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

Wajiha Habib, Rida Hijab Basit,

Sarmad Hussain, Farah Adeeba

مرکز تحقیقات لسانیات

Center for Language Engineering
Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology Lahore
Pakistan

Contents

- Text to Speech (TTS) System
- Unit Selection Technique
- Significance of Optimal Speech Corpus for Unit Selection TTS system
- An Overview of Existing Methodology to Develop Speech Corpus
- Proposed Methodology for Extraction of Urdu Speech Corpus
- Results
- Future Work

Text to Speech (TTS) System

- A text-to-speech (TTS) system converts an arbitrary given text text into speech[1].
- Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database.
- Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output.

Unit Selection Technique

- 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.

Significance of Optimal Speech Corpus for Unit Selection TTS system

- The quality of data-driven text to speech system depends on the quality of its database.
- 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.
- An optimal speech corpus, having maximum number of target units and minimum size is required.

An Overview of Existing Methodology to Develop Speech Corpus

- **The greedy algorithm** begins with an empty initial cover . The first chosen sentences constitute a partial covering which increases. Total covering is achieved at the end of the algorithm [4,10].

- **The spitting algorithm** begins with a ”full” cover, that is the whole set of sentences; covering is thus total, the algorithm does not have to look for missing units. Sentences are removed one by one until no sentence could be removed without damaging the total covering [4,10].

- **The pair exchange algorithm** is a bit different from the others because its aim is more to improve the cover than to build it. At initial state, the cover contains a given number of sentences chosen arbitrarily . An out-of-cover sentence and an in-cover sentence are chosen. The two sentences are exchanged temporarily; if the covering is better, that is more units are covered and the cover is smaller, the change is kept, otherwise it is rejected [10].

Greedy Algorithm

- Greedy algorithm is an iterative approach that aims to maximize the coverage of target units while selecting minimum number of sentences from the input corpus.
- 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.

An Overview of Existing Methodology to Develop Speech Corpus using Greedy Algorithm

Two factors play significant role in implementation of greedy algorithm

- Choice of target unit
- Sentence scoring criteria

Effects of Selected Target Unit

Target Unit	Pros	Cons
Phoneme	Limited Corpus	Fails to cater the co-articulatory effects between adjoining phonemes [2]
Diphone [3,4,5]	1. Complete language coverage 2. Small database 3. Caters co-articulatory effects	Lack of full context
Tri-phone[6,7,8]	Coverage of all phones in all contexts	Full coverage is impractical
Syllable[9]	Basic requirement for a tonal syllabic language like Chinese.	Full coverage is impractical

Criteria Employed for Scoring the Sentence

- High number of units in sentence , Sentence length, Multiple occurrences of the unit, Coverage of rare units[10]
- Unique diphone[11]
- Maximum syllable level information[9]
- Minimum match score[3]

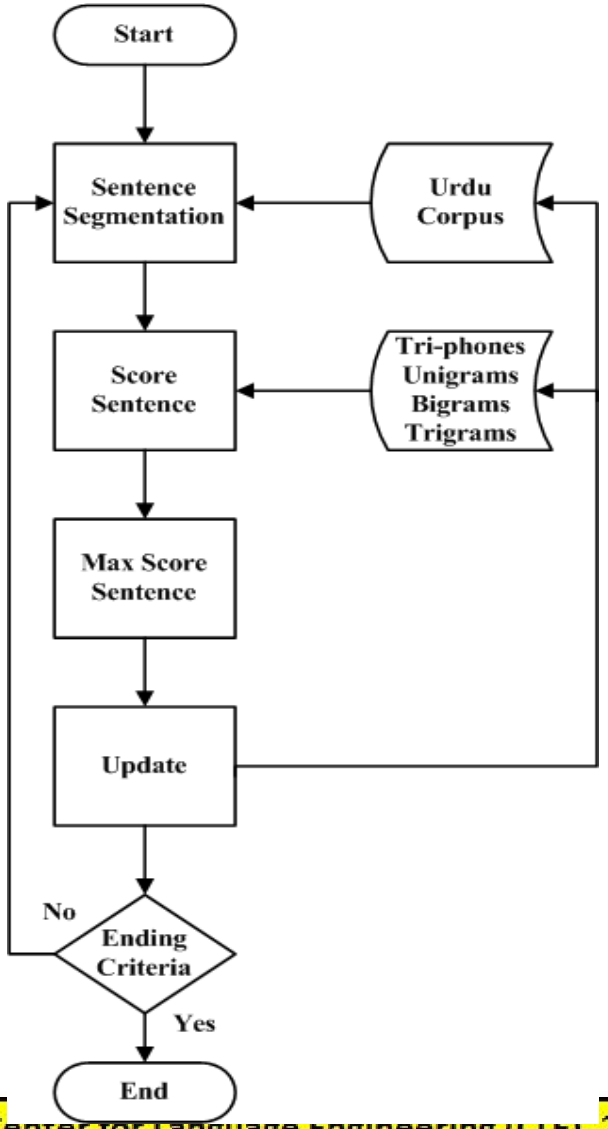
Speech Corpus for Urdu Language

- Phonetically rich wordlist was extracted from Urdu corpus.
- Greedy algorithm has been used to extract those words from the corpus which give maximum coverage of high frequency tri-phones.
- Sentence were manually fabricated from that wordlist [6].

Proposed Methodology

- 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.
- The algorithm assigns scores to all the sentences in a corpus according to the number of uncovered units in the sentence.

Flow Diagram for Proposed greedy Algorithm

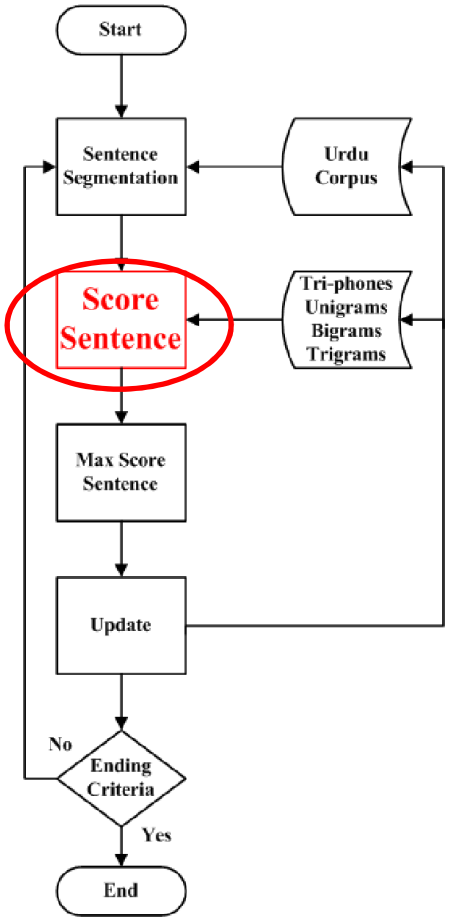


Sentence Scoring Criteria

- A sentence is considered optimal if it has maximum distinct units and a small length. This have been represented using a formula which is as follows:

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

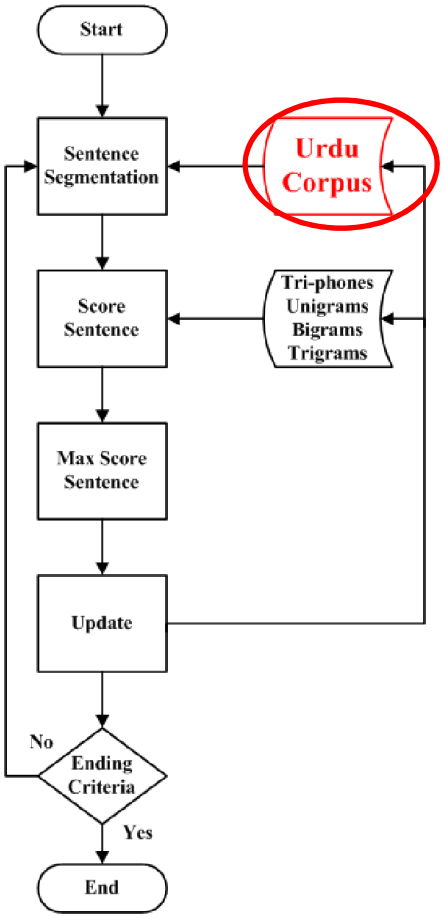
- Here, *N* refers to the number of uncovered units and *w* refers to the weight of respective units.



Corpus Description

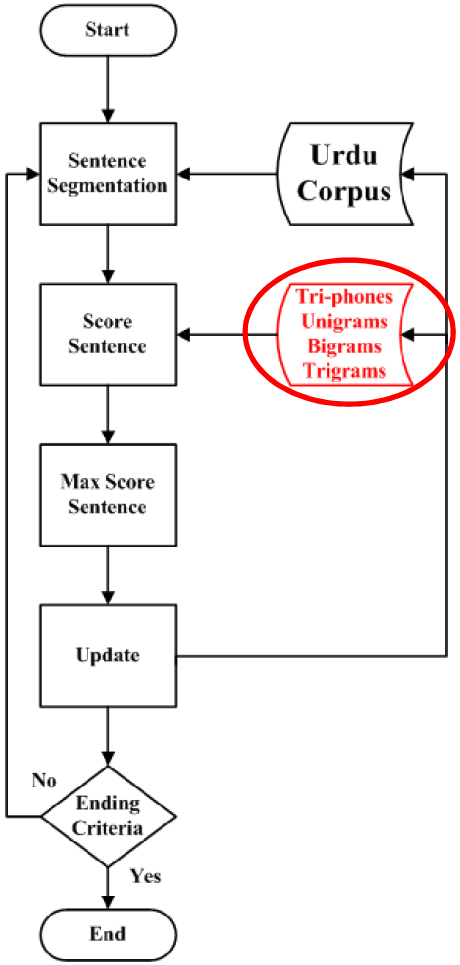
Three different corpora have been used for speech corpus extraction to ensure diversity.

- 37M Word Corpus [12]
- CLE Urdu Digest Corpus [13]
- Urdu News Corpus



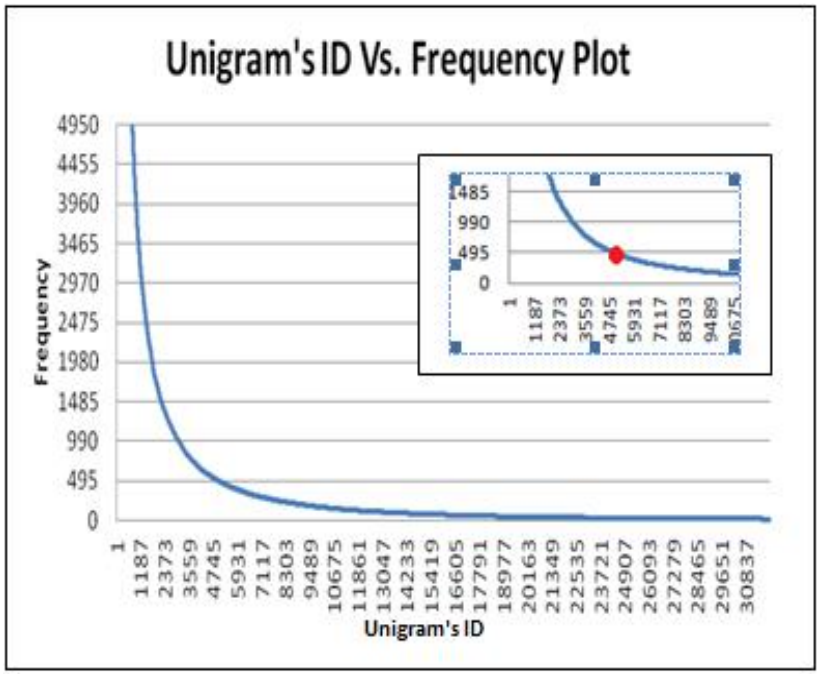
Target List Generation

- The 37 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.



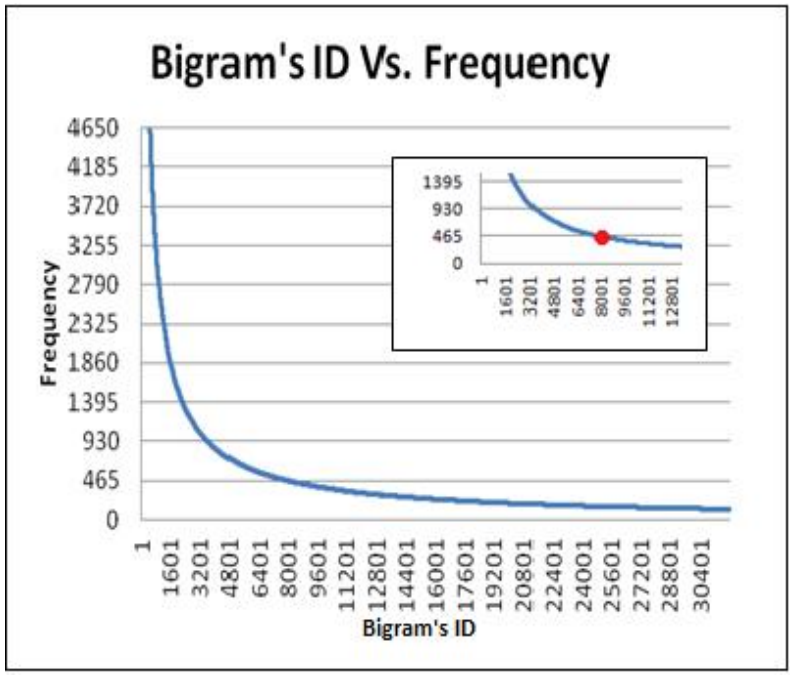
Target List Generation

- 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.



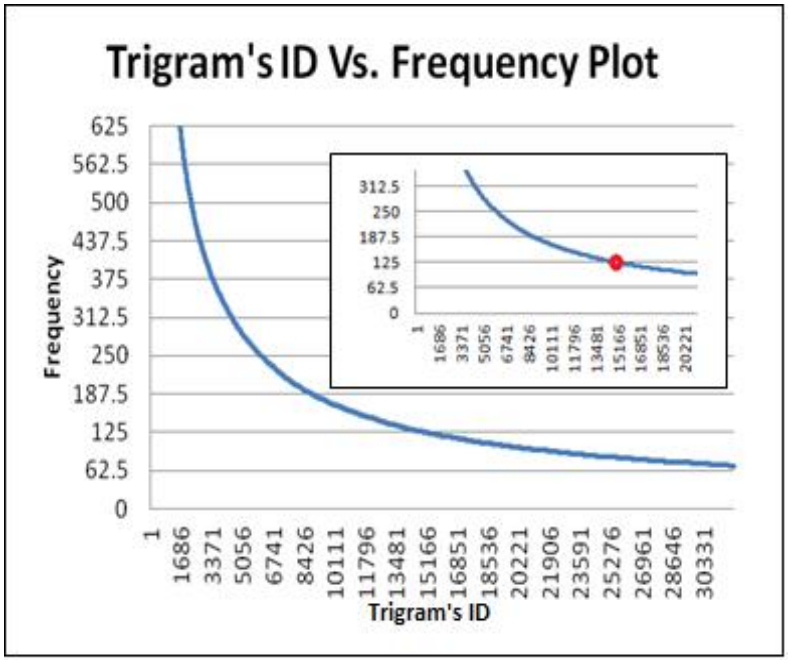
Target List Generation

- The threshold value for bigram list is 465.



Target List Generation

- The threshold value for trigram list is 125.



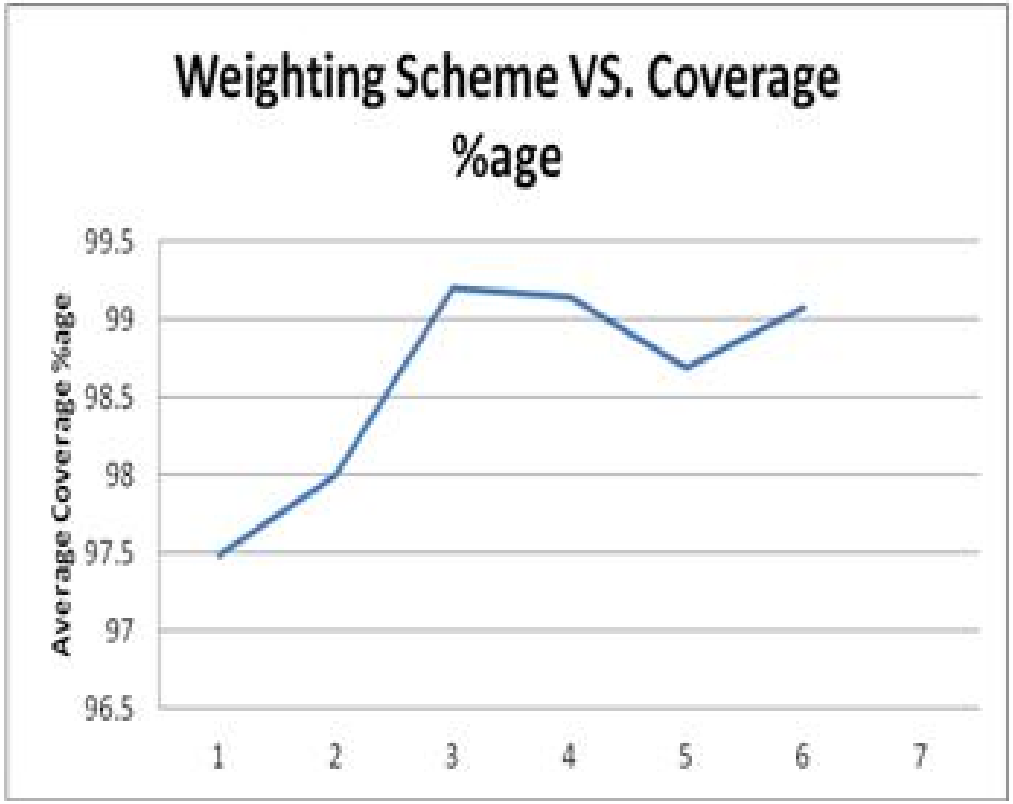
Weighting Scheme

- A unit with higher contribution must be given the larger weight.
- x weight was given to word unigram
- 1/7x weights was given to tri-phones assuming that a single word contains 7 tri-phones (5 phones) on average.
- Word bigrams were given weight 2x as it consists of two words.
- Experiments were 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.
- 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.

Results of Different Weighting Schemes on 37M Word Corpus

$w_{tri-phone}, w_{word}, w_{bigram}, w_{trigram}$	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

Coverage result for different weighting schemes



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 37 million word corpus whereas 1.5 hour speech corpus has been obtained from 1M Urdu Digest and news corpus.

Results of greedy algorithm for different corpora

Corpus Description	37M 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

Future Work

- 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.

Thank You!

References

1. Raj, A. A., Sarkar, T., & Pammi, S. C. Text Processingfor Text-to-SpeechSystemsin Indian Languages.
2. Harris, Cyril M. "A study of the building blocks in speech." The Journal of the Acoustical Society of America 25.5 (1953): 962-969.
3. B. Bozkurt, O. Ozturk, and T. Dutoit, "Text design for TTS speech corpus building using a modified greedy selection", in proc. *Eurospeech* , 2003.
4. 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.
5. 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.

References

6. Raza, A., Sarmad Hussain, Huda Sarfraz, Inam Ullah, and Zahid Sarfraz. "Design and development of phonetically rich Urdu speech corpus",in proc. *COCOSDA*,2009
7. Suyanto, "Modified Least-to-Most Greedy Algorithm to Search a Minimum Sentence Set" in proc. *TENCON*, Hong Kong, 2006
8. 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.
9. 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
10. 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.

References

11. 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.
12. 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.
13. 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.