems that were integrated with screenreaders. ORCA under linux, and NVDA2010 under windows were integrated. Initially there was lot effort spent on collecting data from appropriate voice talents. Some salient features include accommodation for the agglutinative nature of Indian languages. Selection of optimal text was crucial to this phase. It is very common to find a large number of words that are concatenated to make new words. Such words should not be present in the training data, as these words are not articulated properly. Finally, Chandamama stories were taken to build the corpus. Data was collected from a voice talent. The voice talent had to be carefully chosen so that the voice was amenable for modification. Labelling was performed semi automatically, where signal processing cues were provided for manual labelling. Syllable-based unit selection synthesis based systems were developed for Tamil, Telugu, Hindi, Marathi, Bengali and Malayalam. Use of syllable as a unit for synthesis was the novelty of the proposed work. The quality was found to be much better than diphone based systems.

These systems were integrated with ORCA/ NVDA. Visually challenged persons were trained to use these systems to interact with computer systems. Training workshops were conducted at a number of different places. Figure 1 shows a photograph of a workshop in progress at IIT Madras. A participatory design approach was used to design the interfaces [1]. Over a 5-year period about 180 visually challenged students were trained to use Microsoft Office, excel, email and Internet in the vernacular.

B. PHASE II: 2012–2017

The Phase I systems were large footprint and were therefore limited to desktops with adequate RAM and hard disk. The entire voice would require anywhere between 700MB-1GB of space. In the meanwhile, this period saw the proliferation of smartphones. The objective was modified to include "small footprint synthesisers" for 13 Indian languages. During Phase I, it was observed that most vernacular websites included quite a lot English text. During Phase I, whenever such text was found, the voice was replaced by an American/British English voice which sounded unnatural. This issue is addressed in Phase II, where every language also provides a bilingual version to enable code switching between the native tongue and English by the same voice. Owing the small footprint requirement, statistical parametric speech synthesis systems were built around the hidden Markov model based speech synthesis systems (HTS). Although HTS systems are capable of generalisation for unseen sequences, during training appropriate state-tying and clustering should be performed. Hidden Markov model (HMM) based classification and regression trees are available for English which organise the HMMs based on context, rules of languages. Also the graphemic representations vary from language to language. For scalability, a common label set (a superset of sounds across the 13 languages) (Figure 2) and a common question set (CQS) were developed first. A subset of the CQS is given below:

QS "LL-Low_Vowel" {a^*, aa^*} QS "LL-IVowel" {i^*, ii^*}
QS "LL-OVowel" {oi^*, o^*, oo^*, ou^*} QS "LL-UVowel" {u^*, uu^*}
QS "LL-Glide" {y^*, w^*, yq^*, zh^*}

Next, a common parser that works for all the 13 languages was developed. The parser was developed such that the "language dependent" and "language independent" components were separated [2]. The labels obtained in Phase I was still erroneous. This issue was addressed by using both machine learning and signal processing in tandem. Signal processing was used to determine stop consonant, fricative and nasal boundaries. The machine learning algorithm used these boundaries as sentinels during training[3], [4]. Figure 3 shows a part of a Hindi speech utterance labeled using this approach. TTS systems using this procedure were developed for 13 different Indian languages and the corresponding Indian English flavours. Degradation mean opinion score (DMOS) listening test shows an average relative improvement of 14.8% with this approach

Different applications were developed to make the TTS services available on the Android platform :

(1) Tamil TTS app
(2) Hindi TTS app and (3) Indic TTS app - for text-tospeech synthesis of 13 Indian languages (4) Tamil learning app and (5) Safe pregnancy app. Given that today vocabulary independent TTS is available, the consortium has partnered with IndusOS to integrate TTS with the operating system for smart phones. The system for IndusOS is capable of "translating and reading SMSes in the tongue of the user." All the data and systems have been released under GPL license for use by both commercial and research establishments.

Specific MoUs have been signed with Wipro, Samsung, Timbre Media, Shinano Technologies, Inferon online services, etc. We havalso assisted TCS Ignite team - Chennai and Digital Impact Square - a TCS foundation initiative. The following companies have signed the online license agreement: (1) Amity Software Systems Limited (2) Amazon (3) Australian Survey Research (4) Crosscode Technologies Private Limited (5) datametica (6) Dheeyantra Research Labs (7) eVenturers (8) Evferest IT Services Pvt. Ltd. (9) GMX (10) gnani.ai (11) Hungama Digital Media Entertainment Pvt.Ltd. (12) ICFOSS (13) Idea (14) Manorama Social Mobile Analytics Cloud (15) Meritnation (16) Microsoft (17) Mihup (18) Mindtree (19) niqotin (20) Nithra Edu Solutions India Pvt Ltd (21) Pratham (22) Process 9 (23) Reliance Industries Limited (24) Samsung (25) TCS(26) Tevatel (27) Timbre Media (28) Trinity Unicepts Pvt. Ltd (29) Wipro (30) Yandex. The educational institutes that have signed the online license agreement are listed below:(1) CMU (2) IIT Bombay (3) IISc (4) Gujarat University (5) IIT Bhubaneswar (6) IIT Guwahati (7) IIT Hyderabad (8) UIUC (9) NIT Calicut (10) The Chinese University of Hong Kong (11) Siksha 'O' Anusandhan University (12) NIIT University (13) Shri Guru Gobind Singhji Institute of Engineering & Technology (14) UT Dallas. Table I shows the download statistics of the online Indic TTS resources (website: www.iitm.ac.in/donlab/tts/).

DOWNLOAD TYPE	COUNT
Database	2450
CLS	402
Unified Parser	375
Hybrid Segmentation	262
Pruning	148
Voices	760
Android Applications	570
Synthesis Documents	1433

III. TTS IN INDIAN LANGUAGES – THE FUTURE

The drawback of the current day TTS systems is lack of prosody. Prosody of speech utterances can vary from context to context. For example, the prosody used in news has to be very different from that of story-telling.

Naturalness of speech in different tasks is related to prosody. Some efforts in this direction are in progress at the various participating institutions. Other applications include dynamic IVRs in different languages (basically a system that understands the anxiety of user and provides appropriate responses), health care (where a person can get his/her prescription/lab report read out), weather reports (for farmers, people on the road), a Google map enabled with Indic voices can be of great use to taxi companies like Uber and Ola to help navigation in the vernacular. Information about traffic congestion can be also be announced on the radio (with local content). Last but not the least, persons with cerebral palsy, visual challenged can be brought into the main stream with TTS technologies in the vernacular. All of these systems require smart designs on smart phones for speed of responses. Android based platforms that are cheap and fast are required. This suggests that most of the tasks must be done in hardware. Building TTS hardware can go a long in enabling robust and cost effective technologies. Further, extensive effort on dialogue design must be under-taken to suit the Indian environment where all of us polyglots, for example, multilingual dialogue design.

The authors are from Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai

REFERENCES

[1] Kurian, Anila Susan and Narayan, Badri and Madasamy, Na-garajan and Bellur,Ashwin and Krishnan, Raghava and G,Kasthuri and Vishwanath, Vinodh M. and Prahallad, Kishore and Murthy, Hema A., "Indian Language Screen Readers and Syllable Based Festival Text-to-Speech Synthesis System," Proceedings of the Second Workshop on Speech and Language Pro-cessing for Assistive Technologies, Association for Computa-tional Linguistics, pp.63–72,July, 2011, Edinburgh, Scotland, UK. http://www.aclweb.org/anthology/W11-2307.

[2] Arun Baby, N L Nishanthi, Anju Thomas and Hema A Murthy, "A unified parser for developing Indian language text to speech synthe-sizers," TSD, Volume 9924 of the series Lecture Notes in Computer Science pp 514-521.

[3] S Aswin Shanmugam and Hema A Murthy, "A Hybrid Approach to Segmentation of Speech Using Group Delay Processing and HMM Based Embedded Reestimation,"Proc. INTERSPEECH 2014, pp.1648-1652.

[4] Arun Baby, Jeena J Prakash, Rupak Viswanathan, and Hema A Murthy, "Deep Learning Techniques in Tandem with Signal Processing Cues for Phonetic Segmentation for Text to Speech Synthesis in Indian Languages," INTERSPEECH 2017, Stockholm, Sweden, Aug 20th-24th.