VFAST Transactions on Software Engineering
https://vfast.org/journals/index.php/VTSE@ 2026, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 14, Number 2, April-June 2026 pp:81-95

A DeepSpeech2-Inspired Convolutional Recurrent Framework for Low-Resource Urdu Speech Recognition

Syed Azeem Inam[1*], **Syeda Nazia Ashraf**[2], **Hassan Hashim**[1], **Syeda Wajiha Naim**[3], **Muhammad Ahmed Ansari**[4], **Ahmed Raza Khanzada**[1]

[1]Department of Artificial Intelligence and Mathematical Sciences, Sindh Madressatul Islam University, Karachi, Pakistan; [2]Department of Computer Science, Sindh Madressatul Islam University, Karachi, Pakistan; [3]Department of Software Engineering, Sindh Madressatul Islam University, Karachi, Pakistan; [4]Cloud Solutions for IT & Communication Co, Riyadh, Saudia Arabia

Keywords: *Urdu automatic speech recognition, Speech-to-text transcription, Low-resource language processing, DeepSpeech2-inspired CRNN, Connectionist Temporal Classification, Common Voice Urdu*

Journal Info:
Submitted:
April 7, 2026
Accepted:
April 18, 2026
Published:
April 25, 2026

Abstract Low-resource language automatic speech recognition is a difficult task due to small annotated corpora, large speaker and phonetic diversity, and the lack of strong end-to-end metrics. The case of Urdu is especially significant because of the high number of speakers and the inability to provide high-performing open automatic speech recognition systems to date. The study presents an end-to-end Urdu speech-to-text model, built upon a DeepSpeech2-inspired convolutional recurrent neural network, which integrates a spectrogram-based acoustic modeling, bidirectional gated recurrent units, and Connectionist Temporal Classification to learn alignment-free transcription. This model was trained and tested on the Urdu subset of the Mozilla Common Voice corpus with 58,119 training utterances and 6,458 validation utterances and evaluated on a held-out test set. The proposed system has shown to converge consistently during training with a validation Word Error Rate of 21.29% and loss of 5.87 at epoch 478, and a final test Word Error Rate of 17.05, Sentence Error Rate of 34.72, and Word Information Loss of 0.41. The proposed model achieved better performance on the same evaluation setting compared with a reduced recurrent baseline, a transformer-based baseline, and a wav2vec2-style baseline, whose WERs were 23.84%, 19.62%, and 18.31%, respectively. Analysis of ablation also indicated that convolutional feature extraction, as well as deep bidirectional temporal modeling, are essential to performance, and error analysis revealed phonetic confusion, dialectal variation, noise, and high-speed speech as the most prevalent causes of recognition error. These results demonstrate that a well-tuned convolutional recurrent model can provide a competitive solution for Urdu automatic speech recognition under low-resource conditions and offers a reproducible reference point for future studies.

***Correspondence author email address:** syed.azeem@smiu.edu.pk
DOI: 10.21015/vtse.v14i2.2392

1 Background

Researchers have shown a strong desire to create systems that mimic human communication abilities since the first Automatic Speech Recognition (ASR) system. These systems recognize and interpret language using algorithms to do speech analysis [1]. As of the time of this study, numerous ASR systems have been created for a wide range of languages, from regional to foreign,

with English being the language that has received the greatest research in terms of system implementation [2]. To achieve accurate recognition and speech representation, many ASR algorithms require training with massive amounts of annotated speech data [3]. However, concern arises for languages that do not have an excessive amount of annotated data and lack resources to train the ASR system, such as the Urdu language, which is spoken by more than 350 million people all over the globe [4]. It is the national language of Pakistan and is one of the most spoken languages in India [5]. Inclusive of this, the number of speakers of the Urdu language is continuously increasing in certain regions such as the USA, UK, and Canada, mainly due to the South Asian diaspora [6]. It is a challenging language for researchers of linguistics due to its inherent grammatical form and vocabulary adopted from Arabic, Persian, and the native languages of South Asia, and because of its Arabic-Persian script [6].

Hidden Markov Models (HMM), Support Vector Machines (SVM), and Decision Trees (DT) have been used by researchers in speech recognition for a long time; however, the researchers of the current era are utilizing the power of deep learning (DL) architectures to bolster the performance of the ASR systems [7]. DL has been implemented in a wide range of research studies inclusive of speech processing applications in the past decades mainly because of its nature of enhancing the capabilities of computers to mimic human behavior more effectively [8–10]. Although it has been in existence for about 60-70 years, its utilization in the ASR system was not until recently, when it revolutionized the development of the ASR system [11], as ASR systems have always been the reason for new and advanced development in machine learning (ML), but the promising advancements in the ASR system have still not aligned with the expectations and are not on the level of human performance [12]. The system involving speech-to-text translation mainly uses a hybrid system of either cascading or end-to-end models [13]. Although cascade systems excel in offline speech translation, they are less effective for real-time applications, where end-to-end models give better performance [14]. Whereas some researchers have utilized the power of Finite-State Automata (FSA)[15], some researchers incorporated Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) for extracting features for an ASR System [7].

In the present study, we have proposed a modified DL algorithm primarily based on the famous DeepSpeech2 algorithm to design an ASR system for Urdu speech-to-text. Currently, comparatively few studies have reported end-to-end ASR systems specifically designed for Urdu, which motivates the present work as a reproducible baseline for future investigation. The major contributions of this work are as follows:

- To create a DL architecture that is inspired by DeepSpeech2 and uses bi-directional GRU layers to extract contextual details from the audio stream.
- To demonstrate that the availability of pre-segmented data does not affect the effectiveness of our suggested DL model in producing desired outcomes.
- Using the suggested architecture in an ASR system for data-poor and resource-poor languages.
- Implement the proposed architecture for Urdu as a case study and evaluate its performance under a controlled and reproducible experimental setting.

The study consists of four sections. In Section 2, we have discussed the related work in the utilization of DL algorithms for designing an ASR system. Section 3 elaborates on the architecture of the developed model along with the procedure of its implementation. Furthermore, in Section 4, we have discussed the results of our study, and finally, in Section 5, we have presented the conclusion of the present study.

2 Literature Review

From voice activation systems to the education sector [16], the application of ASR systems is enormous, mainly due to the advancements in machine translation [17–21] and sequence-to-sequence models [22, 23] that have contributed greatly to the initial improvement of ASR systems [24]. Also, due to the considerable efforts of the researchers to overcome the challenges of developing ASR systems for low-resource languages. Recently, Mehta et al. introduced a recognition system for the Arabic language [25], Dua et al. presented a comprehensive study on the development of ASR systems for the Gujarati language [26], and Gupta et al. proposed a

DL model for the recognition of non-Indian languages that includes, Chinese, Dutch, French, Finnish, Japanese, Spanish, Greek, German, and Persian [27] (see Table 1).

Although ASR systems of today are advanced, can recognize a wide range of acoustics, and can handle the complexities within the spoken languages, the most pivotal development in the ASR system was the introduction of the HMM, which deviated the researchers from the study of conventional pattern recognition and consequently strengthened the accuracy of ASR systems [28, 29]. The researchers also integrated the text encoders with the pre-existing end-to-end speech translation systems, so that the models can be trained effectively on both the labeled data and the unlabeled data; however, they still need an extensive amount of data to be trained to give proper results [30]. In [31], the authors proposed a method called Large-Scale Collaborative Data-Driven Discovery (LCDD) for increasing precision by extracting global trends from intricate local trends. Berard et al. proposed an end-to-end translation system for speech-to-text that was trained using an audiobook dataset [32]. In a related study, the researchers in [33] developed another translation system for speech-to-text, which does not require any source language text at the same time of training and decoding, consequently, resolving many of the challenges encountered while dealing with under-resourced situations. Adam et al. proposed DL techniques for speech recognition and synthesis with the prospect of clear and naturally generated speech [34]. Researchers in [35] proposed a Text-to-Speech (TTS) system that could handle longer forms of text. Higuchi et al. discussed the Non-Autoregressive (NAR) model for an end-to-end ASR system [36], whereas the researchers in [37] combined the techniques of digital signal processing and natural language processing (NLP) to design a TTS system specially designed for visually impaired individuals.

Nadeem et al. worked on the detection of a language by utilizing a hybrid approach where they integrated the CNN with the Long Short-Term Memory (LSTM) algorithm and measured the performance of their model using accuracy as their evaluation metric [38]. In [39], the researchers utilized DL approaches for designing a system that can convert text into speech, especially for low-resource languages. The researchers in [7] utilized a Deep Neural Network (DNN) to extract the abstract features from raw audio data directly. [40] proposed an architecture of a Recurrent Encoder-Decoder DNN for translating speech from one language into text from another language. Ren et al. proposed SimulSpeech, which is an advanced cascade system with simultaneous ASR and Neural Machine Translation [41]. Möller et al. used a two-step approach where, firstly, they used a log-likelihood for comparison of the features and then extracted the parameters of the speech signal [42]. Hayashi et al. proposed a toolkit for the text-to-speech system based on the Kaldi ASR system [43]. Kim et al. presented a joint CTC-attention-based model inside the multi-task learning framework [44]. Furthermore, Tang et al. proposed an attention-based sequence-to-sequence model by proposing the co-training of ASR and Speech Translation (ST) with the denoising autoencoder and machine translation, with the limitation that it required an enormous amount of data and would not be suitable for low-resource languages [45].

Researchers in [46] and [47] proposed a sequence-to-sequence neural network. For the performance metric of speech-to-text Automatic Speech Recognition (ASR) systems, the researchers in [48] suggested the usage of precision and recall for measuring the performance of the speech summarization system due to their ability to retain and capture essential information. However, when dealing with low-resource languages, researchers in [49] recommended the Word Error Rate (WER) as the best predictor for the ASR system as it effectively accounts for the factor of language relatedness, size of the pre-training data, and data augmentation. In addition to WER, the researchers have also recommended the utilization of Connectionist Temporal Classification (CTC) loss for better convergence [50–52], especially when using a DL architecture of ASR systems.

The researchers in [53] discuss the unavailability of large amounts of transcribed spontaneous speech data and propose a HHMM-based LVASR. This study gives a Word Error Rate of 58.4% and lacks the incorporation of joint loss optimization. Correspondingly, research presented in [54] developed a Subspace Gaussian Mixture Model (SGMM) with a promising WER of 9.64%; however,

[1]Values are not directly comparable because datasets, evaluation protocols, and resource conditions differ.

Table 1. Contextual comparison with selected ASR paradigms from the literature[1]

Model Name	Model Type	Architecture Components	Feature Eng. Req.	End-to-End	Dataset Suitability	CTC Loss	WER (%)	Loss Opt.	Robustness to Variability	Computational Efficiency	Key Limitations	Key Strengths
Proposed CRNN (DeepSpeech2-Inspired)	End-to-End Deep Learning	Conv2D + ReLU + BatchNorm + Bidirectional GRU + CTC Decoder	Low	Yes	Low-resource	Yes	Validation (21.29); Final test (17.05)	Yes (5.87)	High	Medium	Moderate WER compared to heavily engineered systems; limited by dataset size (~58K samples)	End-to-end; no pre-segmentation needed; stable convergence; handles Urdu variability; CTC-based loss optimization; low feature engineering
HMM-based LVASR [64]	Classical	HMM + GMM + Hand-crafted Features (MFCC/PLP)	High	No	High-resource	No	58.40	No	Low	Medium	Very high WER; requires extensive manual feature engineering; no end-to-end optimization; poor generalization	Well-established theoretical framework; interpretable model structure
Subspace GMM (SGMM) [65]	Classical / Statistical	SGMM + Subspace Projections + Hand-crafted Features	High	No	High-resource	No	9.64	No	Medium	Low	No end-to-end learning; requires extensive feature engineering; poor scalability to new domains; computationally expensive training	Lowest reported WER; strong acoustic modeling with subspace projections
CNN-LSTM Hybrid [66]	Hybrid Deep Learning	CNN + LSTM + CTC/Cross-Entropy	Moderate	Partial	Moderate-resource	Yes/No	25–35	Yes	Medium	Medium	Requires moderate feature engineering; LSTM training can be slow; limited bidirectional context	Captures local and temporal features; moderate complexity; applicable to multiple languages
Seq2Seq (Encoder-Decoder) [67]	End-to-End Deep Learning	Encoder RNN (LSTM/GRU) + Decoder RNN + Beam Search	Low	Yes	Moderate-resource	No	20–30	Yes	Medium	Low	Exposure bias during inference; requires large training data; computationally expensive decoding	End-to-end learning; flexible output length; joint acoustic-language modeling
Attention-based ASR (LAS) [68]	End-to-End Deep Learning	Encoder (BiLSTM/BiGRU) + Attention Mechanism + Decoder	Low	Yes	Moderate to High-resource	No	10–20	Yes	High	Low	High computational cost; requires large datasets for optimal performance; complex training	Superior context learning; flexible alignment; state-of-the-art for high-resource languages
Transformer-based ASR [69]	End-to-End Deep Learning	Multi-Head Self-Attention + Positional Encoding + Feed-Forward Layers	Low	Yes	High-resource	Yes/No	8–15	Yes	High	Low	$O(n^2)$ complexity; requires very large datasets; high GPU memory; not ideal for low-resource settings	Captures long-range dependencies; parallelizable training; state-of-the-art performance on large corpora

like LVASR, it fails to consider the overall loss optimization, requiring tedious hand engineering for obtaining the overall good performance. In the present study, WER and CTC-based optimization are adopted to provide a consistent evaluation and training framework for Urdu end-to-end ASR.

3 Methodology

3.1 Research design

The state of the art in ASR of low-resource languages like Urdu is a difficult technical challenge because annotated corpora are limited, phonetic and speaker diversity are

high, and large-scale aligned speech-text data are scarce. To address these issues, the current research creates an end-to-end Urdu ASR model, inspired by DeepSpeech2, with the help of a CRNN. It aims to train a direct association between acoustic observations and textual transcriptions that do not use phoneme-level associations, manually-crafted pronunciation dictionaries, or independently trained acoustic and language networks. The general methodological design is an experimental pipeline that is reproducible and involves dataset preparation, text normalization, acoustic feature extraction, neural sequence modeling, optimization, and multi-level evaluation. Unlike traditional hybrid ASR pipelines, the proposed framework embraces an end-to-end learning paradigm where the feature representation learning and transcription modeling stages are learned together. This design is especially applicable to low-resource environments, where linguistic resources are manually engineered and are either scarce or unavailable. In addition to model development, the model includes a formal validation plan comprising controlled baseline comparison, component-wise ablation analysis, and error diagnostics. The increased flexibility of this design allows interpretation of performance improvements in a more rigorous way and minimizes the chance of including improvement due to experimental conditions that are not controlled (see Table 2).

Table 2. Overall methodological pipeline

Stage	Description	Objective
Dataset preparation	Speech-text pairing and partitioning	Ensure reproducibility and prevent leakage
Text preprocessing	Normalization and token standardization	Enable consistent evaluation
Acoustic processing	FFT-based spectrogram extraction	Capture time-frequency characteristics
Model architecture	CRNN with convolutional and BiGRU layers	Learn hierarchical acoustic representations
Training	CTC-based optimization	Enable alignment-free sequence learning
Evaluation	WER, SER, and WIL	Quantify transcription accuracy
Validation design	Baseline comparison, ablation, and error analysis	Strengthen scientific rigor

3.2 Dataset and experimental protocol

The experiments were made on the Urdu subset of the Mozilla Common Voice corpus, consisting of speech recordings along with human-transcribed transcriptions of a diverse group of speakers. This corpus was chosen as it offers an openly available and reproducible reference point to Urdu ASR, as well as capturing the practical variability of speech conditions, such as differences in pronunciation, speaker diversity, recording quality, and environmental noise. The dataset was split into training, validation, and test subsets, which were mutually exclusive. The training set had 58,119 utterances, and the validation set had 6,458 utterances. A reserved set (held-out test set) was used in the final evaluation. This separation was preserved throughout all experiments such that model development, hyperparameter selection, and reporting final performance were methodologically independent. In partition spaces where metadata was available, spillage was minimized between speakers to minimize information leakage and pessimistic estimates of performance. This is especially significant in speech recognition, where speaker similarity across splits can be artificially enhanced by generalization. Moreover, the protocol of the experiment clearly differentiates between monitoring at the validation stage and the final test evaluation. Checkpoint selection and training diagnostics were only based on validation data, but all primary results reported in Section 4 were based on the held-out test set (see Table 3).

Table 3. Dataset configuration

Parameter	Value
Dataset	Mozilla Common Voice (Urdu subset)
Training utterances	58,119
Validation utterances	6,458
Test set	Held-out, speaker-disjoint where possible
Task	Urdu speech-to-text transcription
Recording conditions	Multi-speaker, heterogeneous environments

3.3 Text normalization and acoustic preprocessing

Both reference and predicted transcriptions were normalized to text before being assessed to avoid

the potential of error measures being distorted by superficial formatting differences. The pipeline of normalization included lowercasing, tokenization, and removal of punctuation marks. The use of lowercasing minimized the orthographic inconsistency, tokenization streamlined lexical division, and punctuation elimination meant that the evaluation was based on acoustic recognition and not non-acoustic written clues. On the acoustic side, the raw speech signals were transformed into time-frequency representations by Fast Fourier Transform (FFT) processing. The adoption of spectrogram characteristics as the main input representation was due to its ability to retain both spectral and temporal dynamics, thus providing an appropriate foundation to end-to-end neural speech modeling. Spectrograms provide a more structured representation of a raw waveform input compared to it, and are advantageous, especially in low-resource training regimes. Padded batching and prefetching were added to the data pipeline to support utterances of varying lengths. This design enhanced the efficiency of computations while retaining longer sequences that could otherwise be cut off during training (see Table 4).

Table 4. Preprocessing pipeline

Stage	Operation	Purpose
Text normalization	Lowercasing	Remove case variability
Text normalization	Tokenization	Standardize word boundaries
Text normalization	Punctuation removal	Avoid non-acoustic bias
Audio processing	FFT transformation	Convert waveform to frequency domain
Feature extraction	Spectrogram	Capture acoustic structure
Data handling	Padded batching	Support variable-length input
Data handling	Prefetching	Improve training efficiency

3.4 Acoustic feature representation

The suggested framework is based on spectrogram-based acoustic features, which offer a succinct and informative description of the speech signal. Depending on the FFT setup, the dimensionality of the input is defined, resulting in a feature space of 193. This representation enables the model to retain important spectral data and be computationally tractable. The hybrid convolutional-recurrent architecture used in this study is well-fitted to spectrogram features. The convolutional front-end learns local spectral patterns, such as formant-like patterns and short-time phonetic patterns, and the recurrent encoder learns longer-term dependencies in the utterance. This combination is particularly helpful with Urdu ASR, where both contextual and speaker-related variation can have a significant impact on transcription accuracy (see Table 5).

Table 5. Acoustic feature settings

Parameter	Value
Feature type	Spectrogram / Mel-spectrogram
Signal transform	FFT-based
Input dimension	193
Sequence representation	Time series of spectral vectors

3.5 Model architecture

The proposed ASR model is based on a DeepSpeech2-inspired CRNN architecture (see Figure 1), combining convolutional feature extraction with a recurrent sequence model into a single end-to-end model. The model starts with the spectrogram input features that are filtered through two 2D convolutional layers. These layers serve as a front-end acoustic encoder, which extracts local time-frequency patterns as being pertinent to phonetic discrimination. Convolutional layers are then preceded by batch normalization and ReLU. ReLU activation enhances non-linear discrimination of features and speeds up optimization because of its ability to stabilize optimization by reducing internal covariate shift, whereas Batch normalization does the same. The feature maps are then reshaped into a sequence representation to be used in recurrent modeling, after convolutional processing. The sequence representation is then fed through five layers of bidirectional Gated Recurrent Units. The bi-directional recurrence allows the model to have contextual dependencies of the past and future frames, which proves useful in continuous speech recognition where phonetic interpretation is often contingent on the context. The dropout regularization is implemented following recurrent layers to minimize overfitting and enhance generalization in the

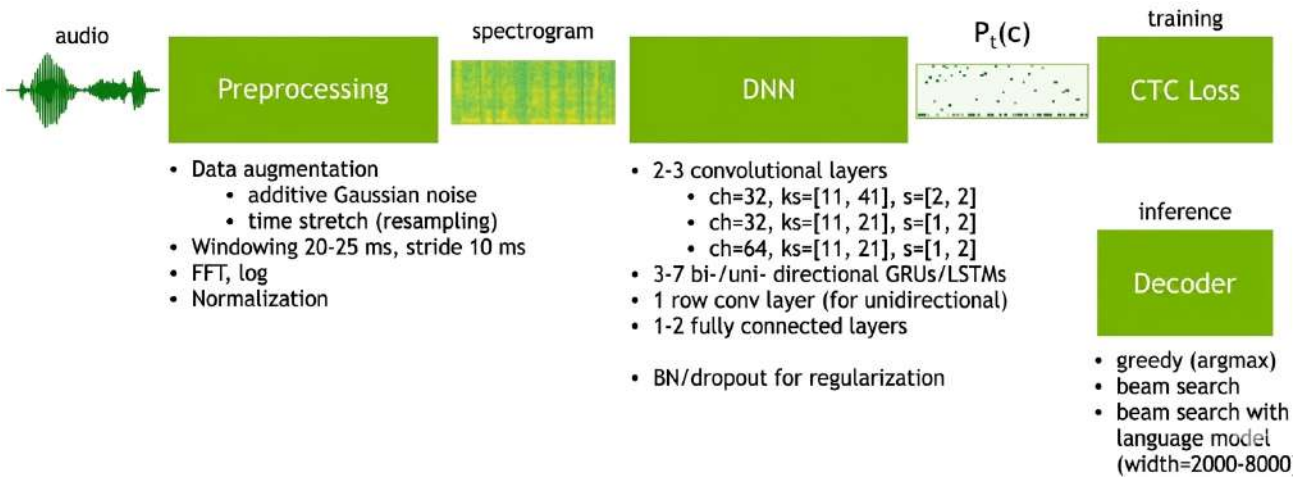

Figure 1. DeepSpeech2 Architecture

low-resource setting. Lastly, an encoded sequence representation is refined with a fully connected layer using ReLU activation, before classification. The output layer uses SoftMax activation, which outputs character-level posterior probabilities, which include the blank symbol needed in CTC-based decoding (see Table 6).

Table 6. Model architecture

Component	Configuration
Input	Spectrogram features
Convolutional layers	2 × Conv2D + BatchNorm + ReLU
Recurrent layers	5 × Bidirectional GRU
Regularization	Dropout
Dense layer	Fully connected + ReLU
Output	SoftMax, character-level
Training objective	CTC loss

3.6 Training Strategy

Connectionist Temporal Classification (CTC) loss was used to train the model, and it allows learning without alignment between speech sequences of the input and speech transcriptions of the output. This is especially beneficial in Urdu ASR since frame-level speech-text correspondences are not common in low-resource collections. CTC does not necessitate explicit segmentation but rather enables the model to discover valid alignments implicitly in training. The Adam optimizer with a learning rate of 1x10-4 was used to optimize. The training was performed in mini-batches of padded sequences and monitored with the help of validation-based checkpoints and TensorBoard logging. This configuration allowed the stability of optimization and maintained complete traceability of training behavior. Experiments were implemented with TensorFlow and Keras on a two-GPU system with two NVIDIA RTX 3090 GPUs. Validation metrics were only applied to training diagnostics and model selection, but final model evaluation was done on the held-out test set (see Table 7).

Table 7. Training configuration

Parameter	Value
Loss function	CTC
Optimizer	Adam
Learning rate	1x10-4
Framework	TensorFlow / Keras
Hardware	2 × RTX 3090 GPUs
Monitoring	TensorBoard
Checkpointing	Enabled

3.7 Evaluation metrics and comparative framework

WER was the main measure of model performance and is the most commonly used measure of ASR quality. WER was supplemented by Sentence Error Rate (SER) and Word Information Lost (WIL), which, in turn, allowed a more detailed evaluation of the recognition performance on lexical and sentence levels. All the models provided as baselines were tested with equal experimental conditions, such as dataset split, text

normalization pipeline, and scoring protocol. The comparative framework included three baselines, including a simplified structural alternative, known as a reduced recurrent model, and two transformer-based ASR models (including recent attention-driven architectures) and a wav2vec2-style model (including self-supervised pretraining-based recognition). The proposed CRNN was taken as the target model that was compared to these baselines (see Table 8).

Table 8. Baseline framework

Model	Type	Role
Reduced RNN	Shallow recurrent	Structural baseline
Transformer ASR	Attention-based	Modern baseline
wav2vec2	Self-supervised	Pretrained baseline
Proposed CRNN	Hybrid convolutional-recurrent	Target model

4 Results and discussion

4.1 Training dynamics and convergence

The proposed model exhibited stable convergence under the CTC optimization framework. As training progressed, validation WER decreased consistently, accompanied by a gradual reduction in loss. This trend shows that the model was able to learn increasingly discriminative acoustic and temporal representations using the training corpus. The improvement rate decreased after around 400 epochs, indicating that the model had reached a performance plateau. The last epoch that was tracked was epoch 478, with a validation WER of 21.29% and a loss of 5.87, which showed convergence without any indication of instability or extreme overfitting. The continuous improvement in both WER and loss substantiates the stability of the proposed training strategy and proves that the CRNN architecture is capable of coping with the acoustic variability in the corpus of Urdu (see Table 9, Figure 2 and 3).

4.2 Final test performance

The final model was evaluated on the held-out test set to obtain an unbiased estimate of generalization performance. The results of the test prove that the suggested system can be used to reach a competitive level of recognition in the conditions of multi-speaker and non-homogeneous recording in real-life conditions. The model attained a WER of 17.05%, an SER of 34.72%, and a WIL of 0.41. Combined, these findings suggest that the system obtained a large share of the target lexical content with a reasonable sentence-level reliability of a low-resource Urdu ASR system. The smaller test-set WER compared to the validation trajectory indicates that the chosen checkpoint was able to generalize well to the held-out evaluation set (see Table 10 and Figure 4).

Table 9. Validation performance across epochs

Epoch	WER (%)	Loss
100	24.87	6.12
200	22.94	5.98
300	21.76	5.91
400	21.41	5.88
478	21.29	5.87

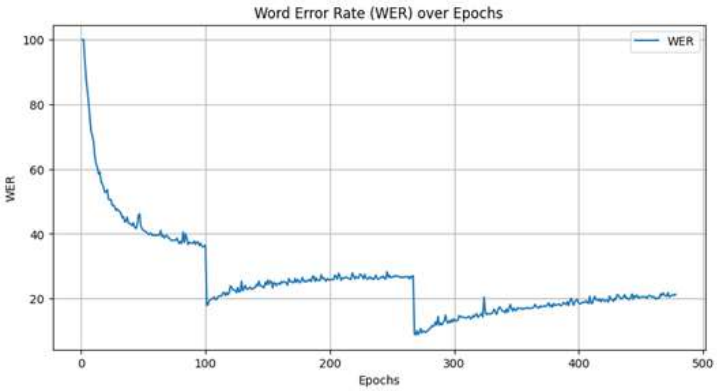

Figure 2. Training of WER and Epochs

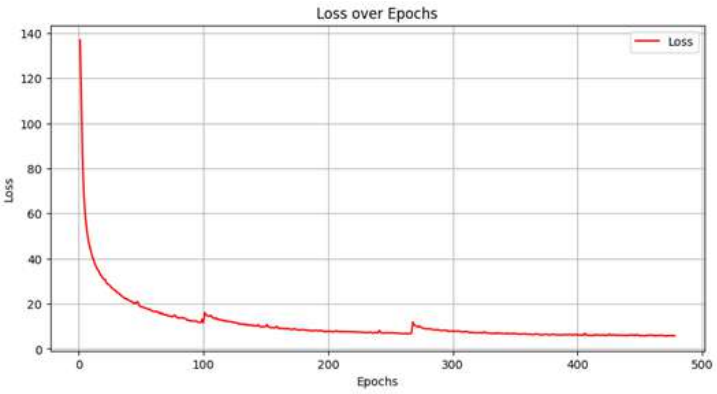

Figure 3. Training of Loss and Epochs

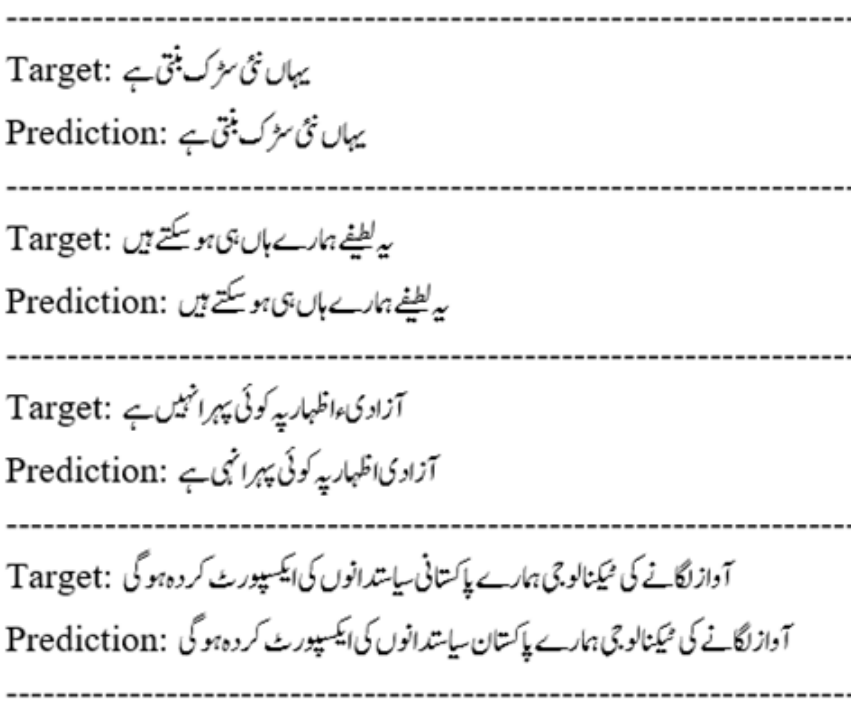

Figure 4. Performance of our proposed model

Table 10. Final test performance

Metric	Value
WER (%)	17.05
SER (%)	34.72
WIL	0.41

4.3 Comparative analysis

The controlled baseline comparison was conducted to determine whether the performance improvements of the proposed CRNN are explained by the beneficial characteristics of the architecture instead of the conducive training conditions. The comparison shows that the proposed model performed better than all the baselines tested with the same experimental protocol. The lower recurrent base gave a WER of 23.84%, which suggests that a shallower recurrent design cannot resolve the temporal complexity of Urdu speech. This result was significantly improved by the transformer-based baseline (WER of 19.62%) and the wav2vec2-style baseline (WER of 18.31%), which also indicates the advantage of the new representation learning. However, the suggested CRNN demonstrated the best results of WER of 17.05%, implying that a set of convolutional front-end processing and a bidirectional recurrent model is still beneficial in this low-resource context. The results suggest that the suggested architecture provides a good trade-off between the local spectral modeling and the long-range contextual encoding. Notably, the gains are determined by like-for-like comparison, as opposed to cross-study aggregation, which makes the conclusions methodologically more justified (see Table 11).

Table 11. Baseline comparison

Model	WER (%)
Reduced RNN	23.84
Transformer	19.62
wav2vec2	18.31
Proposed CRNN	17.05

4.4 Ablation and component analysis

A component-wise ablation study was performed to get a better understanding of the contribution of individual architectural components. The complete CRNN architecture was always more effective than any of the reduced versions, which confirms that the effectiveness of the

model is due to the interplay of the key constituent parts of the architecture but not to a single significant constituent. The elimination or reduction of the convolutional front-end lowered the performance to a WER of 19.48 %, which shows that local acoustic feature extraction is a significant factor in recognition performance. A decrease in recurrent depth led to a WER of 20.91%, indicating that deep temporal modeling is also needed. Substituting the bidirectional GRUs with unidirectional ones raised WER to 19.76 %, which highlights the importance of bidirectional context. Similarly, minimizing dropout deteriorated WER to 18.92 %, indicating that regularization is a contributing factor to generalization. Overall, the findings of the ablation prove the statement that the combination of convolutional feature extraction, deep bidirectional recurrence, and regularization is the strength of the model. The loss seen in all mutated forms is a good indication that the gains reported are structurally based (see Table 12).

Table 12. Ablation study

Variant	WER (%)
Full CRNN	17.05
Conv-reduced	19.48
Shallow RNN	20.91
UniGRU	19.76
Reduced dropout	18.92

4.5 Error analysis and interpretation

To go beyond aggregate measures, the error analysis was conducted in detail to establish the prevailing sources of recognition failure. The error types distribution indicates that the most frequent error types were substitutions, which constituted 52.3% of the total error types, deletions (28.7%), and insertions (19.0%). The preponderance of substitutions indicates that the model typically relates coarse phonetic organization but cannot fine-tune lexical discrimination, especially in cases where the acoustically similar Urdu phonemes play a role. The most frequent errors involved deletions and were more prevalent with fast or coarticulated speech, where the time limits are less clear. The noisy or acoustically ambiguous parts were commonly related to less frequent insertion errors. Qualitative analysis also indicated that dialectal differences and noise of the environment had a significant impact on recognition stability. Phonetic confusion was found to be the main cause of recognition failure, whereas fast speech heightened the chance of deletion patterns. These findings show that, despite the model learning a strong coarse-grained speech-to-text mapping, it remains limited by the fine phonetic ambiguity and variation in real-world recordings. These results are in line with the established difficulties of low-resource multilingual ASR (see Table 13 and Table 14).

Table 13. Error distribution

Error type	Percentage
Substitution	52.3%
Deletion	28.7%
Insertion	19.0%

Table 14. Error characteristics

Category	Observation
Phonetic confusion	Dominant error source
Dialect variation	Noticeable performance degradation
Noise	Increased recognition errors
Fast speech	More frequent deletion errors

4.6 Discussion

Experimental findings indicate that the proposed DeepSpeech2-inspired CRNN shows good and consistent performances in Urdu ASR in realistic low-resource settings. The validation pathway supports the fact that the convergence is stable, the test results held out to prove the good performance of the proposed architecture are good, and the controlled comparative analysis demonstrates that the proposed architecture is better than other recurrent, transformer-based, and self-supervised baselines in the same experimental context. The ablation study also verifies that the performance attained is not by chance, but rather an outcome of the interplay of convolutional acoustic modeling, deep bidirectional temporal encoding, and dropout-based regularization. Simultaneously, the error analysis also shows that there are significant limitations, especially phonetic confusion, dialectal variability, and noisy speech. In this respect, the proposed system provides a competitive baseline that establishes a reproducible

reference point for future advances in Urdu automatic speech recognition. Further development can be made by enhanced front-end denoising, dialect-aware modeling, data augmentation, and the combination of larger self-supervised speech encoders. The proposed CRNN, within the context of the current research, however, can position a methodologically sound and empirically competitive baseline to end-to-end Urdu speech-recognition.

5 Conclusion

The current work provides a complete end-to-end automatic speech recognition model of Urdu built upon a DeepSpeech2-style convolutional recurrent neural network that has been trained on the Urdu Mozilla Common Voice. The proposed system uses a combination of spectrogram-based acoustic features, convolutional front-end processing, bidirectional recurrent sequence modeling and Connectionist Temporal Classification to learn direct speech-to-text mappings without the need of phoneme-level alignment or manually generated linguistic resources.

According to the experimental results, the proposed framework offers competitively-trained behavior and stable training within realistic low-resource conditions. The model had a validation WER of 21.29 at epoch 478 and an end-of-training held-out test WER of 17.05, surpassing the reduced recurrent and transformer-based and wav2vec2-style baselines tested using the same protocol. The ablation experiment showed that the entire architecture is as powerful as a combination of convolutional feature extraction, deep bidirectional recurrent encoding, and dropout-based regularization, and not any of the design choices.

The model is also limited as explained in the error analysis. The main failures were caused by the substitution errors, and the further degradation was related to phonemes of acoustically similar Urdu, dialect of the language, environmental noises, and rapid or coarticulated speech. Such results imply that, although the model is able to capture coarse phonetic and temporal structure well, additional enhancements will be necessary with more robustness in order to capture fine-grained phonetic ambiguity and recording variability.

On the whole, the research shows that convolutional recurrent end-to-end systems are still very effective in low-resource Urdu ASR as long as they are trained within a controlled and reproducible evaluation setting. In addition to its direct empirical value, the work provides a practical foundation for future studies on Urdu speech recognition, including dialect-aware modeling, noise-robust front ends, larger self-supervised encoders, and broader public evaluation resources.

Data Availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Author Contributions

Syed Azeem Inam: Conceptualization, Methodology, and Writing- Original draft preparation **Syeda Nazia Ashraf**: Data curation and Software. **Hassan Hashim**: Visualization and Investigation. **Syeda Wajiha Naim**: Supervision. Muhammad Ahmed Ansari: Validation and Organization. **Ahmed Raza Khanzada**: Review and Re-investigation

Compliance with Ethical Standards

It is declared that all authors don't have any conflict of interest. It is also declare that this article does not contain any studies with human participants or animals performed by any of the authors. Furthermore, informed consent was obtained from all individual participants included in the study.

Funding Information

Not Applicable.

References

[1] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *J. Acoust. Soc. Am.*, vol. 24, no. 6, pp. 637–642, Nov. 1952, doi: 10.1121/1.1906946.

[2] M. A. Anusuya and S. K. Katti, "Speech recognition by machine, a review," Jan. 2010.

[3] W. Ghai and N. Singh, "Literature review on automatic speech recognition," *Int. J. Comput. Appl.*, vol. 41, no. 8, pp. 42–50, Mar. 2012, doi: 10.5120/5565-7646.

[4] H. Aldarmaki, A. Ullah, S. Ram, and N. Zaki, "Unsupervised automatic speech recognition: A review,"

Speech Commun., vol. 139, pp. 76–91, Apr. 2022, doi: 10.1016/j.specom.2022.02.005.

[5] K. Riaz, "Baseline for Urdu IR evaluation," in *Proceedings of the 2nd ACM Workshop on Improving Non-English Web Searching*, New York, NY, USA: ACM, Oct. 2008, pp. 97–100. doi: 10.1145/1460027.1460045.

[6] A. Daud, W. Khan, and D. Che, "Urdu language processing: A survey," *Artif. Intell. Rev.*, vol. 47, no. 3, pp. 279–311, Mar. 2017, doi: 10.1007/s10462-016-9482-x.

[7] M. Humayoun, H. Hammarström, and A. Ranta, "Urdu morphology, orthography and lexicon extraction," Apr. 2022.

[8] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, Dec. 2018, pp. 112–118. doi: 10.1109/SLT.2018.8639583.

[9] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019, doi: 10.1109/ACCESS.2019.2896880.

[10] A. H. Meftah, Y. A. Alotaibi, and S.-A. Selouani, "Evaluation of an Arabic speech corpus of emotions: A perceptual and statistical analysis," *IEEE Access*, vol. 6, pp. 72845–72861, 2018, doi: 10.1109/ACCESS.2018.2881096.

[11] M. A. Anusuya and S. K. Katti, "Speech recognition by machine, a review," Jan. 2010.

[12] L. Deng, "Deep learning: From speech recognition to language and multimodal processing," *APSIPA Trans. Signal Inf. Process.*, vol. 5, no. 1, 2016, doi: 10.1017/ATSIP.2015.22.

[13] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 5, pp. 1060–1089, May 2013, doi: 10.1109/TASL.2013.2244083.

[14] J. Guo *et al.*, "The HW-TSC's simultaneous speech-to-text translation system for IWSLT 2023 evaluation," in *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 376–382. doi: 10.18653/v1/2023.iwslt-1.35.

[15] J.-X. Zhang *et al.*, "Voice conversion by cascading automatic speech recognition and text-to-speech synthesis with prosody transfer," Sep. 2020.

[16] M. Mohri, F. Pereira, and M. Riley, "Weighted automata in text and speech processing," *arXiv preprint cs/0503077*, 2005.

[17] M. Stinson, S. Stinson, L. Elliot, and R. Kelly, "Relationships between benefit and use of a speech-to-text service, perceptions of courses, and course performance," in *Annual Meeting of the American Educational Research Association*, San Diego, CA, 2004.

[18] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[19] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," Sep. 2016.

[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[21] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[22] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[23] C.-C. Chiu *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 4774–4778.

[24] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, pp. 4960–4964.

[25] Y. Jia *et al.*, "Leveraging weakly supervised data to improve end-to-end speech-to-text translation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2019, pp. 7180–7184. doi: 10.1109/ICASSP.2019.8683343.

[26] S. Mehra, V. Ranga, R. Agarwal, and S. Susan, "Speaker independent recognition of low-resourced multilingual Arabic spoken words through hybrid fusion," *Multimed. Tools Appl.*, Mar. 2024, doi: 10.1007/s11042-024-18804-w.

[27] M. Dua, B. Bhagat, S. Dua, and N. Chakravarty, "A review on Gujarati language based automatic speech recognition (ASR) systems," *Int. J. Speech Technol.*, vol. 27, no. 1, pp. 133–156, Mar. 2024, doi: 10.1007/s10772-024-10087-8.

[28] A. Gupta, R. Kumar, and Y. Kumar, "Hybrid deep learning based automatic speech recognition model for recognizing non-Indian languages," *Multimed. Tools Appl.*, vol. 83, no. 10, pp. 30145–30166, Sep. 2023, doi: 10.1007/s11042-023-16748-1.

[29] R. Shaik and S. Venkatramaphanikumar, "Sentiment analysis with word-based Urdu speech recognition," *J. Ambient Intell. Humaniz. Comput.*, vol. 13, no. 5, pp. 2511–2531, May 2022, doi: 10.1007/s12652-021-03460-x.

[30] S. Shaikh Naziya and R. R. Deshmukh, "LPC and HMM performance analysis for speech recognition system for Urdu digits," *IOSR J. Comput. Eng.*, 2017.

[31] N. F. Khan, N. Hemanth, N. Goyal, P. KR, and P. Agarwal, "Call translator with voice cloning using transformers," in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, IEEE, Apr. 2024, pp. 1–6. doi: 10.1109/I2CT61223.2024.10543304.

[32] J. R. Bellegarda, "Interaction-driven speech input: A data-driven approach to the capture of both local and global language constraints," *ACM SIGCHI Bull.*, vol. 30, no. 2, pp. 102–105, 1998.

[33] A. Berard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Apr. 2018, pp. 6224–6228. doi: 10.1109/ICASSP.2018.8461690.

[34] A. Berard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," Dec. 2016.

[35] E. E. B. Adam, "Deep learning based NLP techniques in text to speech synthesis for communication recognition," *J. Soft Comput. Paradigm*, vol. 2, no. 04, pp. 209–215, 2020.

[36] X. Huang *et al.*, "Whistler: A trainable text-to-speech system," in *Proc. 4th Int. Conf. Spoken Language Processing (ICSLP)*, 1996, pp. 2387–2390, doi: 10.1109/ICSLP.1996.607289.

[37] Y. Higuchi *et al.*, "A comparative study on non-autoregressive modelings for speech-to-text generation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, Dec. 2021, pp. 47–54. doi: 10.1109/ASRU51503.2021.9688157.

[38] I. Isewon, J. Oyelade, and O. Oladipupo, "Design and implementation of text to speech conversion for visually impaired people," *Int. J. Appl. Inf. Syst.*, vol. 7, no. 2, pp. 25–30, Apr. 2014, doi: 10.5120/ijais14-451143.

[39] M. A. Nadeem, S. H. H. Bukhari, M. U. Arshad, S. Naeem, M. O. Beg, and W. Shahzad, "Language detection and localization, for Pakistani languages, in acoustic channels," in *2022 17th International Conference on Emerging Technologies (ICET)*, IEEE, Nov. 2022, pp. 142–147. doi: 10.1109/ICET56601.2022.10004691.

[40] S. R. Mache, M. R. Baheti, and C. N. Mahender, "Review on text-to-speech synthesizer," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 8, pp. 54–59, 2015.

[41] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," Mar. 2017.

[42] Y. Ren *et al.*, "SimulSpeech: End-to-end simultaneous speech to text translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 3787–3796. doi: 10.18653/v1/2020.acl-main.350.

[43] S. Möller, F. Hinterleitner, T. H. Falk, and T. Polzehl, "Comparison of approaches for instrumentally predicting the quality of text-to-speech systems," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[44] T. Hayashi *et al.*, "Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2020, pp. 7654–7658. doi: 10.1109/ICASSP40776.2020.9053512.

[45] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Mar. 2017, pp. 4835–4839. doi: 10.1109/ICASSP.2017.7953075.

[46] Y. Tang, J. Pino, C. Wang, X. Ma, and D. Genzel, "A general multi-task learning framework to leverage text data

for speech to text tasks," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2021, pp. 6209–6213. doi: 10.1109/ICASSP39728.2021.9415058.

[47] R. J. Weiss, R. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, "Wave-Tacotron: Spectrogram-free end-to-end text-to-speech synthesis," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2021, pp. 5679–5683. doi: 10.1109/ICASSP39728.2021.9413851.

[48] A. Tjandra, S. Sakti, and S. Nakamura, "Listening while speaking: Speech chain by deep learning," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, Dec. 2017, pp. 301–308. doi: 10.1109/ASRU.2017.8268950.

[49] A. Nenkova, "Summarization evaluation for text and speech: Issues and approaches," in *Ninth International Conference on Spoken Language Processing*, 2006.

[50] M. C. Stoian, S. Bansal, and S. Goldwater, "Analyzing ASR pretraining for low-resource speech-to-text translation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2020, pp. 7909–7913. doi: 10.1109/ICASSP40776.2020.9053847.

[51] P. Bahar, T. Bieschke, and H. Ney, "A comparative study on end-to-end speech to text translation," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, Dec. 2019, pp. 792–799. doi: 10.1109/ASRU46091.2019.9003774.

[52] P.-H. Le, H. Gong, C. Wang, J. Pino, B. Lecouteux, and D. Schwab, "Pre-training for speech translation: CTC meets optimal transport," in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., in *Proceedings of Machine Learning Research*, vol. 202, PMLR, Jul. 2023, pp. 18667–18685. [Online]. Available: https://proceedings.mlr.press/v202/le23a.html.

[53] Y. Liu, J. Zhu, J. Zhang, and C. Zong, "Bridging the modality gap for speech-to-text translation," Oct. 2020.

[54] A. A. Raza, S. Hussain, H. Sarfraz, I. Ullah, and Z. Sarfraz, "An ASR system for spontaneous Urdu speech," *The Proc. of Oriental COCOSDA*, pp. 24–25, 2010.

[55] S. Naeem *et al.*, "Subspace Gaussian mixture model for continuous Urdu speech recognition using Kaldi," in *2020 14th International Conference on Open Source Systems and Technologies (ICOSST)*, IEEE, Dec. 2020, pp. 1–7. doi: 10.1109/ICOSST51357.2020.9333026.

[56] R. Ardila *et al.*, "Common Voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[57] L. Maison and Y. Estève, "Some voices are too common: Building fair speech recognition systems using the Common Voice dataset," *arXiv preprint arXiv:2306.03773*, 2023.

[58] H. Kwon, D. Park, and O. Jo, "Silent-hidden-voice attack on speech recognition system," *IEEE Access*, 2024.

[59] B. Arendale, S. Zarandioon, R. Goodwin, and D. Reynolds, "Spoken language recognition on open-source datasets," *SMU Data Sci. Rev.*, vol. 3, no. 2, p. 3, 2020.

[60] G. Cámbara, J. Luque, and M. Farrús, "Convolutional speech recognition with pitch and voice quality features," *arXiv preprint arXiv:2009.01309*, 2020.

[61] H. A. Z. Shahgir, K. S. Sayeed, and T. A. Zaman, "Applying wav2vec2 for speech recognition on Bengali Common Voices dataset," *arXiv preprint arXiv:2209.06581*, 2022.

[62] A. Nowakowski and W. Kasprzak, "Automatic speaker's age classification in the Common Voice database," in *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, IEEE, 2023, pp. 1087–1091.

[63] D. Amodei *et al.*, "Deep Speech 2: End-to-end speech recognition in English and Mandarin," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, in *ICML'16*, JMLR.org, 2016, pp. 173–182.

[64] S. Mehra, V. Ranga, and R. Agarwal, "Multimodal integration of mel spectrograms and text transcripts for enhanced automatic speech recognition: Leveraging extractive transformer-based approaches and late fusion strategies," *Comput. Intell.*, vol. 40, no. 6, Dec. 2024, doi: 10.1111/coin.70012

[65] H. A. Alsayadi, A. A. Abdelhamid, I. Hegazy, and Z. T. Fayed, "Arabic speech recognition using end-to-end deep learning," *IET Signal Processing*, 2021, doi: 10.1049/sil2.12057.

[66] A. Rahman, Md. M. Kabir, M. F. Mridha, M. Alatiyyah, H. F. Alhasson, and S. S. Alharbi, "Arabic speech recognition: Advancement and challenges," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3376237.

[67] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, "End-to-end speech recognition: A survey,"

IEEE/ACM Trans. Audio Speech Lang. Process., 2024, doi: 10.1109/TASLP.2023.3328283.

[68] J. Tang, J. Hou, Y. Song, L.-R. Dai, and I. McLoughlin, "Effective exploitation of posterior information for attention-based speech recognition," *IEEE Access*, 2020, doi: 10.1109/ACCESS.2020.3001636.

[69] A. M. Samin *et al.*, "BanSpeech: A multi-domain Bangla speech recognition benchmark toward robust performance in challenging conditions," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3371478.