Design of Speech Corpus for Open Domain Urdu Text to Speech System Using Greedy Algorithm

Wajiha Habib
Center for Language Engineering, KICS-UET Lahore
wajiha.habib@kics.edu.pk

Rida Hijab Basit
Center for Language Engineering, KICS-UET Lahore
rida.hijab@kics.edu.pk

Sarmad Hussain
Center for Language Engineering, KICS-UET Lahore
sarmad.hussain@kics.edu.pk

Farah Adeeba
Center for Language Engineering, KICS-UET Lahore
Farah.adeeba@kics.edu.pk

Abstract

Unit selection speech synthesis is one of the most widely used techniques for high quality text to speech (TTS) systems. A unit selection text to speech system requires a large database of recorded and annotated speech, which contains both phonetic and prosodic variations. Designing phonetically rich and balanced speech corpora with minimum number of utterances is an intricate task. Several optimization methods are used for this purpose and "Greedy algorithm" is one of them. This paper introduces a greedy algorithm, which maximizes the coverage of high frequency unigrams, bigrams and trigrams while selecting minimal number of sentences from input corpus. The algorithm has been applied on different corpora collected from different domains and a speech corpus for Urdu TTS system is designed. A significant coverage of tri-phone has also been achieved.

1. Introduction

Unit selection technique for speech synthesis is a data-driven, concatenative approach. It dynamically selects the longest sequence of phonetic segments from the speech database, matching the characteristics of the target to be synthesized. The elegance of this approach lies in the lesser amount of signal processing required on the final utterance because the prosodic information is already a part of the corpus stored in the inventory. Furthermore, fewer concatenations result in a more natural speech output. However, the quality of data-driven text to speech system depends on the quality of its database.

A unit selection text to speech system requires a large database of recorded and annotated speech, which contains both phonetic and prosodic variations. At run time, appropriate units are selected from the database and they are concatenated to produce the desired utterance. The required memory size for unit selection system is very large. In addition, multilayer annotation of recorded speech is needed, which is a tedious and time consuming task. Hence, there is a need to optimize the speech corpus in such a way that maximum coverage of target units can be achieved with minimum corpus size. Greedy algorithm serves this purpose and has been used for intelligibly reducing the corpus.

This paper proposes a greedy algorithm for designing an optimal speech corpus for unit selection text to speech system. The rest of the paper has been organized as follows: Section 2 carries the literature review of greedy algorithm techniques designed for different languages, Section 3 describes the proposed methodology and Section 4 contains description of the data gathered for extraction of speech corpus. Section 5 describes implementation and evaluation of the proposed algorithm to select optimal speech corpus, Section 6 analyzes the resulting corpus and Section 7 holds conclusion.

2. Literature Review

Different techniques have been used to design a corpus for speech applications. Greedy algorithm is one of those methods employed, to extract an optimal reduced speech corpus from large corpus. It is an iterative approach that aims to maximize the coverage of target units while selecting minimum

number of sentences from the input corpus. The target unit varies from phone to phrase level i.e. phone, diphone, tri-phone, syllable, unigram, bigram and trigram. Selection of target unit for corpus design is based on the domain and needs of the application field. Coverage of larger units results in larger database, which in turn would produce high quality speech whereas smaller size of target units results into smaller database with compromised speech quality.

Phoneme can be used as a target unit for speech corpus design. Phone level coverage results in a limited corpus but phoneme sized chunks fail to cater the co-articulatory effects between adjoining phonemes [19]. Acoustic behavior of a phone is dependent on its previous and next phone. So, speech corpus should contain all the phones in all contexts. Therefore, tri-phone coverage is taken into account [12,14,15]. However, full coverage of tri-phone is impractical due to its very huge number. Diphone is used as the basic unit for corpus selection in [7,9,10] as it is affordable to build a corpus with high coverage of diphones. Diphone is an acoustic chunk from the middle of one phoneme to the middle of the next phoneme. Diphone is a desirable unit in concatenative synthesis because it gives complete language coverage, consumes less memory and as the co-articulation effects are minimal at the center of the phoneme, it caters the co-articulatory effects.

For a tonal syllabic language like Chinese, syllable coverage is the basic requirement for corpus design [16]. A new linguistic unit "vocalic sandwich" is defined in [13]. It is a sequence of phonemes, like vowels and semi-vowels surrounded by two phonemes (consonants). Greedy algorithm searches for those sentences that maximize the vocalic sandwich coverage rate.

In [8,15], multiple occurrences of target unit are acquired to capture all acoustic variations. Target phonetic distribution is focused in [17,18]. The aim is to extract a small database having the same probability distribution of phonetic features as the distribution in total database. The phonetic distribution includes phonemes, diphone patterns etc.

Sentence selection through greedy algorithm is carried out on the basis of calculated score and different criteria are employed for scoring the sentence. Francois et al. use five different techniques for scoring the sentences [8]. These include: high number of units in the sentence, sentence length, multiple occurrences of the unit and coverage of rare units. Kelly et al. score sentences according to the unique diphones they cover [9]. Different weights are assigned to the diphones in the corpus and Okapi formula is used to calculate scores of the sentences [10]. Sentences with maximum score are selected in [8,9,10]. Zhang et al. have selected sentences that provide maximum syllable level information [16]. Sentences are scored according to the information they provide regarding the syllables to be covered and in the end, best score sentence is selected.

A cluster tree from general speech database is built in [11] by clustering similar units (according to some features) together. Based on these features (phonetic, metrical and prosodic context), clusters are split unless a small acoustic distance is obtained. Greedy algorithm is then applied to find best coverage utterances. For scoring a sentence, cluster tree is traversed and sentences are scored accordingly. The best score sentences are selected in the end.

Least to Most (LTM) Greedy algorithm has also been used for corpus reduction [14]. Tri-phones to be covered are sorted in increasing order of their frequency. Least frequent tri-phone is selected along-with others that have the same occurring frequency and a separate list is maintained. Another list contains all the sentences that cover these tri-phones. These sentences are scored and best score sentence is selected. This process is repeated for all the tri-phones. In the end, redundant sentences are removed from the reduced corpus manually.

Minimum match score sentence is considered as the best sentence in [7]. For selecting sentences, the context of diphones is checked in the selected sentences and it is compared with the candidate sentence. It then calculates the match score for the two entities. Low match score indicates that the diphone in the candidate sentence has a different context as compared to the one in the selected sentence, so in this case low cost sentences are selected for the maximum coverage of diphones. The greedy algorithm stops selecting sentences when a certain number of sentences have been selected.

Greedy algorithm can be used to extract a list of words from corpus that provides maximum coverage of basic unit [12,17]. This reduced word list is then used to construct sentences manually.

Semi-automatic algorithm has been used that generates sentences using Finite State Transducers (FST) [13]. States represent vocalic sandwiches whereas arcs represent valid transitions between vocalic sandwiches (bigram sandwiches). This process involves human intervention. The algorithm generates sentences that give maximum coverage but as they are being generated by FST so they can be completely incorrect or senseless. For this purpose, a person with linguistics background must be sitting and operating it. The operator can accept, reject or ask to build another sentence based on the

requirements. This algorithm takes three minutes on average to build a credible sentence.

Speech corpus for Urdu language has also been designed [12]. The resulting corpus consists of sentences which are manually fabricated from the phonetically rich wordlist. Greedy algorithm has been used to extract those words from the corpus which give maximum coverage of high frequency tri-phones.

Manual construction of sentences is a tedious and troublesome task. A better approach should be used to avoid this time consuming and laborious effort. Therefore, an algorithm is devised to select optimal sentences directly from the corpus instead of constructing the sentences manually through the use of the wordlist. The proposed algorithm automates the process of speech corpus selection and produces a sentence based optimal speech corpus for Urdu.

3. Methodology

Diphones and tri-phones are used as basic units for speech synthesis. The greedy algorithm techniques employed, look for maximum coverage of these two units while selecting the minimum sentences from the corpus. The strategy used is good enough for diphone and tri-phone concatenative synthesis but in case of unit selection the size of the unit can be varied. As we are designing speech corpus for unit selection TTS system, we have proposed an algorithm that takes four units of different sizes that need to be covered while constructing a reduced speech corpora. These four units are: tri-phones, word unigrams, word bigrams and word trigrams. 80% of speech corpus has been extracted using top down approach in which coverage of longer high frequency units (word unigrams, word bigrams and word trigrams) has been maximized. Tri-phone coverage has been given less attention because phonetic transcription is required to report the tri-phone coverage but the existing transcription lexicon is in-sufficient to give 80% of Urdu language coverage. A significant coverage of tri-phones has also been achieved in the process. The rest of the tri-phones will be covered using the bottom up approach in remaining 20% of speech corpus.

3.1. Proposed Algorithm

The proposed greedy algorithm takes Urdu corpus and target lists as input. Target lists are the lists of those units, which need to be covered in the reduced corpus. The units consist of tri-phones, word unigram, word bigram, and word trigram. Unique tri-phone list is generated from the phones present in Urdu language. This list is further reduced by collapsing those phonemes that have similar acoustic effect [12]. The remaining three target lists (unigrams, bigrams and trigrams) are generated from the corpus itself. Urdu corpus and these lists are used to run the proposed greedy algorithm. Lists are updated throughout the algorithm whereas selected sentences are removed from the original corpus.

The algorithm assigns scores to all the sentences in a corpus according to the number of uncovered units in the sentence. A flow diagram of selection process is shown in Figure 1.

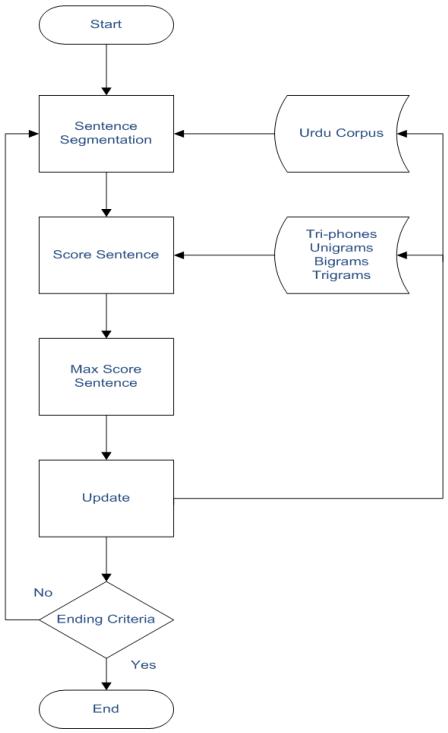

Figure 1. Flow diagram for proposed greedy algorithm

For scoring a sentence, a criterion has been devised. Based on the criterion, a sentence is considered optimal if it has maximum distinct units and a small length. All this have been represented using a formula which is as follows:

$$Score = \frac{(N_{triph} \times w_{triph}) + (N_{uni} \times w_{uni}) + (N_{bi} \times w_{bi}) + (N_{tri} \times w_{tri})}{Length\ of\ Sentence}$$

Here, *N* refers to the number of uncovered units and *w* refers to the weight of respective units which has been decided in the testing phase. Output has been analyzed with different weighting schemes

during the testing phase and the scheme which provided the best coverage is selected.

At each iteration, the algorithm picks the most useful sentence (maximum score sentence) to include in the selected sentence list, removes that sentence from the corpus and updates the lists (tri-phones, unigrams, bigrams, trigrams). These steps are repeated until the lists have been completely covered or the selected sentence score is less than some threshold value or the number of words in reduced corpus reaches some specified value. Devised algorithm generates the following outputs:

- Reduced Urdu corpus giving maximum coverage of tri-phones, high frequency unigrams, high frequency bigrams & high frequency trigrams that occurred in the larger Urdu corpus
- Phone coverage report
- Tri-phone coverage report
- Unigram coverage report
- Bigram coverage report
- Trigram coverage report
- Reduced corpus size

4. Corpus Description

The corpus for generic TTS system should be gathered from a broad range of domains to ensure diversity. Therefore, we have used three different corpora for extraction of reduced speech corpus. One of the Urdu corpora selected for TTS is a typed text corpus that has been taken from Urdu books [21]. It consists of 35 million words. The corpus contains 861 books from different domains i.e. religion, science, biography, poetry, travel, short stories and literature. These books not only cover Urdu characters but also have a coverage of English characters, Arabic, digits, URLs and special symbols.

Another corpus being used is "CLE Urdu digest corpus 1M[1]" which has been collected from Urdu digest [20]. Urdu news corpus of 2.6 million words is the third corpus, which has been used for speech corpus selection. The news corpus has been collected from different Urdu news websites i.e. BBC, Jang etc.. The news corpus is from the year 2005 and covers different sections from the news. These include: business, editorials, news and sports.

The proposed greedy algorithm has been implemented on these three corpora described above. The first corpus of typed Urdu books has been used for testing the proposed greedy algorithm. The weights that produce the best coverage result have been selected and used for obtaining corpus from the other two corpora. Details for testing and evaluation of greedy algorithm have been documented in the following sections.

5. Evaluation and Testing

During the implementation of greedy algorithm, different target lists comprising of those units which need to be covered in the reduced corpus, have been used. Different techniques are used to generate these lists, which will be explained in the following section. Moreover, weight assignment will also be described in detail.

5.1 Target Lists Generation

The 35 million word corpus has been used for generating lists of unique word unigrams, word bigrams and word trigrams along-with their frequency. These lists are sorted on the basis of frequency and the resulting lists are plotted to find the threshold for target lists generation. In Figure 2, unigrams are plotted against their frequencies. After the frequency value 495, a constant behavior is shown by graph. A sub-list is formed consisting of only those unigrams having the frequency greater than or equal to 495.

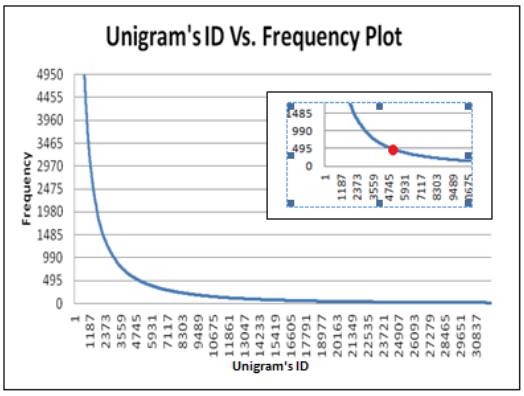

Figure 2. Unigram's frequency plot

Same method is followed for bigrams and trigrams. The threshold value for bigram list is 465 and for trigram list is 125 as shown in Figure 3 and 4 respectively. Based on these threshold values, sub-lists for bigrams and trigrams have been generated. These sub-lists are given as target lists to the greedy algorithm and the coverage of these lists is focused while obtaining the reduced Urdu corpus.

[1]http://cle.org.pk/clestore/urdudigestcorpus1M.htm

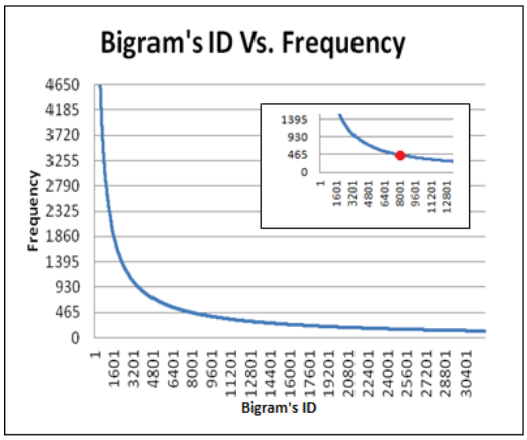

Figure 3. Bigram's frequency plot

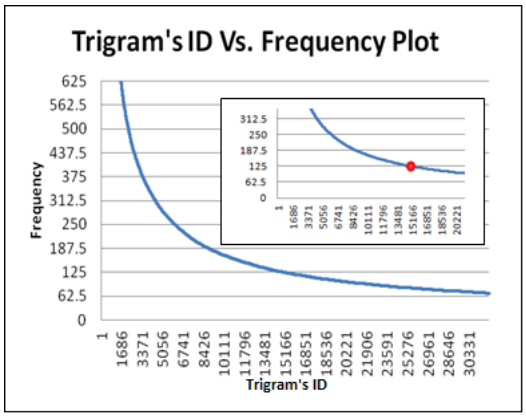

Figure 4. Trigram's frequency plot

5.2. Weight Assignment

An appropriate weighting scheme is required to prioritize the selection of target units. A unit with higher contribution must be given the larger weight. Weighting scheme devised, gives x weight to word unigrams, 1/7x to tri-phones assuming that a single word contains 7 tri-phones (5 phones) on average. Word bigrams have been given weight 2x as it consists of two words. Experiments have been performed on three different weights for word trigrams: 3x, 4x and 5x. 3x as word trigram covers three words, 4x for covering two bigrams and 5x for covering three words and two bigrams. Results have been gathered by testing these weights on a smaller corpus. The best coverage has been achieved assigning 1/7x weight to tri-phones, x weight to words, 2x weight to bigrams & 5x weight to trigrams.

Afterwards this weighting scheme has been applied on 35 million word corpus but the results were not so promising. For the better coverage of unigrams, bigrams & trigrams; weight of tri-phone has been kept constant whereas weights for other three have been tweaked to obtain a reduced corpus with above 90% coverage of unigrams, bigrams and trigrams. The reason, the weight of tri-phone has been kept minimal and tri-phone coverage has not been taken into account, is that phonetic transcription is not available for all the words in the corpus. The words with no available phonetic transcription are transcribed as silence. The stopping criterion is 70,000 words in reduced corpus. Results have been summarized in the Table 1 as given below and the weighting scheme which provided the best average coverage %age has been selected.

Table 1. Coverage results with different weighting schemes

$W_{tri-phone}, W_{word}, W_{bigram}, W_{trigram}$	Unigram Coverage %age	Bigram Coverage %age	Trigram Coverage %age	Average Coverage %age
0.017,0.1,0.3, 0.583	93.8	99.837	98.826	97.488
0.017,0.1,0.2, 0.683	95.52	99.549	98.913	97.994
0.017,0.2,0.3, 0.483	99.86	99.818	97.926	99.199
0.017,0.25,0. 3,0.433	100	99.874	97.559	99.144
0.017,0.15,0. 25,0.583	97.76	99.637	98.679	98.692
0.017,0.18,0. 2,0.603	99.62	98.810	98.780	99.07

Figure 5 shows the average coverage against different weighting schemes. The weights at which the best coverage has been achieved are 0.017 for tri-phones, 0.2 for unigrams, 0.3 for bigrams and 0.483 for trigrams. These weights are tested with different stopping criteria based on sentence score and resultant number of words in reduced corpus.

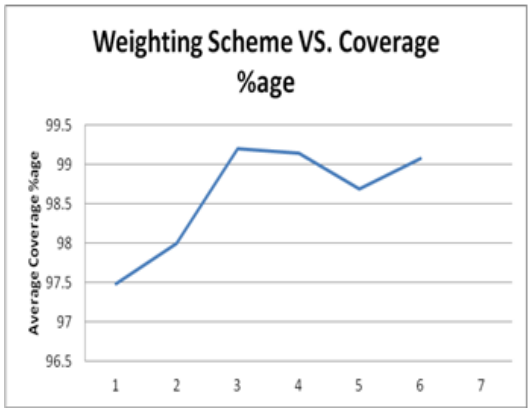

Figure 5. Coverage result for different weighting schemes

6. Finalization of Corpus

Total speech required for TTS system corpus is of 10 hours. Top down approach has been used for extraction of 80% of speech corpus (8 hours of recorded speech). Approximately 6.5 hours of speech corpus (70,000 words) has been obtained from 35 million word corpus whereas 1.5 hour speech corpus has been obtained from 1M Urdu Digest and news corpus. Target lists generation for second and third iteration of greedy algorithm has been done using a different method. For the generation of unigram, bigram & trigram sub-lists from 1M corpus, the unigrams, bigrams and trigrams covered in the reduced corpus obtained from 35 million word corpus are removed from the respective lists of 1M corpus. The subsequent lists are then used to find the target lists by plotting graphs for each of these lists.

After running the greedy algorithm, the reduced corpus from 1M corpus has been merged with the reduced corpus of 35 million word corpus. Now for news corpus same process is repeated as with 1M corpus but this time with the merged corpus (35 million word corpus & 1M corpus). The unigrams, bigrams & trigrams covered in the merged corpus are removed from the lists of news corpus and target lists are generated by plotting graphs for each of the lists (unigrams, bigrams & trigrams).

At the end of this process, 8 hours of speech corpus has been gathered using greedy algorithm, which is then used for recording purposes. Table 2 shows the results of greedy algorithm for the three different corpora.

Table 2. Results of greedy algorithm for different corpora

Corpus Description	35M Corpus	1M Corpus	News Corpus
Unigram Coverage %age	99.86	96.94	100
Bigram Coverage %age	99.818	99.06	95.86
Trigram Coverage %age	97.926	96.92	76.77
Average Coverage %age	99.199	97.64	90.88
Number of Words in Reduced Corpus	70,000	9000	7921

More than 90% coverage of unigrams, bigrams and trigrams target lists has been achieved in 80% of speech corpus. Tri-phone coverage could not be reported at the time of corpus extraction due to incomplete lexicon. The speech corpus extracted through greedy algorithm has been transcribed for tri-phone analysis and 34053 unique tri-phones are found in reduced corpus.

7. Conclusion and Future Work

In this paper, a greedy algorithm has been proposed to extract minimum size of corpora from some reference corpus while maximizing the coverage of target units for text to speech systems. Target units covered include tri-phones, word unigrams, word bigrams and word trigrams. The proposed algorithm is used to create a speech corpus for open domain unit selection Urdu text to speech system.

The corpus obtained in the process is 80% of the whole speech corpus required for TTS system. Minimum attention has been paid to the tri-phone coverage. In development of remaining 20% of speech corpus, tri-phone coverage will be focused.

The selected speech corpus will be used for recording. Those recorded speech files will be annotated and the tagged speech will be used as the database of unit selection Urdu TTS.

8. Acknowledgement

This work has been conducted through the project, Enabling Information Access for Mobile based Urdu Dialogue Systems and Screen Readers supported through a research grant from ICTRnD Fund, Pakistan.

9. References

[1] Taylor, Paul. Text-to-speech synthesis. Cambridge University Press, 2009.

[2] Peterson, Gordon E., William S-Y. Wang, and Eva Sivertsen. "Segmentation techniques in speech synthesis", Journal. Acoustical Society of America. 30.8 (2005), 739-742.

[3] N. Campbell and A. W. Black, "Prosody and the Selection of Source Units for Concatenative Synthesis". In J. van Santen, R. Sproat, J. Olive, and J. Hirschberg, "Progress in Speech Synthesis". Springer Verlag, 1995.

[4] Tokuda, K., Masuko, T., Yamada, T.: "An Algorithm for Speech Parameter Generation from Continuous mixture HMMs with Dynamic Features". In: Proc. of Eurospeech (1995).

[5] Suendermann, David, Harald Höge, and Alan Black. "Challenges in speech synthesis." Speech Technology. Springer US, 2010. 19-32.

[6] Zen, H., Toda, T.: "An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005",inproc. of *Inter speech* 2005, Lisbon, pp. 93–96 (2005).

[7] B. Bozkurt, O. Ozturk, and T. Dutoit, "Text design for TTS speech corpus building using a modified greedy selection", in proc. *Eurospeech* , 2003.

[8] François, H. and Boëffard, O., "The Greedy Algorithm and its Application to the Construction of a Continuous Speech Database", in proc. *LREC, Las Palmas de Gran Canaria*, Spain, 2002.

[9] Kelly, A., A. Ní Chasaide, H. Berthelsen, C. Campbell, and C. Gobl. "Corpus Design Techniques for Irish Speech Synthesis",in proc.*China-Ireland International Conference on Information and Communications Technologies*, NUI Maynooth, Ireland. 2009.

[10] Wei, Zhang, Liu Yayu, Deng Ye, and Pang Minhui. "Automatic Construction for a TTS Corpus with Limited Text" *In Measuring Technology and Mechatronics Automation (ICMTMA), 2010 International Conference on*, vol. 1, pp. 707-710. IEEE, 2010.

[11] Black, Alan W., and Kevin A. Lenzo. "Optimal data selection for unit selection synthesis" *In 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*. 2001.

[12] Raza, A., Sarmad Hussain, Huda Sarfraz, Inam Ullah, and Zahid Sarfraz. "Design and development of phonetically rich Urdu speech corpus",in proc. *COCOSDA*,2009

[13] Cadic, Didier, Cédric Boidin, and Christophe d'Alessandro. "Towards Optimal TTS Corpora." in proc. *LREC*. 2010.

[14] Suyanto, "Modified Least-to-Most Greedy Algorithm to Search a Minimum Sentence Set" in proc. *TENCON*, Hong Kong, 2006

[15] François, H. and Boëffard, O., "Design of an Optimal Continuous Speech Database for Text-To-Speech Synthesis Considered as a Set Covering Problem ", in proc. *Eurospeech*, Aalborg, Denmark, 2001.

[16] Zhang, Jianhua Tao Fangzhou Liu Meng, and Huibin Jia. "Design of Speech Corpus for Mandarin Text to Speech"The Blizzard Challenge 2008 workshop, Oct.2008

[17] Montero, Juan Manuel, Ricardo de Córdoba, José A. Vallejo, Juana M. Gutiérrez-Arriola, Emilia Enríquez, and José Manuel Pardo. "Restricted-domain female-voice synthesis in Spanish: from database design to ANN prosodic modeling," in proc. INTERSPEECH, pp. 621-624. 2000.

[18] Vorapatratorn, Surapol, Atiwong Suchato, and Proadpran Punyabukkana. "Automatic online text selection for constructing text corpus with custom phonetic distribution" in proc. *Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference* on, pp. 6-11. IEEE, 2012.

[19] Harris, Cyril M. "A study of the building blocks in speech." The Journal of the Acoustical Society of America 25.5 (1953): 962-969.

[20] Urooj, S., Hussain, S., Adeeba, F., Jabeen, F. and Parveen, R. "CLE Urdu Digest Corpus", in proc. of Conference on Language and Technology 2012 (CLT12), Lahore, Pakistan.

[21] Adeeba F., Akram Q. Khalid H., Hussain S., "Urdu Books N-gram Corpus", in proc. of *Conference on Language and Technology 2014* (CLT14), Karachi, Pakistan.