Is word-to-phone mapping better than phone-phone mapping for handling English words?

Naresh Kumar Elluru
Speech and Vision Lab
IIIT Hyderabad, India
nareshkumar.elluru@research.iiit.ac.in

Anandaswarup Vadapalli
Speech and Vision Lab
IIIT Hyderabad, India
anandaswarup.vadapalli@research.iiit.ac.in

Raghavendra Elluru
Speech and Vision Lab
IIIT Hyderabad, India
raghavendra.veera@gmail.com

Hema Murthy
Department of CSE
IIT Madras, India
hema@iitm.ac.in

Kishore Prahallad
Speech and Vision Lab
IIIT Hyderabad, India
kishore@iiit.ac.in

Abstract

In this paper, we relook at the problem of pronunciation of English words using native phone set. Specifically, we investigate methods of pronouncing English words using Telugu phoneset in the context of Telugu Text-to-Speech. We compare phone-phone substitution and word-phone mapping for pronunciation of English words using Telugu phones. We are not considering other than native language phoneset in all our experiments. This differentiates our approach from other works in polyglot speech synthesis.

1 Introduction

The objective of a Text-to-Speech (TTS) system is to convert a given text input into a spoken waveform. Text processing and waveform generation are the two main components of a TTS system. The objective of the text processing component is to convert the given input text into an appropriate sequence of valid phonemic units. These phonemic units are then realized by the waveform generation component. For high quality speech synthesis, it is necessary that the text processing unit produce the appropriate sequence of phonemic units, for the given input text.

There has been a rise in the phenomenon of "code mixing" (Romaine and Kachru, 1992). This is a phenomenon where lexical items of two languages appear in a single sentence. In a multilingual country such as India, we commonly find Indian language text being freely interspersed with English words and phrases. This is particularly noticeable in the case of text from web sources like blogs, tweets etc. An informal analysis of a Telugu blog on the web showed that around 20-30% of the text is in English (ASCII) while the remaining is in Telugu (Unicode). Due to the growth of "code mixing" it has become necessary to develop strategies for dealing with such multilingual text in TTS systems. These multilingual TTS systems should be capable of synthesizing utterances which contain foreign language words or word groups, without sounding unnatural.

The different ways of achieving multilingual TTS synthesis are as follows (Traber et al., 1999; Latorre et al., 2006; Campbell, 1998; Campbell, 2001).

1. **Separate TTS systems for each language:** In this paradigm, a seperate TTS system is built for each language under consideration. When the language of the input text changes, the TTS system also has to be changed. This can only be done between two sentences/utterances and not in the middle of a sentence.

2. **Polyglot speech synthesis:** This is a type of multilingual speech synthesis achieved using a single TTS system. This method involves recording a multi language speech corpus by someone who is fluent in multiple languages. This speech corpus is then used to build a multilingual TTS system. The primary issue with polyglot speech synthesis is that it requires development of a combined phoneset, incorporating phones from all the languages under consideration. This is a time consuming process requiring linguistic knowledge of both languages. Also, finding a speaker fluent in mul-

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 196–200,
Sofia, Bulgaria, August 4-9 2013.

tiple languages is not an easy task.

3. **Phone mapping:**
   This type of multilingual synthesis is based upon phone mapping, whereby the phones of the foreign language are substituted with the closest sounding phones of the primary language. This method results in a strong foreign accent while synthesizing the foreign words. This may not always be acceptable. Also, if the sequence of the mapped phones does not exist or is not frequently occurring in the primary language, then the synthesized output quality would be poor. Hence, an average polyglot synthesis technique using HMM based synthesis and speaker adaptation has been proposed (Latorre et al., 2006). Such methods make use of speech data from different languages and different speakers.

In this paper, we relook at the problem of pronunciation of English words using native phone set. Specifically, we investigate methods of pronouncing English words using Telugu phoneset in the context of Telugu Text-to-Speech. *Our motivation for doing so, comes from our understanding of how humans pronounce foreign words while speaking. The speaker maps the foreign words to a sequence of phones of his/her native language while pronouncing that foreign word. For example, a native speaker of Telugu, while pronouncing an English word, mentally maps the English word to a sequence of Telugu phones as opposed to simply substituting English phones with the corresponding Telugu phones*. Also, the receiver of the synthesized speech would be a Telugu native speaker, who may not have the knowledge of English phone set. Hence, approximating an English word using Telugu phone sequence may be more acceptable for a Telugu native speaker.

We compare phone-phone substitution and word-phone mapping (also referred to LTS rules) for the pronunciation of English words using Telugu phones. We are not considering other than native language phoneset in all our experiments. This differentiates our work from other works in polyglot speech synthesis.

2 Comparison of word-phone and phone-phone mapping

Table 1 shows an example of the word *computer* represented as a US English phone sequence, English phone-Telugu phone mapping and English word-Telugu phone mapping, along with the corresponding IPA transcription. The English word-Telugu phone mapping is not a one to one mapping, as it is in the case of English phone-Telugu phone mapping. Each letter has a correspondence with one or more than one phones. As some letters do not have a equivalent pronunciation sound (the letter is not mapped to any phone) the term _epsilon_ is used whenever there is a letter which does not have a mapping with a phone.

	Computer
US English Phones	/k ax m p y uw t er/ [k ə m p j u t ɝ]
phone-phone mapping	/k e m p y uu t: r/ [k e m p j u: ʈ r]
word-phone mapping	/k a m p y uu t: a r/ [k a m p j u: ʈ a r]

Table 1: English word *computer* represented as US English phone sequence, US English phone-Telugu phone mapping and English word-Telugu phone mapping

To compare word-phone (W-P) mapping and phone-phone (P-P) mapping, we manually prepared word-phone and phone-phone mappings for 10 bilingual utterances and synthesized them using our baseline Telugu TTS system. We then performed perceptual listening evaluations on these synthesized utterances, using five native speakers of Telugu as the subjects of the evaluations. The perceptual listening evaluations were setup both as MOS (mean opinion score) evaluations and as ABX evaluations. An explanation of MOS and ABX evaluations is given in Section 4. Table 2 shows that results of these evaluations.

MOS		ABX		
W-P	P-P	W-P	P-P	No. Pref
3.48	2.66	32/50	4/50	14/50

Table 2: Perceptual evaluation scores for baseline Telugu TTS system with different pronunciation rules for English

An examination of the results in Table 2 shows that manually prepared word-phone mapping is preferred perceptually when compared to manual phone-phone mapping. The MOS score of 3.48 indicates that native speakers accept W-P mapping for pronouncing English words in Telugu TTS.

For the remainder of this paper, we focus exclusively on word-phone mapping. We propose a method of automatically generating these word-phone mapping from data. We experiment our approach by generating a word-phone mapping which maps each English word to a Telugu phone sequence (henceforth called EW-TP mapping). We report the accuracy of learning the word-phone mappings both on a held out test set and on a test set from a different domain. Finally, we incorporate this word-phone mapping in our baseline Telugu TTS system and demonstrate its usefulness by means of perceptual listening tests.

3 Automatic generation of word-phone mapping

We have previously mentioned that letter to phone mapping is not a one to one mapping. Each letter may have a correspondence with one or more than one phones, or it may not have correspondence with any phone. As we require a fixed sized learning vector to build a model for learning word-phone mapping rules, we need to align the letter (graphemic) and phone sequences. For this we use the automatic epsilon scattering method.

3.1 Automatic Epsilon Scattering Method

The idea in automatic epsilon scattering is to estimate the probabilities for one letter (grapheme) G to match with one phone P, and then use string alignment to introduce epsilons maximizing the probability of the word's alignment path. Once the all the words have been aligned, the association probability is calculated again and so on until convergence. The algorithm for automatic epsilon scattering is given below (Pagel et al., 1998).

Algorithm for Epsilon Scattering :
/*Initialize $prob(G, P)$ the probability of G matching P*/
1. for each $word_i$ in training_set
count with string alignment all possible G/P association for all possible epsilon positions in the phonetic transcription
/* EM loop */
2. for each $word_i$ in training_set
alignment_path = $argmax \prod_{i,j} P(G_i, P_j)$
compute $prob_{new}(G, P)$ on alignment_path
3. if($prob \neq prob_{new}$) go to 2

3.2 Evaluation and Results

Once the alignment between the each word and the corresponding phone sequence was complete, we built two phone models using Classification and Regression Trees (CART). For the first model, we used data from the CMU pronunciation dictionary where each English word had been aligned to a sequence of US English phones (EW-EP mapping). The second model was the EW-TP mapping.

Once both the models had been built, they were used to predict the mapped phone sequences for each English word in the test data. For the purposes of testing, we performed the prediction on both held out test data as well as on test data from a different domain. The held out test data was prepared by removing every ninth word from the lexicon.

As we knew the correct phone sequence for each word in the test data, a ground truth against which to compute the accuracy of prediction was available. We measured the accuracy of the prediction both at the letter level and at the word level. At the letter level, the accuracy was computed by counting the number of times the predicted letter to phone mapping matched with the ground truth. For computing the accuracy at the word level, we counted the number of times the predicted phone sequence of each word in the test data matched with the actual phone sequence for that word (derived from the ground truth). We also varied the size of the training data and then computed the prediction accuracy for each model. We did so in order to study the effect of training data size on the prediction accuracy.

Tables 3, 4 show the accuracy of the models. An examination of the results in the two tables shows that incrementally increasing the size of the training data results in an increase of the prediction accuracy. The native speakers of Indian languages prefer to speak what is written. As a result there are fewer variations in word-phone mapping as compared to US English. This is reflected in our results, which show that the word level prediction accuracy is higher for EW-TP mapping as compared to EW-EP mapping.

Training set size	Held-out(%)		Testing(%)	
	Letters	words	Letters	words
1000	92.04	39	81.43	16.6
2000	94.25	44.98	82.47	17.5
5000	94.55	47	84.40	25.1
10000	95.82	59.86	89.46	44.7
100000	94.09	56.37	93.27	55.10

Table 3: Accuracy of prediction for English word - English phone mapping

Training set size	Held-out(%)		Testing(%)	
	Letters	words	Letters	words
1000	92.37	28	82.22	18.8
2000	94.34	45.45	83.79	25.1
5000	95.89	68.2	88.40	42.7
10000	96.54	71.67	94.74	70.9

Table 4: Accuracy of prediction for English word-Telugu phone mapping

4 Integrating word-phone mapping rules in TTS

For the purpose of perceptual evaluations we built a baseline TTS systems for Telugu using the HMM based speech synthesis technique (Zen et al., 2007).

To conduct perceptual evaluations of the word-phone mapping rules built from data in 3.2, we incorporated these rules in our Telugu TTS system. This system is henceforth refered to as T_A. A set of 25 bilingual sentences were synthesized by the Telugu TTS, and ten native speakers of Telugu performed perceptual evaluations on the synthesized utterances. As a baseline, we also synthesized the same 25 sentences by incorporating manually written word-phone mapping for the English words, instead of using the automatically generated word-phone mapping rules. We refer to this system as T_M.

The perceptual evaluations were set up both as MOS (mean opinion score) evaluations and as ABX evaluations. In the MOS evaluations, the listeners were asked to rate the synthesized utterances from all systems on a scale of 1 to 5 (1 being worst and 5 best), and the average scores for each system was calculated. This average is the MOS score for that system. In a typical ABX evaluation, the listeners are presented with the the same set of utterances synthesized using two systems A and B, and are asked to mark their preference for either A or B. The listeners also have an option of marking no preference. In this case, the listeners were asked to mark their preference between T_A and T_M. The results of the perceptual evaluations are shown in Table 5.

MOS		ABX Test		
T_M	T_A	T_M	T_A	No. Pref
3.48	3.43	51/250	38/250	161/250

Table 5: Perceptual results comparing systems T_M and T_A

An examination of the results shows that perceptually there is no significant preference for the manual system over the automated system. The MOS scores also show that there is not much significant difference between the ratings of the manual and the automated system.

5 Conclusions

In this paper we present a method of automatically learning word-phone mapping rules for synthesizing foreign words occurring in text. We show the effectiveness of the method by computing the accuracy of prediction and also by means of perceptual evaluations. The synthesized multilingual wave files are available for download at `https://www.dropbox.com/s/7hja51r5rpkz5mz/ACL-2013.zip`.

6 Acknowledgements

This work is partially supported by MCIT-TTS consortium project funded by MCIT, Government of India. The authors would also like to thank all the native speakers who participated in the perceptual evaluations.

References

A.W. Black and K. Lenzo. 2004. Multilingual Text to Speech synthesis. In *Proceedings of ICASSP*, Montreal, Canada.

N. Campbell. 1998. Foreign language speech synthesis. In *Proceedings ESCA/COCOSDA workshop on speech synthesis*, Jenolan Caves, Australia.

N. Campbell. 2001. Talking foreign. Concatenative speech synthesis and the language barrier. In *Proceedings Eurospeech*, pages 337–340, Aalborg, Denmark.

J. Latorre, K. Iwano, and S. Furui. 2006. New approach to polygot speech generation by means of an HMM based speaker adaptable synthesizer. *Speech Communication*, 48:1227–1242.

V. Pagel, K. Lenzo, and A.W. Black. 1998. Letter to sound rules for accented lexicon compression. In *Proceedings of ICSLP 98*, volume 5, Sydney, Australia.

Suzzane Romaine and Braj Kachru. 1992. *The Oxford Companion to the English Language*. Oxford University Press.

C. Traber, K. Huber, K. Nedir, B. Pfister, E. Keller, and B. Zellner. 1999. From multilingual to polyglot speech synthesis. In *Proceedings of Eurospeech 99*, pages 835–838.

H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, and K. Tokuda. 2007. The HMM-based speech synthesis system version 2.0. In *Proceedings of ISCA SSW6*, Bonn, Germany.