Building a Better Indian English Voice using "More Data"

Rohit Kumar, Rashmi Gangadharaiah, Sharath Rao,
Kishore Prahallad, Carolyn P. Rosé, Alan W. Black

Language Technologies Institute
Carnegie Mellon University, Pittsburgh, PA, USA
{ rohitk, rgangadh, skrao, skishore, cprose, awb } @ cs.cmu.edu

Abstract

We report our experiments towards improving an existing publicly available Indian English voice using additional data. The additional data was used to create new duration and pronunciation models as well as to convert the existing voice to create a more Indian sounding voice. Two experiments along the above lines are reported. In the first experiment, we found that changing the pronunciation models has the potential to improve an existing Indian English voice. We conducted a second experiment to validate this finding. The second experiment shows the potential value in carefully investigating the separate effects of the different components of a pronunciation model in order to understand their unique contributions to improving an Indian English voice.

1. Introduction

English is the official language of India. Over 200 million people use Indian English. In this paper, we refer to the English used in news telecasts as Indian English. The English used in India, although originally acquired by native Indian speakers during the course of the British rule, is known to have undergone transformations along various dimensions of the language including its phonology, morphology, syntax and word usage [1]. While borrowing models from American or British English may be the right way to bootstrap Indian Language systems, it is essential that changes in the above mentioned aspects of Indian English are modeled appropriately in these systems.

Our motivation for this work is two fold. First, we want to develop a better Indian English voice. Second, we want to study whether additional data can be used either to improve a given Indian English voice or to build newer voices with very little data. We hypothesize that additional data can be used to improve multiple models used in any text to speech system (TTS). In particular we focus on three key components of a TTS, i.e., the duration model, the pronunciation model, and the voice data used to build the synthesis model.

The remainder of the paper is organized as follows. Section 2 discusses the design and results of the first experiment. Section 3 describes the second experiment along with our findings. Discussion of the results from both the experiments is found in Section 4 which is followed by conclusions and next steps.

2. Experiment 1: The new models

In the first experiment, we used additional data to create new duration, pronunciation and synthesis models. We experimentally evaluate their separate effects on two different response variables.

2.1. Data

We start with two baseline voices (KSP and BDL) distributed as a part of the CMU Arctic [2] set of voices. Both of these voices include recordings of 1132 optimally selected sentences. KSP is the voice of a native Indian who is a fluent speaker of Indian English. BDL is the voice of a standard American English speaker. Both KSP and BDL are male speakers.

The additional data we used is comprised of an Indian English pronunciation lexicon and speech recorded by five male Indian English speakers. Each of the five speakers recorded 100 sentences of the CMU Arctic set. These utterances were originally recorded for the ConQuest project to build acoustics models for an Indian English speech recognition system. Hence the recording was done in an office space unlike the CMU Arctic KSP and BDL voices which were recorded in a recording booth. Given the number of utterances per speaker and the quality of the recordings, the additional data by itself was not suitable for building high quality synthesis voices. Hence we use this data for building new duration models as well as for conversion as described later in this section.

2.1.1. Indian English Pronunciation Lexicon

The Indian English pronunciation lexicon was built specifically for this project. It is comprised of 3489 words derived from the 1132 CMU Arctic sentences and the 200 sentences from the SCRIBE Project [3]. An American English phoneme set was used to represent the pronunciation of these words in Indian English. Despite the differences between the American and Indian English, an American English phoneme set was used to represent the pronunciations in the Indian English lexicon because it allows us to bootstrap the Indian English dictionary from existing letter to sound rules as described ahead.

We used the CMU Dictionary [4] and a set of letter to sound rules built from the dictionary to generate American English pronunciations for the 3489 words. These pronunciations were then corrected by the authors to match the Indian English pronunciations. During corrections, if a desired phoneme was unavailable in the phoneme set, the nearest available phoneme (in terms of minimal mismatch of articulatory descriptors) was chosen.

After the manual corrections, the new Indian English phoneme sequences were syllabified and stress marked using a set of rules derived from characteristics of Indian Languages as discussed below.

The basic units of the writing system in Indian languages are referred to as "Aksharas". The properties of Aksharas are as follows: (1) An Akshara is an orthographic representation

of a speech sound in an Indian language; (2) Aksharas are syllabic in nature; (3) The typical forms of Akshara are V, CV, CCV and CCCV, thus have a generalized form of C*V; (4) An Akshara always ends with a vowel (which includes nasalized vowels); [5]. In view of these points, given a sequence of phones, one can consistently mark syllable boundaries at vowels. This heuristic is typically followed in building TTS systems for Indian languages [6]. At the same time, a simple set of rules are followed to assign stress to the syllables. A primary stress level is associated with the first syllable and to the other syllables which have non-schwa vowels. A secondary stress is associated with the rest of the syllables which have schwas. Assuming that Indian English speakers tend to borrow syllabification and stress assignment characteristics from their native languages, we wanted to investigate how the use of these rules would affect the quality of an Indian English TTS.

On analyzing the new Indian English Pronunciation lexicon we observed that only 918 (26.3%) words needed any correction at all. At the phoneme level only a 7.2% change was observed. The majority of these changes were phoneme substitutions. The most common substitution included vowel substitutions (like /aa/ → /ao/ e.g. h**o**stilities). Also, several common consonant substitutions like /z/ → /s/ and /w/ → /v/ were observed.

2.2. The New Models

We created 15 different voices using different combinations of converted voices, duration models and pronunciation models. We used the FestVox framework [7] to build all of these different models and voices.

2.2.1. The converted voices

We used the speech from two of the 5 speakers in the additional data to convert the KSP and BDL utterances. A converted set of utterance is represented as a 2-tuple <SOURCE, TARGET>. The SOURCE refers to the original speaker whose utterances are being converted. SOURCE can be KSP and BDL in our case. TARGET refers to the speaker to which SOURCE is being converted. One of the two target speakers we used from the additional data is a North Indian (NIE) speaker, and the other is a South Indian (SIE) speaker. Also it may be noted that KSP is a South Indian speaker too.

We use a GMM based Spectral conversion method [8] to create the converted voices. The 5 converted voices are <KSP, NIE>, <KSP, SIE>, <BDL, NIE>, <BDL, SIE> and <KSP, KSP> respectively. The <KSP, KSP> converted voice is used to compare the new voices with the existing Indian English voice and can be assumed to have the lowest distortion due to conversion.

2.2.2. The duration models

The duration models predict the duration of a phoneme during synthesis. The models are trained on phoneme segments obtained by automatically segmenting the given utterances. We use a publicly available Ergodic HMM based segmenter distributed with FestVox.

The baseline duration model was built using the 1132 utterances of the KSP voice. The experimental duration model in this case was built using the 1132 utterances of the KSP voice and the 500 utterances from the additional data. We refer to the experimental duration model as KSP++ which we contrast with the baseline duration model, namely KSP. Both the duration models are built using correlation and regression trees (CART) and are based on phonetic and syllabic features of the segment as well as its context.

2.2.3. The pronunciation models

A pronunciation model converts a given word to its pronunciation. The pronunciation of a word is comprised of the phoneme sequence corresponding to the sounds of the word and the syllabification of the phoneme sequence. Each syllable also carries information about its stress. A typical pronunciation model is comprised of a dictionary and a set of letter to sound (LTS) rules. The LTS rules may either be hand crafted or learnt from the dictionary. Given a word, a pronunciation model typically does a lookup in the dictionary. In case the dictionary does not contain the pronunciation of

Table 1. Results of the first Experiment (sorted by Mean Intelligibility)

Converted Voice	Duration Model	Pronunciation Model	Intelligibility		Indian-ness	
			Mean	Std. Dev	Mean	Std. Dev
KSP, KSP	KSP	IE	4.9	1.79	5.92	1.41
KSP, KSP	KSP++	IE	4.87	1.79	5.37	1.86
KSP, KSP	KSP++	CMU	4.48	1.83	5.3	1.89
KSP, SIE	KSP++	IE	4	1.78	5.4	1.69
KSP, SIE	KSP	IE	3.85	2.07	5.02	1.8
BDL, SIE	KSP++	CMU	3.78	1.97	2.77	2.15
KSP, NIE	KSP	IE	3.7	2.22	4.73	2.25
KSP, NIE	KSP++	IE	3.48	1.93	4.38	2.12
BDL, NIE	KSP++	CMU	3.48	2.07	2.53	1.79
KSP, SIE	KSP++	CMU	3.4	2.16	4.47	2.14
BDL, SIE	KSP++	IE	3.18	2.05	2.55	2.06
BDL, NIE	KSP	IE	3.17	2.06	2.83	1.63
BDL, NIE	KSP++	IE	3.13	1.88	3.17	1.7
KSP, NIE	KSP++	CMU	3.1	2	4.75	1.76
BDL,SIE	KSP	IE	2.87	2.04	2.72	1.78

the given word, the LTS rules are used to generate the pronunciation of the word.

We use two different pronunciation models in the first experiment. The baseline pronunciation model (CMU) is built from the CMU Dictionary consisting of over 105,000 words. The experimental pronunciation model which we refer to as IE, is built from the Indian English pronunciation lexicon of 3489 words described earlier. The LTS rules for both the models have been trained using CART [9].

2.3. The pilot experiment

To study the effect of (1) the different source and target voices, (2) the duration models and (3) the pronunciation models, we created 15 different festival [10] compatible voices. All voices are built to use a Unit Selection Synthesizer [11]. Table 1 lists the 15 different voices in terms of the models and converted voice they use.

In the first experiment, these 15 voices were subjectively evaluated for two different perceived measures: Intelligibility and Indian-ness. 15 subjects were asked to listen to 60 utterances and score each utterance for both the measures independently on a scale 0 to 7. For Intelligibility, they were instructed to score a zero if they did not understand even a single word of the utterance and to score a 7 if the utterance was perfectly understandable. For Indian-ness, they were instructed to score a 0 if the utterance did not sound like an Indian speaker at all and to score a 7 if the utterance sounded perfectly like an Indian speaker. Subjects were instructed to evaluate both the measures independent of each other.

15 subjects participated in this evaluation under controlled conditions. All subjects used the same equipment (laptop, speakers) and performed the listening task in the same office. All subjects are of Indian origin and are graduate students at Carnegie Mellon University. They have not been outside India for more than 4 years. The subjects were 21 to 27 years old.

The 60 utterances given to the subjects were composed of 4 utterances from each voice in random order in order to avoid ordering effects.

2.4. Preliminary evidence and directions

Table 1 enumerates the average scores for each of the voice on both the measures along with the corresponding standard deviations. The voice built from the KSP → KSP conversion performed best among all the other voices. The KSP Source voice was scored significantly higher than the BDL voice on both the measures. Further, the KSP voice as a target was significantly better than NIE. SIE was not significantly different from either KSP or NIE as a target voice. The <KSP, KSP> converted voice performed better than all the other converted voices because the distortion caused by conversion was minimal for that pair. However SIE not being significantly different from KSP shows the potential for creating new voices using a baseline voice and very little speech data from a target voice in the case where the source and target speakers have similar characteristics. Both SIE and KSP are South Indian English speakers of comparable age and educational background.

There was no effect of the duration model on either of the outcome measures. We found that both the duration models selected exactly the same sequence of units per utterance despite generating different targets. We understand that this is because of the low cost associated with duration mismatch as well as the restricted diversity of units in the inventory. The units matching the targets generated by both the duration models turn out to be the same in all cases.

Comparing across all the 15 experimental voices, we found no significant difference between the two pronunciation models. However, if we restrict our attention to the data from the <KSP, KSP> converted voice, we then see a significant difference in the average Intelligibility between the pronunciation models ($p=0.008$) when we included a variable in the model indicating for each judgment which sentence was spoken to account for variance caused by differences in the words included across sentences. A similar effect was observed for the voices based on the <KSP, SIE> converted voice ($p=0.044$).

Based on the evidence that <KSP, KSP> was the best of the converted voices and that <KSP, SIE> was among the better ones of the converted voices, ranking second according to the average intelligibility scores, we hypothesize that the improvements due to the IE pronunciation model were observable only in the good voices which were least distorted due to voice conversion. Based on this reasoning, we decided to further investigate the effect of the experimental pronunciation model using high quality voices like the unconverted CMU Arctic KSP voice.

3. Experiment 2: The field study

In the follow up experiment, we decided to focus on studying the contribution of the pronunciation model towards building a better Indian English voice. Unlike the first experiment, we conducted the second study in India.

In this experiment, we wanted to compare the two pronunciation models from the first experiment, CMU and IE, with high quality voices which have been built without any degradation due to voice conversion. We start with CMU Arctic KSP data and use two different synthesis techniques supported by Festival [10] to build the high quality voices: A unit selection approach referred to as CLUNITS [11] and a statistical parametric synthesis technique called CLUSTERGEN [12].

3.1. Three pronunciation models

To further study the contribution of the various components of the Indian English pronunciation model we introduce an intermediate pronunciation model derived from the CMU Dictionary. The intermediate pronunciation model (referred to as CMU+IESyl) was built by applying the Indian English syllabification and stress assignment rules to the baseline CMU Dictionary.

The intention of using this intermediate model was to study the individual contributions of two macro components of the Indian English pronunciation model i.e. the pronunciation (letter to sound rules) and the rules for syllabification and stress assignment. While CMU and CMU+IESyl pronunciation models can be compared to study the effect of the syllabification and stress assignment rules, the contrast between CMU+IESyl and IE pronunciation models can be used to study the contribution of the modified pronunciations for Indian English.

Table 2. Results of the field Experiment

Synthesis Technique	Pronunciation Model	Intelligibility		Naturalness	
		Mean	Std. Dev	Mean	Std. Dev
CLUNITS	CMU	3.83	1.18	3.37	1.1755
CLUNITS	CMU+IESYL	3.76	1.2	3.33	1.2368
CLUNITS	IE	3.88	1.16	3.48	1.1853
CLUSTERGEN	CMU	2.80	1.36	2.21	1.3597
CLUSTERGEN	CMU+IESYL	2.82	1.38	2.24	1.3972
CLUSTERGEN	IE	2.92	1.38	2.23	1.3737

3.2. Experimental Design

We built 6 different voices using all combinations of the 3 pronunciation models (CMU, CMU+IESyl, IE) and the 2 synthesis techniques (CLUNITS, CLUSTERGEN). All voices were built on the CMU Arctic KSP data.

Duration models were trained on the same data for all the voices. However, it must be noted that as the phoneme sequence for several words would be different for the different pronunciation models, the duration models will not be exactly the same for all the voices. We think that this is acceptable as building the duration model does not need any new knowledge engineering into the voice since they are built fully automatically given the KSP utterances and automatically generated segment labels. Table 2 enumerates the 6 voices.

23 participants evaluated all the 6 voices on two different measures: Intelligibility and Naturalness. Both these measures are similar to those used in the first experiment. We choose the term Naturalness instead of Indian-ness in this study as the participants in this study are resident in India. In this study a scale of 0 to 5 was used for both the outcome measures. The instructions for scoring each of the measures were similar to those in the first experiment.

The subjects used a web based interface to evaluate upto 6 sets of 30 utterances. Most of the subjects completed all the 6 sets in their evaluation. All subjects were 20 to 27 years old students at IIIT Hyderabad, India. Each set contained the same 30 sentences, 5 synthesized by each of the 6 voices. However in every set the 5 sentences synthesized by each voice were different. Further each set was randomized to avoid any ordering effects.

For our analysis, we consider a session to be the duration a single participant spends on evaluating one of the 6 sets. 128 sessions were completed among the 23 participants and in total 3840 utterances were evaluated.

3.3. Results

The results from the second experiment are shown in Table 2. We find a significant effect of the pronunciation model on the Intelligibility measure considering the session as a random factor in the analysis. $F(2, 3710) = 3.24$, $p < 0.04$. The IE pronunciation model proves to be better than the CMU+IESYL pronunciation model, although the effect size is very small ($p < 0.05$, effect size = 0.079).

In order to contrast between the different components of the 3 pronunciation models, we compared the CMU+IESYL and the IE pronunciation models. We found the Indian English pronunciation lexicon had a small but significant effect on Intelligibility as compared to the CMU dictionary when both of them use the same syllabification rules and stress marks.

On comparing the CMU and CMU+IESyl pronunciation models, we found no effect of the syllabification and stress marking rules in improving the intelligibility of Indian English. This observation leads us to conclude that the new pronunciation lexicon contributes to improving the Indian English voice. These studies also highlight that modifications in pronunciation lexicon provide better improvement in intelligibility than use of modified stress and syllable patterns on baseline CMU dictionary.

We also observe that the CLUNITS synthesis performs better than the CLUSTERGEN technique on both the measures ($p < 0.001$, effect size for intelligibility=0.71 and effect size for Indian-ness=0.85) for all the three pronunciation models.

4. Discussion

There have been other efforts in building an Indian English TTS. An Indian-accent TTS [13] uses a pronunciation model which does a morphological analysis to decompose a word and then looks up the pronunciation of the constituents in a dictionary containing about 3000 lexical items. If the pronunciation of any constituent is not found in the dictionary, it uses a set of hand crafted letter to sound rules [14] to obtain the pronunciation. [15] describes a method to build non-native pronunciation lexicons using hand-crafted rules in a formalism capable of modeling the changes in pronunciation from a standard (UK/US) pronunciation to a non-native pronunciation. [16] also describes a formalism and a set of rules for letter to sound transformation. However, unlike [14] and [15], [16] also discusses rules for syllabification as a part of pronunciation modeling.

Unlikely the above mentioned, we use automatic methods to derive the letter to sound rules. None of the mentioned work discusses stress assignment which we consider as an integral part of pronunciation modeling.

In this paper we have evaluated the contribution of pronunciation modeling in an Indian English TTS. This work reports our current finding and lays out directions for further investigation into the roles of pronunciation model and its components in building an Indian English TTS.

We believe the mismatch between pronunciations in the CMU Dictionary and the Indian English syllabification and stress assignment rules caused the CMU+IESyl pronunciation model to under perform. We are interested in improving the syllabification and stress assignment rules used for Indian Languages to be suitable for use with Indian English pronunciation modeling. Also, we would like to study the use of a larger manually modified pronunciation lexicon to improve the IE pronunciation model.

5. Conclusions

We conducted two experiments to evaluate new models for improving an existing Indian English voice. We found that voice conversion can be a useful technique for creating new voices with little data from an existing voice, particularly when the new voice and the existing voice share qualitative characteristics.

We also find that an Indian English pronunciation model can be the key to building a better Indian English voice. We experimented with a small manually corrected lexicon and found that it helps in improving the intelligibility of the voice. Further it may be noted that the Indian English lexicon was bootstrapped from American English letter to sound rules and only 26.3% words needed corrections. This can be an efficient technique for creating a non-native pronunciation lexicon.

While a better pronunciation lexicon is crucial in building a good pronunciation model, it may be worthwhile to further investigate the individual roles of syllabification and stress assignment. Also, the use of new phoneme set designed to incorporate the peculiarities of an Indian English phonology can be part of the next steps.

6. Acknowledgements

We thank our collaborators Raghavendra E. and Bhaskar T. at IIIT Hyderabad in helping us conduct the second study. Also we thank fellow graduate students at CMU and students at IIIT Hyderabad for participating in the experiments.

7. References

[1] Balridge, J. "Linguistic and Social characteristics of Indian English", *Language in India, Vol. 2, 2002*.

[2] Kominek, J. and Black, A. W., "The CMU Arctic speech databases", *5th ISCA Speech Synthesis Workshop, Pittsburgh, PA*, 2004.

[3] "SCRIBE – Spoken Corpus of British English," http://www.phon.ucl.ac.uk/resource/scribe/, 1990

[4] Carnegie Mellon University, "The CMU pronunciation dictionary", http://www.speech.cs.cmu.edu, 2000

[5] Prahallad, L., Prahallad, K. and GanapathiRaju, M., "A Simple Approach for Building Transliteration Editors for Indian Languages", *Journal of Zhejiang University Science, vol. 6A, no.11, pp. 1354-1361, Oct 2005.*

[6] Prahallad, K., Kumar, R., Sangal, R., "A Data-Driven Synthesis Approach for Indian Languages using Syllable as a basic unit," *International Conference on NLP, Mumbai, India, 2002.*

[7] Black, A. W. and Lenzo, K. A., "Building Synthetic Voices – for Festvox 2.1," *2007*, http://festvox.org/bsv/.

[8] Toda, T., Black, A. W., Tokuda, K., "Spectral Conversion Based on Maximum Likelihood Estimation Considering Global Variance of Converted Parameter," *Intl. Conf. on Acoustics, Speech and Signal processing, Philadelphia, Pennsylvania, 2005*.

[9] Black, A. W., Lenzo, K. A., Pagel, V. "Issues in Building General Letter to Sound Rules," *3rd ESCA Workshop on Speech Synthesis, pp. 77-80, Australia, 1998.*

[10] Black, A. W., and Taylor, P. A., "The Festival Speech Synthesis System: System documentation," *Technical Report HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Scotland, UK, 1997.*

[11] Black, A. W. and Taylor, P. A., "Automatically clustering similar units for unit selection in speech synthesis," *Proceedings of Eurospeech97, vol. 2 pp.601-604, Rhodes, Greece, 1997.*

[12] Black, A. W., "CLUSTERGEN: A Statistical Parametric Synthesizer using Trajectory Modeling," *Interspeech 2006 - ICSLP, Pittsburgh, PA., 2006.*

[13] Sen, A. and Samudravijaya, K., "Indian accent text-to-speech system for web browsing," *Sadhana, Vol. 27, Part 1, pp. 113-126 February, 2002*.

[14] Sen, A., "Pronunciation Rules for Indian English Text-to-Speech system," *ISCA Workshop on Spoken Language Processing, Mumbai, India, 2003*.

[15] Kumar, R., Kataria, A., Sofat, S., "Building Non-Native Pronunciation Lexicon for English Using a Rule Based Approach," *International conference on NLP, Mysore, India, 2003*.

[16] Mullick, Y. J., Agrawal, S. S., Tayal, S., Goswami, M., "Text-to-phonemic transcription and parsing into mono-syllables of English text," *Journal of the Acoustical Society of America, Volume 115, Issue 5, pp. 2544, 2004.*