Research Article

End-to-End Speaker-Independent ASR for Pashto: Resource Development and System Implementation

Muhammad Akbar Ali Khan[1*], Arbab Waseem Abbas[2], Waseem Ullah Khan[2*], Sahibzada Abdur Rehman Abid[3], Ilham Hamid[4]

[1]Department of R&D Khyber College of Dentistry, Peshawar; [2]Institute of Computer Sciences & Information Technology, The University of Agriculture, Peshawar; [3]Khyber Pakhtunkhwa Board of Investment & Trade (KP-BOIT); [4]Hash Computing Pvt. Ltd., Peshawar, Pakistan

Abstract: This paper examines the creation of a continuous speech recognition system for Pashto and presents the work as a response to a clear gap in existing research. The study explains why earlier efforts focused on digits or isolated words and links this limitation to the lack of a continuous speech database and a phonetic dictionary. To the best of author's knowledge, no earlier research has been undertaken on the continuous voice recognition of Pashto. This might be due to the unavailability of the Pashto continuous speech database and the Pashto phonetic dictionary. This paper includes the development of a medium size corpus along with a phonetic dictionary that supports practical training needs. Feature extraction utilized Mel Frequency Cepstral Coefficients and Perceptual Linear Prediction features, while classification employed a Hidden Markov Model method, adhering to accepted practices in speech processing and accommodating Pashto phonology. The system scored a recognition rate of 74.25 percent on broadcast speech, a number that shows functional competence given the variability prevalent in such audio.

Received: October 22, 2024; **Accepted:** December 08, 2025; **Published:** December 24, 2025
***Correspondence:** Muhammad Akbar Ali Khan (akbar.muhammad@kcd.edu.pk) and Waseem Ullah Khan (engrwaseem.aup@gmail.com)
Citation: Khan, M.A.A., A.W. Abbas, W.U. Khan, S.A.R. Abid and I. Hamid. 2025. End-to-end speaker-independent ASR for Pashto: Resource development and system implementation. *Journal of Engineering and Applied Sciences,* 44: 37-49.
DOI: https://dx.doi.org/10.17582/journal.jeas/2025.44.37.49
Keywords: Pashto ASR, Pashto Continuous speech recognition, Pashto speech corpus, Speaker-independent ASR, Low-resource language, HTK toolkit, Feature extraction, Broadcast speech recognition

Introduction

The development of machines that can mimic human behavior, particularly the capability to perceive spoken messages known as Automatic Speech Recognition (ASR) has intrigued the scientists and engineers for decades. The research on ASR started in the 1930's with the analysis of the speech signal (Douglas and Brousseau, 1992), along with the production of synthetic speech. The research in ASR progressed from the development of simple machines which could recognize a small set of vocabulary to the complex systems that can recognize fluently spoken natural languages. In the modern world, human's interaction with the machines has increased drastically and this dependence on the machines is increasing further with each passing day. Speech is the most suitable modality to interact with and communicate with machines and therefore the development of speech-based interfaces has gained much currency in

recent years. Though much progress has been made in speech-based human-computer interaction (HCI), but it is yet limited both in terms of its operation and scope. The earlier work on ASR is limited to the controlled environment and for the deployment of such interfaces in the real environment still needs further research (Miciak, 2008). Moreover, the focus of earlier research was mainly the development of ASR systems in the English language and a few other languages of the world such as French (Brousseau *et al*., 1995), Chinese (Lee *et al*., 1993) and German.

Pashto language is spoken in parts of South and Central Asia including, Pakistan, Afghanistan, Iran and Pashtuns Diaspora across the globe (Shriberg, 2005). It is one of the two official languages of Afghanistan and a major language in two of the four provinces of Pakistan i.e. Khyber Pakhtunkhwa and Balochistan. Besides it is also the native language of the Federally Administered Tribal Areas of Pakistan. Although the exact number of Pashto speakers is not known, however the estimated number of Pashto ranges from 45 million to 50 million (Hejazi *et al*., 2005). Despite the vast number of speakers, very little research on Pashto speech recognition has yet been carried out and therefore Pashto is considered to be one of the less researched and limited resources languages (Kathol *et al*., 2005).

While recent advances in deep learning (e.g., end-to-end ASR (Amodei *et al*., 2015)) have revolutionized high-resource languages, Pashto's lack of large datasets and dialectal diversity remain barriers. The proposed work bridges this gap by:

- Providing the first publicly available phonetic dictionary for Pashto ASR.
- Demonstrating feasibility of continuous recognition in noisy broadcast environments.
- Establishing a baseline for future research on low-resource ASR.

This research undertakes the development of Pashto continuous speech ASR system. The Pashto ASR has been implemented using the Cambridge University's Hidden Markov Model toolkit (HTK). As HTK cannot recognize the Pashto script, therefore the Pashto text was first converted to the roman script.

The rest of paper is organized as follows. Section-2 of this paper presents the language analysis, transliteration of Pashto text and transcription of Pashto speech under the title of data preparation. In section-3 the speech database developed for the Pashto ASR is described briefly. The most important phonetic dictionary development is discussed in section-4. The methodology used for the development of Pashto ASR is discussed in detail in section-5. In section 6 results and analysis are discussed in detail. Section-7 concludes the paper and identifies future work directions. Transliteration of Pashto script and its related issues are discussed in section-2. Section-3 describes the transcription of Pashto speech used for the development of Phonetic Dictionary. Section-4 describes HMM technique and HTK tool used to perform the experiment. Section-5 describes the Phonetic Dictionary created for the experiment and the speech Database created for experiment. The methodology used for the development of Pashto ASR is discussed in detail in section-6. Section-7 concludes the paper.

Related Work

Automatic Speech Recognition (ASR) for Pashto remains an understudied domain compared to widely researched languages like English, Mandarin, or Arabic. Existing research on Pashto ASR can be categorized into isolated word/digit recognition, phonetic resource development, and preliminary continuous speech recognition. This section reviews prior work and identifies gaps addressed by our study.

Isolated Word and Digit Recognition

Early efforts in Pashto ASR focused on small-vocabulary tasks, such as digit recognition. Khan *et al*. Khan *et al*. (2015) developed an HMM-based system for Pashto digits (0–9), achieving 92 percent accuracy in controlled environments. However, their work did not generalize to continuous speech or larger vocabulary. Similarly, Ahmad *et al*. (2018) employed Deep Neural Networks (DNNs) for isolated word recognition, reporting 85 percent accuracy on a 50-word lexicon. While these studies demonstrated feasibility for constrained tasks, they lacked scalability to real-world applications requiring fluent speech recognition.

Phonetic and Corpus Development for Pashto

The absence of standardized phonetic resources has hindered Pashto ASR progress. Zahid *et al*. (2017)

compiled a 5-hour read-speech corpus with 10 speakers but did not implement a full ASR system. Khan and Usman (2019) proposed a SAMPA-based phonetic transcription scheme, yet their work remained theoretical without integration into an ASR pipeline. These studies highlighted the need for large-scale, annotated speech datasets and practical transcription methodologies gap our work addresses by introducing a TIMIT-compatible scheme and a 1242-word phonetic dictionary.

Continuous Speech Recognition Attempts

Few studies have explored continuous Pashto ASR. Gul *et al*. (2020) conducted a preliminary study using HMMs on a 1-hour dataset, achieving only ~50% accuracy due to limited training data and simplistic modeling. Their results underscored the challenges of coarticulation and speaker variability in continuous speech. In contrast, our work improves upon this by:

- Utilizing a larger, phonetically diverse broadcast corpus (29 minutes, 6 speakers).
- Optimizing HMM training with MFCC/PLP features, yielding 74.25–78.16% accuracy.
- Introducing a speaker-independent system, unlike prior speaker-dependent approaches (Rahman and Ali, 2016).

Comparative Analysis and Research Gaps

Table 1 summarizes key differences between prior work and our contributions.

Table 1. *Comparison of Pashto ASR research.*

Study	Scope	Method	Accuracy	Limitations
Khan *et al*. (2015)	Isolated digits	HMM+ MFCC	92%	Small vocabulary, clean speech
Ahmad *et al*. (2018)	Isolated words	DNN	85%	Limited to 50 words
Gul *et al*. (2020)	Continuous speech	Basic HMM	~50%	Tiny dataset, low accuracy
This study	Continuous speech	HMM+ MFCC/ PLP	74–78%	First robust system for broadcast Pashto

Language Analysis

The native language of the Pashtun people is Pashto, which is predominantly spoken in the Khyber Pakhtunkhwa and Balochistan provinces of Pakistan, as well as in the majority of Afghanistan and some regions of India. It is one of Afghanistan's two official languages, the other being Dari. Pashto is also known by several names, including Pashtu, Pashto, Pushto, and Afghani. The earliest records of Pashto writing can be traced back to the sixteenth century. In 1936, it was officially designated as the language of Afghanistan (Parthasarathy and Coker, 1992). Approximately forty to sixty million individuals are estimated to speak Pashto (Allen, 1987).

Pashto, being one of the least explored languages, is missing essential resources such as a standard database, a phonetic dictionary, programming platforms, and support for interfacing with various language processing tools and packages that are necessary for conducting research in Pashto language processing. Additionally, the fundamental concepts of the Pashto language are discussed, and aspects such as the Pashto character set, Pashto phonetic structure, and other important issues pertaining to Pashto ASR development are covered. The distinctions between the Pashto language and the English language are also briefly explored.

Like any other spoken languages, Pashto has its special character set. The Pashto character set is a modification of the Arabic one, with a few more letters to indicate those phonemes which are peculiar to the Pashto language and do not have a corresponding symbol in the Arabic character set. The language uses a script derived from the Persian one, which is in turn modified from that of Arabic. Thus, most of the alphabets of the Pashto language ultimately trace back to the Arabic language.

Baizid Roshan (931-980) organized the early Pashto character set from Arabic, creating the earliest known Pashto character set (Hunt and Black, 1996). Since that point, it has been subject to numerous revisions by different scholars, with significant contributions from Akhund Darviza, Khan Abdul Samad Khan, Khushal Khan Khatak, Qalandar Momand and Wazir Mohammad Gul Khan.

The existing set of Pashto characters have been revised by the conference of Pashto Tolana (summit), after which they have been reviewed and approved by the experts of Khyber Pakhtunkhwa and Afghanistan.

Table 2. *Pashto Character Set*

S. No	Alphabet	S. No	Alphabet	S. No	Alphabet
1	ا	16	ر	31	ق
2	ب	17	ړ	32	ک
3	پ	18	ز	33	ګ
4	ت	19	ژ	34	ل
5	ټ	20	ږ	35	م
6	ث	21	س	36	ن
7	ج	22	ش	37	ڼ
8	ځ	23	ښ	38	و
9	چ	24	ص	39	ه
10	څ	25	ض	40	ي
11	ح	26	ط	41	ې
12	خ	27	ظ	42	ى
13	د	28	ع	43	ۍ
14	ډ	29	غ	44	ئ
15	ذ	30	ف		

Pashto Characters Set

The core collection of symbols in a language is referred to as the alphabet or character set of that language. In Pashto, this is referred to as Pashto Haroof-e-Tahajee. The character set of the Pashto language has forty-four letters (Tegey and Robson, 1996). Table 2 presents the character set.

The Pashto language includes all 28 characters from the Arabic alphabet. Furthermore, there are a number of letters in the Pashto character set that are not part of any additional Arabic script. Characters such as چ, پ, and ژ are also present in the character sets of languages like Persian and Urdu. For example, some symbols that indicate retroflex consonants are /ʈ/, /ɖ/, and /ɳ/. These characters written in Pashto are like the standard Arabic characters 'nun', 'te', 'dāl', and 're' with slight variations. The difference is that 'skəṇay', 'paṇḍak', or 'ğaṛwanday' is attached beneath these characters. This additional portion resembles a little circle, as indicated in table 2.2.1, notably characters numbered 5, 14, 17, and 37. The 20th and 23rd letters appear similar to the 16th and 21st letters, respectively, in table 2.2.1, having a dot above and below. The 8th and 10th characters, which represent /d͡z/ and /t͡s/, look like ح, the 11th character of table 2.2.1, with the distinction that the 8th character has a 'hamza' (ء) above it, however, the 10th character contains three Pashto-specific dots above it. Furthermore, the Pashto language includes ى, ي, ۀ, and ى for further diphthongs and vowels.

Character Orientation at Different Positions

The vast majority of the forty-four characters in the Pashto language exhibit various orientations based on their position within the words (Titze and Martin, 1998). For instance, a character located at the beginning of a word will have one orientation, while one situated in the middle will adopt a different orientation, and one at the end will display yet another orientation. Additionally, certain characters continue to have the same orientation regardless of their position, but others will only appear in select areas and not in all three points.

Most of these characters change their orientation at the Beginning, Middle, and Ending positions. Table 3 illustrates each of these places as well as the character orientations.

Phonetic Division of Characters into Vowels and Consonants

Like English, the Pashto language also incorporates the notion of vowels and consonants, and like English, Pashto character set is also phonetically divided into two groups: vowel sound letters make up one group, while consonant sound characters make up the other (Titze and Martin, 1998). The vowel sound letters are presented in Table 4, and the consonant sound letters are specified in Table 5.

Table 3: Pashto Characters Orientation at Three Different Positions.

S. No	Contextual form				Pashto Unicode (Hex)
	Isolated	End	Middle	Beginning	
1	ا , آ	ـا	-	-	0627, 0622
2	ب	ـب	ـبـ	بـ	0628
3	پ	ـپ	ـپـ	پـ	067E
4	ت	ـت	ـتـ	تـ	062A
5	ټ	ـټ	ـټـ	ټـ	067C
6	ث	ـث	ـثـ	ثـ	062B
7	ج	ـج	ـجـ	جـ	062C
8	چ	ـچ	ـچـ	چـ	0686
9	ح	ـح	ـحـ	حـ	062D
10	خ	ـخ	ـخـ	خـ	062E
11	څ	ـڅ	ـڅـ	څـ	0685
12	ځ	ـځ	ـځـ	ځـ	0681
13	د	ـد	-	-	062F
14	ډ	ـډ	-	-	0689
15	ذ	ـذ	-	-	0630
16	ر	ـر	-	-	0631
17	ړ	ـړ	-	-	0693
18	ز	ـز	-	-	0623
19	ژ	ـژ	-	-	0698
20	ږ	ـږ	-	-	0696
21	س	ـس	ـسـ	سـ	0633
22	ش	ـش	ـشـ	شـ	0634
23	ښ	ـښ	ـښـ	ښـ	069A
24	ص	ـص	ـصـ	صـ	0635
25	ض	ـض	ـضـ	ضـ	0636
26	ط	ـط	ـطـ	طـ	0637
27	ظ	ـظ	ـظـ	ظـ	0638
28	ع	ـع	ـعـ	عـ	0639
29	غ	ـغ	ـغـ	غـ	063A
30	ف	ـف	ـفـ	فـ	0641
31	ق	ـق	ـقـ	قـ	0642
32	ک	ـک	ـکـ	کـ	06A9
33	ګ	ـګ	ـګـ	ګـ	06AB
34	ل	ـل	ـلـ	لـ	0644
35	م	ـم	ـمـ	مـ	0645
36	ن	ـن	ـنـ	نـ	0646
37	ڼ	ـڼ	ـڼـ	ڼـ	06BC
38	و	ـو	-	-	0648
39	ه	ـه	ـهـ	هـ	0647
40	ي	ـي	ـيـ	يـ	064A
41	ې	ـې	ـېـ	ېـ	06D0
42	ی	ـی	-	-	06CC
43	ۍ	ـۍ	-	-	06CD
44	ئ	ـئ	ـئـ	ئـ	0626

Table 4: *Vowel Sound Letters*

S. No	Alphabet	S. No	Alphabet
1	ا	5	ي
2	و	6	ى
3	ه	7	ىٔ
4	ي	8	ئ

Table 5: *Consonant Sound Letters*

S. No	Alphabet	S. No	Alphabet
1	ا	23	ښ
2	ب	24	ص
3	پ	25	ض
4	ت	26	ط
5	ټ	27	ظ
6	ث	28	ع
7	ج	29	غ
8	څ	30	ف
9	چ	31	ق
10	ځ	32	ک
11	ح	33	ګ
12	خ	34	ل
13	د	35	م
14	ډ	36	ن
15	ذ	37	ڼ
16	ر	38	و
17	ړ	39	ه
18	ز	40	ي
19	ژ	41	ي
20	ږ	42	ى
21	س	43	ىٔ
22	ش	44	ئ

Concept of Capital and Small in Letters

Pashto possesses several qualities that assist language processing more effectively than the English language. One such feature is the absence of differentiation between capital and small letters. In contrast to English, Pashto does not utilize capital and small letters. This distinction can be illustrated through examples from both languages.

The English text and its corresponding translation in Pashto are depicted above in Figure 1. It is evident that the Pashto language lacks the concept of capital and small letters, which is present in the English language.

All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

د بشر ټول افراد آزاد نړۍ ته راځي او د حیثیت او حقوقو له پلوه سره برابر دی۔ ټول د عقل او وجدان خاوندان دی او یو له بل سره د وروری په روحیې سره باید چلند کړی۔

Fig 1: *PASHTO Language Script*

Acronyms in Pashto Language

From a technical perspective, acronyms are typically words created by combining the initial letters or segments of other words. For instance, UNO represents the United Nations Organization, while UK stands for the United Kingdom. Acronyms serve to abbreviate the names of organizations, countries, institutions, associations, procedures, and similar entities, facilitating their frequent use in everyday conversations.

According to the author's knowledge, there are no acronyms utilized in the Pashto language. The words are pronounced as they are. Nevertheless, some English words that Pashto speakers use in their daily lives have been integrated into the Pashto language. These English-derived acronyms in Pashto subsequently function as ordinary isolated words.

Hard Space vs. Soft Space

The spacing between letters and words in the Pashto language is a major orthographic problem. In various languages, including English, Chinese, German, and French, spaces are utilized only to distinguish words from one another. In contrast to these languages, the Pashto language employs two separate sorts of spaces: hard space and soft space. Soft space is placed between individual letters, whilst hard spaces are utilized to divide full Pashto words from one another.

Diacritic Marks

The Pashto language uses several diacritical markings, just like other languages. There are four diacritic markings in the Pashto language, which are referred to as 'zwar', 'zer', 'pesh', and 'zwarakay'. Table 6 lists the Pashto symbols for these marks along with their Unicode.

Following this concise overview of the features of the Pashto language in relation to language processing and speech recognition research, the primary concerns regarding the establishment of a standard character set, as well as the relevant transliteration and transcription schemes addressed in this research project, are outlined in the subsequent sub-section.

Table 6. Pashto Language Diacritic Marks

S. No	Name	Symbol	Unicode
1	Zwar	◌َ	064E
2	Zer	◌ِ	0650
3	Pesh	◌ُ	064F
4	Zwarakay	◌ٔ	0653

Transliteration of the Pashto Text

As HTK is designed primarily for the English language, it cannot be used directly with the Pashto text written in the Arabic script. To transform the Pashto text into a form compatible with the HTK it needs to be transliterated into the Roman script. In order to accomplish this task, a standard transliteration scheme should either exist or needs to be developed in which each alphabet symbol of the source language, i.e. Pashto, has a one-to-one mapping to a Romanized symbol. To the best of authors' knowledge, no agreed upon transliteration scheme yet exists for the Pashto languages. In this work, the transliteration scheme proposed in [18] was used with some minor modifications. The transliteration scheme is shown in Table 7.

The above transliteration scheme, though work well for the creation of training and test data for the speech recognition research, however the transliteration from English to Pashto may have some issues, such as,

- It uses the transliteration sh for ش which may map back to ش or ﺳ
- It maps س, ص, and ث to s that will be difficult to map back from the other side

Besides these limitations the proposed scheme works perfectly for the mapping in one direction i.e. transliterating Pashto text into the Roman script, needed for the data preparation for speech recognition research.

Microsoft Visual Studio was used to implement the transliteration technique in order to automatically obtain the transliterated text for the Pashto files. The code for both the front-end and back-end design was handled in Visual C#.

Table 7: *The Transliteration Scheme Used*

Grapheme	Transliteration	Grapheme	Transliteration
ا	a	ص	s
ب	b	ض	z
پ	p	ط	t
ت	t	ظ	z
ټ	T	ع	H
ث	s	غ	gh
ج	j	ف	f
چ	ch	ق	q
څ	dz	ک	k
ځ	ts	ګ	g
ح	h	ل	l
خ	x	م	m
د	d	ن	n
ډ	D	ڼ	N
ذ	z	و	W
ر	r	ه	h
ړ	R	ی	y
ز	z	ي	I
ژ	zh	ې	E
ږ	g	ۍ	@y
س	s	ئ	@i
ش	sh	آ	A
ښ	kh	ك	k

Transcription of the Pashto Speech

Transcription refers to the process of providing a phonetic symbol to each of the sounds observed in the speech. The International Phonetic Alphabet (IPA) provides a standard set of phonemes for this purpose. The IPA charts contain all phones that can be pronounced by humans using different places and manners of articulations. However, in many cases, the use of IPA symbols is either inconvenient, or they conflict with the reserved symbols of the HTK. An alternative transcription scheme is provided by the Speech Assessment Methods Phonetic Alphabet (SAPMA). A transcription scheme based on SAPMA referred as 'Pashto SAPMA' was initially developed for the Pashto language as shown in Table 8, however when used with the HTK, the SAPMA based transcription also faced issues similar to those of IPA.

Finally, to cope with the issues of capitalization, con-

trol symbols and HTK reserved symbols, a new transcription scheme, in line with the TIMIT was developed (Zue *et al*., 1989). The revised scheme along with the corresponding IPA symbols is given in Table 9.

Table 8: *Pashto SAMPA*

IPA	Transcription	IPA	Transcription
ʔ	?	f	f
b	b	q	q
p	p	k	k
t	t	ɲ	N
ʈ	t`	l	l
d ʒ	dZ	m	m
t ʃ	tS	n	n
dz	dz	ɳ	n`
ts	ts	w	w
x	x	h	h
d	d	I	i, I
ɖ	d`	ee	ee
z	z	əy	@y
r	r	əi	@i
ɻ	r\`	ɑ	A
ʒ	j, Z	ə	@
g	g	ʊ	O
s	s	oo	oo
ʃ	S	E	E
χ	X	e	e
ɣ	G	a	a

The transliterated files obtained in the previous section were transcribed using this scheme. The xtrans-windows-1.1 tool was used for the transcription process (XTrans Tool). The transcription was carried by listening to the audio of each word again and again and taking into consideration all the phonemes heard from the audio file.

Very strong co-articulation effect was observed in the continuous Pashto speech, and it was found that in many cases it was very hard to recognize the word boundaries due to the overlap of the final phone of one word with the initial phone of the following word. As a result, one word appearing with different surrounding words has different pronunciations and thus has multiple entries in the phonetic dictionary.

Speech Corpus

The initial phase in creating a speech corpus/database involves choosing appropriate content. To facilitate continuous speech recognition, the database must ensure that the recorded speech file encompasses all phonetic units of the language. Consequently, a Pashto broadcast audio file was selected as the content, categorized as read speech within the speech corpus. The audio file utilized for developing the corpus was spoken by six individuals: four males and two females. All speakers were adults and native to the language. The six speakers from Ashna Television, who provided the content, are detailed in Table 10.

Table 9: *Updated Transcription Scheme*

IPA	Transcription	IPA	Transcription
ʔ	ak	f	f
b	b	q	q
p	p	k	k
t	t	ɲ	ng
ʈ	tt	l	l
d ʒ	jh	m	m
t ʃ	ch	n	n
dz	dz	ɳ	nn
ts	ts	w	w
x	x	h	h
d	d	I	ih
ɖ	dd	ee	ee
z	z	əy	axy
r	r	əi	axi
ɻ	rr	ɑ	aa
ʒ	j	ə	ax
g	g	ʊ	uh
s	s	oo	oo
ʃ	sh	E	eh
χ	xx	e	e
ɣ	gh	a	a

Table 10: *Database SPEAKERS' Description*

S. No	Speaker Name Code	Gender	Description
1.	SSA	Male	News Caster
2.	SSL	Female	News Caster
3.	AUZ	Male	News Reporter
4.	GZ	Male	News Reporter
5.	RD	Female	News Reporter
6.	SZ	Male	Politician

The unit for sound typically employed for continuous speech ASR includes Phonemes, which are the smallest yet meaningful units of sound in speech. While a standard set of Pashto phonemes is yet to be estab-

lished (Naeem, 2006), the phoneme list utilized in this research is a modified version of the list referenced by (Olson, 2008). The phoneme listing along with how frequently they appear in the audio speech is shown in Table 11. For a continuous speech database, it is vital that the material is phonetically diverse, meaning it should contain all phonemes with an appropriate number of occurrences in both the training and test datasets.

The audio data used for the development of speech corpus is the News broadcast from the Ashna television, Washington, USA. The total duration of the audio file is 29 minutes and 57 seconds and is split into 235 sentences. The average numbers of words in each sentence are 20.5. The total numbers of words in the database are 4818 whereas the number of distinct words is 1242. The numbers of speakers in the database are six among which four are male and two are female, all of them being adult. The Adobe Audition version 1.0 software was utilized for the purpose of splitting and editing the broadcast audio file. The audio file was opened in Adobe Audition version 1.0 and was edited and split while in zoom-in mode to facilitate easier editing and splitting. The utterance of the designated length was played back and verified for the spoken words, resulting in the creation of new word files. The audio file was divided into sentences and prompts using a sampling rate of 44100 Hz. The sentences and prompts were subsequently saved in .wav format.

Phonetic Dictionary Creation

A phonetic dictionary, commonly known as a pronouncing dictionary, serves as a reference that illustrates how words are pronounced according to their sounds. The development of a phonetically balanced dictionary is essential for automatic speech recognition in any language, and it should ideally include every word in that language. Nevertheless, for the Pashto language, establishing such a dictionary is quite challenging because of the lack of standardized vocabulary and pronunciation.

Phonetic dictionary is one of the integral parts of speech recognition. The phonetic dictionary is created for 1242 words. X-Trans-windows-1.1 software has been used for the phonetic description of the audio file. X-Trans software is a multi-lingual, multi-platform transcription toolkit. It can be used for multiple languages for transcription.

Table 11: *Phonemes, count in training data, test data and total*

S. No	Phoneme	Training count	Test count	Total count
1	aa	441	145	586
2	sp	1240	36	1276
3	ak	129	46	175
4	B	126	45	171
5	D	199	69	268
6	S	207	78	285
7	eh	90	28	118
8	F	37	16	53
9	Ih	337	120	457
10	N	372	147	519
11	R	402	138	540
12	ax	297	104	401
13	T	255	83	338
14	Jh	63	22	85
15	A	915	366	1281
16	L	326	105	431
17	K	191	70	261
18	I	5	3	8
19	J	116	43	159
20	W	282	119	401
21	oo	183	65	248
22	M	270	100	370
23	U	139	55	194
24	uh	184	74	258
25	q	40	14	54
26	z	92	31	123
27	dd	32	14	46
28	ee	287	123	410
29	tt	47	21	68
30	gh	47	17	64
31	dz	39	19	58
32	y	2	0	2
33	rr	87	32	119
34	ng	28	12	40
35	g	103	14	117
36	sil	2	2	4
37	p	115	45	160
38	e	9	3	12
39	x	88	33	121
40	sh	76	36	112
41	ch	30	16	46
42	h	105	46	151
43	ek	1	0	1
44	axy	22	8	30
45	axi	16	8	24
46	nn	4	2	6
47	xx	22	5	27
48	ts	24	10	34

After the phonetic transcription the phonetic dictionary was created in alphabetical order. The dictionary has 1242 words and so can be considered as a large vocabulary dictionary.

Pashto ASR System Development

After the development of Speech corpus and creation of Pashto Phonetic Dictionary needed for the development of acoustic model and language Model for the Pashto continues speech ASR system, experiments were performed on these data for the performance of ASR.

Data Segmentation

The speech broadcast speech file was segmented into 235 prompts/sentences using the Adobe Audition Software version 1.0. The data acquired is subsequently partitioned into training and test sets. The prompts were approximately divided into three ratios, specifically one (3:1) for the training and test data, meaning that 174 prompts were allocated for training while the remaining 61 prompts were designated for testing the recognizer. The distribution of prompts was made in such a way so as to ensure that all the phonemes of the test prompts must be in the training prompts. Table 10 shows the statistics of the phoneme's occurrences in both training and test data. The phoneme no 32 in table shows a rare occurrence case, occurred twice in training and was not available in the test prompts.

Feature Extraction and Training

The training data is used for developing both the acoustic and language models. Features suitable for the ASR were extracted from the training data and models were developed for each speech unit. MFCC based features which make the defective standard features for ASR research and widely used in English language ASR were used in this research. Hidden Markov Models (HMM) based speech recognition toolkit of Cambridge University, i.e. HTK is used for the training and testing of the recognizer. The block diagram of the ASR system showing different stages of the recognizer and utilized in this research work is shown below in Figure 2.

MFCC features were derived from the training data with the aid of HTK, and models for the phonemes, including their context-dependent bi-phones and triphones, were created based on these features following the standard HTK training methodology. The initial 13 features derived from the MFCC coefficients were selected, together with their first and second derivatives, leading to a feature set of 39 dimensions. These characteristics were used to create a five-state left-right HMM with three emitting states.

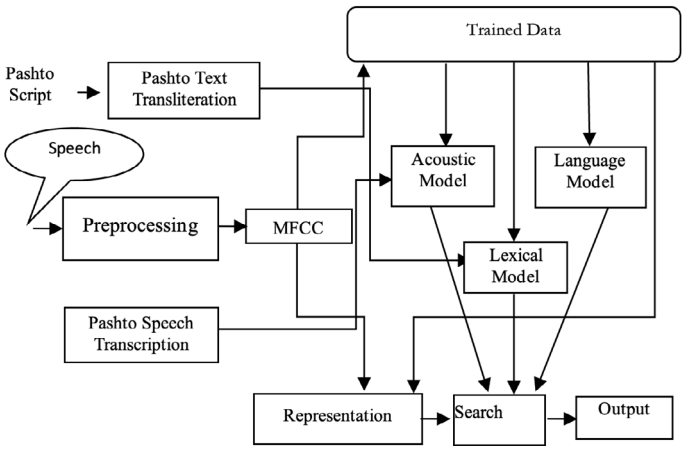

Figure 2*: Block Diagram of Pashto ASR System*

The coding parameters selected for obtaining the MFCC features were as described in Table 12.

Table 12: *MFCC Parameters and Values*

S. No	Parameter Name	Value
1	Source Format	WAV
2	TARGETKIND	MFCC_0_D_A
3	TARGETRATE	100000.0
4	WINDOWSIZE	250000.0
5	PREEMCOEF	0.97
6	NUMCHANS	26
7	CEPLIFTER	22
8	NUMCEPS	12
9	SAVECOMPRESSED	TRUE
10	SAVEWITHCRC	TRUE
11	USEHAMMING	TRUE

Recognition Results and Analysis

The recognition performance of the developed ASR system for the MFCC features and HMM classifier is given below. The performance of the system was checked for both the percentage of correctly recognized word and accuracy, and the results for different values of insertion penalty '–p' and grammar scale factor '–s' were obtained. The best performance of 74.25% words correct recognition was obtained for

p = 7-9 and s = 12-13. The comprehensive outcomes of various ASR experiments are presented in Table 12, alongside the symbols illustrated in Table 13. 'N' represents Total Utterances, 'H' denotes Correctly Recognized Utterances, 'D' indicates Deletion error, 'S' signifies Substitution error, and 'I' refers to Insertion error.

The MFCC feature gives acceptably good results for the Pashto ASR particularly for the broadcasted speech with uncontrolled environment and relatively high background noise, used in these experiments.

The ASR experiments were also conducted on a different set of features, i.e. Perceptive Linear Prediction (PLP) features. The best performance obtained from the PLP based features was 78.16% correct word recognition. The PLP results for other parameters are shown in Table 14.

Conclusion

This paper outlines the evolution of Pashto Continuous Speech Recognition using an HMM classifier. Both MFCC and PLP features have been incorporated. At first a speech database was built from a broadcast speech. The Pashto script was transliterated into English text format to enable its application using the available tools for ASR research. A word list was constructed from the transliterated file, and a phonetic dictionary was compiled. Furthermore, transliteration and transcription techniques for Pashto were developed as part of this endeavor. To make the experiments easier, a large phonetic vocabulary was put together.

As the inaugural study on continuous ASR for the Pashto language, the research detailed in this paper establishes a foundational benchmark for subsequent comprehensive investigations into various aspects of Pashto language processing and ASR. The tools created in this research, including transliteration and transcription schemes, have the potential for further refinement, enabling their application on additiona l speech recognition platforms and software packages. The results of Pashto Continuous Speech

Table 13: *PASR Results for MFCC Technique*

-p value	-s valve	N	H	D	S	I	% Correct	Accuracy
-10.0	20.0	1227	851	252	124	12	69.36	68.38
-6.0	12.0	1227	866	220	141	19	70.58	69.03
0.0	10.0	1227	900	166	161	28	73.35	71.07
5.0	15.0	1227	895	192	140	22	72.94	71.15
6.0	13.0	1227	908	166	153	27	74.00	71.80
7.0	12.0	1227	911	159	157	29	74.25	71.88
8.0	12.0	1227	911	159	157	29	74.25	71.88
8.5	12.5	1227	911	159	157	29	74.25	71.88
9.0	13.0	1227	911	159	157	29	74.25	71.88
10.0	20.0	1227	869	221	137	20	70.82	69.19

Table 14: *PASR Results for PLP Technique*

-p value	-s value	N	H	D	S	I	% Correct	Accuracy
5.0	15.0	1227	947	130	150	26	77.18	75.06
0.0	10.0	1227	947	113	167	28	77.18	74.90
10.0	20.0	1227	951	143	133	22	77.51	75.71
7.0	12.0	1227	959	102	166	32	78.16	75.55
-6.0	12.0	1227	935	138	154	21	76.20	74.49
6.0	13.0	1227	951	108	168	27	77.51	75.31
8.0	12.0	1227	956	100	171	35	77.91	75.06
-10.0	20.0	1227	919	186	122	18	74.90	73.43
9.0	13.0	1227	958	99	170	29	78.08	75.71
8.5	12.5	1227	959	99	169	32	78.16	75.55

Recognition (PCSR) can be further improved by incorporating more training data through a bootstrapping process and by utilizing the existing dictionary. The development of a more extensive vocabulary speech database will enhance the training of models, especially the context-dependent bi-phone and tri-phone models. Future research will also explore the application of additional feature extraction techniques, such as wavelets and wavelet packets, along with the implementation of various classifiers. The transcription presented in this study is grounded in the Yousafzai dialect, with the potential for expansion to encompass other dialects.

Acknowledgments

We would like to express our profound appreciation to Dr. Fatima Tuz Zuhra, assistant professor at the University of Peshawar's Computer Science department, for her significant assistance in creating the automated program that converted Pashto script into the appropriate text format. Her expertise and dedication were instrumental in advancing this critical component of our research.

Novelty Statement

The novelty lies in developing the first medium-scale Pashto continuous-speech corpus and phonetic dictionary, enabling the creation of the first HMM-based continuous Pashto ASR system.

Author's Contribution

Muhammad Akbar Ali Khan: Conceptualization, methodology, data management, writing-original draft
Arbab Waseem Abbas: Conceptualfzation, methodology, formal analysis, writing-original draft, project administration
Waseem Ullah Khan: Conceptualization, methodology, data management, writing-original draft
Sahibzada Abdur Rehman Abid: Methodology, formal analysis, data management, project administration
Ilham Hamid: Conceptualizatiori; methodology,-formal analysis, data management, project administration

Generative AI and AI-assisted technology statement
No generative AI and AI-assisted tools were used in the preparation of this manuscript.

Conflict of interest
The authors have declared no conflict of interes.

References

Douglas, B.P. and J.M. Brousseau (1992). The design for the Wall Street Journal-based CSR database. in Proc. DARPA SLS Workshop, Stroudsburg, PA, USA.

Miciak, M. (2008). Character recognition using Radon transformation and principal component analysis in postal applications. in Proc. Int. Multiconference on Computer Science and Information Technology (IMCSIT), Wisla, Poland, pp. 495–500.

Brousseau, J. *et al.* (1995). French speech recognition in an automatic dictation system for translators: The Transtalk project. in Proc. Eurospeech.

Lee, L.-S. *et al.* (1993). Golden Mandarin (I)—A real-time Mandarin speech dictation machine for Chinese language with very large vocabulary. IEEE Trans. Speech Audio Process., vol. 1, no. 2, pp. 158–179.

Shriberg, E. (2005). Spontaneous speech: How people really talk and why engineers should care. in Proc. INTERSPEECH, Lisbon, Portugal, pp. 1781–1784.

Hejazi, S.A. and Kazemi, R.S. (2008). Ghaemmaghami, and S. Sharif. Isolated Persian digit recognition using a hybrid HMM–SVM. in Proc. Int. Symp. Intelligent Signal Processing and Communications Systems (ISPACS), Bangkok, Thailand, pp. 1–4.

Kathol, A., K. Precoda, D. Vergyri, W. Wang, and S. Riehemann. (2005). Speech translation for low-resource languages: The case of Pashto. in Proc. INTERSPEECH, pp. 2273–2276.

Khan, M. *et al.* (2015). Automatic speech recognition of Pashto digits using HMMs. in Proc. IEEE ICET.

Ahmad, N. *et al.* (2018). Pashto isolated word recognition using DNNs. in Lecture Notes in Computer Science, vol. 11015, Springer.

Zahid, S. *et al.* (2017). Development of a Pashto speech corpus. in Proc. LREC.

Khan, A. and M. Usman. (2019). Phonetic transcription for Pashto ASR. ACM Trans. Asian Low-Resour. Lang. Inf. Process.

Gul, K. *et al.* (2020). Preliminary continuous

Pashto ASR. in Proc. IEEE ISSPIT.

Rahman, T. and S. Ali. (2016). Speaker-dependent Pashto ASR. Int. J. Comput. Sci. Inf. Secur., vol. 14.

Amodei, D. *et al*. (2015). Deep Speech: End-to-end ASR. arXiv preprint arXiv:1512.02595,.

Parthasarathy, S. and C.H. Coker. (1992). Automatic estimation of articulatory parameters. Comput. Speech Lang., vol. 6, pp. 37–75,

Allen, J.M.S. and D Hunnicutt (1987). Klatt, From Text to Speech: The MITalk System. Cambridge, MA, USA: MIT Press,

Hunt, A.J. and Black A.W. (1998). Unit selection in a concatenative speech synthesis system using a large speech database. ATR Interpreting Telecommunications Research Labs., Kyoto Japan.

History of Pashto. PashtoWeb. [Online]. Available: http://www.pashtoweb.com/HistoryPashtoE.htm

Tegey, H. and B. Robson. (1996). A Reference Grammar of Pashto. Washington, DC, USA: Center for Applied Linguistics.

Titze, I.R. and D.W. Martin. (1998). Principles of voice production. J. Acoust. Soc. Amer., vol. 104, p. 1148.

Zue, V., S. Seneff and J. Glass. (1989). Speech database development: TIMIT and beyond. in Speech Input/Output Assessment and Speech Databases.

X Trans Tool. Philadelphia, PA. USA: Linguistic Data Consortium. [Online]. Available: https://www.ldc.upenn.edu/language-resources/tools/xtrans

Naeem, H.U. (2006). New Puxto Primer (The 21st Century Updated & Augmented Puxto Phonetic Alphabet).

Olson, R. (2008). Speaking Afghan Pashto, Interlit Foundation.