Design and development of phonetically rich Urdu speech corpus

Agha Ali Raza
ali.raza[*]
NUCES[**]

Sarmad Hussain
sarmad.hussain[*]
NUCES

Huda Sarfraz
huda.sarfraz[*]
NUCES

Inam Ullah
inam.ullah[*]
NUCES

Zahid Sarfraz
zahid.sarfraz[*]
NUCES

Abstract

Phonetically rich speech corpora play a pivotal role in speech research. The significance of such resources becomes crucial in the development of Automatic Speech Recognition systems and Text to Speech systems. This paper presents details of designing and developing an optimal context based phonetically rich speech corpus for Urdu that will serve as a baseline model for training a Large Vocabulary Continuous Speech Recognition system for Urdu language.

1. Introduction

Center for Research in Urdu Language Processing (CRULP; www.crulp.org) at NUCES is currently working on a project entitled *Telephone-based Speech Interfaces for Access to Information by Non-literate Users* in collaboration with Carnegie Mellon University. The goal of this project is to investigate the use of speech interfaces for users to access online health related information in Pakistan. This will be achieved by developing a telephone based dialogue system consisting of an Urdu Speech Recognition system and a Text to Speech system that can interact with the health workers to answer their queries.

One key component of this system a Large Vocabulary Automatic Speech Recognition (LVASR) system for Urdu. This system requires the construction of a *phonetically rich* and *balanced* corpus for recognition of continuous and spontaneous speech in Urdu. Once the training corpus is recorded, it has to be labeled. The system will be based on Hidden Markov Models, using Sphinx 3 [3] trainer and Sphinx 4 ([4], [5]) decoder.

This paper describes the process employed in the design and development of the phonetically rich Urdu speech corpus, the initial step in the development of the Urdu LVASR. The next section briefly reviews similar work done for other languages and the phonetic characteristics of Urdu. Sections 3 and 4 and 6 describe the word and sentence based corpus development process in detail. Sections 5 and 7 analyze the resulting corpus and Section 8 concludes the results.

2. Background and Literature Review

Urdu, the national language of Pakistan, is spoken by around a 100 million people around the globe [1]. Phonetically, it is a rich language with a large inventory of consonants (Appendix A), and numerous long nasal, long non-nasal and short vowels (Figure 1). Urdu also has some diphthongs [6]. It is written in Arabic script in Nastalique style using an extended Arabic character set [7]. The character set includes basic and secondary letters, aerab (or diacritical marks), punctuation marks and special symbols ([8], [9]). However, everyday-Urdu is written only using the letters, which primarily represent just the consonantal content, and the use of diacritics, which represent the vowels in Urdu, is optional. Though this does not cause any difficulty for the native speaker, the absence of vowel marks makes the job of letter to sound mapping more difficult computationally [10]. As a result, Urdu corpora obtained from sources like newspapers etc. are generally phonetically transcribed using lexical lookup, though manual review is necessary for cases where multiple pronunciation are possible for same written form.

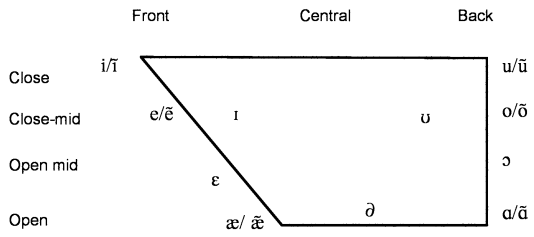

Figure 1 Urdu Vowels

A lot of work has been done on the development of speech resources for many languages of the world. These resources have been developed both for TTS (e.g. [15]) and ASR systems (e.g. [12, 13, 16, 17]). The main goal in the development of speech corpora is phonetic coverage [22], which allows them to represent the phonetic structure of the target language. Speech

[*] @nu.edu.pk

[**] FAST - National University of Computer and Emerging Sciences, Pakistan (www.nu.edu.pk)

978-1-4244-4400-7/09/$25.00 ©2009 IEEE

corpora have been developed for various tasks, including: (a) isolated word corpora, e.g. Lithuanian [21], (b) continuous speech, e.g. Indian Languages [13], Hindi [16] and Greek [17], and (c) continuous and spontaneous speech, e.g. Dutch [14], Mandarin [20], and Russian [22].

The second criterion for the speech corpus development is the phonetic balance, i.e. the phonetic content should occur in the same proportions as in the language, to properly train the statistical models, as discussed for Russian [22], Amharic [12], and Mexican Spanish [23]. The phonetic richness can simply be phone-based [20] or context-based. The context-based methods take into consideration either a single immediate context, using diphone-based methods [13] or both beginning and ending context, using triphone-based methods [16, 22]. However, an analysis in [24] shows that the triphone richness may not improve the accuracy of speech recognizer significantly but it requires much more data.

There is also difference in approaches towards gathering the data for the speech corpora. Most of the automatic approaches utilize some kind of a greedy algorithm to maximize the number of sound units (half-phones, phones, diphones or triphones) in minimal data set [13, 15, 24]. Still other make phonetically balanced sentences by comparing the phonetic composition with a language model by using perplexity [23]. This set is made richer by adding spontaneous speech data, e.g. from interviews [22] or recorded free speech. Still other approaches may include collection of text which represents the phonetic richness and proportion of a language [23].

3. Methodology

The goal of this work is to develop a sentence based corpus for Urdu, automating the design task as much as possible using existing language resources. A fundamental criterion is to cover all possible phone combinations that are used in Urdu. The resulting phonetically rich corpus can serve to provide the baseline acoustic models for continuous LVASR for Urdu. However it will not necessarily be phonetically balanced. In order to convert it into a balanced corpus, recordings of actual interviews, using everyday speech will be done and transcribed. This will not only serve to balance the corpus but also cater for the spontaneous speech modeling requirement.

The first step is to find all possible phones that are used in Urdu Speech. A more practical approximation is to find all possible phonemes that exist in Urdu and then try to construct a word based corpus with the goal of covering all those phonemes. The word list is then manually converted into sentences. The Urdu corpus that is used for this purpose has been developed at CRULP [19] and consists of 18 million words of Urdu. This data is gathered from various domains. The corpus is not fully diacritized and hence cannot be mapped completely to phonemes using simple letter to sound rules.

It must be mentioned here that an approach could have been to pick phonetically rich sentences directly from the corpus instead of making a word list and then converting them into sentences. However, this sentence list would not have been *minimal*, a criterion that can be controlled while making a word list. It would have been more natural representation of the language, even if redundant, but it has not been possible because the corpus is not diacritized.

3.1. Corpus Analysis and Development of Lexicons

The word-based corpus was analyzed and found to consist of around 50,000 unique words. A list of these words was formed and phonemically transcribed in two passes. In the first pass an automatic transcription was done using the letter to sound rules. However, due to the lack of diacritics this resulted in partial transcription only. In the second pass the words were manually phonemically transcribed. As a result a phonetic lexicon is obtained which gives word to phoneme set mappings (henceforth referred to as *phonetic lexicon*). Standard SAMPA representation [25] is used for phonetic and phonemic transcriptions.

Next a word frequency analysis of the corpus is done to find the frequency of occurrence of all the 50,000 words in the corpus. This analysis gives another lexicon containing word to frequency mappings (henceforth referred to a *word-frequency lexicon*).

3.2. Phonetically rich word list

The primary goal of the development of a phonetically rich corpus was to ensure that it represents all sounds that occur in Urdu. This allows it to serve as a baseline for training an Automatic Speech Recognition system. However since the targeted ASR system is to be used for continuous and spontaneous speech, therefore a simple occurrence of all phonemes will not suffice for the following reasons.

3.2.1. Effects of phonetic context. The acoustic properties of a phone are not localized and are affected by the acoustic properties of the neighboring phones, i.e. the phonetic context.

3.2.2. Across word effects. In continuous speech words run into each other hence the last phone of a word maybe affected by the initial phone(s) of the

following word. Hence across-word influences will form a part of the phonetic context as well.

3.2.3. Spontaneous Speech. The system must be trained for spontaneous speech in which the words are not carefully articulated. This often results in shorted words with missing phones or modified versions of target phones. Hence, it was required that the system should be trained by a model coming from free speech as well. However, this should be in addition to the speech read from phonetically rich text, as spontaneous speech is not guaranteed to be phonetically rich.

The problems mentioned in Sections 3.2.1 and 3.2.2, are discussed in detail in Section 4. The problem of spontaneous speech is discussed in Section 5.

4. Tri-phoneme based phonetically rich corpus

As mentioned in Section 3.2.1, simple phonemic enrichment cannot guarantee that the resulting word set will be a representative of the acoustic properties of all the phones. Hence, context must be added to the phone set. For this we made sequence of three phonemes (henceforth referred to a *tri-phoneme*) the basic unit of the acoustic model for a particular sound. We define a tri-phoneme to consist of three phonemes $\{P_1\ P_2\ P_3\}$, where P_2 is the target phoneme and P_1 and P_3 act as the phonetic context. Now in order to represent the acoustic properties of all sounds, the dataset should contain all possible tri-phonemes that can occur in the language.

As the phonemic inventory of Urdu language comprises of 62 phonemes (excluding silence), there can be a total of 250,047 potential tri-phoneme combinations (including silence as a phoneme). In order to find the tri-phoneme combinations that actually exist in Urdu the phonetic lexicon is analyzed for tri-phonemes and their frequency of occurrence. This analysis is done from two different perspectives.

The first analysis is done to find all the unique tri-phonemes that occur within words. It is assumed for this analysis that all words are followed and preceded by silence, which is dealt with as a separate phoneme. The analysis shows that the corpus contained 18,294 unique in-word tri-phonemes.

Next, an analysis of the across-word tri-phonemes is performed. However, for this analysis, instead of finding all the *existing* across-word tri-phonemes, all the *potential* across word tri-phonemes are found (including the in-word tri-phonemes). This is done due to the flexible syntactic structure of Urdu which allows many possible arrangements of words in a sentence. This is achieved by assuming that every word (preceded by silence) can be followed by any other word (followed by silence). This list is found to contain around 85,000 unique tri-phonemes. This number should be interpreted as the upper limit of the number of tri-phonemes in Urdu.

4.1. Word list construction

In order to allow utility in an ASR system the word list used for training the speech model must consist of a minimal word set that can maximize the number of tri-phonemes. In order to compute the word set a modified version of the Set Covering algorithm [11] is used. A decision has to be made whether to maximize the number of across-word tri-phonemes or the in-word ones. If the list is constructed on the basis of across-word tri-phonemes the word set will consist entirely of word-pairs instead of words. This will make the job of converting these into sentences very difficult (and for some combinations of words it will not even be possible).

Moreover, even if the sentences are generated from a wordlist based only upon in-word tri-phonemes, they will necessarily contain many of the across-word tri-phonemes and a post analysis can reveal the shortcomings. These can in turn be compensated by generating a supplementary sentence list.

4.2. Minimal word list

The Set Covering algorithm (Figure 2) is used to generate the minimal list of words that contains all the in-word tri-phonemes. The list that is generated contained 8200 words (for the condition in Figure 2a). However, the major problem with the list is that it consists primarily of words that are long, unfamiliar or borrowed words from other languages such as:

ماہرینِ تابکاریات، استادالاساتذہ، کسٹمر ریلیشن شپ مینجمنٹ، بین البینکی شرح منافع

Such a wordlist cannot be effectively used for the ASR system as the native speakers of Urdu will not be able to fluently go through the sentences made from such words. This will prevent the aspects of continuity and spontaneity to be present in the recordings. Besides it would make the job of making the sentences far too hard. However, it must be noted that this is the smallest list for the given corpus that can be generated which covers all tri-phonemes.

4.3. High frequency minimal word list

The solution to the problem presented in the previous section is to give weight to the frequency of occurrence of the words in the corpus as well. Hence the condition of the Set Covering algorithm can be modified as shown in Figure 2b. This way the weights w_f and w_n can be adjusted to get a minimal list of

common (high frequency) words of Urdu. The list(s) thus obtained contain more words than the one generated in 4.2 but fulfill the requirements of familiarity of words.

Different values of weights are tried however the problem of uncommon words continues for most of the experiments. Finally, priority is given to frequency of occurrence of the word in the corpus and subsequent weight to the number of tri-phonemes that it adds to the set (as shown by the condition in Figure 2c). The final wordlist generated contained 11,884 high frequency words.

The major problem with this wordlist is the high number of words. Considering an average sentence length of 8 words, we would end up with 1486 sentences, which is too much to be practically read out by a few speakers. And for the ASR repetition is also desired for this data.

```
; Inputs
;   X is the input corpus
;   C is the condition
; Output
;   O is the output list of words
Greedy-Set-Cover(X, C)
1.  U = X
2.  O = φ
3.  while U ≠ φ
4.        do  select  word  W  from  U  that
    maximizes C
5.           U = U - W
6.           O = O + {W}
7.  endwhile
8.  return O
```

a
```
C = |tri-phonemes(W) ∩ tri-phonemes(U)|
```

b
```
   let,  N  =  |tri-phonemes(W)  ∩  tri-
phonemes(U)|
   let, F = Frequency(W)
   then, C = w_f F + w_n N
```

c
```
   let,  N  =  |tri-phonemes(W)  ∩  tri-
phonemes(U)|
   let, F = Frequency(W)
   then,
   if N = φ
      C = MIN
   else
      C = F
   Endif
```

Figure 2: Greedy Algorithm for Phone Set Covering. Parts a, b and c show the different conditions imposed

4.4. Reduced high frequency minimal word list

To reduce the size, the wordlist is carefully analyzed and several rules were devised based on the phonetic structure of Urdu and the acoustic properties of the phones. Following are the major rules that are formulated to reduce the size of the in-word tri-phoneme list. The reduction is at the cost of losing some context. However, this is done to so that there is minimal compromise.

4.4.1. Voiced/voiceless unaspirated stops in context positions. When voiced or voiceless unaspirated stops occur before or after the target phone, their acoustic context has same effect on the target phone, as long as they have same *place* of articulation. Hence, the affect on the target remains quite minimal (especially spectrally) by the variation in the voicing property (in case of unaspirated stops). So, we collapsed all the voiced/voiceless unaspirated stops at the same place occurring at context positions to the voiced version. This reduces the tri-phoneme set significantly.

4.4.2. Aspirated/unaspirated stops at tri-phoneme ends. The aspirated/unaspirated stops occurring at the end of tri-phonemes affect the target phone similarly. Therefore these two types can be merged (with some compromise).

4.4.3. Removing low frequency tri-phonemes. Next a frequency analysis of the tri-phoneme list is performed. The goal was to find the frequency of occurrence of each tri-phoneme in the corpus. All tri-phonemes occurring more than 10 times are selected for inclusion in the list. At a later stage other tri-phonemes can be added if required.

As a result of applying the above constraints 9,436 tri-phonemes are removed from the tri-phoneme list, hence leaving behind 8,858 tri-phonemes to cover for recording. Using the algorithm of Section 4.3. High frequency minimal word list the final wordlist generated after removing the tri-phonemes that fall into the above mentioned categories contains 5,681 unique high frequency words. This is comparable with the most optimal list generated earlier using the greedy set cover method which had mostly unfamiliar 4,390 words.

5. Word list analysis

The wordlist generated as a result of this exercise was analyzed to confirm that it does indeed cover most of the required tri-phonemes. It was confirmed that it contains all the 62 phonemes, and 8,858 tri-phonemes. Overall it contains 10,133 unique tri-phonemes. This is because every new word added in the set cover algorithm may also add more tri-phonemes which are not in the list but are part of the selected word. This also adds some measure of phonetic balance to the word list. Figure 3 shows the logarithmic plot of frequency of occurrence of each tri-phoneme in the corpus (the curve above) and its frequency of occurrence in the word list (shown below).

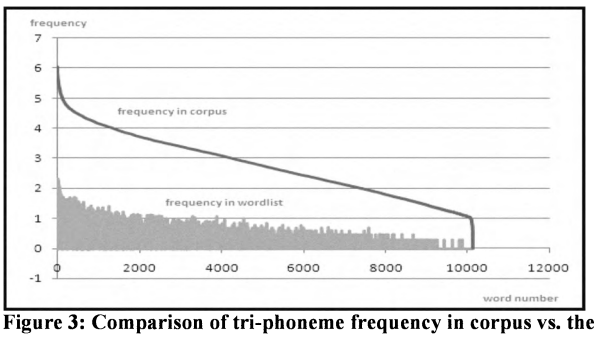

Figure 3: Comparison of tri-phoneme frequency in corpus vs. the tri-phoneme frequency in the wordlist

6. Sentence generation

The 5681 words generated as described in Section 4, are used by a team of language experts to construct sentences. The aim is to construct sentences that are grammatically correct and sound natural to native Urdu speakers. The guidelines followed during sentence generation are given below:

- Each sentence consists of at least five words
- Sentences with commas are avoided, in order to avoid sentences including lists of items
- Native Urdu speakers should be able to utter the sentence without much difficulty
- The word list has no diacritical marks, so if any words are detected which are ambiguous in pronunciation, sentences are constructed for all variations in the pronunciation, with the appropriate diacritical marks inserted
- Sentences that do not make semantic sense are allowed to be part of the set as long as they are grammatically correct, and easy to read fluently, but should be avoided as much as possible

A total of 725 sentences are produced as a result of this exercise. For quality control, each sentence is reviewed by a member of the team not taking part in the sentence construction. Sentences that are found to be difficult or odd for a native Urdu speaker to utter are identified and sent back to the sentence construction process. For example, Figure 4 shows examples of good and bad sentences. The first sentence is good as it is short, easy to read and makes complete sense. The next sentence is graded as average as it is slightly difficult to read smoothly since the initial part is almost a tongue twister. Some of the words may also be unfamiliar for the average Urdu speaker. Otherwise it is correct, grammatically and semantically. The third sentence is only marginally accepted as it is semantically odd, and may cause the reader to react unexpectedly. Grammatically it is correct and over all short. The last example is rejected because it is too long and difficult to make sense of. This makes it almost impossible to read through smoothly.

بڑی بحث اور تجزیہ کے بعد یہ فیصلہ ہوا
کاشف قزلباش تشنج کے سبب معالج کیلئے بہترین معالج کے پاس گیا
شنید ہے کہ بوبی پنچولی کا حسن نزلے کے سبب مرجھا جانے کو ہے
واچ نگر کے کتھک خنازیر شیڈ راؤنڈر میں گوند نچوڑ کراور سوجی چھڑک کر دوروں پر آئے

Figure 4 Examples of sentences

8. Conclusion

The work presented in this paper describes the design and development of a context based phonetically rich speech corpus for Urdu language. The purpose of this corpus is to act as a baseline resource for training acoustic models for Urdu speech and to give as much phonetic coverage as possible, without regard to balance. For this purpose tri-phonemes are used as the basic unit and a broad but practical set is extracted in two ways: by collapsing context which has approximately same spectral effect, and by focusing on high frequency tri-phonemes. A greedy algorithm is then used to derive a minimal list of familiar words from a corpus derived wordlist to cover the tri-phonemes. The set is based upon high frequency words of Urdu to facilitate fluent reading by a native speaker. The word list is converted into sentences manually, and is analyzed to confirm that it satisfies the need. These sentences are being used for recordings by native speakers to ensure that there is phonetic coverage in the speech corpus being developed. This part of the corpus will be focusing on diversity rather than balance. The corpus will be phonetically balanced by recording interview-based spontaneous speech by the participants

Acknowledgement

The work has been funded through a research grant by Higher Education Commission, Govt. of Pakistan.

References

[1] www.ethnologue.com accessed on 20th April, 2009.

[2] D. Jurafsky, J. H. Martin, A. Kehler, K. Vander Linden and N. Ward, Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, 2000

[3] CMUSphinx: The Carnegie Mellon Sphinx Project, http://cmusphinx.sourceforge.net/html/cmusphinx.php

[4] Sphinx-4 - A speech recognizer written entirely in the Java (TM) programming language, http://cmusphinx.sourceforge.net/sphinx4/

[5] Speech at CMU, http://www.speech.cs.cmu.edu/

[6] S. Hussain. 1997. Phonetic Correlates of Lexical Stress in Urdu. Unpublished Doctoral Dissertation, Northwestern University, Evanston, USA.

[7] S. Hussain. 2003. www.LICT4D.aisa/Fonts/ Nafees_Nastalique. Proceedings of 12th AMIC Annual Conference on E-Worlds: Governments, Business and Civil Society, Asian Media Information Center, Singapore.

[8] M. Afzal and S. Hussain. 2001. Urdu Computing Standards: Development of Urdu Zabta Takhti (UZT 1.01). *Proceedings of IEEE International Multi-topic Conference,* Lahore, Pakistan.

[9] S. Hussain, and M. Afzal. 2001. Urdu Computing Standards: Urdu Zabta Takhti (UZT 1.01). *Proceedings of IEEE International Multi-topic Conference,* Lahore, Pakistan.

[10] S. Hussain, Letter to Sound Rules for Urdu Text to Speech System, Proceedings of Workshop on "Computational Approaches to Arabic Script-based Languages", COLING 2004, Geneva, Switzerland (2004).

[11] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein, Introduction to Algorithms, Second Edition, The MIT Press, Massachusetts Institute of Technology, Cambridge Massachusetts, 2001.

[12] S. T. Abate, W. Menzel, and B. Tafila, An Amharic speech corpus for large vocabulary continuous speech recognition, 2005.

[13] G. Anumanchipalli, R. Chitturi, S. Joshi, R. Kumar, S. P. Singh, R. N. V. Sitaram, and S. P. Kishore, Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems.

[14] D. Binnenpoorte, C. Cucchiarini, H. Strik, and L. Boves, Improving automatic phonetic transcription of spontaneous speech through variant-based pronunciation variation modelling, 2004, pp. 681–684.

[15] B. Bozkurt, O. Ozturk, and T. Dutoit, Text design for TTS speech corpus building using a modified greedy selection, 2003.

[16] V. Chourasia, K. Samudravijaya, and M. Chandwani, Phonetically Rich Hindi Sentence Corpus for Creation of Speech Database, Proc. O-COCOSDA, pp. 132–137, 2005.

[17] V. Digalakis, D. Oikonomidis, D. Pratsolis, N. Tsourakis, C. Vosnidis, N. Chatzichrisafis, and V. Diakoloukas, Large Vocabulary Continuous Speech Recognition in Greek: Corpus and an Automatic Dictation System, 2003.

[18] P. A. Heeman, The American English SALA-II Data Collection, 2004.

[19] M. Ijaz and S. Hussain, Corpus Based Urdu Lexicon Development, 2007.

[20] A. Li, F. Zheng, W. Byrne, P. Fung, T. Kamm, Y. Liu, Z. Song, U. Ruhi, V. Venkataramani, and X. X. Chen, CASS: A phonetically transcribed corpus of Mandarin spontaneous speech, 2000.

[21] G. Raškinis, Building medium-vocabulary isolated-word Lithuanian HMM speech recognition system, Informatica, vol. 14, pp. 75-84, 2003.

[22] A. L. Ronzhin, R. M. Yusupov, I. V. Li, and A. B. Leontieva, Survey of Russian Speech Recognition Systems.

[23] L. Villaseñor-Pineda, M. Montes-y-Gomez, D. Vaufreydaz, and J. F. Serignat, Experiments on the Construction of a Phonetically Balanced Corpus from the Web, Lecture notes in computer science, pp. 416-419, 2004.

[24] Y. C. Yio, M. S. Liang, Y. C. Chiang, and R. Y. Lyu, Biphone-rich versus triphone-rich: a comparison of speech corpora in automatic speech recognition, 2005, pp. 194-197.

[25] SAMPA computer readable phonetic alphabet, www.phon.ucl.ac.uk/home/sampa/

Appendix A

	Bilabial		Ldental		Dental		Alveolar		Retroflex		Palatal		Velar		Uvular	Phar	Laryn
Voicing	-	+	-	+	-	+	-	+	-	+	-	+	-	+			
Plosive	p pʰ	b bʰ			t̪ t̪ʰ	d̪ d̪ʰ	t tʰ	d dʰ					k kʰ	g gʰ	q		ʔ
Nasal		m						n						ŋ			
Trill								r									
Flap									ɽ								
Fricative			f	v			s	z			ʃ	ʒ	x	ɤ			h
Lateral								l									
Approximant												j					
Affricates											ʧ ʧʰ	ʤ ʤʰ					

Table 1 Urdu Consonants