INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

On building mixed lingual speech synthesis systems

SaiKrishna Rallabandi, Alan W Black

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA

srallaba@cs.cmu.edu, awb@cs.cmu.edu

Abstract

Codemixing - phenomenon where lexical items from one language are embedded in the utterance of another- is relatively frequent in multilingual communities. However, TTS systems today are not fully capable of effectively handling such mixed content despite achieving high quality in the monolingual case. In this paper, we investigate various mechanisms for building mixed lingual systems which are built using a mixture of monolingual corpora and are capable of synthesizing such content. First, we explore the possibility of manipulating the phoneme representation: using target word to source phone mapping with the aim of emulating the native speaker intuition. We then present experiments at the acoustic stage investigating training techniques at both spectral and prosodic levels. Subjective evaluation shows that our systems are capable of generating high quality synthesis in codemixed scenarios.

1. Introduction

Data driven statistical parametric speech synthesis systems have displayed a continued improvement over the recent past, in terms of speech quality [11, 20]. These improvements can be attributed to developments in the aspects such as speech parameterization [10, 16, 23], modeling[5, 27], post filtering[24, 17, 7] and have led to deployment in consumer grade systems[26]. Currently, such Text to Speech (TTS) systems assume that the input is in a single language and that it is written in native script. However, due to the rise in globalization, phenomenon such as code mixing / code switching are now seen in various types of text ranging from news articles through comments/posts on social media, leading to co-existence of multiple languages in the same sentence. Incidentally, these typically are the scenarios where TTS systems are widely deployed as speech interfaces and therefore these systems should be able to handle such input. Even though independent monolingual synthesizers today are of very high quality, they are not fully capable of effectively handling such mixed content that they encounter when deployed. These synthesizers in such cases speak out the wrong/accented version at best or completely leave the words from the other language out at worst. Considering that the words from other language(s) used in such contexts are often the most important content in the message, these systems need to be able to handle this scenario better.

In the next subsection, we first discuss briefly the issue of codemixing and then highlight the kind of codemixing that we are dealing with.

1.1. Code Mixing

CodeMixing is a phenomenon where linguistic units such as phrases, words and morphemes of one language are embedded into an utterance of another language [9]. This is a common phenomenon in multilingual societies such as in India where English has transitioned from the status of a foreign language to that of a second language. Moreover, due to the diversity in terms of the regionality and the proficiency, the patterns of codemixing found are rather different from one another, often leading to confusion. [18] states that there are, in general, three types of mixing:

- *insertion* or *embedding* of content (lexical items or entire constituents) from English into a regional language.
- *alternation* between structures from both the languages.
- *congruent lexicalization* of material from different lexical inventories into a shared grammatical structure.

[18] also identifies that the lexical items that can be inserted during mixing are adverbarial phrases, single nouns and determiner-nouns. We performed an informal analysis on a Hindi recipe blog on the web and found that while the content was all in ASCII, around 15 percent of the words were English words (though often misspelt), and almost all of them were nouns or adverbs, in line with the observation in [18]. A similar analysis of a Telugu blog showed that around 20-30 percent of the text is in English (ASCII), and again, most of them being nouns. Such mixed text data poses a variety of challenges to the speech synthesis system due to their innate characteristics such as contractions, non-standard pronunciation, and non-standard sentence constructions, etc.

Current approaches handling this scenario fall into three categories: phone mapping, multilingual or polyglot. In phone mapping, the phones of the foreign language are substituted with the closest sounding phones of the primary language, often resulting in a strong accented speech. In a multilingual setting, each text portion in a different language is synthesised by a corresponding monolingual TTS system. This typically means that the different languages will have different voices unless each of the voices is trained on the voice of same multilingual speaker. The polyglot solution refers to the case where a single system is trained using data from a multilingual speaker. Similar approaches to dealing with codemixing have been focused on assimilation at the linguistic level, and advocate applying a foreign linguistic model to a monolingual TTS system. The linguistic model might include text analysis and normalisation, a G2P module and a mapping between the phone set of the foreign language and the primary language of the TTS system [25, 6, 2]. Other approaches utilise cross-language voice conversion techniques [15] and adaptation on a combination of data from multiple languages[14]. Assimilation at the linguistic level is fairly successful for phonetically similar languages [2], and the resulting foreign synthesized speech was found to be more intelligible compared to an unmodified non-native monolingual system but still retains a degree of accent of the primary language. This might in part be attributed to the non-exact correspondence between individual phone sets.

In this paper, we investigate approaches to build mixed-lingual speech synthesis systems based on separate recordings

http://dx.doi.org/10.21437/Interspeech.2017-1244

Table 1: *Overview of Systems with variation in Grapheme to Phoneme Mapping.*

Level	Config	Description	Example
Text	P2P_Mono	Phone to Phone in Monolingual System	Stanford → s t ae n f er d → s tr e nB ph E 9r dr
Text	W2P_Mono	Word to Phone in Monolingual System	Stanford → s tx aa n oo r dx
Text	Translit_Mono	Word to transliteration in Monolingual System	Stanford → native script of the language
Spectral	Separate	Combined phoneset with language tag	Stanford → s_E t_E ae_E n_E f_E er_E d_E
Spectral	Shared	Combined phoneset without language tag	Stanford → s t ae n f er d
Spectral	Word	Language of word as question	Same as in Soft_Tag
Spectral	WC	Tri context for word is used as question	Same as in Soft_Tag
Prosodic	BL	Baseline statelevel Duration Prediction	NA
Prosodic	Ratio	Ratio used for modification	NA
Prosodic	Gauss	Gaussian used for modification	NA
Prosodic	OLA	Only Look ahead features	NA
Prosodic	OPF	Only phonetic features	NA

in the individual languages with the ability to appropriately synthesize the 'embedded' lexical items of English , thereby leading to a more natural output. Specifically, we build on the work done in [26] and investigate various techniques to train Indic voices that can speak both the primary language and also high-quality English, for the common situation in which English text is encountered in a primarily Indian language document. We present systems at three different levels: Text level, acoustic modeling level and prosody modeling and try to answer the questions: (1) What modifications should we do at the G2P level so that the current systems can handle mixed text ? (2) How to train effective acoustic models that can handle mixed phone set ? and (3) What are the changes required at the prosodic level for generating natural sounding prosody? In section 2, we present the formulation and description of approaches at the text, spectral and prosodic levels. In section 3, we explain our experiments followed by evaluation and conclusion in section 4 .

2. Mixed Lingual Systems

In this section, we first present the formulation of our text based approaches and then describe our systems at spectral and prosodic levels.

2.1. Data

We have used speech and text data from 4 languages to build the systems described in this paper: Hindi,Telugu, Marathi and Tamil. For Hindi, we have used the Mono and English parts of the male speaker from speech data released as a part of resources for Indian languages [1]. We noticed that the Hindi utterances were longer and therefore used all the 1,132 prompts from the Arctic set but only the first 600 prompts from the Mono set so that both Hindi and English utterances are of equal duration (approximately an hour each). For the remaining languages, we have used the speech and text data that has been collected for [26]. In all these cases speech data was sampled at 16 kHz and recorded in a high quality studio environment. For Telugu and Tamil we have used the recordings from female speakers and for Marathi, from male speaker thereby ending up with systems from two male and two female speakers overall. For evaluation, we have used the test sentences from multilingual category from [20] for the respective languages.

2.2. Approaches for Grapheme to Phoneme Conversion

Grapheme-to-phoneme conversion (G2P) is one of the first tasks in speech synthesis and essentially is a conversion from a word in orthography to its spoken form or pronunciation. This can be seen, in an oversimplification, as maximizing $Prob(P/G)$ where P is the phoneme sequence and G is the grapheme sequence of a single language. However, in case of 'embedding', the G contains graphemes from both native language as well as English. In this case, a phone to phone mapping is employed to map the phones from English to the native language.

However, in practice, this method results in a strong foreign accent while synthesizing the english words. [8, 21] proposed a method to use a word to phone mapping instead, where an english word is statistically mapped to Indian language phones. This can be seen as maximizing the expression:

$$\prod_{i,j \in S,W} Prob(s_i|w_j) \tag{1}$$

where $w \in W^d$ is a word in source language with a vocabulary(W) of size d. The intuition in this can be seen as

$$\prod_{i,j,k \in S,M,W} Prob(m_k|w_j) * Prob(s_i|m_k) \tag{2}$$

where $m \in M$ was referred to as the mental mapping of the native speaker. In this paper, we take a more direct approach and investigate the use of transliterations as the phoneme internal representation. We hypothesize that there is a single model which has generated both the transliteration and the phoneme itself. This serves as a more concrete mapping problem and can be seen as maximizing the expression

$$\prod_{i,j,k \in S,T,W} Prob(t_k|w_j) * Prob(s_i|t_k) \tag{3}$$

where $concat(t) \in T$ is the transliterated form. We have used this approach previously, [22], but this the first time we are systematically comparing the three possible G2P approaches in such scenarios.

2.2.1. Systems built

The systems we built at this level are mentioned in table 1. We have built a monolingual system (P2P_Mono) with phone to phone mapping as a baseline method. We then built systems

W2P_Mono and Translit_Mono with word to phone and transliteration applied on the English words respectively.

2.2.2. *Pipeline*

We follow a three step procedure. First, we identify the language of each individual word in the sentence. This is using a very simple method- based on the orthography of the word. We then apply the appropriate grapheme to phoneme conversion technique to the English words and obtain pronunciation, taking into account the corresponding postlexical rules. The final step is to generate speech using the sequence of phonemes.

2.3. Approaches for Acoustic Modeling

There are two dimensions in which we can vary the input features for synthesis. First at the phone level itself, choosing to explicitly separate the phones by original language (we add a language id suffix to the phone name), or taking the union of the phones across the languages (e.g. the data for English /t/ and Hindi /t/ are treated as one class). Secondly we provide contextual features to identify the language that the phone actually appears in (e.g. is it in an English or a Hindi word – and also the language id of the surrounding words). In the second case of language contextual tagging, these features may allow pronunciation distinctions between longer phrases in a particular language and isolated words in a codemixed utterance.

2.3.1. *Systems built*

The systems we built at this level are mentioned in table 1. The system 'Separate' uses a combined phoneset, where the phones of English are explicitly separated from the phones of Indic by adding a language tag _E or _I denoting English and Indic respectively. The system 'Shared', as the name indicates, uses a combined phoneset obtained by the union of phones from English and Indic phonesets - if a phone is common in both the sets it is retained as is, and the disjoint phones are added separately. We then build systems 'Word' and 'WC' incorporating contextual features to identify the language. These systems differ in the amount of context used. The system 'WC' uses a tri word context while the system 'Word' uses a single word context.

2.4. Approaches for Modeling Prosody

In bilingual sentences, it was shown that the context of source language will influence the embedded target language words, which will change the original prosody of the target language [28]. In order to establish a mixed-language speech synthesis system based on separate corpora, it is therefore important to consider such influences to generate natural mixed-lingual prosody. From the corpus, we didnot observe much differences in pitch between the monolingual and mixed lingual synthesis systems, as in [28]. We suspect that this might be due to the proficiency level of bilingual speech, which is used on a daily basis in India. However, we did observe marked differences in the duration of the words, specifically at the point of switch from source language to the target language. In this subsection, we present our approaches to build combined prosodic models using separate monolingual speech and text data. We specifically look at two different ways of achieving this: (1) By manipulating the durations predicted by the baseline model and (b) Incorporating extra features while building the model, thereby modifying the prediction model itself.

- **Ratio based Manipulation System** [28] In this system we first obtain the predicted durations from the baseline prosodic model and then transform the durations of English segments to account for the contextual effect that might be caused by the source language. The multiplication factor λ was obtained using the mean durations of the segments during the training stage.

$$dur_{new} = \lambda * dur_{pred} \tag{4}$$

- **Gaussian based mapping system**
  We have observed that there seem to be two separate gaussian distributions followed by the phones from arctic and the indic recording sets. Therefore, we built system which modifies the durations of individual phonemes based on the following:

$$dur_{new} = \lambda * \frac{\sigma_{indic}}{\sigma_{arctic}} * (dur_{pred} - \mu_{arctic}) + \mu_{indic} \tag{5}$$

  where μ and σ indicate the mean and standard deviation of the individual phonemes respectively.
- **System with only look ahead models** - These are the models that donot take the previous context into account while predicting the duration of the current state.
- **System with only phonetic feature based models** - These are models trained without taking the names of phones into account and considering only the context in which they occur.

3. Experiments

All the systems were built using standard Clustergen [3] voice building procedure. Systems P2P_Mono and P2P_Multi were built using phone matching and systems W2P_Mono and W2P_Multi were built using epsilon scattering method [4], the idea in which is to estimate the probabilities for one grapheme G to match with one phone P, and then use string alignment to introduce epsilons maximizing the probability of the words alignment path. We have followed the same procedure outlined in [8]. To transliterate the English words from the Romanized script to the native script as a part of the system Translit_Mono, we have used Brahmi-Net transliteration [13] which considers this problem as a phrase based translation. The sequences of characters from source to the target language are learnt using a parallel corpus trained using Moses [12]. This system supports 13 Indo-Aryan languages, 4 Dravidian languages and English including 306 language pairs for statistical transliteration. The systems at spectral and prosodic levels were built varying the question sets in the clustergen[3] voice building process appropriately.

Table 2: *MOS Scores for Naturalness in Text G2P based experiments*

Lang/Config	P2P_Mono	W2P_Mono	Translit_mono
Hi-Eng	2.7	3.9	3.1
Tel-Eng	3.1	3.8	3.3
Mar-Eng	2.4	3.3	-
Tam-Eng	2.6	3.4	-

Table 3: *Results from Preference Test for Spectral Mapping Experiments among the systems using separate and shared phonesets*

Config	Separate	Shared	No Preference
Hi-En	78/400	**286**/400	36/400
Te-En	56/250	**172**/250	22/250
Ma-En	48/250	**167**/250	35/250
Ta-En	63/250	**144**/250	47/250

Table 4: *Results from Preference Test for Spectral Mapping Experiments among the systems using different levels of word context. Both these systems use shared phonesets*

Config	Word	WC	No Preference
Hi-En	4/50	7/50	**39**/50
Te-En	16/50	**22**/50	12/50
Ma-En	11/50	**26**/50	13/50
Ta-En	13/50	**19**/50	18/50

3.1. Evaluation

Evaluation was performed in the form of listening tests using [19]. We have conducted two types of listening tests: (1) Rating the naturalness in terms of Mean Opinion Score (MOS) on a scale of 1(least natural) to 5(highly natural) and (2) ABX Preference test where the users need to mention their preference towards either of the systems or state that they prefer neither. The systems using grapheme to phoneme based approaches and prosodic mapping have been tested using the former while the rest of the systems, i.e spectral modeling systems have been evaluated using preference tests. All the listening tests involved test sentences generated using the Multilingual test set (ML) from [20].

3.2. Discussion

3.2.1. Front End

The evaluation results for the front end systems are presented in table 2. The word to phone mapping based systems seem to outperform the rest of the systems across all the languages consistently. The transliteration based systems seem to be performing better than the basic phone mapping based systems in the languages they were deployed in, but seem to be lagging behind the word to phone mapping systems. An informal preference test showed that the transliterated system does reduce the accented nature of phone mapping procedure for some words, but the reduction itself was not found to be as much as that obtained by the word to phone based modeling approach. Therefore it appears that using a separate orthographic system might not necessarily result in the best quality synthesis. One reason for this might be that the errors in the transliteration process itself act as barriers hindering the system from reaching its full potential in terms of voice quality.

3.2.2. Spectral Modeling

The evaluation results for the spectral systems are presented in table 3 and 4. Each of the systems was used to generate 50 sentences from the test set which were evaluated students who were native speakers of the respective languages. The systems in Hindi were evaluated by 8 students, leading to 400 observation points while those from remaining languages were evaluated by 5 students resulting in 250 observation points. We observe from table 3 that across all the 4 languages, the systems using a shared phoneset are preferred in a significant manner. From the results in 4, where all the systems were built using the shared phoneset due to the observed preference from 3, it appears that using the word context definitively doesnot deteriorate the system; at the same time it does not lead to a substantial imporvement either. We analysed the sentences that were not preferred by either of the systems in this case and found atleast two interesting characteristics common across the languages: (1) The English parts of the sentences seems to be a bit dull and flattened out compared to the native language counterparts. We hypothesize that this might be due to the manner in which the models were trained: using separate corpora as opposed to a multilingual corpus which has codemixed sentences, leading to a train test mismatch no matter how shared the phonesets are. The model when predicting the English segments might therefore be tending towards the mean of the training observations due to lack of proper context. It might be interesting to see if this artifact can be corrected/addressed by either using a codemixed database or by using some form of adaptation. (2) There seems to be an uncharacteristic gap between the English words which have two parts, ex: shortcut as short PAU cut , download as down PAU load, etc. We have anticipated this behaviour at the point of switch between the languages and therefore explicitly tried to model the context, but we didnot expect this in case of English words. In addition, we have observed that the stress pattern in these instances became a bit wierd.

Table 5: *MOS Scores for Naturalness in prosodic modeling based experiments*

Config	Baseline	Ratio	Gauss	OLA	OPF
Hi-Eng	3.9	3.8	3.8	3.6	3.4
Tel-Eng	3.6	3.7	3.5	3.4	3.5
Mar-Eng	3.7	3.7	-	-	-
Tam-Eng	3.4	3.3	-	-	-

3.2.3. Prosodic Modeling

In this case, none of the systems that we have tried were successful in completely eliminating the seemingly disjoint(and therefore sudden variations in) speeds in the English and the Indic parts of test sentences. This is also clear from the evaluation results for the front end systems are presented in table 5. Surprisingly, none of the systems were able to outperform the baseline system, indicating that incorporating modifications artificially in the duration of the sentences is easily noticeable as unnatural by human evaluators. As a part of ongoing work, we hope to uncover some aspects in this by analysing the set of questions the test sentences pass through before final prediction. There might also be interesting adaptation techniques that can help reduce this.

4. Conclusion

In this paper, we investigated approaches to build mixed-lingual speech synthesis systems based on separate recordings and present systems at three different levels. From evaluations,we have identified interesting issues which occurred as a result of the train-test discrepancy. We are investigating them as an ongoing work and hope to understand them as well as formulate better techniques to handle the codemixed text.

5. References

[1] A. Baby. Resources for indian languages. In *CBBLR workshop, International Conference on Text, Speech and Dialogue,2016*.

[2] L. Badino, C. Barolo, and S. Quazza. Language independent phoneme mapping for foreign tts. In *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[3] A. W. Black. Clustergen: a statistical parametric synthesizer using trajectory modeling. In *INTERSPEECH*, 2006.

[4] A. W. Black, K. Lenzo, and V. Pagel. Issues in building general letter to sound rules. 1998.

[5] A. W. Black and P. K. Muthukumar. Random forests for statistical speech synthesis. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] N. Campbell. Talking foreign. In *Proc. Eurospeech*, pages 337–340, 2001.

[7] L.-H. Chen, T. Raitio, C. Valentini-Botinhao, Z.-H. Ling, and J. Yamagishi. A deep generative architecture for postfiltering in statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(11):2003–2014, 2015.

[8] N. K. Elluru, A. Vadapalli, R. Elluru, H. Murthy, and K. Prahallad. Is word-to-phone mapping better than phone-phone mapping for handling english words? In *ACL (2)*, pages 196–200, 2013.

[9] S. Gella, K. Bali, and M. Choudhury. ye word kis lang ka hai bhai? testing the limits of word level language identification, 2014.

[10] H. Kawahara. Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds (¡ special issue¿ introduction to the amazing world of sounds with demonstrations). *Acoustical science and technology*, 27(6):349–353, 2006.

[11] S. King. The blizzard challenge 2016. 2016.

[12] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics, 2007.

[13] R. P. Kunchukuttan, Anoop and P. Bhattacharyya. Brahmi-net: A transliteration and script conversion system for languages of the indian subcontinent. In *HLT-NAACL. 2015*, 2015.

[14] J. Latorre, K. Iwano, and S. Furui. Polyglot synthesis using a mixture of monolingual corpora. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 1, pages I–1. IEEE, 2005.

[15] M. Mashimo, T. Toda, K. Shikano, and N. Campbell. Evaluation of cross-language voice conversion based on gmm and straight. 2001.

[16] M. Morise, F. Yokomori, and K. Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884, 2016.

[17] P. K. Muthukumar and A. W. Black. Recurrent neural network postfilters for statistical parametric speech synthesis. *arXiv preprint arXiv:1601.07215*, 2016.

[18] P. Muysken. *Bilingual speech: A typology of code-mixing*, volume 11. Cambridge University Press, 2000.

[19] A. Parlikar. TestVox: web-based framework for subjective evaluation of speech synthesis. *Opensource Software*, 2012.

[20] K. Prahallad, A. Vadapalli, S. Kesiraju, H. Murthy, S. Lata, T. Nagarajan, M. Prasanna, H. Patil, A. Sao, S. King, et al. The blizzard challenge 2014. In *Proc. Blizzard Challenge workshop*, volume 2014, 2014.

[21] S. K. Rallabandi, A. Vadapalli, S. Achanta, and S. Gangashetty. Iiit hyderabad's submission to the blizzard challenge. In *Proc. of Blizzard Challenge 2015*.

[22] S. Sitaram, S. K. Rallabandi, and S. R. A. W. Black. Experiments with cross-lingual systems for synthesis of code-mixed text. In *9th ISCA Speech Synthesis Workshop*, pages 76–81, 2015.

[23] F. Soong and B. Juang. Line spectrum pair (lsp) and speech data compression. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.*, volume 9, pages 37–40. IEEE, 1984.

[24] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. A postfilter to modify the modulation spectrum in hmm-based speech synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 290–294. IEEE, 2014.

[25] L. M. Tomokiyo, A. W. Black, and K. A. Lenzo. Foreign accents in synthetic speech: development and evaluation. In *Interspeech*, pages 1469–1472, 2005.

[26] A. Wilkinson, A. Parlikar, S. Sitaram, T. White, A. W. Black, and S. Bazaj. Open-source consumer-grade indic text to speech. In *9th ISCA Speech Synthesis Workshop*, pages 190–195.

[27] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King. Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4460–4464. IEEE, 2015.

[28] Y. Zhang and J. Tao. Prosody modification on mixed-language speech synthesis. In *Chinese Spoken Language Processing, 2008. ISCSLP'08. 6th International Symposium on*, pages 1–4. IEEE, 2008.