JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 09 Issue 01
2025

Research Article
https://doi.org/10.56979/901/2025

Real-Time Voice-to-Voice Translation for Cross-Lingual Communication: Cascade Pipeline and RNN Based Approach

Shanza Bibi[1], Hina Sattar[1*], Laraib Fatima[1], Ayesha Iqbal[1], and Umar Farooq Shafi[2]

[1]Department of Computer Science, Government Sadiq College Women University, Bahawalpur, 63100, Punjab, Pakistan.
[2]Departement of Computer Science, The Islamia University of Bahawalpur, Bahawalpur, 63100, Punjab, Pakistan.
[*]Corresponding author: Hina Sattar. Email:hinasattar@gscwu.edu.pk

Received: April 10, 2025 **Accepted:** May 27, 2025

Abstract: To facilitate smooth conversations, language diversity presents communication challenges, particularly in face-to-face conversations. Real-time voice-to-voice translation for cross-lingual communication bridges these gaps. Most of the population of Pakistan speaks Urdu and is not proficient in English. Language is a major barrier to accessing information and participating in global discourse. This study focused on overcoming the barrier by utilizing machine learning for multilingual voice translation. This system is designed to translate Pakistan's native languages into English, supporting real-time communication. A real-time speech translation system utilizes a two-stage approach. First, the System is trained by combining a custom and pre-trained Wav2Vec 2.0 unlabeled dataset, and achieves 98.76% accuracy. Second, the cascade pipeline is employed to support accurate translation of text from the source into the target language. In the cascade pipeline architecture, each language demonstrates a distinct recognition accuracy, which corresponds to its linguistic prominence and availability of training data. It operates by taking the user's voice as input from a microphone and employs Automatic Speech Recognition (ASR) for speech recognition and to convert speech into text [1]. To convert translated text back to the voice Text-to-Speech (TTS) [2] module is employed. End-to-end pipelines enable effective real-time communication and offer an effective and user-friendly solution for overcoming the language barrier in a multi-lingual environment. This work significantly minimizes the gaps in multilingual communication.

Keyword: Real-Time Voice Translation; Voice-to-Voice Translation; Speech-to-Text Translation; Text-to-Speech Translation; Multi-Languages Translation

1. Introduction

A single individual can't be proficient in all of the languages. Hundreds of languages are spoken worldwide. Nowadays, as we are living in a global village, the language barrier is the main resistance to communication. Humans can't learn 900 languages in the world. In the advanced world of technology, working on voice is not that difficult. Over the past few years of study on Speech translation research. Research on South Asia and China is found. The accuracy of the proposed models is much better, as all of them use Spanish, French, and English datasets, as much research has been done on those languages, and their datasets are easily available. No one focuses on Pakistan's native languages, Urdu, Punjabi, Sindhi, and many more. The world is transforming into a global village, and overcoming language barriers has become increasingly important for faster communication and collaboration across diverse linguistic communities. In Pakistan, a significant population speaks Urdu, Sindhi, Balochi, and Punjabi, which are widely used regional languages. Many

individuals do not frequent in English. This linguistic gap poses challenges in communication, particularly in professional, educational, and technological contexts. This paper aims to address this gap by exploring the existing methodologies, technologies, and challenges in developing RTV2V systems for Urdu, Sindhi, and Punjabi. By doing so, the paper emphasizes the need for localized solutions that cater to the linguistic diversity of Pakistan while leveraging advancements in speech processing and artificial intelligence.

Traditional speech-to-speech relied on a cascade system comprising Automatic Speech Recognition (ASR) [1], Machine Translation (MT), and Text-to-Speech (TTS) modules [2]. These pipelines often suffered from error propagation, where mistakes in ASR adversely affected translation and synthesis, ultimately reducing the naturalness of the output. Similarly, conventional Voice Conversion (VC) techniques primarily utilize Gaussian Mixture Models (GMMs) and Artificial Neural Networks (ANNs), which operate on a frame-by-frame basis and require parallel, time-aligned data. [3] These are effective for limited tasks but struggle to model prosodic and temporal variations, resulting in robotic or monotonous voice outputs and poor speaker generalization.

Sequence-to-sequence (seq2seq) models were introduced to address these limitations, enabling end-to-end training and the modeling of long-range dependencies in speech [4]. These methods improved the naturalness and contextual coherence of speech while removing the need for parallel data. To extract speaker-independent content features, Phonetic Posterior Grams (PPGs) were used, making many-to-one and many-to-many VC possible. The adoption of Variational Auto Encoders (VAEs), Generative Adversarial Networks (GANs), and non-parallel frameworks like StarGANv2-VC facilitated unsupervised, many-to-many VC with speaker similarity, emotional expressiveness, and cross-lingual capabilities, though GANs often required careful training to avoid instability [5].

Automated speech recognition (ASR) is used to perform RTV2V simply with Text-to-Speech STT Synthesis, which is called Voice-to-Text. It picks up audio in the form of sound waves and applies linguistic algorithms to convert the audio input into words, digital characters, and phrases. STT extracts the feature and decodes according to the matching feature, and then generates text. As long as the text is generated from the audio, then Text-to-Speech is used, which is often called a cascade system. Text-to-speech (TTS) converts text into spoken audio, essentially allowing a computer to "read aloud" any given text by generating a synthetic human-like voice output. It utilizes complex algorithms to analyze written text, identify pronunciation rules, and generate the corresponding sound waveforms to create audible speech. The seq2Seq model is adopted to accurately and sequentially translate the voice of one language into the voice of another language. Major contribution to this research are listed below.

- Urdu and English languages voice-to-voice translation system
- Removing noise from the input data
- Cloning of the input speaker's voice during V2V translation
- Ensure the privacy of the user's voice

We provide a detailed analysis of the literature, methodology, results, and implications of our findings. In Section 2, we review the literature. Section 3 describes the proposed methodology. In Section 4, we show the results of the model. In Section 5, Limitations were discussed. The study was finally concluded in 6, which also suggests future avenues for research and conclusion.

2. Literature Work

Advances in technology, research in speech-to-speech translation have gained significant attention in real-time voice conversion (RTVC). This led to a diverse range of models and methodologies aimed at transforming source speech into target speech while preserving naturalness and speaker identity. The primary goal is to minimize language barriers and enable seamless communication, allowing individuals to converse in their native language without the need to learn another. Followings are a compilation of notable research on speech-to-speech translation in 2025.

SimulTron is a real-time speech-to-speech translation (S2ST) model based on Translatotron, enhanced with a conformer encoder, wait-k attention, and a streaming vocoder [6]. It processes speech in three stages:

confirming encoding, decoding, streaming vocoding, and TensorFlow Lite for efficient deployment. It achieved a BLEU score of 51.2 by evaluating Spanish and English translation with a 3-second delay, surpassing Translatotron 1. Low latency and smooth translation were confirmed in real-time testing on a Pixel 7 Pro. It further improved BLEU scores but showed slight MOS degradation. SimulTron balances translation accuracy, speed, and efficiency for mobile applications.

DiffuseST is a direct speech-to-speech translation (S2ST) system that incorporates a diffusion-based synthesizer to preserve a zero-shot speaker [7]. It translates multiple languages into English with a smaller, parameter-efficient architecture. The system uses a tokenizer, acoustic encoder, phoneme decoder, and a novel diffusion synthesizer, and it improves audio quality and speaker similarity by over 23%. DiffuseST used the previous NAT synthesizer in audio quality and speaker cloning to achieve low latency and faster inference. Despite a slight accuracy in pronunciation, excels in speaker preservation, setting a foundation for future

The authors proposed trajectory-based and sub-band modelling techniques that improved output quality and processing efficiency [3]. The emergence of neural network-based models, including Deep Neural Networks (DNN), Artificial Neural Networks (ANNs), not favorable in voice conversion, and autoregressive vocoders like WaveNet, enabled end-to-end systems with speaker generalization. Supporting components like Text-to-Speech (TTS) systems and Optical Character Recognition (OCR) have enhanced accessibility, especially for visually impaired users. Tacotron 2 and its multi-speaker extensions, combined speaker encoders and attention mechanisms, allow for zero-shot speaker adaptation, enabling speech synthesis from short reference samples without requiring explicit text alignment while maintaining audio fidelity.

The flexibility of converting not only the voice characteristics but also the pitch contour and duration of input speech, the seq2seq methodology used [8]. Speaker identity conversion experiments and found that ConvS2S-VC obtained higher sound quality and speaker similarity than baseline methods, and was effective for a many-to-many conversion model. A fully convolutional architecture, simultaneously learning mappings, and conditional batch normalization are the main key features of ConvS2S-VC that were used for parallel computing batch normalization and using a single model to convert multiple speakers [4]. Data collected from multiple speakers is used to train and capture the latent features, and used to switch batch normalization layers by the target speaker

Advances in non-parallel voice conversion (VC) have introduced powerful generative models such as StarGAN v2. A many-to-many VC enables without relying on parallel data or text supervision [5]. StarGAN v2 model incorporates key innovations such as a style encoder, F0-consistent loss via a pre-trained JDC network, and a novel adversarial source classifier loss to enhance the similarity of speaker identity [9]. It converts Plain reading speech into stylistic speech, such as emotional and falsetto speech, by the Style encoder and framework, and generalizes well to tasks like cross-lingual and singing VC, despite being trained on limited English-only datasets. Combining multiple loss functions, including speech consistency, cycle consistency, and F0 consistency, the approach produces natural, speaker-similar outputs at near TTS-level quality and operates in real time when paired with fast vocoders like Parallel WaveGAN.

Communication for low latency is a critical technology, such as conferences and live broadcasts [10]. Real-time speech translation generates translated speech while processing. It is challenging due to the need for accurate translation and optimal timing in speech generation. To overcome existing approaches that rely on cascaded models combining automatic speech recognition (ASR), text-based translation, and text-to-speech synthesis suffer from error propagation and hinder joint optimization, StreamSpeech, as a direct Simul-S2ST model, is proposed [11]. StreamSpeech uses a two-pass architecture, first translating source speech into target text and then converting it into target speech [12]. It utilizes multiple connectionist temporal classification (CTC) decoders, optimized through ASR and speech-to-text translation (S2TT) tasks that enhance policy optimization and alignment learning. The model architecture comprises a streaming speech encoder, a simultaneous text decoder, and a synchronized text-to-unit generation module. The model leverages a wait-and-write policy, where it waits for recognized speech text before generating the corresponding translated speech output [13]. Experiments are conducted on the CVSS-C benchmark and include synthesized target speech from CoVoST 2 that achieves state-of-the-art results in both offline and Simul S2ST tasks [14]. It uses a

two-pass architecture autoregressive ASR model for speech-to-text translation with non-autoregressive (NAR) text-to-unit conversion, effectively balancing reordering needs and alignment consistency [15].

Transformer-based Speech-to-Unit Translation (S2UT) model proposed a system that directly translates speech into discrete speech units, bypassing text transcription [16]. The huBERT-based pre-trained model encodes input speech into continuous representations and, before transformer layers into 1D-convolutional layers by using Downsampling. It uses two decoding strategies; the Stacked Strategy generates per decoding step. Reduced Strategy collapses to speed up processing. Intermediate attention, decoder modules, and CTC decoding are used for multitasking learning. To convert discrete units into a waveform Modified HiFi-GAN vocoder is used. FastSpeech 2 is added for reduced unit output, a duration prediction module. Translation quality is evaluated by BLEU scores, and the S2UT reduced model performs with the S2T+TTS system and bridges 63% of the gap between the cascaded ASR+MT+TTS system and Translatotron. speech quality is evaluated by MOS. The S2UT model trained with source text auxiliary tasks achieves 88% of the performance of the model and is trained with both source and target text.The overview of state of the art is listed below in **Table 1.**

Table 1. Quick review of literature work

Reference	Proposed Work	Methods Used	Limitations
[10]	Introduces StreamSpeech, a real-time Simul-S2ST model with a two-stage architecture that performs speech-to-text translation and speech synthesis via text-to-unit conversion.	Two-pass pipeline (Simul-S2TT+Simul-T2U) Multi-task learning (ASR, S2TT, T2U) Streaming encoder CTC decoders Wait-and-write policy, Unit-level vocoder.	Complex architecture remains modular internally, albeit with a Heavy Multi-Task training cost. Synthesis quality is tied to vocoder performance. Requires fine-tuning latency vs quality Limited language scope.
[16]	Proposes a speech-to-text translation model that bypasses text transcription, using discrete units for direct S2ST.	HuBERT audio encoder: Transformer-based unit generator HiFi-GAN vocoder, FastSpeech 2 duration modeling, Multitask CTC & attention training	Slightly lower BLEU/MOS than cascaded ASR+MT+TTS - Reliance on pre-trained speech encoders. Quality is dependent on unit segmentation.
[6]	Presents SimulTron, an on-device real-time S2ST model optimized for mobile deployment (Pixel 7 Pro) with conformer encoder and streaming vocoder.	Conformer encoder, Wait-k attention, Streaming vocoder (Parallel WaveGAN) TF-Lite deployment BLEU/MOS evaluation on Spanish to English.	Slight MOS degradation. Optimized only on limited hardware, Language pair generalizability is unclear.
[7]	Introduces DiffuseST, a diffusion-based direct S2ST model with a small footprint and improved speaker preservation.	Acoustic encoder + phoneme decoder Diffusion-based mel synthesizer Discrete-unit diffusion, Low-latency NAT backbone	Slight pronunciation accuracy drop Diffusion inference remains relatively slow.
[3]	Introduces GenVC, a self-supervised zero-shot voice conversion capable of	Self-supervised speaker/content disentanglement	Early-stage, based on the arXiv preprint

	converting speech across unseen speakers.	Speaker encoder + universal vocoder.	Unknown real-time performance/capacity.
[8]	Proposes BNE-Seq2seqMoL, a bottle-neck feature + MoL attention seq2seq approach for any-to-many and potentially any-to-any VC.	CTC-attention phoneme recognizer Bottleneck feature extractor Seq2seq synth + MoL location-relative attention Downsampling for long attribute alignment.	Temporal alignment complexity Scalability to unseen speakers needs extension via the speaker encoder.
[5]	Presents StarGANv2-VC, a non-parallel, unsupervised many-to-many VC framework with style control and fast inference.	GAN-based VC (StarGAN v2) Style encoder, F0-consistency network - Adversarial source classifier, cycle/speech/F0/style consistency losses, Parallel WaveGAN vocoder.	Monolingual (English) training only Slight MOS degradation Style extraction consistency challenges.

3. Methodology

The current research work proposes a voice-to-voice translation system without parallel speech data between the source and target languages. We sued two methodologies for voice translation.

1. Real-Time Bilingual V2V Translation Model Using RNN
2. Real-Time Multi-lingual V2V Translation Model using Cloud Cascade Pipelines

3.1. Real-Time Bilingual V2V Translation Model Using RNN

A custom RNN multi-layer feedforward neural network classifier was designed to classify English and Urdu speech categories, performing a binary classification based on Wav2Vec 2.0 embedding [17]. Input layer aligned with the dimensionality of the extracted embedding, followed by a series of fully connected hidden layers with progressively decreasing dimensionalities: 512, 256, 128, 64, 32, and 16 units. Rectified Linear Unit (ReLU) activation function is used in each hidden layer and incorporates a dropout rate of 30% to mitigate overfitting and enhance generalization. The final output layer has two neurons, corresponding to the binary classification task (English vs. Urdu), and does not apply an activation function. The deep and comprehensive structure of the network is specifically designed to extract and encode high-level features from the pre-trained speech embedding effectively. A brief description regarding RNN layers and parameters shown in **Figure 2.**

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Linear-1                  [-1, 512]         393,728
              ReLU-2                  [-1, 512]               0
           Dropout-3                  [-1, 512]               0
            Linear-4                  [-1, 256]         131,328
              ReLU-5                  [-1, 256]               0
           Dropout-6                  [-1, 256]               0
            Linear-7                  [-1, 128]          32,896
              ReLU-8                  [-1, 128]               0
           Dropout-9                  [-1, 128]               0
           Linear-10                   [-1, 64]           8,256
             ReLU-11                   [-1, 64]               0
          Dropout-12                   [-1, 64]               0
           Linear-13                   [-1, 32]           2,080
             ReLU-14                   [-1, 32]               0
          Dropout-15                   [-1, 32]               0
           Linear-16                   [-1, 16]             528
             ReLU-17                   [-1, 16]               0
          Dropout-18                   [-1, 16]               0
           Linear-19                    [-1, 2]              34
================================================================
```

Figure 1. RNN classifier

3.1.1. Data Collection

Data is the power in the world of technology; authentic and reliable data empowers the system. It is the building block of every intelligent technology, whether machine learning/ Deep Learning. In this study combination of wav2vec2_fairseq_base_ls960 [18] and a custom dataset is used to train the model. It is trained on 19 hours of unlabeled speech from the LibriSpeech dataset using the Fairseq framework. It is a self-supervised learning model developed by Facebook AI. It is used to extract rich, high-level acoustic features from both Urdu and English speech samples. The model's robustness to noise and its efficiency on low-resource tasks made it particularly suitable for our bilingual dataset. The custom dataset is recorded by the voice recorder, and each language contains only 101 recorded files, which is not sufficient to train the model. The vector of extracted features was labeled as English or Urdu and combined into a single dataset. The dataset was split into training and testing sets using an 80:20 stratified split to ensure balanced class representation in both subsets.

For real-time voice translation on multi-lingual communication, cloud cascade pipelines are used. As the Google translation library GTL is authentic and supports maximum language datasets and offers an accuracy of 90% and for less common languages, it has an accuracy less than the other. It almost preserves 82% overall meaning of translations and translates sentences in sequence.

3.1.2. Data Preprocessing

English and Urdu speech samples are stored in separate parallel corpus directories. Initially dataset contains .m4a audio files which were converted to .wav format. To maintain consistency and model compatibility, the .m4a files were deleted after conversion. Each audio file was sampled at 16 kHz and standardized to a maximum duration of 3 seconds. Longer files were truncated to ensure fixed input dimensions across the dataset, and shorter files were zero-padded.

3.1.3. Noise Reduction

To improve the quality of the input signal, remove background noise from the audio files. The de-noising enhances the clarity of features extracted from speech signals, which is critical in noisy environments or for languages like Urdu, where phonemes may be subtle.

3.1.4. Extract Features

This study employed Facebook's pre-trained Wav2Vec 2.0 Base model to extract deep acoustic representations from the cleaned audio waveforms [18]. The model transforms the waveform into a contextual embedding. For each audio sample, the mean of output representations of the Wav2Vec 2.0 across the time dimension was computed, resulting in a fixed-size embedding per utterance.

3.2. Real-Time Multi-lingual V2V Translation Model using Cloud Cascade Pipelines

In proposed methodology we used cloud cascade pipelines to support multilingual V2V translation. Cloud cascade pipelines are used for effective and accurate translation of regional languages [19]. Deep Translator works like a wrapper to access Google Translate. ASR+TTS and noise remover are used to effectively translate voice. By using four famous Pakistani languages and one international language, it provides given translations

1. English to Urdu
2. English to Punjabi
3. English to Sindhi
4. English to Balochi

We adopt a pre-trained sequence-to-sequence model that uses ASR [1] for speech recognition, STT to convert voice into text data, and TTS to convert text data into voice. Under the assumption that the representation encoded by the language-agnostic speech encoder is independent of language, we train our model using only target language data without language-parallel data. We can predict the target language from the speech of the source language because the representation encoded from the speech of the source language also contains information about the speech, irrespective of language. The framework of proposed Cloud Cascade pipelines is shown in Figure 2.

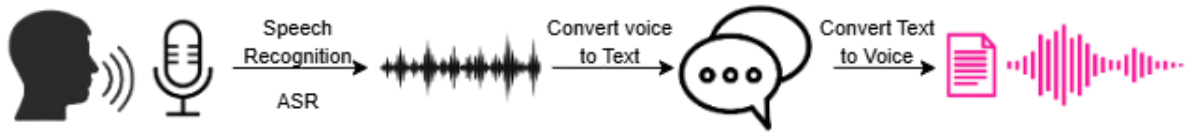

Figure 2. Cloud Cascade Pipeline for multi-lingual V2V Translation

3.2.1. Automated Speech Recognition

Voice data is input through the microphone or headset. The ASR model is used for Speech recognition as it provides the facility to speak into an external microphone, headset, or built-in microphone. After audio capturing, the recordings are pre-processed to improve the accuracy and quality of recognition [1]. This involves removing background noises and irrelevant frequencies, stabilizing the volume level, and converting the audio file into a standard format. The original and denoise waveform and spectrogram of given English and Urdu voices are shown in Figure 3 and Figure 4 respectively.

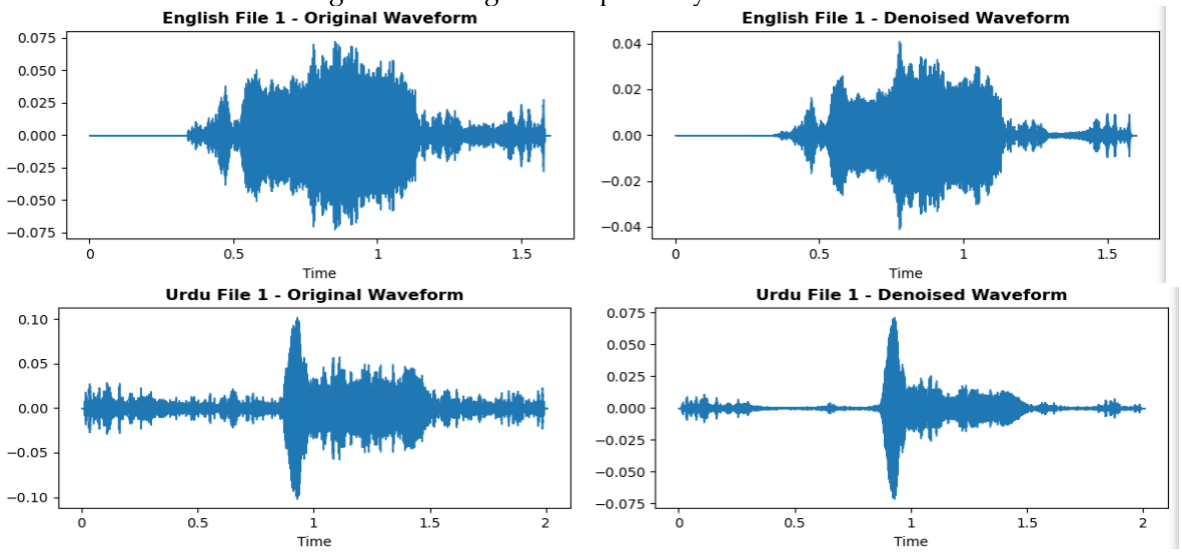

Figure 2. Original and De-Noise waveform of both Urdu and English voices

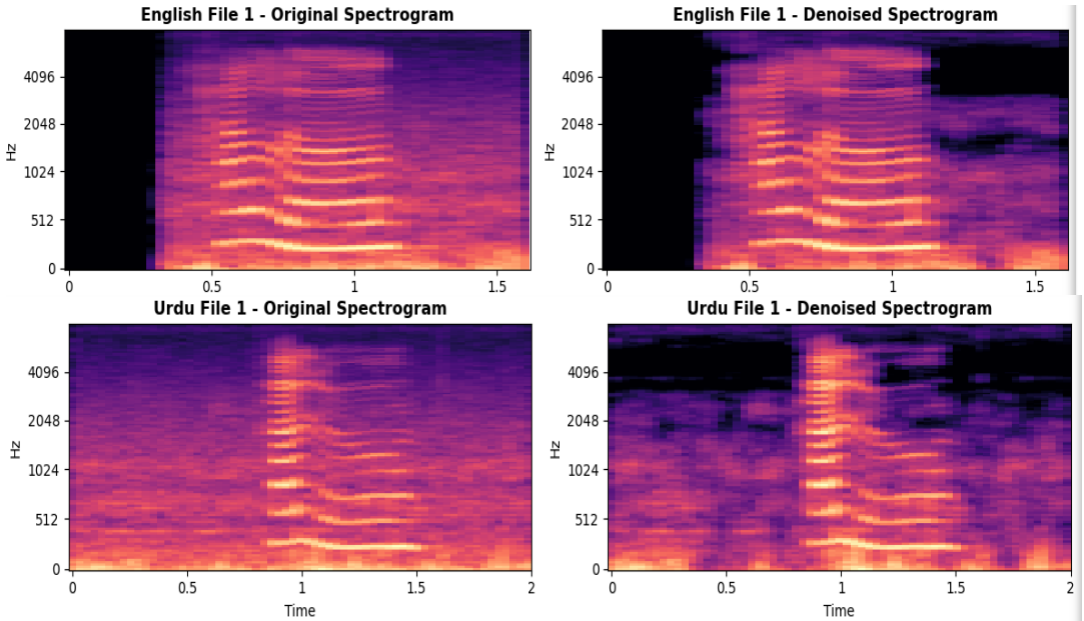

Figure 3. Spectrogram of both the original and de-noised voice of Urdu and English

3.2.2. Feature Extraction

In this study, Mel-Frequency Cepstral Coefficients (MFCCs) are used for feature extraction, which are widely recognized for their effectiveness in speech and audio processing. Each audio file, English or Urdu, first passes through a noise reduction algorithm to remove background noise and improve clarity in the voice. The de-noised signals compute 40 MFCC features that capture essential aspects of the sounds, such as shape and dynamics of the vocal tract, which are crucial for distinguishing between different phonemes and speaker characteristics. The extracted features are normalized by subtracting the mean and dividing by the standard deviation, which enhances the training stability of deep learning models. It effectively transforms raw audio signals into structured, compact numerical representations that preserve linguistic and speaker-relevant information, enabling accurate language classification and real-time translation in subsequent stages of the pipeline.

3.2.3. Translating Text into Speech

For translating speech into text most frequently used module is STT It's sometimes called "read aloud" technology. The purpose of this model is to maintain the user's emotions, feelings, and tone and ensure privacy as data is not stored [2]. After the translation data is removed from the model. Textual information originates from text-based communications or equipment display readouts. Text-based examples of communications that could be spoken via TTS include news, email, rich site summary (RSS), Web, Web logs (blogs), Internet relay chat (IRC), instant messaging (IM), and short message service (SMS). Its synthesized voice types (e.g., male and female) are usually employed to convey various types of information. For example, routine or urgent information can be conveyed in male or female voices, respectively.

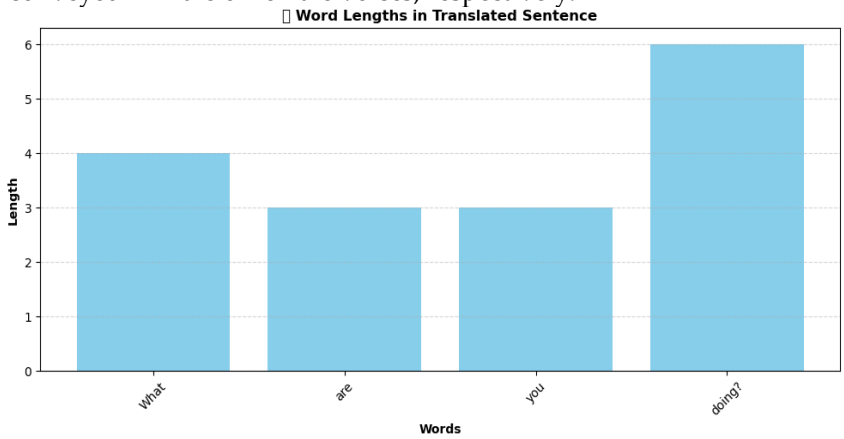

Figure 4. Word length in translated sentence

4. Results and Discussion

This section presents the experimental results obtained from evaluating both custom and Wav2Vec 2.0 Base models. The bilingual v2v translation is carried out with deep learning based RNN model. The classification report of trained RNN is shown in Figure 6. Report showed that RNN model is trained with for both languages English and Urdu with 95% and 86% precision, 86% and 95% recall and with overall 90% accuracy.

Detailed Classification Report:

	precision	recall	f1-score	support
English	0.95	0.86	0.90	21
Urdu	0.86	0.95	0.90	20
accuracy			0.90	41
macro avg	0.91	0.90	0.90	41
weighted avg	0.91	0.90	0.90	41

Figure 6. RNN Classification Report

Epoch-wise plots of both accuracy and loss for the training and validation sets were generated and shown in Figure 7. The model shows steadily increasing training accuracy and decreasing training loss over epochs. However, fluctuations in test accuracy and loss suggest possible overfitting and limited generalization to unseen data. A 2×2 confusion matrix was generated to visualize the distribution of predictions across the two language classes (English and Urdu), offering insight into the types of errors made by the model as shown in Figure 8. The confusion matrix shows that the model accurately classified most English and Urdu samples, with only 4 misclassifications out of 41. It achieved high performance, correctly predicting 18 English and 19 Urdu instances.

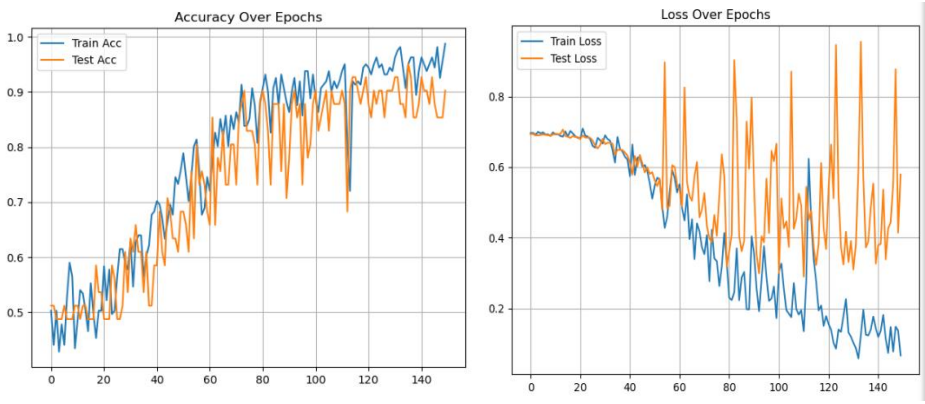

Figure 7. Training and validation accuracy and loss curves for the RNN classifier Embedding model.

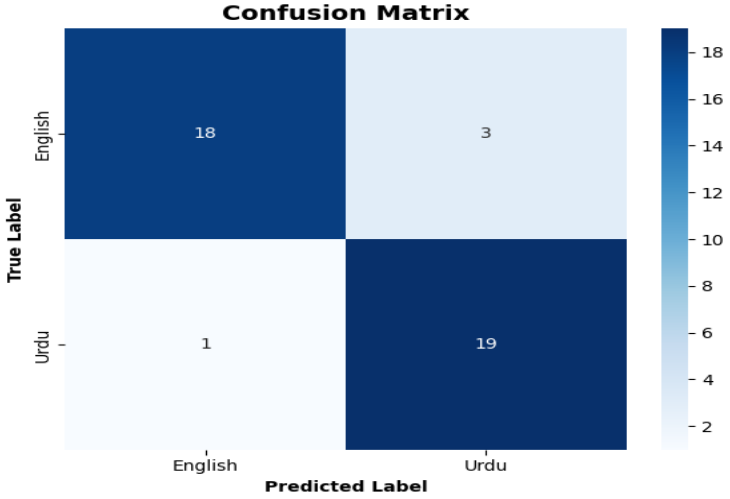

Figure 8. Confusion matrix of Urdu and English showing prediction accuracy per class.

The performance of both bilingual RNN and Multilingual Cloud Cascade pipeline method is analyzed based on experiments, we give multiple inputs to both models to check their performance and we used multiple evaluation metrics, including accuracy, precision and recall to compare both models.

Table 2. Evaluation Metrics of Voice-to-voice translation

Model	Language	Sample	TP True Translation	TN No Output	FP Noisy	FN Wrong Translation	Precision Tp=Tp/Tp+Fp	Recall	Accuracy
RNN classifier	Urdu	20	19	18	1	3	0.8636	0.95	0.9024
	English	21	18	19	3	1	0.8636	0.95	0.9024
	Urdu	95	82	392	13	13	0.8632	0.8632	0.948

Cloud-cascade pipeline	English	109	82	373	18	20	0.8318	0.8165	0.924
	Sindhi	112	96	373	16	16	0.8571	0.8571	0.938
	Punjabi	91	76	392	15	15	0.8352	0.8352	0.936
	Balochi	93	74	387	20	19	0.7872	0.7957	0.922

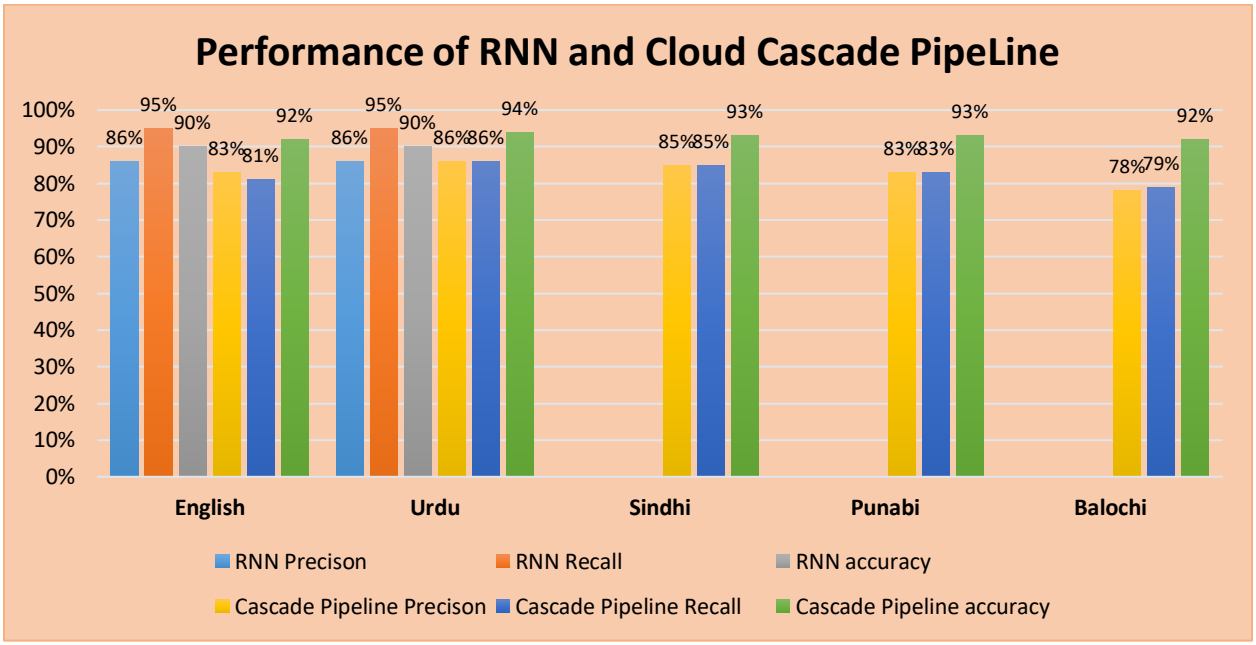

Figure 5. Performance Metrics by Cascade pipeline

The performance of the proposed RNN bilingual V2V translation model, and multilingual cascade pipeline model is illustrated in Figure 9. The RNN model achieved high performance for English and Urdu, with 86% precision, 95% recall, and 90% accuracy, showing strong recognition in limited languages. In contrast, the Cascade Pipeline performed well across all five languages, maintaining precision and recall above 78%, and accuracy above 92% in most cases. This suggests that while RNN excels in binary classification, the Cascade Pipeline offers more consistent performance in multilingual scenarios.

5. Limitation

This model is trained on a combined built-in and custom dataset. Despite the encouraging results, to perform critical operations current system architecture is dependent on an active internet connection, such as Automatic Speech Recognition (ASR) via the Google Speech Recognition engine, and text-to-speech translation and synthesis using gTTS. It relies on cloud-based services may restrict the system's applicability in environments with limited or unstable internet connectivity.

6. Conclusion and Future Work

Dealing with Natural Languages is an interesting and challenging task of Natural Language Processing. Especially in the case of a lack of resources. The experimental findings of this study validate the effectiveness of the proposed cloud-based cascade speech translation system in facilitating real-time multilingual communication, specifically between Urdu and English. I explore voice-to-voice translation by Speech Recognition (ASR), translate voice into text, and then text into voice. Remove noise from the input data voice. The system achieves high translation accuracy and maintains low latency in most use cases. This system provides a reliable and efficient solution for voice-to-voice translation in communication. This work significantly minimizes the gaps in Real-time voice-to-voice communication; however, it introduces a dependency on stable internet connectivity. In the future, improve the system's ability to handle code-switched

speech, regional dialects, and accent variability, which currently pose challenges in recognition and translation accuracy. The system may be extended through the incorporation of fully offline components to enhance robustness and portability, and to include more languages and enhance work from Real-Time Voice-to-Voice translation.

References

1. T. O. YA Fonggi, "Analysis of voice recognition system on Translator for daily use," in Engineering, Mathematics and Computer Science Journal (EMACS), 2021•journal.binus.ac.id, 2021.
2. K. A. K. J. AS Deshpande, "Voice to Voice Language Translation System," in International Journal of …, 2014 - academia.edu, 2014.
3. H. L. X. A. G. L. P. G.-P. K. D. Zexin Cai, "GenVC: Self-Supervised Zero-Shot," in arXiv:2502.04519v1 [eess.AS], 2025.
4. S. M. I. Y. C. S. M. I. D. Songxiang Liu, "Any-to-Many Voice Conversion with," in arXiv:2009.02725v3 [eess.AS] 23 May 2021, 2021.
5. A. Z. N. M. Yinghao Aaron Li, "StarGANv2-VC: A Diverse, Unsupervised, Non-parallel Framework for," in arXiv:2107.10394v2 [cs.SD] 2, 2021.
6. E. N. O. R. Y. D. Y. J. N. B. H. Z. M. R. A Agranovich, "SimulTron: On-Device Simultaneous Speech to Speech Translation," in ICASSP 2025-2025 IEEE International Conference on Acoustics …, 2025•ieeexplore.ieee.org, 2025.
7. X. Y. M. N. J. L. E. D. D. L. N. T. C. S. K. S. N Hirschkind, "Diffusion Synthesizer for Efficient Multilingual Speech to Speech Translation," in arXiv preprint arXiv:2406.10223, 2024•arxiv.org, 2024.
8. H. Kameoka, K. Tanaka, D. Kwaśny, T. Kaneko and N. Hojo, "ConvS2S-VC: Fully Convolutional Sequence-to-Sequence Voice Conversion," in IEEE, 2020.
9. B. Sisman, J. Yamagishi, S. King and H. Li, "An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning," in IEEE, 2020.
10. Q. F. S. G. Z. M. M. Z. Y. F. S Zhang, "Streamspeech: Simultaneous speech-to-speech translation with multi-task learning," in arXiv preprint arXiv:2406.03049, 2024•arxiv.org, 2024.
11. Y. F. S Zhang, "Unified segment-to-segment framework for simultaneous sequence generation," in Advances in Neural Information Processing Systems, 2023•proceedings.neurips.cc, 2023.
12. S. P. I. K. P. C. C. W. Y. C. Y. T. A. L. S. W. H Inaguma, "Unity: Two-pass direct speech-to-speech translation with discrete units," in arXiv preprint arXiv:2212.08055, 2022•arxiv.org, 2022.
13. G. N. K. C. V. L. J Gu, "Learning to translate in real-time with neural machine translation," in arXiv preprint arXiv:1610.00388, 2016•arxiv.org, 2016.
14. A. W. J. P. C Wang, "Covost 2 and massively multilingual speech-to-text translation," in arXiv preprint arXiv:2007.10310, 2020•arxiv.org, 2020.
15. Y. Z. Y. F. Q Fang, "Daspeech: Directed acyclic transformer for fast and high-quality speech-to-speech translation," in Advances in Neural Information Processing Systems, 2023•proceedings.neurips.cc, 2023.
16. P. C. C. W. J. G. S. P. X. M. A. P. Y. A. Q. H. Y. T. J. P. A Lee, "Direct speech-to-speech translation with discrete units," in arXiv preprint arXiv:2107.05604, 2021•arxiv.org, 2021.
17. M. R. A. I. P. B. D. G. SK Hossain, " Enhancing Bangla Local Speech-to-Text Conversion Using Fine-Tuning Wav2vec 2.0 with OpenSLR and Self-Compiled Datasets Through Transfer Learning," in ieomsociety.org, 2024.
18. Y. Z. A. M. M. A. A Baevski, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Advances in neural information processing systems, 2020•proceedings.neurips.cc, 2020.
19. P. P. MDR Athas, "CallTran: Voice Translation for End-to-End Communication over the Internet," in 2024 Second International …, 2024 - ieeexplore.ieee.org, 2024.
20. P. S. Ranaa, "Advancements in Real-Time Voice Conversion Technologies: A," 2024.
21. S. M. I. Y. C. S. M. I. D. Songxiang Liu, "Any-to-Many Voice Conversion with," in arXiv:2009.02725v3 [eess.AS] 23 May 2021, 2021.
22. I. V. V. G. S. M. K. &. J. W. Vadim Popov, "DIFFUSION-BASED VOICE CONVERSION WITH FAST," in arXiv:2109.13821v2 [cs.SD], 2022.
23. Q. F. S. G. Z. M. M. Z. Y. F. S Zhang, "Streamspeech: Simultaneous speech-to-speech translation with multi-task learning," in arXiv preprint arXiv:2406.03049, 2024•arxiv.org, 2024.

24. R. Lavanya, A. K. Gautam and A. Anand, "Real Time Translator with Added Features for Cross Language Communiciation," in IEEE, 2024.
25. M. Nikitha1, "Voice Translator," in International Journal of Innovative Research in Engineering, 2024.