GLOBAL SYLLABLE SET FOR BUILDING SPEECH SYNTHESIS IN INDIAN LANGUAGES

E.Veera Raghavendra†, Srinivas Desai†, B. Yegnanarayana†, Alan W Black‡, Kishore Prahallad†‡

†International Institute of Information Technology - Hyderabad, India.
‡Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
{raghavendra,yegna}@iiit.ac.in, srinivasdesai@research.iiit.ac.in, {awb,skishore}@cs.cmu.edu

ABSTRACT

Indian languages are syllabic in nature where many syllables are found common across its languages. This motivates us to build a global syllable set by combining multiple language syllables to build a synthesizer which can borrow units from a different language when the required syllable is not found. Such synthesizer make use of speech database in different languages spoken by different speakers, whose output is likely to pick units from multiple languages and hence the synthesized utterance contains units spoken by multiple speakers which would annoy the user. We intend to use a cross lingual Voice Conversion framework using Artificial Neural Networks (ANN) to transform such an utterance to a single target speaker.

Index Terms— Speech synthesis, polyglot synthesis, global syllable set.

1. INTRODUCTION

In our earlier paper [1], we have proposed that syllable is a better choice of unit for syllabic/phonetic languages such as Indian languages and also discussed approximating the nearest syllable when required unit is not found. Need for designing global syllable set arises from the fact that, we can cover only a few syllables for each language. As all Indian languages have a common syllabic/phonetic base, one does not use the term "alphabet" to refer the set of letters. Instead, the set is called "Akshara". In all Indian languages, an Akshara is pronounced in same way regardless of its position within a word, unlike in English where the pronunciation varies widely, depending not only on the word but also on the location of the letter within the word. Utilizing these features we have combined multiple language databases into one and created a global syllable database. The advantage of this approach is we can build bigger syllable inventory. If we assume Indian languages have 50 phones including 35 consonants and 15 vowels and theoretically possible syllable combinations in Indian language with *CCV* representation are 18375 (35 * 35 * 15). Table 1 shows individual and global database details of unit collection. In the Table 1, *unique syllables* gives the number of unique syllables in the language and *%of syllables* gives percentage of possible syllables covered in the language. Last row *global* shows the statistics when combined multiple languages.

Table 1. *Unit collection for individual, global databases and % of syllables of covered from maximum possible combinations of syllables in a language.*

Language	Unique Syllables	% of syllables
Telugu	1790	9.74
Hindi	2757	15
Tamil	1892	10.29
Global	**4997**	**27.19**

From Table 1 we can observe that individual languages can cover maximum of 15% of possible syllables where as the global set covers 27.19%. It shows that combining multiple languages gives the good coverage of units and number of syllables common across the languages are 1442 (6439-4997). But informal studies show that the synthetic speech contains multiple voice identities and it affects naturalness. As a result we are transforming all the voices to one speaker and conducted user study for transformed and multiple voice synthesized utterances. The results are discussed in the Section 5.

The rest of the paper is organized as follows. Section 2 discusses the previous work done in the field of multilingual and polyglot synthesis. Section 3 discusses the baseline system with global syllable set. Section 4 discusses the framework for cross lingual voice conversion. Section 5 discusses the experiments on Telugu, Hindi and Tamil and perceptual and subjective evaluations.

2. PREVIOUS WORK

A multilingual synthesis [2] uses a common set of rules and algorithms to synthesize speech in multiple languages. Thus, a collection of language specific synthesizers does not qualify as a multilingual system. Ideally, all language specific information should be stored in data tables, and all algorithms should be shared by all languages. It is hard to achieve such an ideal system. The issue is that researchers tend to optimize

their methods for one language at a time. As a result, their algorithm often contain parameters that are sufficient to cover the language they have dealt. Min Chu et. al [3] have developed a multilingual TTS for English and Mandarin. Same speaker has been used to create the speech database for two languages and common rules have been implemented for text processing. Soft Prediction Only (SPO) technique is applied to normalize the pitch for both languages as English is a stress and Mandarin is a tonal language. Prosodic Constraint Oriented (PCO) approach has been used for unit selection during the synthesis. Traber et al. [4] proposed a distinction between polyglot and multilingual systems. They defined "polyglot systems" as those that can synthesize several languages using the same voice with appropriate pronunciation, and "multilingual systems" as those that have to change the synthesis process and output voice to synthesize different languages. In [5], a HMM-based method is proposed to combine monolingual corpora from several languages to create a single polyglot average voice. With this synthesizer they can synthesize any of the languages included in the training data with the same output voice and speech quality by means of supervised MLLR adaptation [6]. Latorre et al. [7] discusses a method for approximating the sounds of languages not included in the polyglot training data. The sounds are approximated from one language to another by means of the similarity between the articulatory features of source and target phones. These features are derived from the IPA representation of the phones. When there is no similar articulatory features found between the source and target, an ad-hoc assignment was done using a linguistic expert.

In this paper we are proposing a global syllable set for building speech synthesis system in Indian languages. The use of global syllable set is similar to definition of polyglot synthesis, as we interested to use same set of syllables to generate voices in multiple languages. The distinction with our approach and polyglot [5], [7] is that they have applied *cross lingual voice adaptation* to create an average voice. In our case, we are applying *cross lingual voice conversion* to transform to global syllable set to sound as a single speaker.

3. BASELINE SPEECH SYNTHESIS SYSTEM

In our previous work we have designed a syllable based synthesizer [1] based on approximate matching of syllables. To build a global syllable set we updated the phoneset with all possible syllables from each language, Telugu, Hindi and Tamil. The lexical parser has also been modified accordingly, to generate appropriate syllables. Once the syllables are obtained, the text is synthesized using the approach described in [1].

To evaluate the synthesizer which is based on global syllable set, we have conduced subjective and objective evaluations in comparison with a diphone based synthesizer. We selected a set of 10 sentences from Telugu news bulletin. Ten subjects who participated in these perceptual tests do not have any experience in speech synthesis. Each listener rated each synthesized utterance. We used three evaluation metrics: mean opinion score on a scale from 1 (worst) to 5(best) for individual utterances, standard deviation (SD) is computed on each listener mean scores; and AB ranking tests where the listener had to choose whether the baseline synthesized utterance or the version using voice conversion was better. As a part of objective evaluations Mel-Cepstral Distortion (MCD) [8] are calculated between original and synthesized utterances. Lower the MCD value the better in speech synthesis.

The results shown in Table 2 indicate the syllable based synthesizer based on global syllable set performs poorer than diphone based synthesizer for Telugu. However, in all our previous studies we have observed that syllable performs better than the diphone in Indian languages [1]. This indicates that multiple voice identities in the synthesized speech annoys the user. Thus a major issue in the use of global syllable set is to minimize the perceptual differences obtained due to multiple voices. To address this issue we propose an approach of voice conversion.

Table 2. *Global Syllable (GSyl) Vs Diphone.*

Test	Telugu		
	GSyl	Diphone	Similar
AB-Test	26/100	43/100	31/100
MOS	2.98	3.22	-
SD	0.58	0.63	-
MCD	6.057	5.992	-

4. FRAMEWORK FOR CROSS LINGUAL VOICE CONVERSION

In this work, voice conversion is done across languages. Generally, voice conversion requires parallel database. But, it is difficult to find parallel database across the languages. This section explains how to get the parallel database across the languages and transform multiple voices to one speaker. This technique consists of two stages. In the first one, parallel database is created using target speaker TTS. In the second, source speakers are transformed to target speaker by means of ANN. In this paper we have limited our experiments to three languages: Telugu, Hindi and Tamil. Telugu and Hindi languages share common phoneset where as Tamil has less number of phones and it is subset of Telugu phoneset except two phones.

4.1. Generation of Parallel Database

To create parallel database between source and target, a set of sentences from each training speaker: Telugu, Hindi and Tamil, are selected.

Let S be the set of text sentences from source speakers.

$$s = \{s_1, s_2, ..., s_n\} \quad (1)$$

Since, the phoneset and pronunciation of all the languages are similar in Indian languages [9]. These sentences could be synthesized using target speaker TTS. Let T be the set of synthesized utterances using target TTS.

$$T = \tilde{S} = \{\tilde{S}_1, \tilde{S}_2, ..., \tilde{S}_n\} \quad (2)$$

and S be the set of original utterances from source speakers.

$$S = S_1, S_2, ..., S_n \quad (3)$$

4.2. Training Voice Conversion Models

When the target sentences are synthesized using target speaker TTS, the number of frames (MCEP vectors) may not be same with source speaker utterances. To make the frames equal, target speech frames are aligned to source speech using dynamic time warping (DTW) algorithm [10]. Artificial neural networks are used to capture the relationship between source and target features [11]. We used feedforward neural networks (FFNN) with back propagation learning law [12] to adjust the weights of the neural network which minimizes the mean squared error between the source and target features. Architecture of the network, learning rate, momentum, number of iterations and training error details are given in the experiments section.

Once the model is obtained, transformation is performed on the multiple voice identity utterance. MCEPs and F0 features are extracted using fixed frame advance of 5ms. F0 is extracted with ESPS function "get_f0" and synchronized to the MCEPs. Source speaker MCEPs are transformed to target speaker MCEPs. 25 coefficient MCEPs are combined with the linear transformed F0 [13] to give a 26 feature vector for every 5 ms. Then the speech is reconstructed from the 26 feature vector using the MLSA filter [14].

5. EXPERIMENTS

5.1. Speech Database Used

The quality of the unit selection voices depends to a large extent on the variability and availability of representative units. It is crucial to design a corpus that covers all speech units and most of their variations in a feasible size. The speech databases used for Telugu, Hindi, and Tamil are recorded by female speakers and the number of utterances, words and unique words details are given in the Table 3. All sentences are recorded in a professional studio and the sentences are read in relaxed reading style, which is between "formal reading style" and "free talk style", in moderate speed. Recordings are performed in a soundproof room with close-talking microphone.

5.2. Characteristics of the phonetic sets

The Telugu and Hindi phoneset consists of 50 phones, including 15 vowels and 35 consonants. Tamil phoneset has 41 phones including 15 vowels and 26 consonants. Tamil has 2 different phones compared to Telugu and Hindi phoneset. These phones are manually mapped to nearest phones in Telugu based on their articulatory features.

Table 3. *Language database details.*

Language	No.Of. Sentences	No.Of. Words	Unique Words
Telugu	1631	27303	8026
Hindi	585	14398	14398
Tamil	2392	33945	7817

5.3. Preparation of Parallel Database

We experimented with source as Telugu, Hindi and Tamil and target as Telugu speaker. A set of 40 sentences from each language is taken and Tamil phones are mapped to Telugu phoneset when required. Later the source sentences are synthesized using Telugu TTS with Festival framework [15].

5.4. ANN models for Voice Conversion

The training dataset contains 120 utterances consisting of 223466 segments, around 18.16 minutes of speech including less than 0.5 seconds of silence at the beginning and ending of each utterance. The durations of individual languages are 7.32, 5.45 and 5.39 for Telugu, Hindi and Tamil respectively. For modeling source and target, we employed 4 layer FFNN whose general structure is shown in [11]. The first layer is the input layer which consists of linear elements. The second and third layers are hidden layers. The fourth layer is the output layer which represents the target speaker. Activation functions at first and fourth layer are linear and at second and third layer are non linear.

Table 4 shows the various parameters used for transforming source speakers to Telugu speaker.

Table 4. *Parameters for ANN modeling.*

Type	**Parameters**
Architecture	25 L 50 N 50 N 25 L
Learning Rate	0.01
Momentum	0.3
Epochs	200
Error on Training Data	0.0802576

5.5. Synthesis

During the synthesis the input sentence is broken into words and language tag is assigned to each word. The tag would be helpful for applying language specific rules for converting word to phonetic form. Once the syllables are obtained from the given text, it is synthesized using global syllable set as explained in Section 3. Then the synthesized utterance is transformed to Telugu speaker as explained in Section 4.2.

In order to evaluate the utterances synthesized using transformed voice (referred as Global + VC), we conducted listening tests in comparison with utterances synthesized from baseline system (referred to as Global) as described in Section 3. Please note that the ten subjects participated in this perceptual study are different from the subjects participated in Table 2. Different subjects participated for different experiments to avoid any bias the subjects might hold. Table 5 show the MOS, Standard deviation (SD) and MCD scores for both the techniques.

Table 5. *MOS,SD and MCD scores for global syllable set with voice conversion (Global + VC) and global syllable set (Global).*

Test	MOS		SD		MCD	
	Global + VC	Global	Global + VC	Global	Global + VC	Global
Telugu	2.33	2.84	0.82	0.75	5.729	6.057
Hindi	2.64	3.01	0.61	0.40	5.889	6.213
Tamil	1.96	2.46	0.91	1.01	5.813	6.154

The perceptual scores shown in Table 5 indicate that subjects prefer multi speaker voice identity utterances than the transformed voice but the MCD scores show that transformed voice is better than multi speaker voice. Since the multi speaker voice utterance is generated using prerecorded segments and it sounds as human voice. Instead, the transformed voice is reconstructed using MLSA. Though the voice is intelligible and consistent but it is not as human sounding as unit selection voices. This might have created a bias in the subjects towards multi speaker voice. To validate our hypothesis, we conduced perceptual studies with ten subjects only on transformed voice for Telugu and the average MOS score and standard deviation are 3.225 and 0.20 respectively. The comparison of 3.225 with MOS score of 2.84 observed for Telugu in Table 5 shows that our hypothesis is valid and the quality of the transformed voice is better than multi speaker voice.

6. CONCLUSION

In this paper, we have discussed the need for designing global syllable set using multiple Indian languages. To avoid the multiple voice identities in the synthesized speech all the voices are transformed into a single speaker. We have built Telugu, Hindi and Tamil synthesizers using global syllable set. We conducted subjective and objective evaluations to evaluate these synthesizers between multiple voice identity utterances and transformed synthesized utterances.

7. ACKNOWLEDGMENTS

We would like to thank Venkatesh Keri, Sachin Joshi and Ramakrishna for useful discussion and suggestions during this work. We also thank all the people of LTRC and graduate students of IIIT Hyderabad for their participation in the perceptual tests.

8. REFERENCES

[1] E. V. Raghavendra, B. Yegnanarayana, A. Black, and K. Prahallad, "Building sleek synthesizers for multi-lingual screen reader," *in Proceedings of Interspeech*, pp. 1865–1868, September 2008.

[2] R. Sproat, "Multilingual text-to-speech synthesis: the bell labs approach," *Kluwer Academic Publisher*, 1998.

[3] M. Chu, H. Peng, Y. Zhao, Z. Niu, and E. Chang, "Microsoft mulan - a bilingual TTS systems," *in Proceedings of ICASSP 2003*, vol. 1, pp. I–264– I–267, April 2003.

[4] C. Traber, K. Huber, K. Nedir, B. Pfister, E. Keller, and B. Zellner, "From multilingual to polyglot speech synthesis," *in Proceedings of Eurospeech*, vol. 2, pp. 835–838, September 1999.

[5] J. Latorre, K. Iwano, and S. Furui, "Polyglot synthesis using a mixture of monolingual corpora," *in Proceedings of ICASSP*, vol. 1, pp. 1–4, 2005.

[6] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," *The Third ESCA/COCOSDA Workshop on Speech Synthesis*, pp. 273–276, November 1998.

[7] J. Latorre, K. Iwano, and S. Furui, "Cross-language synthesis with a polyglot synthesizer," *in Proceedings of Interspeech-2005*, pp. 1477–1480, September 2005.

[8] T. Toda, A. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech," *in Proceedings of 5th ISCA Speech Synthesis Workshop*, pp. 31–36, June 2004.

[9] P. Lavanya, P. Kishore, and G. Madhavi, "A simple approach for building transliteration editors for indian languages," *Journal of Zhejiang University Science*, vol. 6A, no.11, pp. 1354–1361, October 2005.

[10] Y. Stylianou, O. Cappe, and E. Moulines, "Statistical methods for voice quality transformation," *in Proceedings of Eurospeech*, pp. 447–450, Sept. 1995.

[11] S. Desai, E.V. Raghavendra, B. Yegnanarayana, A. Black, and K. Prahallad, "Voice conversion using artificial neural networks," *submitted at IEEE workshop on Spoken Language Technologies*, December. 2008.

[12] B. Yegnanarayana, "Artificial neural networks," *Prentice-Hall, New Delhi*, 1999.

[13] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through phoneme-based linear mapping functions with STRAIGHT for mandarin," *in Proceedings of Fourth International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 4, pp. 410–414, August 2007.

[14] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," *in Proceedings of ICASSP*, vol. 8, pp. 93–96, April 1983.

[15] A. Black, P. Taylor, and R. Caley, "The festival speech synthesis system," *http://festvox.org/festival*, 1998.