Proceedings of the 23rd Conference of the Oriental COCOSDA
November 5-7, 2020, Yangon, Myanmar

Enhancing Large Vocabulary Continuous Speech Recognition System for Urdu-English Conversational Code-Switched Speech

Muhammad Umar Farooq, Farah Adeeba, Sarmad Hussain, Sahar Rauf, Maryam Khalid

Center for Language Engineering,
Al-Khawarizmi Institute of Computer Science,
University of Engineering and Technology, Lahore.
{umar.farooq,farah.adeeba,sarmad.hussain,sahar.rauf,maryam.khalid}@kics.edu.pk

Abstract

This paper presents first step towards Large Vocabulary Continuous Speech Recognition (LVCSR) system for Urdu-English code-switched conversational speech. Urdu is the national language and lingua franca of Pakistan, with 100 million speakers worldwide. English, on the other hand, is official language of Pakistan and commonly mixed with Urdu in daily communication. Urdu, being under-resourced language, have no substantial Urdu-English code-switched corpus in hand to develop speech recognition system. In this research, readily available spontaneous Urdu speech corpus (25 hours) is revised to use it for enhancement of read speech Urdu LVCSR to recognize code-switched speech. This data set is split into 20 hours of train and 5 hours of test set. 10 hours of Urdu BroadCast (BC) data are collected and annotated in a semi-supervised way to enhance the system further. For acoustic modeling, state-of-the-art DNN-HMM modeling technique is used without any prior GMM-HMM training and alignments. Various techniques to improve language model using monolingual data are investigated. The overall percent Word Error Rate (WER) is reduced from 40.71% to 26.95% on test set.

Index Terms: Urdu-English code-switching, Urdu speech recognition, under-resourced language

1. Introduction

Code Switching (CS), spontaneous use of two or more languages in a single conversation, is a prevalent linguistic phenomenon in multi-cultural societies or the countries where native and official languages are different. The dominant language is usually referred as *matrix language* and the secondary language is termed as *embedded language*. CS renders a monolingual Natural Language Processing (NLP) system clueless about language and muddles the context when system confronts the embedded language. Therefore, CS is very challenging for most of the monolingual NLP tasks such as Automatic Speech Recognition (ASR), Part of Speech (POS) tagging, Machine Translation (MT) and summarization. An increasing research interest is observed in developing CS speech recognition systems [1, 2, 3] since most of the off-the-shelf systems are monolingual.

The most difficult challenge in developing models for new language pairs is the annotated data sparsity for code-switched speech. It makes both acoustic and language modeling a Gordian knot. The problem is exacerbated in case of low resource languages which have even very small monolingual data. Urdu is the national language and lingua franca of Pakistan which is spoken by more than a hundred million speakers in Pakistan, India, Bangladesh and the regions of Europe [4]. English, being official language of Pakistan, is commonly mixed with national language *Urdu* in daily communication. Though English is rich but Urdu is an under resourced language with small available monolingual data. Various code-switched speech recognition systems including English-Mandarin [5], Frisian-Dutch [6], English-Malay [7], French-Arabic [8] and Hindi-English [9] have been studied but no such system for English-Urdu code-switched speech exists so far.

Over the years, limited efforts have been made to develop resources and speech technologies for Urdu. A recent research [10] focused to fill this gap and a LVCSR was developed for Urdu language. A neural network was trained on 300 hours of read speech (from 1586 speakers) Urdu data which yielded a WER of 13.5% on test set. The test set was 9 hours of unseen speech data (from 62 speakers). From previous to latest researches were restrained to limited vocabulary [11], isolated words [12] or small set of speakers [13].

Sarfarz et al. [14] designed and developed an Urdu speech corpus of 44.5 hours. 25 hours of this corpus consisted of conversational speech. It was based on interview speech from various speakers which hinted that it incorporated Urdu-English code-switching naturally. Rather than adopting dilatory process of designing and collection of CS speech data, the aforementioned corpus was acquired to train the system. However, the English words were forced transliterated in Urdu during annotation of speech data. So, the corpus is reworked in this research to make the text corpus code-switched. Furthermore, Urdu BC news data is collected from online audio/video sources and annotated in a semi-supervised way. Most of the data is fetched from *YouTube* and radio shows covering entertainment, political and current affairs domains. 10 hours of Urdu spontaneous CS speech is collected and added to train the acoustic model.

For acoustic model training, efforts are being made to replace widely used DNN-HMMs [1, 15] with end-to-end training [16, 17, 18, 19, 20]. The aim of such researches is to expedite ASR building by avoiding manual development of large lexicons and corpus for language models. However, end-to-end models' performances are yet behind that of DNN-HMMs [16, 17, 18]. Most of the neural networks in DNN-HMMs are trained using alignments and context dependency tress from GMM-HMM training [21]. However, a novel acoustic modeling strategy [22] is used to train the acoustic model which trains the network in flat start manner and doesn't rely on any previous information.

In this paper, read speech Urdu LVCSR system is enhanced to recognize code-switched and spontaneous speech. Urdu spontaneous speech corpus is reworked to make it usable for

978-1-7281-9896-5/20/$31.00 ©2020 IEEE

Table 1: *Statistics of Urdu corpora*

Corpus	No. of speakers	Duration (hours)	No. of utterances
Training corpora			
Urdu LVCSR corpus *(Read speech)*	1586	300	213677
Business corpus*(Read speech)*	445	115	67928
Urdu CS *(Spontaneous speech)*	62	20	17919
Urdu BC *(Spontaneous speech)*	52	10	4601
Total	**2145**	**445**	**304125**
Testing corpus			
Urdu CS *(Spontaneous speech)*	20	5	3912

code-switched speech. Urdu broadcast data is collected and annotated to enhance the system accuracy.

2. Corpus

Baseline system is trained on available read speech corpus and language model developed in [10]. Corpus covered some erstwhile Urdu corpora, proper nouns and fabricated personal information carried in sentences. Furthermore, a fraction of corpus was from Urdu news channels' websites and tweets. A few sentences (779) were added to cover the most frequent English words mixed in Urdu. Speech corpus was collected from 1586 speakers (ages ranging from 18-50) using USB microphone, USB headset, hands-free and laptop microphone. All recordings were sampled at 16 KHz. Vocabulary size for this corpus was 199K and a large vocabulary continuous read speech ASR was trained on this data [10].

115 hours from business corpus are also added in it to make the system further stable for read speech. This corpus covers e-commerce websites, telecommunication domain, online news websites and general categories. Business domain data in Urdu carries more English words than other domains. The intention behind adding business read speech data to existing ASR is to enhance DNN's learning of Urdu phonemes pronounced as part of English words.

Sarfarz et al. [14] developed a speaker independent spontaneous Urdu speech corpus of 25 hours. This corpus is primarily made up of spontaneous interview speech and thus contains frequent Urdu-English code switching. Since the focus of corpus development was spontaneous Urdu ASR, all English words were forced transliterated in Urdu. In this research, the same corpus is obtained and reworked to use it for enhancing read speech Urdu LVCSR [10] to recognize Urdu-English code-switched speech. This corpus is termed as *Urdu CS corpus* throughout the paper. Though the details of the corpus can be found in [14], however the process of corpus development (concerned with the scope of this research) is briefly described here.

For speech corpus collection, a speaker recruitment process was designed to select native Urdu speakers without any speech impediments. Featured speakers' ages ranged from 20 to 55 years. Recording sessions were conducted in office rooms, student labs and sometimes in homes. Thus, the corpus carried the external background noise. Recording was done through a microphone and over a telephone line simultaneously. Microphone (Logitech USB mic) was resting on the table close to the interviewee's mouth. Since this research intends to enhance readily available wide-band Urdu LVCSR, so only microphone speech is used in this paper. The data was recorded and digitized at 16KHz sampling rate, and 16-bit Pulse-Code Modulation (PCM) with mono channel.

A series of questions was asked to each interviewee during the session. For an interview, five sets of questions were designed for volunteer interviewees which covered daily routine, past experience, hobbies, interests and diversified topics. Speakers' were not restrained for proper articulation or being monolingual. 25 hours of effective speech data was released, recorded from 82 speakers. The corpus was segregated into train and test sets keeping both the sets gender balanced. The data was manually annotated on sentence level with transliteration of English words into Arabic script as well. The details of the corpus are tabulated in Table 1.

This corpus is revised manually to transcribe English words into Latin script. Like Hindi, Urdu speakers commonly mix English words or phrases in their daily communication (sometimes referred as *Pinglish*). Manifold English words are so commonly used in spoken Urdu that their alternative Urdu words are becoming extinct. Moreover, intra-word switches make linguists perplexed about language tag of the word. The problem is exacerbated in case of intra-word switching in compound words (with one word in English and the other in Urdu, example is given in Table 2). So, in order to use the Urdu spontaneous corpus for code-switched ASR, the challenge is to define a criterion to decide the language tag of a word. After a perusal of such muddling words, a simple rule is defined:

- Each candidate word is transliterated into Latin script.
  - If the word is present in Cambridge English dictionary [23], it is transcribed in Latin script and in Arabic script otherwise.

So, the aforementioned rule is followed while revising the spontaneous Urdu corpus to make it Urdu-English spontaneous speech corpus.

After revisiting it, the corpus is investigated to analyze the types of code-switching it carries, number of switches in an utterance and number of English words in a switch. It is noticed that the corpus contains all types of code-switching such as intra-word, intra-sentential and inter-sentential (examples are tabulated in Table 2).

Out of 21831 utterances, 10617 sentences contain code-switching with an average number of 3.5 switches per utterance and 1.3 English words during English turn. However, in case of inter-sentential switching, a maximum of 13 English words are observed in the corpus. Corpus contains 3769 English and 6785 Urdu unique words.

Though the addition of 20 hours of spontaneous CS data reduces WER on evaluation set but it is a very small amount to add in 415 hours of read speech. So, using this ASR, Urdu BC data is segmented and transcribed to grow spontaneous CS corpus. Transcriptions against each file are verified and rectified by linguists. Speakers, in BC interviews and talk shows, sometimes switches to other native local languages such as Punjabi as well but these segments are dropped out. 10 hours of BC data are added in the ASR so far. Out of 4601 utterances of Urdu BC

Table 2: *Examples of types of code-switching from Urdu-English CS corpus*

Types of code-switching	**Example**
Intra-word	سگریٹ نوشی sɪgrəæt noːʃiː *Smoking*
Intra-sentential	decision لیا تو اس time پہ I was very happy dɪsɪjən lɪjaː t̪oː ʊs t̪aːɪm peː aːiː vaːz væːriː hæːpiː *(When I) took the decision, I was very happy*
Inter-sentential	اس کي باوجود ہم لوگوں ني یہ تجربہ کیا that was very dangerous leːkɪn ɪs keː baːvəʤuːd̪ həm loːgõː neː jeː t̪əʤərbaː kɪjaː d̪æːt vaːz væːriː deːnʤərəs *In spite of that we tried out, that was very dangerous*

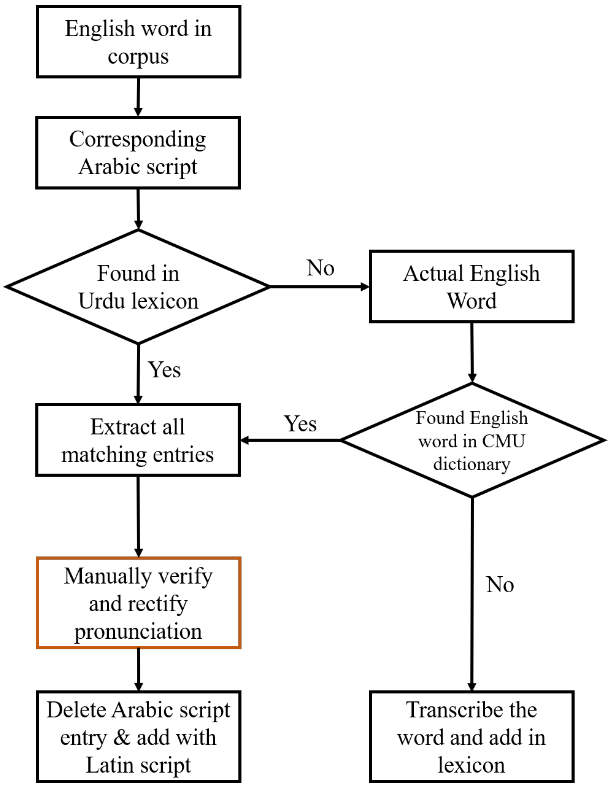

Figure 1: *Process of preparing Pakistani accented English lexicon*

corpus, 3206 sentences contain code-switching with an average number of 5.7 switches per utterance and 2.1 English words during English turn. Urdu BC corpus contains 3693 English and 4901 Urdu unique words. This corpus is termed as *Urdu BC corpus* throughout the paper.

3. Experimental Setup

Acoustic, language and pronunciation models are the building blocks of a speech recognition system. This section briefly describes the development of a bilingual lexicon and building of acoustic and language models.

3.1. Lexicon

Urdu LVCSR [10] was developed with a vocabulary size of 199K words which included the 106K words from readily available Urdu lexicon [24]. This lexicon embodied the complete CISAMPA mapped CMU pronunciation dictionary [23]. However, automatic mapping mapped the words closely to the native accented pronunciation which is sometimes different from Pakistani English accented pronunciation. This limitation makes this lexicon, in its actual form, futile for this project. And re-transcribing the whole English lexicon is a much time-consuming process. Lexicon improvement is expedited through a devised scheme, shown in Figure 1.

Most of the transliterations of the English words exist in Urdu lexicon. For all such words, lexicon entries can be automatically generated by using same pronunciation and changing Arabic scripted word with Latin one. However, this solution poses its own problem, which is; a number of Urdu words are identical to the transliterations of various English words. For example, the English words *fail* and *feel* are transliterated identically in Urdu as فیل /fiːl/, which itself has different meaning in Urdu (*Elephant:*derived from Arabic). Changing the Urdu word to *fail* or *feel* in lexicon means deletion of a valid Urdu entry. A solution may be to keep both, original and altered, entries. But on other hand, *University* is transliterated as یونیورسٹی /juːnɪvərsətiː/, which is commonly used in Urdu for same meaning and keeping the Urdu entry means a redundant term in lexicon. This challenge bridles the fully automatic process and requires manual engagement.

The process is carried out in three hierarchical steps. Transliterations of English words are extracted from Urdu lexicon and the pronunciations are manually verified. Linguists have to just assign binary marks to the entry on two levels that are; if the pronunciation of transliteration is correct for English word and is the Urdu entry removable from lexicon. Remaining words with no entry of their transliteration in Urdu lexicon are checked in CMU dictionary and correct pronunciations of such words are added as alternatives in lexicon by linguists. The words, which does neither exist in Urdu lexicon nor in CMU dictionary, are transcribed manually.

3.2. Acoustic Modeling

To enhance read speech Urdu LVCSR for code-switched spontaneous speech, 20 hours from Urdu CS spontaneous and 10 hours from Urdu BC spontaneous speech corpora are progres-

Table 3: *Nomenclature for acoustic models*

	Description	Duration (H)
AM1	Read speech acoustic model of Urdu LVCSR [10]	415
AM2	$AM1$+20 hours Urdu spontaneous CS speech corpus	435
AM3	$AM2$+10 hours Urdu BC speech corpus	445

Table 4: *Nomenclature for language models*

	Description	Word count (M)	Perplexity
LM1	Corpus of Urdu LVCSR [10]	154	408.82
LM2	$LM1$+code-switched copy of LM1	215	487.74
LM3	$LM2$+copies of Urdu CS data (only train set) text corpus	240	156.96
LM4	$LM3$+English corpus	372	480.37
LM5	$LM4$+copies of Urdu BC data text corpus	404	159.80

sively augmented in 415 hours of read speech training data. Acoustic model is trained through a Time Delay Neural Network (TDNN) without using any prior alignments from GMMs (and hence termed as end-to-end training). Context-dependency is addressed using a left biphones full tree which means a separate HMM model for each biphone pair[1]. However, manifold biphones are never seen during training and network learns to ignore them. Model is trained in a single stage without any interposed alignments and tree building. 2-states HMM topology is used with no restriction on self-loops. For training, 40 Mel-Frequency Cepstral Coefficients (MFCCs) are extracted from data using a window size of 25ms and a shift of 10ms. Only speaker normalization is applied to features to have zero mean and a unit variance. Acoustic model has been detailed in [22]. Kaldi [25] toolkit is used for experimentation.

3.3. Language Modeling

Urdu LVCSR used a language model built on 154M Urdu words crawled from myriad websites. However, language model for code-switched spontaneous speech is challenging since this phenomenon is more conspicuous in spoken language than the written one. Language model is improved gradually by adding data from annotated Urdu CS spontaneous speech, Pakistani English news data crawled from various news websites and spontaneous speech data from Open National American Corpus (ONAC) [26]. The step by step upswing is showed in experimental results. SRI Language Modeling (SRILM) toolkit [27] is used for building trigram language model.

4. Experimental Results

Various acoustic and language models are evaluated on test set of Urdu CS corpus. It contains 56277 words out of which 5135 are of code-switched English. Initially, it is evaluated on Urdu LVCSR system (trained on 415 hours of read speech) which yields an WER of 40.71%. Acoustic and language models are then gradually tuned to improve performance of speech recognition system on spontaneous CS speech. Various acoustic and language models are experimented to optimize the performance on spontaneous and code-switched speech. Nomenclature used in results (Table 5) is given in Table 3 and Table 4 for acoustic and language models respectively.

Test set is evaluated on three acoustic models trained with gradual increment in training speech data. Similarly, five language models are experimented with gradual addition of text corpus. After evaluating data on the acoustic and language model of Urdu LVCSR, language model is improved. While reworking on Urdu CS data, linguists maintained a Urdu to English mapping list. All such mapped words are searched in 154 Million corpus and a copy is prepared changing these Urdu words with their English mappings. This copy is termed as *code-switched copy of LM1* in Table 4. The idea is to auto-generate a code-switched corpus which improves WER slightly. But the enhancement is not much significant because it doesn't generate natural code-switching corpus (which actually occurs in test set). So, some replicas of Urdu CS training set text corpus are added into language model which significantly reduces the perplexity of language model and results in considerable WER improvement. Though the test set is unseen, but still this addition worked a lot due to natural code-switching.

To cover the inter-sentential code-switching, English text corpus was added into language model. This helps to estimate the probabilities of English 3-grams in case of English sentences. English corpus (157 Million words) included ONAC spoken data (2.97 Million words), *Librispeech* data (131.76 Million words) and crawled data of various Pakistani news websites (22.51 Million words). Though the improvement in overall WER is very slight, but it impacted the improvement of confused English words in case of long English sentences. Perplexity of language model on test corpus is though increased by adding too much English data. Finally, Urdu BC data is added into system (both acoustic and language models) which decreases WER to 26.95%. Reduction in perplexity is attributed to coverage of natural code-switching corpus of broadcast data.

5. Conclusion

In this paper, the first step towards large vocabulary spontaneous and Urdu-English code-switched speech recognition system is presented. An available spontaneous Urdu speech cor-

Table 5: *%WER on various acoustic and language models*

Acoustic model	Language model	% WER		
		Total	English	Urdu
AM1	LM1	40.71	98.96	34.86
AM2	LM1	39.4	97.32	33.58
AM2	LM2	38.19	83.73	33.62
AM2	LM3	29.57	67.40	25.77
AM2	LM4	28.46	55.67	25.73
AM3	LM5	26.95	51.29	24.50

[1]Dataset has 84 Urdu phonemes, 2 silence and one hesitation phones. So, there are total 87*86*2=14969 untied HMM states when using 2-state topology

pus of 25 hours from 82 speakers is revised to make it usable for enhancement of read speech Urdu LVCSR for Urdu-English code-switched speech recognition. State-of-the-art DNN-HMM acoustic model is trained. Furthermore, 10 hours of Urdu broadcast data from diverse categories is collected, annotated and incremented in existing train set. WER is reduced from 41% (on 415 hours of read speech) to 26.95%.

6. References

[1] E. Yılmaz, S. Cohen, X. Yue, D. A. van Leeuwen, and H. Li, "Multi-Graph Decoding for Code-Switching ASR," in *Proc. Interspeech 2019*, 2019, pp. 3750–3754.

[2] Z. Zeng, Y. Khassanov, V. T. Pham, H. Xu, E. S. Chng, and H. Li, "On the End-to-End Solution to Mandarin-English Code-Switching Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2165–2169.

[3] K. Taneja, S. Guha, P. Jyothi, and B. Abraham, "Exploiting Monolingual Speech Corpora for Code-Mixed Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2150–2154.

[4] Omniglot, "Urdu," [Online]. Available. https://omniglot.com/writing/urdu.htm, accessed: 2020-05-03.

[5] Z. Zeng, Y. Khassanov, V. T. Pham, H. Xu, E. S. Chng, and H. Li, "On the End-to-End Solution to Mandarin-English Code-Switching Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2165–2169.

[6] E. Yılmaz, H. van den Heuvel, and D. van Leeuwen, "Code-switching detection using multilingual dnns," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 610–616.

[7] B. H. A. Ahmed and T. Tan, "Automatic speech recognition of code switching speech using 1-best rescoring," in *2012 International Conference on Asian Language Processing*, 2012, pp. 137–140.

[8] D. Amazouz, M. Adda-Decker, and L. Lamel, "Addressing code-switching in french/algerian arabic speech," in *Proc. Interspeech 2017*, 2017, pp. 62–66.

[9] A. Pandey, B. M. L. Srivastava, R. Kumar, B. T. Nellore, K. S. Teja, and S. V. Gangashetty, "Phonetically balanced code-mixed speech corpus for hindi-english automatic speech recognition," in *LREC*, 2018.

[10] M. U. Farooq, F. Adeeba, S. Rauf, and S. Hussain, "Improving Large Vocabulary Urdu Speech Recognition System Using Deep Neural Networks," in *Proc. Interspeech*, 2019, pp. 2978–2982.

[11] J. Ashraf, N. Iqbal, N. S. Khattak, and A. M. Zaidi, "Speaker independent urdu speech recognition using hmm," in *2010 The 7th International Conference on Informatics and Systems (INFOS)*, 2010, pp. 1–5.

[12] M. Qasim, S. Nawaz, S. Hussain, and T. Habib, "Urdu speech recognition system for district names of pakistan: Development, challenges and solutions," in *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, 2016, pp. 28–32.

[13] A. A. Raza, A. Athar, S. Randhawa, Z. Tariq, M. B. Saleem, H. Bin Zia, U. Saif, and R. Rosenfeld, "Rapid collection of spontaneous speech corpora using telephonic community forums," in *Proc. Interspeech 2018*, 2018, pp. 1021–1025.

[14] H. Sarfraz, S. Hussain, R. Bokhari, A. A. Raza, I. 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-COCSDA*, 2010.

[15] M. Prasad, D. van Esch, S. Ritchie, and J. F. Mortensen, "Building Large-Vocabulary ASR Systems for Languages Without Any Audio Training Data," in *Proc. Interspeech 2019*, 2019, pp. 271–275.

[16] N.-Q. Pham, T.-S. Nguyen, J. Niehues, M. Müller, and A. Waibel, "Very Deep Self-Attention Networks for End-to-End Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 66–70.

[17] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, and R. T. Gadde, "Jasper: An End-to-End Convolutional Neural Acoustic Model," in *Proc. Interspeech 2019*, 2019, pp. 71–75.

[18] M. Li, Y. Cao, W. Zhou, and M. Liu, "Framewise Supervised Training Towards End-to-End Speech Recognition Models: First Results," in *Proc. Interspeech 2019*, 2019, pp. 1641–1645.

[19] N. Moritz, T. Hori, and J. L. Roux, "Unidirectional Neural Network Architectures for End-to-End Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 76–80.

[20] Y. Belinkov, A. Ali, and J. Glass, "Analyzing Phonetic and Graphemic Representations in End-to-End Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 81–85.

[21] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi," in *Interspeech 2016*, 2016, pp. 2751–2755.

[22] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free mmi," in *Proc. Interspeech 2018*, 2018, pp. 12–16. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1423

[23] "The CMU pronouncing dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict, accessed: 2020-05-03.

[24] F. Adeeba, T. Habib, S. Hussain, Ehsan-ul-haq, and K. S. Shahid, "Comparison of urdu text to speech synthesis using unit selection and hmm based techniques," in *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, 2016, pp. 79–83.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Vesel, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[26] "Open american national corpus," http://www.anc.org/data/oanc/contents/, accessed: 2020-05-03.

[27] A. Stolcke, "Srilm — an extensible language modeling toolkit," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, 2004.