A MULTI-GENRE URDU BROADCAST SPEECH RECOGNITION SYSTEM

Erbaz Khan[1], Sahar Rauf[1], Farah Adeeba[2] and Sarmad Hussain[1]

[1] Center for Language Engineering,
Al-Khawarizmi Institute of Computer Science,
University of Engineering and Technology, Lahore.
firstname.lastname@kics.edu.pk

[2] Department of Computer Science,
University of Engineering and Technology, New Campus, KSK.
firstname.lastname@uet.edu.pk

ABSTRACT

This paper reports the development of a multi-genre Urdu Broadcast (BC) corpus and a Large Vocabulary Continuous Speech Recognition (LVCSR) system. BC speech corpus of 98 hours from 453 speakers is collected and annotated. For acoustic modeling, Time-delay Neural Network (TDNN) is developed with prior Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) training and alignments. For the language model, 3-gram, 4-gram and Recurrent Neural Network (RNN) based models are developed on a text corpus of 188 million words. The developed models are tested on 4.3 hours of unseen BC multi-genre speech dataset and the best Word Error Rate (WER) 18.59% is achieved using RNN based Language Model (LM). Moreover, a detailed word error analysis is carried out to compare the errors made by humans and the Automatic Speech Recognition (ASR) System. The results showed a similar behavior of word misrecognitions by both humans and ASR.

Index Terms — BC, multi-genre, Urdu, corpus, speech recognition

1. INTRODUCTION

In the past decades, electronic media has undergone huge growth. Many hours of Urdu speech data are broadcasted through different channels especially Radio, TV and YouTube on daily basis. This huge amount of data is a big source of information as well as entertainment. This data can be further divided into different genres and each genre represents its own jargon. However, since this data is in unstructured form, it is hard to extract any important insights from it. Manually annotating thousands of hours of data is near to impossible but an ASR can automate the process of generating the audio transcriptions. Recognized text obtained from the ASR, then, can be used for generating analytics which in turn could be used to get numerous useful insights. Moreover, a BC ASR can also prove to be very effective for people with hearing disabilities as they would be able to watch the shows with auto generated subtitles.

To develop an Urdu Automatic Speech Recognition (ASR) system for BC speech, annotation of BC data is required. Annotation of BC data is a strenuous task that requires continuous effort from the annotators. Moreover, BC data is a fusion of gender information, cross talks, dialectical variations, speech types, and requires complete speech alignment with the text. Thus, a lot of research work has been done on improving broadcast speech recognition for different languages. An Arabic ASR for BC speech is developed with multiple deep learning models [1] and the best WER of 42.25% is achieved on 6.2 hours of test data. In [2], three different acoustic models, GMM-HMM, Subspace GMM (SGMM) and Deep Neural Network (DNN), are developed along with a trigram language model for German language ASR and a WER of 35.3% is achieved on spontaneous testing data of 44 minutes and 19.9% is achieved on planned test data of 1.1 hours. In [3] and [4], Hungarian and English ASR systems are developed, in both of which HMMs based acoustic models and n-grams based language models are investigated. WER of 24% and 34.5% are obtained for Hungarian and English systems respectively. An LVCSR for Hindi language is developed using GMM and DNN based models and the WER of 11.5% is achieved on a 1-hour test set [5].

The one indispensable component of any Artificial Intelligence (AI) system is the data. A speech recognition system, similarly, needs a speech corpus to train the system. Thus, researchers have been collecting broadcast speech data for different languages. 15 hours of multi-genre data from YouTube and 1200 hours of data recorded from broadcast news channels is used to develop Arabic ASR [1]. A French language corpus collected from TV and Radio broadcasts is used in [6]. A huge German video training corpus of 900 hours containing 2,705 recordings is developed in [2]. In [3], a Hungarian ASR is developed using 380 minutes of audio data. A Multi-Genre TV broadcast

978-1-6654-0870-7/21/$31.00

data is used to develop an ASR in [4]. The data is comprised of a training set of 349 hours of audios with their subtitles, a development set of 8 hours of audios with manual transcription and a text corpus of 640 million words of TV subtitles. A large-vocabulary Hindi broadcast news data is collected from the All-India radio website and a total of 5.5 hours of Hindi data is collected in [5].

An Urdu-English code-switched LVCSR is developed for microphonic speech [7]. The system included 300 hours of Urdu read speech corpus [8], 115 hours of business read speech data, and 10 hours of Urdu broadcast data. For acoustic modeling, TDNN is developed without using alignments obtained from GMM-HMM, whereas for language modeling trigram model is developed. The best WER of 26.95% is achieved. However, no significant work has been done to develop a large vocabulary BC speech corpus and speech recognition system for Urdu language.

In this study, the development of a multi-genre Urdu BC corpus and ASR is reported. Section 2 elaborates the corpus design, collection, annotation and verification process. Experiments performed and developed models are discussed in Section 3. In Section 4, testing data details are provided. In Section 5, results obtained from the system are discussed, whereas, Section 6 concludes the whole study.

2. BROADCAST CORPUS

A multi-genre BC corpus is developed for the ASR. This section details the important characteristics of the corpus.

2.1. Corpus Selection

The corpus is selected from different BC channels including Radio, TV, and YouTube. The BC shows' content is comprised of talk shows, news shows, and interviews mainly divided into two main types; conversational speech (talk and news shows in which speakers discuss different topics) and interviews (questions and answers). These main types are further divided into four different genres: politics, health and science, entertainment, and current affairs. The data includes 453 unique speakers with 333 males and 120 females. A total of 98 hours of BC data is aligned and annotated. The data distribution among the three main channels is described in Table 1. Furthermore, the data includes 31% conversational speech and 69% interviews.

Table 1: Description of BC training data

Channels	No. of channels	No. of shows per channel	No. of episodes	Percentage Duration
Radio	1	1	56	31
YouTube	6	7	35	33
TV	10	19	46	36

2.2. Data Preprocessing

After the selection of data, the data is converted to *.wav* format with properties that channel should be mono, bitrate 256 kbps and sampling rate 16 kHz. The *.wav* audio is then passed to an already developed ASR [7], which transforms speech into text. Through this process, partial transcription and segmentation of the data at the sentence level is obtained.

2.3. Data Annotation

Table 2: Data tags and their description

Data tags	Symbol	Tag's description
Hesitation marker	:ه:	This symbol is used to denote the hesitations of the speaker's speech
Mispronunc-iation marker	+	This symbol denotes the mispronunciation of a word
Partial speech	-	This symbol is used when a speaker wants to utter a word but could not speak completely. So, this symbol is attached with the incomplete word
Idiosyncratic words	*	This symbol is used to denote the words which are developed by the speaker himself and are not part of the standard dictionary. The slangs are also denoted by this symbol
Speaker's noise	<cough> <laugh> <lipsmack> <sneeze>	These tags denote the very important part of the speech which frequently occurs in a spontaneous speech
Hard-to-understand speech	(( ))	The hard-to-understand speech is marked within these brackets. These brackets can be left blank in case of un-intelligible speech or transcription can be added in case of semi-intelligible speech

Foreign language	<foreign lang="Arabic" > </foreign>	This tag is used to identify the foreign language. As we already transliterated the English words into Urdu so in our case the foreign language is any language other than Urdu and English

In order to address the transcription challenges of BC data, XTrans [9] [10] (a speech annotation tool was used. XTrans is a multilingual tool that efficiently aligns the transcription to the audio and also marks the data at different levels. It is used to mark the information of multiple speakers, their genders, and backgrounds. It also marks the sentence type i.e. question, statement and incomplete.

While annotating the BC data, the role of the annotator was to carefully listen to the sentence or segment and adjust its boundaries with the audio if it was not properly aligned in pre-processing. The annotator had to re-type the data in case of substitutions, deletions, or additions and also had to transliterate the data from English to Urdu. Then, he/she had to mark the speaker information as speaker's name (Standard name mentioned on Google), gender (male, female, child, and other), and speaker's dialect (native/other) and then marked the sentence type: question, statement and incomplete. In our data, Urdu spoken in Punjab was considered as the native language. A complete process of data annotation is elaborated in Figure 1.

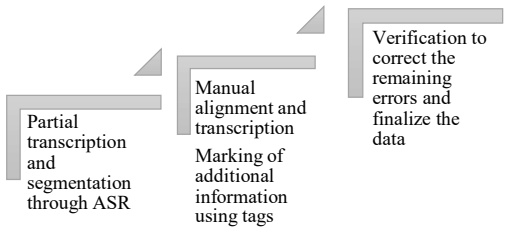

Figure 1 - Annotation process for BC data

The annotator also had to check that the duration of the segment must not exceed 15 seconds. Along with all this information, additional tags were also used to elaborate the data. The description of these tags is given in Table 2.

2.4. Data Verification and Annotation Issues

After the annotation, the data was verified by expert linguists to assure the good quality of the data. The linguists had to listen to complete data and correct the data where possible and also provide feedback to the annotators for future corrections. The main annotation issues found in the data were as follows:

- In some cases, the annotator misinterprets the silence with the closure duration of voiceless stops or plosives like /p,k,t/ etc. and therefore marks the boundary there, which results in incomplete words at the start or end. The solution is to listen to the word carefully and try to differentiate between the proper silence and consonants.
- Fast speech or low pitch at the end of the sentence sometimes cause deletion of the words.

Table 3: Frequently deleted and inserted words by human

Deleted words	POS tag	Inserted words	POS tag
ہے	Tense Auxiliary	کے	Postposition
یہ	Pronoun	اور	Conjunction
سے	Postposition	کہ	Conjunction
اور	Conjunction	ہے	Tense Auxiliary
میں	Pronoun	کی	Postposition
اس	Pronoun	جس	Pronoun
آپ	Pronoun	جو	Pronoun
یا	Conjunction	کر	Conjunction
تو	Conjunction	تھے	Tense Auxiliary
نہیں	Negation Adverb	وہ	Pronoun

- Table 3 is elaborating the 10 most frequent word deletions and insertions by the humans with their Part of Speech (POS) tags. The analysis shows that the pronouns, conjunctions, tense auxiliaries and postpositions are the most frequently deleted and inserted categories.
- Another error is to ignore the nasalization of words like ہے /hæː/ 'is' is frequently used instead of ہیں /hæ̃ː/ 'are' and similarly بات /bɑːt̪/ 'talk' is used instead of باتیں /bɑːt̪ẽː/ 'talks' etc.

After verification of the data, sentences containing mispronunciation marker, partial speech, idiosyncratic words, speaker's noise, hard-to-understand speech, and foreign language tags are removed. Further, overlapping and music segments are also discarded. After this post-processing step, the remaining 71 hours of data is added to the system. The percentage distribution of the genres in this remaining data is described in Table 4.

Table 4: Percentage distribution of genres in training data

Genres	% Distribution of Genres
Entertainment	45
Politics	28.75
Health and science	18.75
Current affairs	7.5

3. ASR EXPERIMENTS

In this section, all the experiments performed to develop different models are discussed.

3.1. Acoustic Modeling

Acoustic model is used to get a relationship between speech sound and the linguistic units, which, in our case, are phonemes. To train the system, 40 Mel Frequency Cepstral Coefficients (MFCCs) are extracted from the audio data using 25 milliseconds window and 10 milliseconds shift.

Acoustic model is developed using Kaldi toolkit [11]. In previous studies, it is shown that conventional GMM-HMM models are easily outperformed by modern deep learning architectures [7], thus Time-delay Neural Network (TDNN) [12] is developed using the alignments obtained from GMM-HMM models. For GMM training, speaker independent Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) are investigated. Further, Speaker Adapted Training (SAT) is done and its alignments are used for the TDNN's training.

Neural network training is done with high resolution MFCCs and i-vector of dimensionality 100 for each sample. TDNN consists of 7 hidden layers. First layer is fixed affine layer whereas rest of the hidden layers are TDNN with number of neurons equal to 1024. Rectified Linear Unit (ReLU) batch norm is used as activation function. Lattice Free Maximum Mutual Information (LF-MMI) objective functions is also used in the TDNN.

3.2. Language Modeling

A large vocabulary multi-genre corpus of 154 million words developed in [8] is increased to 188 million words by adding new data transcriptions and manually generated sentences. Also, the corpus used in [8] is in code-switching text so all the English words are transliterated to Urdu before training the LM. 3-gram and 4-gram are developed using SRI Language Modeling [13] (SRILM) toolkit.

Furthermore, RNN based LM is also developed and the results are obtained via lattice rescoring algorithm.

3.3. Lexicon

The lexicon reported in [7], with the addition of 940 new words is used. The current size of the lexicon is equal to 200k entries. The older lexicon contained both English and Urdu words, however all the English words are transliterated into Urdu orthography for this study. In some cases where alternate pronunciations are present, standard Urdu orthography is used. For example, word America is written as امریکا /əmri:kɑ:/ 'America' in Urdu, in case of English pronunciation of the word alternate transcription امریکا /əmærəkɑ:/ 'America' is also added to the system.

4. TESTING DATA

To gauge the performance of the system, 4.3 hours of unseen testing data is developed using the same techniques as discussed in Section 3. Testing data consists of 25 speakers out of which 14 are male and 11 are female. This data is collected from 5 different broadcast channels and YouTube encompassing multiple genres like entertainment, health and science, current affairs and politics. Duration of each category is given in Table 5.

5. RESULTS AND DISCUSSIONS

5.1. Baseline System

Prior to performing these experiments, the system with only 437 hours of microphonic speech [7] is tested on the same test data. TDNN along with 3-gram LM is used and WER of 36.38% is achieved on the overall test data.

5.2. Improved System

TDNN based acoustic model in combination with 3-gram, 4-gram and RNN based LM is tested on the test set. Table 5 shows the %WER obtained from all the experiments.

Based on the results of the above experiments, it is clear that RNN based model outperforms the n-gram models. When compared to the baseline system, it is evident that the addition of 71 hours of BC data in our baseline system reduced the WER by almost a half.

Table 5: Percentage WER of TDNN based acoustic model in combination with different language models

Category	Duration (min)	3-gram	4-gram	RNNLM
Politics	70	17.22	17.01	13.83
Health Science	55	23.54	24.02	19.40
Entertainment	81	26.28	26.42	21.92
Current Affairs	52	17.54	17.48	15.69
Combined data	258	21.59	21.71	**18.59**

Despite the fact that the entertainment genre is 45% of the entire data, still the %WER is higher for it. The reason being, entertainment shows include background noise/music, and informal speech i.e. talking while laughing, and speaking at higher rate. This makes it difficult for humans to analyze and annotate the speech and consequently, harder for the ASR to understand it properly.

Detailed error analysis is also carried out on the recognized output of our best performing model i.e., RNN-LM. Table 6 shows the most frequently inserted, deleted words whereas Table 7 shows most frequent pair of words in the substitution error.

Following observations are drawn based on the most frequent errors by the ASR:

- Comparing to Table 3 it is noticeable that humans' and ASR's word errors are overlapping, indicating a similar behavior of word misrecognition in both humans and ASR. For example, ہے /hæ:/ 'is' is more frequently deleted, inserted and substituted word in both cases.

Table 6: Frequently deleted and inserted words by ASR

Deletions	POS tag	Insertion	POS tag
م:	Hesitation	ہے	Tense Auxiliary
ہے	Tense Auxiliary	م:	Hesitation
ہیں	Tense Auxiliary	اس	Pronoun
کے	Postposition	تو	Conjunction
اور	Conjunction	کی	Postposition
سے	Postposition	اور	Conjunction
اس	Pronoun	کہ	Conjunction
میں	Pronoun	ان	Pronoun
کر	Conjunction	ہیں	Tense Auxiliary
آپ	Pronoun	یہ	Pronoun

Table 7: Frequently substituted words by ASR

Substitutions	POS tag
ہے -> ہیں	Tense Auxiliary -> Tense Auxiliary
کہ -> کے	Postposition -> Conjunction
نہ -> نا	Negation -> Negation
ہیں -> ہے	Tense Auxiliary -> Tense Auxiliary
پر -> پہ	Postposition -> Postposition
کے -> کہ	Conjunction -> Postposition
ہوں -> ہو	Tense Auxiliary -> Tense Auxiliary
کی -> کے	Postposition -> Postposition
ہوں -> م:	Hesitation -> Tense Auxiliary
کے -> کی	Postposition -> Postposition

- One prominent difference is the marking of hesitation م: as the results show that the system frequently misrecognizes the hesitation م: in all three types of error. This error is less frequent in humans, indicating that humans mark hesitations more easily as compared to the ASR.
- ہے /hæ:/ 'is' and ہیں /hæ̃:/ 'are' are tense auxiliaries and occur at the end of the sentences. Due to less intensity at the end of the sentence, there is a possibility that the system may not correctly recognize these words.
- In spontaneous speech, the words are not fully articulated sometimes by the speakers which could be the reason for the deletion of some words by the system.
- The most frequent error type is substitution. The most common reason for this type of error is the poor quality of respective audio segment. Another reason could be the substitution of homophones that the system recognizes the word's pronunciation right but represents it with different orthography as کہ ->کے in which کہ /ke:/ 'that' is a conjunction and کے /ke:/ 'for ' is a postposition.

6. CONCLUSION

A large vocabulary Urdu BC speech recognition system is developed. The multi-genre BC corpus of 98 hours is collected and annotated from Radio, TV, and YouTube mainly comprised of conversations and interviews. For acoustic modeling, TDNN is trained with prior GMM-HMM alignments. Three different language models are developed i.e., 3-gram, 4-gram and RNN based LM. Test results show that RNN-LM outperforms the n-gram models. Word error analysis shows a significant similarity in the errors made by humans and ASR. In the future, more BC data will be added to improve the system's performance and to decrease the gender imbalance. Data augmentation techniques like speed and

volume perturbation will also be explored. From the applied point of view, a BC talk show analytics extraction system will also be developed using the developed ASR.

7. REFERENCES

[1] M. Najafian, W. -N. Hsu, A. Ali and G. James, "Automatic Speech Recognition of Arabic Multi-Genre Broadcast Media," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Okinawa, Japan, 2017.

[2] M. Stadtschnitzer, J. Schwenninger, D. Stein and J. Koehler, "Exploiting the large-scale German Broadcast Corpus to boost the Fraunhofer," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014.

[3] A. Roy, L. Lamel, T. Fraga da Silva, J. -L. Gauvain and I. Oparin, "Some Issues Affecting the Transcription of Hungarian Broadcast Audio," in *INTERSPEECH*, Lyon, France, 2013.

[4] I. Bada, J. Karsten, D. Fohr and I. Illina, "Data Selection in the Framework of Automatic Speech Recognition," in *ICNLSSP-International Conference on Natural Language, Signal and Speech Processing*, Casablanca, Morocco, 2017.

[5] P. Jyothi and M. Hasegawa-Johnson, "Improved Hindi Broadcast ASR by Adapting the Language Model and Pronunciation Model Using A Priori Syntactic and Morphophonemic Knowledge," in *INTERSPEECH*, Dresden,Germany, 2015.

[6] M. Garnerin, S. Rossato and L. Besacier, "Gender Representation in French Broadcast Corpora and Its Impact on ASR Performance," in *ACM Workshop AI4TV*, Nice, France, 2019.

[7] M. U. Farooq, F. Adeeba, S. Hussain, S. Rauf and M. Khalid, "Enhancing Large Vocabulary Continuous Speech Recognition System for Urdu-English Conversational Code-Switched Speech," in *O-COCOSDA*, Yangon, Myanmar, 2020.

[8] M. U. Farooq, F. Adeeba, S. Hussain, S. Rauf and M. Khalid, "Improving Large Vocabulary Urdu Speech Recognition System Using Deep Neural Networks," in *INTERSPEECH*, Graz, Austria, 2019.

[9] M. L. Glenn, S. Strassel and H. Lee, "XTrans: a speech annotation and transcription tool," in *INTERSPEECH*, 2009.

[10] "XTrans," [Online]. Available: www.ldc.upenn.edu/language-resources/tools/xtrans. [Accessed 14 June 2021].

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian and P. Schwarz, "The Kaldi," in *ASRU*, 2011.

[12] A. Waibel, "Phoneme Recognition Using Time-Delay Neural Networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 37, no. 3, pp. 328-339, 1989.

[13] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition1989," in *Proceedings of the IEEE*, 1989.

[14] D. Jurafsky and J. H. Martin, Speech and Language Processing (2nd Edition), Upper Saddle River, NJ: Prentice-Hall, Inc., 2009.

[15] J. Baker, L. Deng, S. Khudanpur, C. Lee, J. Glass and N. Morgan, "Historical development and future directions in speech recognition and understanding," in *MINDS Report of the Speech Understanding Working Group, NIST*, 2006.

[16] Adeeba, F., Akram, Q., Khalid, H. and Hussain, S., "CLE Urdu Books N-grams," in *Conference on Language and Technology*, Karachi, 2014.

[17] H. Sarfraz, S. Hussain, R. Bokhari, A. A. Raza, Inam Ullah, Z. Sarfraz, S. Pervez, A. Mustafa, I. Javed and R. Parveen, "Speech corpus development for a speaker independent spontaneous Urdu speech recognition system," in *O-COCOSDA*, Kathmandu, Nepal, 2010.

[18] A. Stolcke et al., "SRILM-an extensible language modeling toolkit," in *INTERSPEECH*, 2002.