Bridging the Bandwidth Gap: A Mixed Band Telephonic Urdu ASR Approach with Domain Adaptation for Banking Applications

Ayesha Khalid[1*] **Farah Adeeba**[2*] **Najm Ul Sehar**[1] **Sarmad Hussain**[1]
[1]Center for Language Engineering, Al-Khawarizmi Institute of Computer Science,
University of Engineering and Technology, Lahore
[2]Department of Computer Science,
University of Engineering and Technology, New Campus
ayesha.khalid@kics.edu.pk, farah.adeeba@kics.edu.pk
najm.sehar@kics.edu.pk, sarmad.hussain@kics.edu.pk

Abstract

The accuracy of Automatic Speech Recognition (ASR) systems is influenced by the quality and context of speech signals, particularly in telephonic environments prone to errors like channel drops and noise, leading to higher Word Error Rates (WER). This paper presents the development of a large vocabulary Urdu ASR system for telephonic speech, based on a corpus of 445 speakers from diverse domains. The corpus, annotated at the sentence level, is used to train and evaluate GMM-HMM and chain Time-Delay Neural Network (TDNN) models on a 10-hour test set. Results show that the TDNN model outperforms GMM-HMM. Mixing narrowband and wideband speech further reduces WER. The test sets are also evaluated for the pre-trained model Whisper for performance comparison. Additionally, system adaptation for the banking domain with a specialized lexicon and language model demonstrates the system's potential for domain-specific applications.

1 Introduction

Human speech is a primary mode of communication, making speech recognition a vital area of research. While Automatic Speech Recognition (ASR) systems have made significant strides, approaching human-level performance in controlled environments, telephonic speech (narrowband, NB) remains a persistent challenge compared to studio-quality (wideband, WB) speech. The limitations in bandwidth, coupled with noise and distortion in telephony, degrade ASR performance. This issue is especially pronounced for resource-limited languages like Urdu, where training separate models for NB and WB speech can be difficult due to the scarcity of large-scale corpora.

This study aims to improve NB speech recognition by developing a Mixed-Band (MB) acoustic model using Deep Neural Networks (DNNs), which combines both NB and WB data. Two primary strategies for building MB acoustic models are Bandwidth Extension (BWE) and direct data mixing. While BWE has been widely explored to enhance speech quality and intelligibility (Prasad and Kumar, 2016; Nagel and Disch, 2009; Pulakka and Alku, 2011; Liu et al., 2009), its impact on improving recognition accuracy for mixed-band speech remains limited. In this work, directly mixing WB data with NB data to improve ASR performance for Urdu NB telephonic speech is proposed.

Developing robust ASR systems requires large corpora for both acoustic and language model training. Extensive corpora exist for languages such as English (Godfrey and Holliman, 1997; Post et al., 2013), Mandarin (Liu et al., 2006; Deng et al., 2019), Korean (Bang et al., 2020), and Polish (Ziółko et al., 2018). For instance, the Switchboard corpus (Godfrey and Holliman, 1997) provides 260 hours of conversational English, while the HKUST Mandarin Telephone Speech Corpus (Liu et al., 2006) offers 200 hours of telephonic data covering diverse dialects. A 969-hour corpus was developed for spontaneous Korean speech (Bang et al., 2020), and a 64-hour corpus for Polish telephony (Ziółko et al., 2018) spans domains like street and commands.

However, Urdu, spoken by 70 million people as a first language, remains under-resourced in terms of speech corpora for natural language processing. Notable efforts include the development of a 44.5-hour Urdu ASR system (Sarfraz et al., 2010), which combines microphone and telephone speech from 80 speakers, and domain-specific ASR systems for recognizing district names (Qasim et al., 2016; Rauf et al., 2015), achieving lab and field accuracies of 95.6% and 87.21%, respectively. Additionally, a Punjabi-accented banking-domain telephonic corpus (Mumtaz et al., 2018) with 400 speakers was recorded at 8 kHz and manually annotated.

Farooq et al. (Farooq et al., 2019) developed a large vocabulary continuous speech recognition (LVCSR) system for Urdu, collecting 300 hours of training data from 1,648 speakers and achieving a word error rate (WER) of 13.50% on 9.5 hours of testing data. Despite these efforts, Urdu still lacks a large-scale telephonic speech corpus, particularly for domain-specific applications like banking.

In this study, this gap is addressed by developing a telephonic Urdu speech corpus tailored to the banking domain, specifically for debit card activation. To accelerate corpus development, semi-automatic methods were used, combining microphone-recorded WB speech with telephonic NB data to build acoustic models. These models were then tested on telephonic speech to evaluate performance improvements in real-world applications.

2 Related Work

Automatic Speech Recognition (ASR) for telephonic data in specialized domains, such as banking, presents unique challenges due to varied acoustic environments and domain-specific vocabulary. Key strategies to enhance ASR performance in these scenarios include bandwith mismatch and acoustic modelling and domain adaptation.

2.1 Bandwidth Mismatch and Acoustic Modeling

A significant challenge in telephonic ASR is the mismatch between narrowband (NB) and wideband (WB) speech. Early work by Seltzer and Acero (Seltzer and Acero, 2006) introduced a GMM-HMM-based EM algorithm for mixed-band (MB) acoustic modeling, though improvements in recognition accuracy were limited. Later approaches, such as You and Xu (You and Xu, 2014), leveraged deep neural networks (DNNs) to reduce bandwidth mismatches, showing more promise. Mac et al. (Mac et al., 2019) demonstrated that upsampling NB data yields better results than downsampling WB data in DNN-based acoustic models. Recent advancements using generative adversarial networks (GANs) (Shi, 2023) have shown a 3.65% improvement in accuracy when recognizing mixed-band audio, highlighting the potential of GANs in addressing bandwidth mismatches.

2.2 Domain Adaptation for Telephonic Speech

Domain adaptation plays a crucial role in improving telephonic ASR, especially in specialized domains like banking. The Density Ratio Approach (DRA) (Takagi et al., 2023) adapts language models using large-scale text corpora to enhance recognition in live ASR. Additionally, External Off-Policy Acoustic Catalogs (Chan et al., 2023) have been used to reduce WER by leveraging external audio embeddings, speeding up domain adaptation. Data Distribution Matching (Shinohara and Watanabe, 2023) optimizes training by selecting subsets of training data that closely match the target domain, ensuring minimal WER increases without enlarging the model.

Building on these advancements, Ahmad et al. (Ahmad et al., 2024) proposed a progressive approach that adapts to out-of-domain telephonic speech by sequentially training student models, each using the previous student model as a teacher. This multi-stage training significantly reduced WER in unsupervised adaptation scenarios. In their experiments on Switchboard data, they achieved a 9.8%, 7.7%, and 3.3% absolute WER reduction after each stage of training, underscoring UDA's value in telephonic ASR with limited labeled data.

While significant progress has been made in adapting ASR systems to telephonic speech, challenges remain, particularly in maintaining high performance across diverse telephonic applications and scenarios, such as in the banking domain. Continuing advancements in domain adaptation techniques will be key to overcoming these hurdles.

3 Corpus Design

To develop a robust ASR system for a specific domain and language, a comprehensive text corpus covering essential vocabulary is required. This corpus records speech data for training and testing the ASR system, with the training corpus building the acoustic model and the testing corpus evaluating performance. Linguists systematically developed the text corpus by crawling data from websites using HTTrack[1] and sentences were extracted from HTML files using a Python script, manually rephrasing and supplementing sentences that averaged 14 words, with no repetitions. A linguist verified 10% of the text for syntactic and grammatical accuracy. The following subsections

[1] https://www.httrack.com/

Domain Name	No. of Sentences
Banking	1900
Business	10900
E-commerce	3000
Telecommunication	2000
News	18062
Miscellaneous	417
Total	**36279**

Table 1: Composition of Training Corpus

explain domain coverage in the corpora.

3.1 Training Corpus Design

A rich text corpus was carefully designed to collect Urdu speech data across various domains. The training corpus consists of 36,279 unique sentences, encompassing banking, business, e-commerce, telecommunications, online news, and general categories. Initially, 6,000 sentences were sourced from the phonetically balanced Urdu text-to-speech corpus.

For the banking domain, 1,000 sentences were crafted from brochures on loans, insurance, and banking services, supplemented by translations from English where necessary. An additional 900 sentences were rephrased from online bank websites, covering banking cards, Islamic banking, and payment partners.

In the business domain, 10,900 sentences focused on vocabulary related to cryptocurrency, taxation, and Pakistan's economy. For e-commerce, 3,000 sentences were collected from websites, primarily from pictorial information and menu cards.As shown in Appendix Table 13, the categories covered include a wide variety of products, from electronic devices to groceries.

For telecommunications, 2,000 sentences were generated from service provider websites, addressing apps, bundles, and services, with translations from English rules and regulations into Urdu. Additionally, 18,062 sentences were crawled and manually rephrased from news websites, alongside 417 sentences covering miscellaneous categories. The information on these categories is shown in Appendix Table 14. Table 1 outlines the corpus composition by domain.

3.2 Testing Corpus Design

To evaluate ASR system performance, a phonetically rich and balanced test corpus was developed by expert linguists, covering domains such as business, telecommunications, e-commerce, addresses, dates, and CNIC numbers. The test corpus was designed to be entirely distinct from the training data.

Domain Name	No. of Sentences
Newspaper	1300
Telecommunication	2000
E-commerce	2000
Addresses	1000
Dates	500
Telephone Numbers	250
CNIC Numbers	250
Urdu Digest(Mumtaz et al., 2018)	500
Total	**7800**

Table 2: Composition of Testing Corpus

For the business domain, 1,300 unique sentences were rephrased from online newspapers. The telecommunications and e-commerce domains each contributed 2,000 sentences from various websites. In the addresses domain, 1,000 unique postal addresses were manually rephrased, modifying elements like street and house numbers, along with different address formats used in Pakistan.

To cover dates, 500 sentences were created using carrier sentences, incorporating dates in various formats. These formats are shown in Appendix Table 15 and the carrier sentences are shown in Appendix Table 2.

For telephone number coverage, 250 original numbers were collected, and to ensure privacy, the 11 digits were shuffled multiple times to generate new numbers. These were then incorporated into carrier sentences for data recording.

Similarly, for CNIC numbers, 250 original CNICs were collected, and their 13 digits were shuffled to create new numbers. These were used in carrier sentences, and speakers recorded them for testing .

Additionally, 500 sentences from Urdu Digest 1M corpus(Mumtaz et al., 2018) were included in the testing corpus.

Table 2 presents the composition of the testing corpus and the proportion of each domain contributing to it. In order to cover corpus related to dates carrier sentences were used.

4 Speech Corpus Collection

To collect the speech data, designed text sentences were recorded from multiple speakers. Details of speaker selection, recording process and data alignment are given in subsequent sections.

4.1 Speaker Selection

To record the above-mentioned designed text corpus, the first step was the speakers scrutiny. Both genders (males and females) ranging between 18-50 years of age were selected. The speakers minimum education must be intermediate (Grade 12). Gender balance was maintained and speakers whose mother tongue was Urdu or Punjabi were selected. For a speaker to qualify for the actual recording, he/she had to read 20 sentences correctly. The speakers who can speak Urdu fluently and have no speech disfluency or pronunciation issues were selected for the recording session. Speaker uniqueness was ensured by maintaining a list of CNIC numbers of all speakers who participated in the recordings. Moreover, speakers date of birth, mother tongue, district, telephone number, mobile phone network, and mobile phone brand was also documented.

4.2 Recording Process

The speech corpus was recorded in an indoor environment using a variety of devices, including laptop microphones, USB headsets, and mobile phones (both default microphones and hands-free). Each speaker was asked to use their mobile phone to ensure diverse coverage across brands like Samsung, Nokia, HTC, and Huawei. The recording was conducted in a continuous read speech manner.

4.2.1 Training Corpus Recording

Simultaneous recordings were made on a laptop and a mobile phone, guided by a resource person who assigned each speaker a unique ID and list number. The text corpus was segmented into lists of 20 sentences. For wideband recordings, speakers utilized laptop microphones, USB headsets, or hands-free devices, with speaker details documented in advance. The recording process employed an automated utility that displayed sentences one at a time, allowing speakers to rehearse before recording. The completed recordings were saved as WAV files at a 16 kHz frequency, accompanied by a text file containing sentence IDs and timestamps. Post-recording, the data was sent to the Urdu ASR system for decoding, generating a status file detailing word errors. The goal for each speaker was to achieve at least 150 correctly decoded sentences, with an average of 350 sentences read in 2.5 to 3 hours.

4.2.2 Testing Corpus Recording

The testing procedure mirrored that of the training corpus. Speakers recorded lists of 20 unique sentences using the same microphone utility. Each recorded list was automatically saved in WAV format at 16 kHz, along with a text file documenting the start and end timestamps for each sentence. No ASR was used during testing, so the status file contained only timing information. Each speaker aimed to record 120 unique sentences and was required to re-record any sentences not captured correctly, typically completing the task in about 1 hour. Figure 1 depicts the recording process.

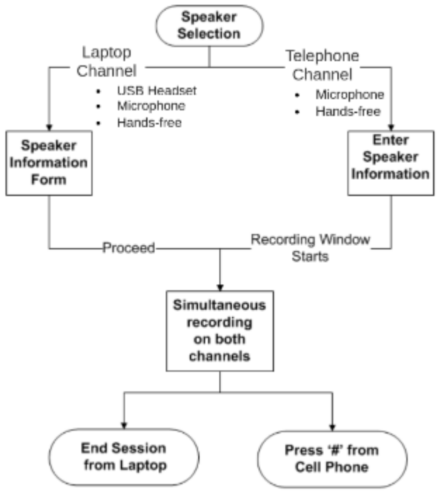


Figure 1: Recording Process

4.3 Data Alignment and Verification

This section discusses the alignment and verification of training and testing data.

4.3.1 Data Alignment and Segmentation

To segment telephonic audio into sentences, laptop recordings were used as references due to distortions from signal drops or weather issues. Manual alignment was necessary despite simultaneous recording, using Audacity[2], an open-source editor. The aligned telephonic audio file is then exported to the system. A Python script segmented the training audios into individual sentences, organizing them into folders for each speaker, with subfolders for laptop and telephonic data, further categorized by ASR-decoded results (0WE, 1WE, 2WE, and residual). Corresponding transcriptions were stored in text files. For testing, a similar script created speaker-specific folders for both channels, containing segmented audio files.

[2] https://www.audacityteam.org/download/

4.3.2 Data Verification

Data verification is crucial for preparing audio data for ASR development. While laptop-recorded data, decoded by the backend ASR, required no manual verification, telephonic recordings were carefully reviewed due to potential issues like signal drop or attenuation. Expert linguists manually verified 10% of the telephonic audios that matched the 0WE category from the laptop channels decoding results, excluding significantly distorted files from ASR training. All audio files from both channels in the testing data were fully verified since no backend ASR decoding was applied. Deliberate modifications (such as inserting or removing Urdu prepositions and conjunctions) were introduced in the text transcriptions for accuracy checks, with any errors leading to repeat verifications by the linguist team.

4.3.3 Corpus Statistics

The presented corpus comprises Urdu read-speech collected through two recording channels: laptop and telephone. Details on the number of speakers and duration are in Table 3, while mother tongue information is in Table 4.

Speakers from various districts of Pakistan, including Lahore, Faisalabad, Gujranwala, and others, are represented, with age coverage ranging from 18 to 50 years. Age distribution in the training and testing data is detailed in Table 5.

Recording on telephone channels utilized mobile phones from brands such as Samsung, Huawei, Nokia, and iPhone, covering networks like Ufone, Jazz, Warid, Zong, and Telenor. The percentage coverage for each mobile network is in Table 6. Additional statistics on the cleaned and verified corpus are presented in Table 7. In addition, Appendix Table 16 clearly compares this paper's corpus and the other existing telephonic Urdu corpora.

5 Experimental Setup

The speech recognition system for the Urdu language is developed using Kaldi[3] , which is an open-source speech recognition toolkit written in C++ and licensed under the Apache License V2.0. It provides adaptable code which can be modified and extended as per requirement. The important features of Kaldi include code-level integration with Finite State Transducers (FSTs) which compiles against the OpenFst toolkit (using it as a library). It also provides extensive linear algebra support through a matrix library that wraps standard BLAS and LAPACK routines (Praveen Kumar et al., 2020) . The toolkit was set up on Ubuntu 18.04 LTS operating system. The machine specifications include an octa-core 2.8 GHz processor of Intel(R) Core (TM) i7 with 32 GB RAM and hosting two NVIDIA graphic cards GeForce GTX 1080[4] having 8 GB memory each. This setup supports the successful implementation of state-of-the-art recipes provided by the Kaldi toolkit for the development of a speech recognition system.

5.1 Feature Extraction

The raw speech signal is complex and unsuitable for direct input into a speech recognition system. To facilitate training, feature extraction is employed to represent the speech signal parametrically. Kaldi offers various scripts for feature extraction, including Mel Frequency Cepstral Coefficients (MFCC), Perceptron Linear Predictive (PLP), and Vocal Tract Length Normalization (VTLN).

For speech recognition system, MFCC extraction was used via the Kaldi recipe, processing 25ms frames with a 10ms shift. After removing the DC offset, the signal is multiplied by a Hamming window, followed by energy calculation in mel-bins, Fast Fourier Transform (FFT), and power spectrum computation. Finally, a cosine transform is applied to the logarithm of the energies. Although the default number of cepstral coefficients is 13, it was configured to 40 to enhance deep neural networks. After calculating MFCC features, Cepstral Mean and Variance Normalization (CMVN) was applied for improved robustness in speech recognition.

5.2 Phonetic Lexicon

The lexicon is essential for speech recognition, defining word pronunciations. For Urdu ASR, a 34K-word phonetic lexicon was created, covering all unique words from the text corpus. Transcriptions follow the Case Insensitive Speech Assessment Method Phonetic Alphabet (CISAMPA)[5] symbols. Linguists mapped each word to standard orthography using dictionaries like Oxford Urdu Dictionary[6] and Urdu Lughat[7]. Alternate pronunci-

[3] https://kaldi-asr.org/

[4] https://www.nvidia.com/en-sg/geforce/products/10series/geforce-gtx-1080/

[5] https://www.cle.org.pk/Downloads/ling_resources/phoneticinventory/UrduPhoneticInventory.pdf

[6] https://languages.oup.com/oxford-global-languages/

[7] http://udb.gov.pk

Data Type	Telephone Data Duration (hours)	No. of Male Speakers	No. of Female Speakers
Training Data	111.5 h	240	205
Testing Data	10.93 h	30	30

Table 3: Number of Speakers and Duration of Speech Data

Mother Tongue	Number of Speakers (Training)	Percentage (Training)	Number of Speakers (Testing)	Percentage (Testing)
Urdu	213	47.86%	28	46.67%
Punjabi	232	52.14%	32	53.33%

Table 4: Speaker Coverage Based on Mother Tongue

Range of Speaker Ages	No. of Speakers (Training)	Percentage (Training)	No. of Speakers (Testing)	Percentage (Testing)
< 20	178	40.00%	6	10.00%
20 24	240	53.93%	27	45.00%
25 29	25	5.62%	10	16.67%
> 30	2	0.45%	17	28.33%

Table 5: Age Distribution of Speakers

Mobile Network	Training Data	Testing Data
Ufone	26.69%	24.24%
Jazz	45.76%	40.91%
Warid	10.80%	16.67%
Zong	10.80%	6.06%
Telenor	5.95%	12.12%

Table 6: Percentage Coverage of Mobile Networks

ations exist for some words, as shown in Appendix Table 17.

5.3 Language Model

SRI Language Modeling (SRILM) toolkit (Praveen Kumar et al., 2020) is used to build a trigram language model. A very large amount of data has been crawled from various Urdu websites including books, magazines, news, and blogs. The corpus is then tokenized on the basis of sentences. Moreover, the designed text corpus for training is also appended after replicating it hundreds of times. The final corpus contains 120 million words which were used for the development of the language model for Urdu ASR.

5.4 Acoustic Modeling

The baseline ASR system was built using a Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) approach. Monophone training and alignment were followed by speaker-independent Linear Discriminant Analysis (LDA) for tri-phone training (tri1) with 2000 decision tree leaves and 11,000 Gaussians. In tri2, Maximum Likelihood Linear Transform (MLLT) increased these to 2500 leaves and 15,000 Gaussians. Speaker Adapted Training (SAT) retained this configuration.

Using SAT alignments, Deep Neural Networks (DNN), specifically Chain TDNN models, were trained with Lattice-Free Maximum Mutual Information (LFMMI) to improve efficiency and transcription quality. High-resolution MFCCs and 100-dimensional i-vectors were computed with 1024 Gaussians. The TDNN had one convolutional layer, seven hidden layers with 625 neurons, ReLU-batch norm activation, and a learning rate that decreased from 0.001 to 0.0001.

5.5 Mixing of WB Speech Data in NB Data

Increasing training data enhances speech recognition performance, as shown in the literature (Seltzer and Acero, 2006; Mac et al., 2019). Mixed-bandwidth training improves telephonic speech recognition. Along with the developed corpus, 174 hours of WB data from (Farooq et al., 2019) (hands-free recorded, 16 kHz) were added. The 111.5 hours of telephonic data were upsampled from 8 kHz to 16 kHz, yielding 285.5 hours of audio at 16 kHz. Results, before and after adding WB data, are detailed in Section 6.

5.6 Testing Setup and Performance Measure

For evaluation of the system, 10.9 hours of telephonic speech data is used to test the primary system which covers multiple domains. Percentage word error rate (WER) is used as the performance measure. It is equal to the number of errors made by ASR divided by the total words in the test data multiplied by 100.

	Training Data	**Testing Data**
Number of Utterances	65,227	6,358
Average Duration per Utterance (in seconds)	6.2	6.2
Average Number of Words per Utterance	13	15
Average Number of Utterances per Speaker	164	106
Total Frequency of Words in Transcription Files	823,819	92,659

Table 7: Corpus Details

6 Performance Evaluation Experiments and their Results

Two models, a GMM-HMM model, and a chain TDNN, are evaluated using the test data. The GMM-HMM model is trained on telephonic training data while Chain TDNN is trained on two datasets, telephonic training data and telephonic+hands-free data from (Farooq et al., 2019). The experiments showed that Chain TDNN outperformed the GMM-HMM model with 24% less WER. The third experiment i.e., the training of Chain TDNN using telephonic (NB) and hands-free (WB), further reduced WER by 3%. However, pre-trained Whispers Large variant performed similar to GMM-HMM with only 4% reduction in WER. Table 8 presents the experimental results carried out.

7 Adaptation of ASR for Banking Domain

In order to tailor the system for the banking domain, language data and lexicon was adapted for a speech-based debit card activation dialog system. This included a language model for debit card numbers (DCN), last four digits (DCLFD), date of birth (DoB), and debit card expiry dates (DCED). The system records client responses via telephone, and ASR decodes them; if the information is correct, the card activates; otherwise, after one retry, the call is transferred to a human operator.

Banking domain corpora from (Mumtaz et al., 2018) was utilized for adaptation, which comprised 10 million entries for DCN, 300K for DCLFD, 815K for DCED, and 2.6 million for DoB. A combined lexicon and corpus were created by merging these categories, with added confirmation words like "sahi" (Correct), "ghalat" (Incorrect), "han" (Yes), and "nahi" (No). Table 9 outlines the corpus and lexicon. The system was tested on individual and combined datasets, comparing System-3 with the Whisper pre-trained model. Although Whisper underperformed due to limited vocabulary, domain-specific models mentioned in this paper achieved better results.

Experiments with the combined language model and lexicon were performed in four configurations: (1) general corpus and lexicon, (2) banking lexicon only, (3) banking corpus only, and (4) both banking lexicon and corpus. Table 10 shows that domain-specific lexicons and corpora significantly improved results, while Table 11 presents the %WER from System 2 and System-3 for each experiment.

Table 12 analyzes phone recognition errors in the banking dataset while Appendix Table 18 shows percentage of Urdu Phonemes in testing and training data. Consonants like /m, l, j/ were frequently misrecognized, while medial vowels such as /[æ]/ and /[e]/, the diphthong /[eɐː]/, and the nasal vowel /[æ]/ faced significant misrecognition challenges due to limited test data coverage. Notably, the dental plosive /[t̰]/ was particularly challenging to identify despite good dataset coverage, while the vowel /[ə]/ was often misrecognized as /[ɑː]/. In conclusion, domain-specific models outperform general pre-trained models like Whisper, emphasizing the need for targeted improvements in phonetic recognition for banking applications.

8 Conclusion

In this paper, a narrowband speech recognition system for Urdu is developed using a mix of wideband speech data. A multi-domain narrowband and wideband speech corpus is recorded from multiple speakers. Based on this data, a large vocabulary continuous speech recognition system for telephonic channels is developed. Various acoustic modeling techniques are investigated, with Chain TDNN outperforming other models. The Chain TDNN, trained on a mixture of narrowband and wideband corpora, outperforms other models and the pre-trained Whisper model. The developed ASR is adapted for the banking domain using a domain-specific lexicon and language model, achieving promising results.

System	Model	Training Data Channel	Duration (hrs)	%WER
System-1	GMM	Telephonic	111.5	54.92
System-2	Chain TDNN	Telephonic	111.5	30.36
System-3	Chain TDNN	Telephonic+Hands-free	285.5	27.56
Whisper (Radford et al., 2023)	Variant-Large	Pre-trained Model	680,000	50

Table 8: Word Error Rate on Different Models Trained on Different Training Data

Category	Corpus	Lexicon
DCN	10,000,000 entries [16- digit numbers]	29 entries [0-9 digits]
DCLFD	300,000 entries	157 entries [0-100 digits]
DCED	815,000 entries (13 different patterns)	201 entries [0-100 digits + Month Names]
DoB	2,679,107 entries (13 different patterns)	201 entries [0-100 digits + Month Names]
Combined	17,679,107	209 entries [digits + month names + confirmation words]

Table 9: Banking Text Corpus and Lexicon Details

Category	No. of Utterances	Duration (Minutes)	WER on Combined LM and Lexicon	WER on Respective LM and Lexicon	WER using Whisper-Large	Phone Error Rate
DCN	83	17	1.72	1.65	19.0	1.83
DCLFD	298	21	5.30	5.43	80.0	4.37
DCED	792	54	3.44	3.40	51.0	4.16
DoB	340	29	3.42	3.32	49.0	3.32
Combined	1513	121	3.87	3.45 (Avg)	49.75	3.42 (Avg)

Table 10: Detailed Test Results

Lexicon and LM used	% WER on System-2	% WER on System-3
General lexicon + general corpus	86.08	91.39
Banking lexicon + general corpus	8.08	7.56
General lexicon + banking corpus	3.14	2.84
Banking lexicon + banking corpus	2.47	1.65

Table 11: Performance of System-2 and System-3 on Banking Domain Data

Sr. no	Urdu Consonants (IPA)	% Phone Error Rate	Sr. no	Urdu Vowels (IPA)	% Phone Error Rate
1	م (m)	10.0	1	اَے (æ)	21.9
2	ل (l)	14.1	2	ا،آ (ea:)	8.2
3	ی (j)	11.9	3	هِ (e)	7.7
4	ت،ط (t̪)	8.2	4	یَں (ɑẽː)	7.1
5	ٹھ (t^{h})	4.6	5	ی (i:)	5.9
6	س،ص،ث (s)	4.4	6	یں (ẽː)	4.8
7	ک (k)	3.5	7	ء (ə)	4.5
8	کھ (k^{h})	3.0	8	آ (ɑː)	2.7
9	پ (p)	2.7	9	و (o:)	2.2
10	ب (b)	2.3	10	ُ (ʊ)	2.0
11	ح،ه (h)	2.3	11	آئی (aːi:)	1.6
12	ن (n)	2.2	12	ِ (ɪ)	1.4
13	چ (tʃ)	2.0	13	آں، اں (ɑ̃ː)	1.4
14	ر (r)	1.8	14	ے (eː)	0.8
15	د (d)	1.8			
16	چھ (tʃ^{h})	1.7			
17	و (v)	1.1			
18	ذ،ض،ظ،ز (z)	0.7			

Table 12: Phone Error Rate on Banking Domain Data

9 Limitations

The development of the mixed-band (MB) acoustic model and Urdu speech recognition system acknowledges several limitations. Although Urdu telephonic speech corpus presented in this paper represents an improvement over prior datasets, it may not fully capture the variety of Urdu accents and dialects, and its domain specificity to the banking sector constrains generalizability to other applications. The speaker demographic, while balanced in gender and comprising native Urdu and Punjabi speakers, is restricted to individuals aged 18-50 with at least an intermediate education level, excluding younger and older populations as well as those with lower educational backgrounds, thereby limiting the model's applicability across a broader audience. Additionally, the recordings were conducted in controlled indoor environments, which may not accurately represent real-world acoustic variability. Lastly, the ASR system is specifically designed for debit card activation, and its effectiveness in other domains remains to be evaluated.

Ethical Considerations

All researchers involved in this study have adhered to the ACM Code of Ethics and conducted their work responsibly. Our aim is to advance speech recognition for Urdu, enhancing technology access in underserved communities. Data collection was conducted with informed consent, primarily utilizing publicly available or simulated telephonic interactions, with all personally identifiable information (PII) anonymized during transcription and annotation. Telephone numbers and 13-digit CNIC numbers were anonymized by shuffling digits to ensure privacy. Our initiative seeks to enhance technological resources for the broader Urdu-speaking community without exploiting individuals based on language, geography, or social standing. We recognize the potential risks associated with the misuse of ASR technology, including surveillance or unauthorized data collection. However, our research strictly focuses on developing models for the benefit of users, not for intrusive purposes. The ethical implications of using such systems in various applications must be carefully considered, and we urge continued dialogue on the responsible deployment of speech recognition technologies to ensure they serve the needs of diverse Urdu-speaking populations without exploitation or marginalization.

References

Rehan Ahmad, Muhammad Umar Farooq, and Thomas Hain. 2024. Progressive unsupervised domain adaptation for asr using ensemble models and multi-stage training. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11466–11470.

Jeong-Uk Bang, Seung Yun, Seung-Hi Kim, Mu-Yeol Choi, Min-Kyu Lee, Yeo-Jeong Kim, Dong-Hyun Kim, Jun Park, Young-Jik Lee, and Sang-Hun Kim. 2020. Ksponspeech: Korean spontaneous speech corpus for automatic speech recognition. *Applied Sciences*, 10(19):6936.

David M Chan, Shalini Ghosh, Ariya Rastrow, and Björn Hoffmeister. 2023. Domain adaptation with external off-policy acoustic catalogs for scalable contextual end-to-end automated speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yu-Chih Deng, Yih-Ru Wang, Sin-Horng Chen, and Chen-Yu Chiang. 2019. Recent progress of mandrain spontaneous speech recognition on mandrain conversation dialogue corpus. In *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Muhammad Umar Farooq, Farah Adeeba, Sahar Rauf, and Sarmad Hussain. 2019. Improving large vocabulary urdu speech recognition system using deep neural networks. In *INTERSPEECH*, pages 2978–2982.

Jianqing Gao, Jun Du, and Enhong Chen. 2018. Mixed-bandwidth cross-channel speech recognition via joint optimization of dnn-based bandwidth expansion and acoustic modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(3):559–571.

Jianqing Gao, Jun Du, Changqing Kong, Huaifang Lu, Enhong Chen, and Chin-Hui Lee. 2016. An experimental study on joint modeling of mixed-bandwidth data via deep neural networks for robust speech recognition. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 588–594. IEEE.

Alexandru-Lucian Georgescu, Horia Cucu, Andi Buzo, and Corneliu Burileanu. 2020. Rsc: A romanian read speech corpus for automatic speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6606–6612.

John J Godfrey and Edward Holliman. 1997. Switchboard-1 release 2. *Linguistic Data Consortium, Philadelphia*, 926:927.

Chuping Liu, Qian-Jie Fu, and Shrikanth S Narayanan. 2009. Effect of bandwidth extension to telephone speech recognition in cochlear implant users. *The Journal of the Acoustical Society of America*, 125(2):EL77–EL83.

Yi Liu, Pascale Fung, Yongsheng Yang, Christopher Cieri, Shudong Huang, and David Graff. 2006. Hkust/mts: A very large scale mandarin telephone speech corpus. In *Chinese Spoken Language Processing: 5th International Symposium, ISCSLP 2006, Singapore, December 13-16, 2006. Proceedings*, pages 724–735. Springer.

Khoi-Nguyen C Mac, Xiaodong Cui, Wei Zhang, and Michael Picheny. 2019. Large-scale mixed-bandwidth deep neural network acoustic modeling for automatic speech recognition. *arXiv preprint arXiv:1907.04887*.

Long Mai and Julie Carson-Berndsen. 2022. Unsupervised domain adaptation for speech recognition with unsupervised error correction. *arXiv preprint arXiv:2209.12043*.

Benazir Mumtaz, Sahar Rauf, Hafsa Qadir, Javairia Khalid, Tania Habib, Sarmad Hussain, Rukhsana Barkat, et al. 2018. Urdu speech corpora for banking sector in pakistan. In *2018 Oriental COCOSDA-International Conference on Speech Database and Assessments*, pages 9–14. IEEE.

Frederik Nagel and Sascha Disch. 2009. A harmonic bandwidth extension method for audio codecs. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 145–148. IEEE.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved speech-to-text translation with the fisher and callhome spanish-english speech translation corpus. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*.

N Prasad and T Kishore Kumar. 2016. Bandwidth extension of speech signals: A comprehensive review. *International Journal of Intelligent Systems and Applications*, 8(2):45–52.

PS Praveen Kumar, G Thimmaraja Yadava, and Haradagere Siddaramaiah Jayanna. 2020. Continuous kannada speech recognition system under degraded condition. *Circuits, Systems, and Signal Processing*, 39:391–419.

Hannu Pulakka and Paavo Alku. 2011. Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2170–2183.

Muhammad Qasim, Sohaib Nawaz, Sarmad Hussain, and Tania Habib. 2016. 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)*, pages 28–32. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Sahar Rauf, Asima Hameed, Tania Habib, and Sarmad Hussain. 2015. District names speech corpus for pakistani languages. In *2015 International Conference Oriental COCOSDA Held Jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 207–211. IEEE.

Huda Sarfraz, Sarmad Hussain, Riffat Bokhari, Agha Ali Raza, Inam Ullah, Zahid Sarfraz, Sophia Pervez, Asad Mustafa, Iqra Javed, and Rahila Parveen. 2010. Speech corpus development for a speaker independent spontaneous urdu speech recognition system. *Proceedings of the O-COCOSDA, Kathmandu, Nepal*, 24.

Michael L Seltzer and Alex Acero. 2006. Training wideband acoustic models using mixed-bandwidth training data for speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):235–245.

Lijuan Shi. 2023. A unified mixed-bandwidth asr framework with generative adversarial network. In *Proceedings of the 2023 4th International Conference on Control, Robotics and Intelligent System*, pages 160–165.

Yusuke Shinohara and Shinji Watanabe. 2023. Domain adaptation by data distribution matching via submodularity for speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. IEEE.

Tatsunari Takagi, Yukoh Wakabayashi, Atsunori Ogawa, and Norihide Kitaoka. 2023. Domain adaptation using density ratio approach and ctc decoder for streaming speech recognition. In *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pages 1–5. IEEE.

Zhao You and Bo Xu. 2014. Improving wideband acoustic models using mixed-bandwidth training data via dnn adaptation. In *INTERSPEECH*, pages 2204–2208.

Bartosz Ziółko, Piotr Żelasko, Ireneusz Gawlik, Tomasz Pędzimąż, and Tomasz Jadczyk. 2018. An application for building a polish telephone speech corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

A Supplementary Material

This appendix contains additional tables and figures that support the findings in the main document.

Table 13 provides an overview of the categories prevalent in the e-commerce domain. This table captures various sectors, including electronics, home appliances, fashion, and groceries, reflecting the diversity of products and services available online. The categories are organized to facilitate understanding of the marketplace's structure and the various offerings within each segment.

E-Commerce Domain Categories
Electronic devices, gadgets, home appliances, Features of different appliances, home (bedding, blankets), Automobiles (bikes, spare parts, loaders), Electronic accessories, health and beauty, Babies and toys, groceries, pets, Home and lifestyle, food items, food deals, Ingredients, cities, cuisines, Food outlets, transport, organizations, Medical equipment, books/magazines, Miscellaneous products, jobs, Electric appliances, electronic gadgets, Features of products, features of OLX, Pets/livestock, fashion, transport, Musical instruments, medical instruments, Popular cities, less popular cities, Locations, property type, purpose, Cars (Japanese, Chinese etc.), car brands, Bikes and parts

Table 13: Categories Covered in E-Commerce

In Table 14, a comprehensive list of general domain categories is presented. This table encompasses a wide range of topics, from agriculture to technology, showcasing the breadth of subjects that can be explored. Each category represents significant fields of study or interest, underlining the multifaceted nature of the domains covered.

General Domain Categories
Plants, agriculture, sports, titles, education fields, country names, political parties, services and utilities, professions, election process, judiciary, administrative domains, trade, medical terms, diseases, entertainment films, cities, religions, army, nationalities, places, vegetables, flowers, crimes, relations, rivers, castes, weapons, grocery items, colors, festivals, events, body parts, birds, animals, metals, utensils, clothes, fruits, jewelry, Islam, dishes, furniture, literature-text types, musical instruments, journalism, weather, emotions, nuts, drugs, insects, time of day, direction, shapes, technology, banking, time measurement units, area measurement units, distance measurement units, weight measurement units, parts of house, transport, vehicles, situation/calamities/disasters

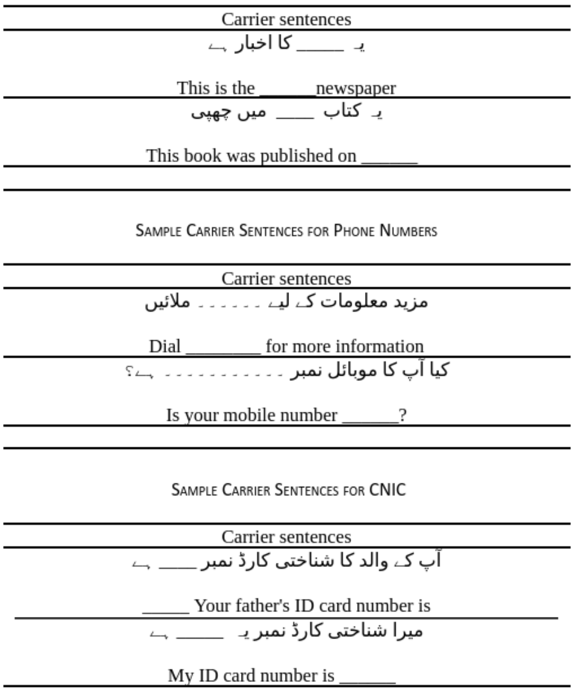
Table 14: Categories Covered in General Domain

Table 15 presents the various date formats incorporated into the testing corpus design for the ASR system. These formats were carefully selected to reflect the diverse ways dates can be expressed in natural speech, which is crucial for evaluating the model's performance in real-world scenarios.

Date Format	**Example**
DD/MM/YYYY	16/10/2018
DD-MM-YYYY	16-10-2018
Month (English) DD, Year	October 16, 2018
DD Month (Urdu) Year	15 2018
DD.MM.YYYY	17.07.1988
DD Month YYYY	19 SEP 2016
DDth Month, YYYY	16th February 2014

Table 15: Different Formats for Dates

Carrier sentences in Figure 2 help to evaluate the system's performance in handling both general language and numerical information in everyday conversations.



Carrier sentences
یہ _____ کا اخبار ہے This is the ______newspaper
یہ کتاب ____ میں چھپی This book was published on ______

SAMPLE CARRIER SENTENCES FOR PHONE NUMBERS

Carrier sentences
مزید معلومات کے لیے - - - - - - ملائیں Dial ________ for more information
کیا آپ کا موبائل نمبر - - - - - - - - - - - - ہے؟ Is your mobile number ______?

SAMPLE CARRIER SENTENCES FOR CNIC

Carrier sentences
آپ کے والد کا شناختی کارڈ نمبر _____ ہے _____ Your father's ID card number is
میرا شناختی کارڈ نمبر یہ _____ ہے My ID card number is ______

Figure 2: Sample Carrier Sentences for Dates,Phone Numbers and CNIC

Table 17 presents a selection of Urdu words that have multiple pronunciations, alongside their English translations and phonetic transcriptions. This table aims to illustrate the linguistic diversity and variability present within the Urdu language, particularly in terms of pronunciation.

Table 16 clearly presents that the corpus presented in this paper stands out for its extensive domain coverage, large number of speakers, and large amount of speech data emphasizing its uniqueness in breadth and scale.

Sr. no	Telephonic Urdu Corpora	Total No. of Speakers	Duration (hours)	Domains Covered	Speech Type
1	Speech Corpus Development for a Speaker Independent Spontaneous Urdu Speech Recognition System(Sarfraz et al., 2010)	82	44.5	Phonemically rich sentences, everyday text (date, numbers, proper names, hobbies, interests, past experiences, television, cricket)	Read and spontaneous speech
2	District Names Speech Corpus for Pakistani Languages(Rauf et al., 2015)	300	12	District names	Read speech
3	Urdu Speech Corpus for Travel Domain(Qasim et al., 2016)	60	–	City names, days, time, and numbers	Read speech
4	Urdu Automatic Speech Recognition for Telephonic Data: A Mixed Band Corpus Development Approach with Domain Adaptation for Banking Applications	445	111.5	Banking, e-commerce, investments, trading, telecommunication	Read speech

Table 16: Corpus Comparison with existing Urdu Corpora

A.1 Phonetic Coverage of Training and Testing Data

A detailed phonetic analysis of the training dataset has been conducted to ensure adequate coverage of all Urdu phonemes. Table 18 elaborates on the Urdu phonemes coverage in the training and banking domain test dataset. The training corpus comprises 56% consonants and 44% vowels, and the test dataset comprises 62% consonants and 38% vowels. /r/ consonant is the most frequent consonant in both training and testing datasets, and /ɑː/ vowel is the most frequent in both training and test data. Aspirated consonants /t^{h}/, /k^{h}/, etc., and nasalized vowels /ĩː/, /æ̃ː/, etc., are less frequent in both datasets. /t^{h}/ is exceptional as it has more frequency of occurrence in test data of the banking domain.

Words having alternate pronunciations	English Translation	Transcriptions
مصنف	Author	M U S A N N A F M U S A N N I F
اونچائی	Height	U U N T S AA II UUN TS AA II
صالح	Pious	S AA L AY S AA L AY H S AA L AYH H S AA L AYH H AY

Table 17: Examples of Words Having Alternate Pronunciations

Frequency of occurrence of Urdu phonemes in training data				Frequency of occurrence of Urdu phonemes in testing data			
Urdu Consonants (IPA)	**%**	**Urdu Vowels (IPA)**	**%**	**Urdu Consonants (IPA)**	**%**	**Urdu Vowels (IPA)**	**%**
ر (r)	6.0	ا، آ (ɑː)	9.4	ر (r)	8.1	ا، آ (ɑː)	8.9
ک (k)	5.6	َ، ء (ə)	7.6	ن (n)	8.1	َ، ء (ə)	8.5
ح،ہ (h)	4.6	ے (e:)	5.8	س،ص،ث (s)	7.3	ے (e:)	5.4
ن (n)	4.5	ی (i:)	5.0	ت،ط (t̪)	6.8	ی (i:)	5.0
س،ص،ث (s)	4.3	ِ (ɪ)	4.2	چ (tʃ)	4.8	و (o:)	3.5
م (m)	3.9	ُ (ʊ)	3.2	ک (k)	3.5	َ (ɪ)	3.2
ت،ط (t̪)	3.4	وُ (u:)	2.1	پ (p)	2.9	ں، آں (ɑ̃ː)	2.3
ل (l)	3.3	و (o:)	1.2	د (d)	2.8	وَ (ɔː)	2.2
ب (b)	2.2	یں (ẽː)	1.0	ٹھ (tʰ)	2.4	ُ (ʊ)	0.9
د (d)	2.1	وَ (ɔː)	0.9	ف (f)	2.2	(æ)	0.3
پ (p)	2.0	ےَ (æ:)	0.7	چھ (tʃʰ)	2.2	آئی (a:i:)	0.3
ی (j)	1.7	ہ (e)	0.5	ب (b)	1.6	یں (ẽː)	0.3
و (v)	1.3	یَں (ɑẽː)	0.4	و (v)	1.3	ا،آ (eaː)	0.3
ذ، ض، ظ، ز (z)	1.2	ُ (æ)	0.4	ح،ہ (h)	0.8	وُ (u:)	0.2
گ (g)	1.2	ہ (o)	0.4	ذ، ض، ظ، ز (z)	0.7	ہ (e)	0.1
ج (dʒ)	1.2	وں (õː)	0.3	ل (l)	0.6	ےَ (æ:)	0.1
ٹ (t)	1.1	اں، آں (ɑ̃ː)	0.3	ی (j)	0.4	یَں (ɑẽː)	0.1
ف (f)	1.0	وُں (ũː)	0.2	ٹ (t)	0.4		
ش (ʃ)	0.8	یں (ĩː)	0.1	م (m)	0.3		
چ (tʃ)	0.7	آءِ (aːɪ)	0.1	کھ (kh)	0.3		
ق (q)	0.6	ا،آ (eaː)	0.0	ج (dʒ)	0.3		
خ (x)	0.5	آئی (aːi:)	0.0	ڑ (r)	0.2		
ڈ (d)	0.5			گ (g)	0.2		
تھ (tʰ)	0.4			تھ (tʰ)	0.1		
چھ (tʃʰ)	0.3						
بھ (bʰ)	0.3						
کھ (kʰ)	0.3						
ڑ (r)	0.2						
ٹھ (tʰ)	0.2						

Table 18: Frequency of occurrence of Urdu phonemes in training and testing data