2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0 | 979-8-3503-7378-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/OTCON60325.2024.10687437

Overcoming the barrier of emotion in interlingual transcription: A case study of Malayalam to English transcription using convolutional neural networks

Sandra John
Department of Computing
Sir David Bell Building
Digby Stuart College
University of Roehampton
London, UK

Abdul Khader Abdul Vahid
Department of Computing
Sir David Bell Building
Digby Stuart College
University of Roehampton
London, UK

Lakshmi Radhakrishnan Nair
Department of Computing
Sir David Bell Building
Digby Stuart College
University of Roehampton
London, UK

Mohammad Farhan Khan
Department of Computing
Sir David Bell Building
Digby Stuart College
University of Roehampton
London, UK

Fakhreldin Saeed
Department of Computing
Sir David Bell Building
Digby Stuart College
University of Roehampton
London, UK

***Abstract*—Speech-to-text translation systems have demonstrated that the quality of training data has a significant effect on the system's performance. Therefore, a lack of high-quality speech data could make it more difficult for the systems to use the direct speech-to-text architecture to effectively learn the mapping process between audio and text. Direct speech-to-text translation systems, unless they make use of feature engineering resources or methodologies, have struggled to attain comparable performance as their counterparts. The effectiveness of using higher dimensional feature space with respect to direct voice translation systems is compared in this paper. Investigations have revealed that using a spectrogram matrix constituting high dimensional feature space has a more robust tendency to deal with real-world scenarios like emotions and accents in the speech dataset.**

Keywords—Machine translation, speech-to-text translation, data pre-processing, spectrograms

I. Introduction

The advancement of neural networks (NN) has greatly influenced the development of end-to-end speech-to-text translation systems [1]–[3]. The foundation of end-to-end speech-to-text translation systems lies in the direct translation of source speech to target text without involving any additional cascading pipeline in the system [4], [5]. In contrast, the conventional speech-to-text translation techniques lie in an interconnected cascade pipeline of automatic speech recognition with a machine translation process. The end-to-end model has a simpler architecture than the cascaded pipeline, however, it is difficult to obtain a large amount of synchronised speech-to-text data for the training process, especially data for interlingual transcription [6].

Interlingual transcription is the process of translating speech from one language into text from another language. To accomplish the complex process of interlingual transcription, a sophisticated pipeline of intelligent models is required which can automatically recognise the speech and perform a textual machine translation [7]. In a real-time interlingual transcription, several issues can impact the accuracy of the overall system. Examples include the accent of the speaker, noise in the background, sampling rate, etc.

In general, the machine transcription process uses the speech data coupled with the transcriptions, which can be used for translating the speech to the target language [8]. As the process directs, most of the research has focused on generating large amounts of data to train the transcription models. However, the main limitation of generating the data for the transcription process leads to the enormous amount of manual tagging of the data. In addition, the issue like the accent of the speaker increases the additional burden on the necessary standard tagging process.

The availability of high-quality noise-free data is a major obstacle in the development of a machine transcription pipeline. Most of the well-known automated speech translation systems are developed using convolutional neural network (CNN) or deep learning (DL) algorithms which are resource intensive. The models that have been trained over the data which is acquired in controlled environments, such as recording studios give better performance. However, such controlled models usually fail to acquire reasonable transcription in real-world scenarios, where the speech data suffers from background noise. The real-world situation makes it difficult for the controlled CNN/DL models to capture contextual under-

979-8-3503-7378-3/24/$31.00 ©2024 IEEE

standing and map speech representations to their translated equivalents.

A well-known Google speech-to-text algorithm uses large deep learning models to efficiently transcribe English texts. However, there is no known satisfactory algorithm available to perform an interlingual transcription such as from non-English to English conversion. Berard *et al.* [9] developed a long short-term memory network for French speech-to-English text conversion by using synthetic speech data. The likely reason behind opting for synthetic speech data is to overcome the barrier of noise and accent. Due to these reasons, Berard et al. model is not able to function in real-world scenarios. A similar methodology has been adopted by Bansal *et al.* [10] by using pseudo text to perform Spanish speech-to-English text conversion.

In contrast to Bansal *et al.*, a more realistic Spanish speech-to-English text conversion model has been proposed by Duong *et al.* [11] which has used a telephonic conversion as a source of data and treated speech as a sequence of 39-dimensional perceptual linear prediction vector. The algorithm proposed by Ghadage and Shelke [12] has used a simple approach to address the data insufficiency issue by training a model by using the Mel-frequency cepstral coefficient feature extraction technique, and minimum distance classifier along with a support vector machine algorithm for transcribing Marathi speech-to-English text.

It is important to highlight that speech-to-text translation systems have shown that the eminence of training data significantly impacts the system's functionality. Thus, insufficient high-quality speech data may deteriorate the systems' ability to learn the audio-to-text mapping process using the direct speech-to-text architecture. Without the assistance of feature engineering techniques or resources, direct speech-to-text translation systems have had difficulty achieving expected performance levels. Hence, this paper compares the efficacy of direct speech translation systems with higher dimensional feature space, and its role in dealing with real-world issues such as emotions and the accent of the speaker.

The reminder of this paper is organized as follows: The approach adopted for acquiring speech data along its cleaning process, and an overview of 1D/2D CNNs are discussed in Section II. The pre-processing steps involved in this study to enhance the accuracy of the CNN models are reported in detail in Section III, followed by a case study and concluding remarks in Sections IV and V respectively.

II. Materials and Methods

A. Speech data acquisition

The proposed study employs a dataset of Malayalam audio recordings that were collected through a mobile device without using any dedicated specialised mic. A targeted data collection strategy has been adopted which records spoken days from native Malayalam speakers and the data was stored in open-source (.ogg) format. For uniformity, a maximum audio length of two seconds per recording is considered, which we found is enough to speak the name of any of the days in Malayalam by native speakers. The dataset is comprised of 871 speech samples of male and female that was strategically collected so that it can be heterogeneous in nature, containing the best possible range of Malayalam accents mixed with emotions taken from various age groups of volunteers ranging between 16-45 years. Table I represents the pronunciation of the name of all English days in Malayalam language and its respective category/class ID with total sample size, along with its descriptive statistics in Fig. 1.

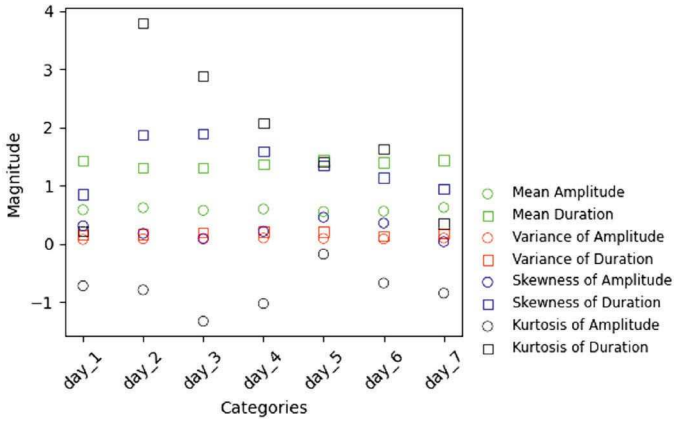

Fig. 1. Descriptive statistics of each data category in terms of amplitude and duration.

B. Data cleaning

Cleaning speech data for machine learning (ML) applications is crucial for several reasons. The first and foremost is that it enhances the quality of the data, ensuring that the ML models receive input that is free from noise, distortions, and irrelevant information. Clean data leads to more accurate and reliable model predictions, as the models can focus on learning meaningful patterns within the data. Additionally, clean audio data facilitates the extraction of relevant features, such as spectral characteristics or temporal patterns, which are essential for effective model training and prediction. To enhance the accuracy of speech transcription two types of enhancement procedures have been adopted, that are silence removal and outlier elimination. Both the procedures are discussed in the following two subsections.

1) Silence removal (SiR): Silence removal plays a crucial role in audio transcription by enhancing the accuracy and efficiency of the transcription process. Silence in audio recordings frequently denotes inactive or silent intervals; these can include pauses, background noise, or non-speech sounds. By eliminating these silent stretches before transcription, the computational load is lessened, and transcription algorithm efficiency is enhanced by concentrating on significant speech content [13]. Hence, in this work, silence removal is adopted as main pre-processing task that is adopted to reduce the enhance the transcription efficiency. To remove the silence part in the dataset, the amplitude of the beginning and end of the speech is clipped if the voice reaches 20% of the mean amplitude of each spoken word. Fig. 2 illustrates the removal of silence

TABLE I
SAMPLE SIZE OF DATASE BEFORE AND AFTER OUTLIER ELIMINATION: ENGLISH TO MALAYALAM PRONUNCIATION

Days	Days in Malayalam	Category ID	No. of samples collected	No. of samples after outlier elimination
Monday	Thinkal	day_1	102	98
Tuesday	Chovva	day_2	89	86
Wednesday	Budhan	day_3	76	74
Thursday	Vyaazham	day_4	103	99
Friday	Velli	day_5	181	173
Saturday	Shani	day_6	141	134
Sunday	Njaayar	day_7	179	165

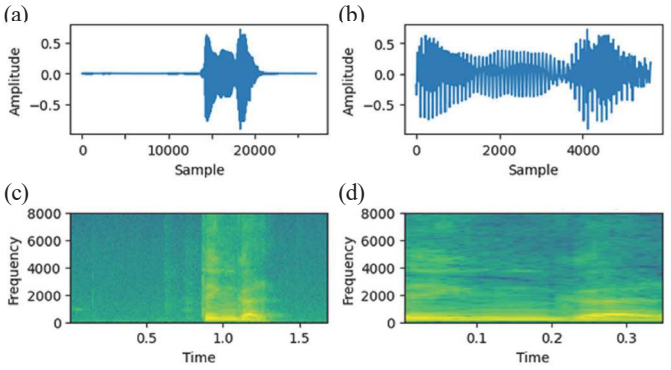

Fig. 2. Impact of silence removal on time duration and spectrogram. (a) Original speech sample, (b) Trimming of original speech sample by removing the silence part, (c) Spectrogram of original speech sample, and (d) Spectrogram of trimmed speech sample.

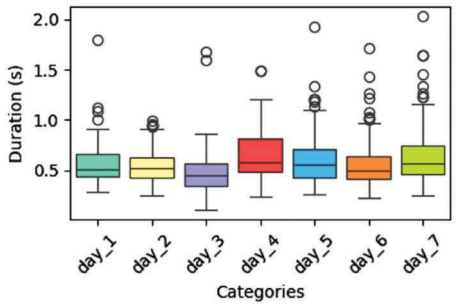

Fig. 3. Boxplots of each category after removing silence from original dataset. The outliers represented by small circles are falling outside 1.5 times the interquartile range of time duration.

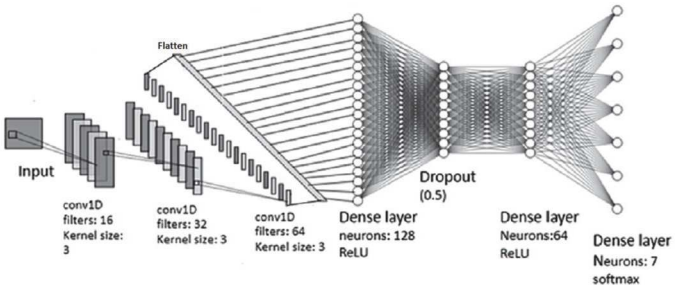

Fig. 4. Architecture of 1D CNN considering temporal speech vector as input voice sample. Each convolutional layer is followed by a max-pooling layer with a pool size of 2.

(trimming) from the speech samples and its impact on time duration and spread of frequencies within spectrogram.

2) Outlier elimination (OE): In general, ML algorithms are designed to learn patterns and relationships within the data, but outliers can skew these patterns, resulting in models that are comparatively less representative of the true underlying distribution. Eliminating outliers help improving the generalisation ability of ML models, ensuring that they perform well on unseen data by reducing the impact of anomalies that are not representative of typical scenarios [14]. As a result, in this work, trimmed speech samples that fall outside of 1.5 times the interquartile range, that is, above and below the quartile of the data distribution are considered outliers and are eliminated from the dataset [15]. Fig. 3 illustrates the boxplot for each category after removing silence for the data, in which small circles represents the outliers in terms of speech duration, and Table I describes the number of remaining speech samples after outlier elimination for each category.

C. 1D/2D convolutional neural networks

The 1D convolutional neural networks (1D CNNs) play a crucial role in the classification of speech data by effectively extracting local features from sequential information. The utilisation of convolutional filters enables the identification of patterns associated with acoustic features as they traverse the input speech signal. This capability allows 1D CNNs to demonstrate translation invariance, ensuring the identification of relevant features regardless of their specific timing within the input signal [16].

The hierarchical feature representation achieved through multiple convolutional layers proves beneficial in capturing intricate structures inherent in speech data, such as phonetic combinations and intonations. Furthermore, the temporal hierarchies and long-term dependencies present in speech signals are addressed, and pooling layers contribute to temporal abstraction, emphasising significant patterns while reducing dimensionality. The adaptability of 1D CNNs to various speech-related tasks, including speaker identification, emotion recognition, and speech command recognition, highlights their versatility [17]. Fig. 4 illustrates the architecture of 1D CNN that is adopted in this study which is comprised of series of convolutional layers followed by fully connected and output layers.

On the other hand, 2D Convolutional Neural Networks (2D CNNs) are instrumental in classifying speech data by effectively leveraging their ability to capture spatial and

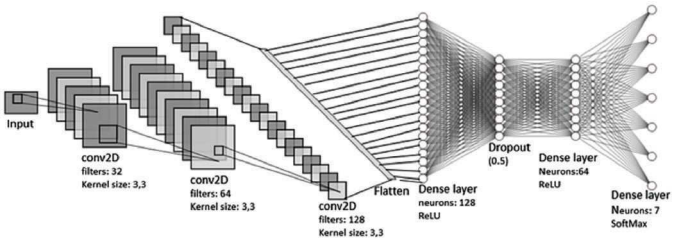

Fig. 5. Architecture of 2D CNN considering spectrogram matrix as input voice sample. Each convolutional layer is followed by a max-pooling layer with a pool size of 2×2.

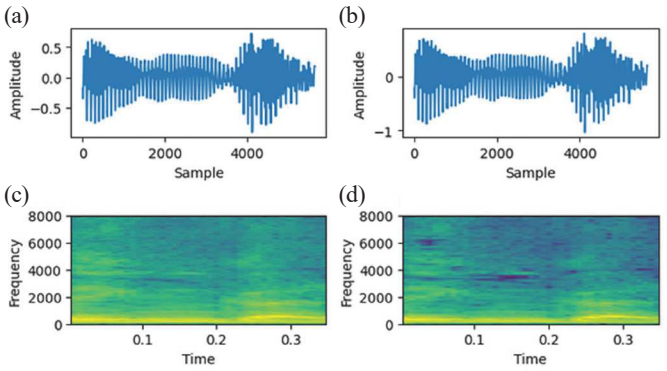

Fig. 6. Impact of speech normalisation on the spectrogram matrix. (a) Silence-removed speech sample, (b) Amplitude normalisation of silence-removed speech sample, (c) Spectrogram of original speech sample, and (d) Spectrogram of silenced-removed amplitude-normalised speech sample.

temporal features simultaneously. Through the utilisation of convolutional filters applied to spectrograms or other time-frequency representations of the audio signal, 2D CNNs excel in identifying patterns associated with both the frequency content and temporal evolution of the speech signal. This capability allows them to extract rich feature representations that capture both spectral and temporal characteristics of the speech data [18]. Fig. 5 illustrates a similar type of architecture of 2D CNN that is adopted in this study comprising series of convolutional layers followed by fully connected and output layers.

In contrast to 1D CNNs, which primarily focus on the temporal dimension of the data, 2D CNNs consider both the time and frequency axes, enabling them to capture a more comprehensive set of features. Additionally, 2D CNNs exhibit translation invariance, ensuring that relevant features can be identified across different positions within the spectrogram. This property allows them to effectively handle variations in speech patterns and speaker characteristics [19], [20].

III. Results and Discussion

As stated earlier, the original dataset was comprised of 871 speech samples of male and female that was strategically collected so that it can be heterogeneous in nature, containing the best possible range of Malayalam accents with different emotions so that it reflects the inherent richness of spoken language. The integration of emotions in the dataset increases the complexity of dataset which can dealt by using suitable pre-processing steps. To investigate the impact of speech emotions without performing any of the data cleaning procedure (namely silence removal and outlier elimination), the original speech vector and spectrogram matrix are directly feeded into 1D and 2D CNN models respectively.

Tables II and III demonstrate the highest attainable performance of the CNN models when different type of datasets are feeded as an input. Observing second row of Tables II and III, it can asserted that emotions are easily manageable when it is converted to spectrogram matrix and feeded into 2D CNN model. In addition the third and fourth row in Tables II and III represent the impact of data cleaning steps on the performance of the CNN models. In the case of 1D and 2D CNNs, the accuracy has been improved by performing the data cleaning steps; which suggest the importance of removing silence and outliers from the dataset. To further investigate the impact of preprocessing steps in the performance of the CNN models, two speech enhancement techniques have been adopted. The first one is amplitude normalisation, while another one is spectral reduction. Both the enhancement steps are discussed in the following sub-sections.

A. Amplitude normalisation (AN)

Amplitude normalisation ensures consistency in the amplitude range across different audio samples, preventing biases in model training caused by variations in signal strength. The ML models often rely on input features being on a similar scale to learn effectively, hence normalising the amplitude helps achieving this by scaling all audio samples to a common range, typically in range of [-1, 1]. Additionally, amplitude normalisation facilitates comparison and integration of audio data from diverse sources or recording environments, as it standardises the representation of speech data regardless of their original amplitude range.

Fig. 6 illustrates the amplitude normalisation of silence-removed speech samples and its impact on the spread of frequencies within spectrogram. Observing Figs. 6(c) and (d), it can be asserted that the high magnitude frequencies lower than 2×10^3 Hz are very mildly affected compared to higher frequency components, suggesting the importance of spectrogram signature of spoken words in lower frequency range. In contrast, the fifth row of Tables II and III represents the importance of amplitude normalisation for both the cases when speech is considered as a vector and spectrogram matrix. As expected the performance of 1D CNN has further improved due to removal of bias which was present earlier in the form of signal strength. However its impact on 2D CNN is not visible due to the existence of similar signatures at lower frequency levels (refer Figs. 6(c) and (d)).

B. Spectral reduction (S_PR)

Spectral reduction of audio data is also one of the important procedure in developing a more generalised ML models. Spectral reduction helps in reducing the dimensionality of the data while preserving relevant information to best possible extent.

TABLE II
PERFORMANCE METRICS OF 1D CNN BY CONSIDERING INPUT DATA AS SPEECH VECTOR

Dataset preprocessing	Preprocessing category	Accuracy	Precision	Recall	F1 Score
Original	No preprocessing	39.08	50.93	36.13	37.37
Silence removal (S_iR)	Data cleaning	49.43	50.24	46.76	47.06
S_iR + Outlier elimination (OE)	Data cleaning	49.70	60.36	45.69	47.62
S_iR + OE + Amplitude normalisation (AN)	Data enhancement	53.89	57.29	50.32	51.63
S_iR + OE + AN + Spectral reduction (S_PR)	Data enhancement	66.47	70.42	63.74	65.50

TABLE III
PERFORMANCE METRICS OF 2D CNN BY CONSIDERING INPUT DATA AS SPECTROGRAM MATRIX

Dataset preprocessing	Preprocessing category	Accuracy	Precision	Recall	F1 Score
Original	No preprocessing	97.70	99.78	99.77	99.77
Silence removed (S_iR)	Data cleaning	98.27	98.77	98.74	98.73
S_iR + Outlier elimination (OE)	Data cleaning	100.00	99.39	99.29	99.30
S_iR + OE + Amplitude normalisation (AN)	Data enhancement	99.40	99.55	99.52	99.52
S_iR + OE + AN + Spectral reduction (S_PR)	Data enhancement	98.86	99.24	99.21	99.21

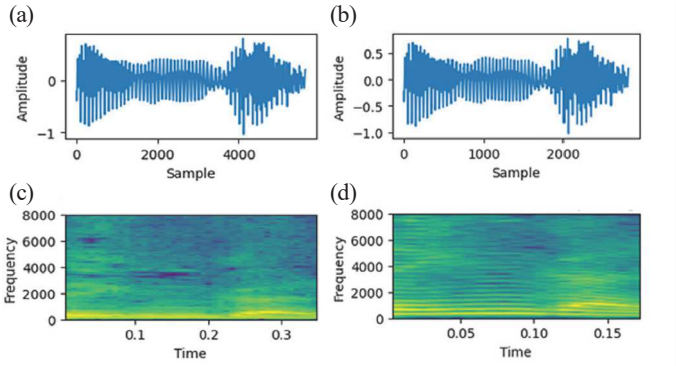

Fig. 7. Impact of spectral reduction on the spectrogram matrix. (a) Silence-removed amplitude-normalised speech sample, (b) Spectral reduction of silence-removed amplitude-normalised speech sample, (c) Spectrogram of original speech sample, and (d) Spectrogram of silence-removed amplitude-normalised spectrally-reduced speech sample.

Speech data typically contain a large number of frequency components, especially in the case of high-resolution spectrograms. Spectral reduction techniques such as principal component analysis help in capturing the most discriminative features while discarding redundant or less relevant information. By reducing the dimensionality of the data, spectral reduction may enhance the efficiency of ML algorithms, as models can be trained more quickly and require less computational resources. By reducing the spectral dimensionality, models become more robust to variations and noise in the input data, leading to better performance on unseen data.

Observing Figs. 7(c) and (d), it can be asserted that the high magnitude frequencies lower than 2×10^3 Hz are significantly affected compared to higher frequency components, implying the disruption in the important part of spectrogram where most of high magnitude signature of spoken words are present. In contrast, the final row of Tables II and III represents the impact of spectral reduction for both the cases when speech is considered as a vector and spectrogram matrix. Considering the case of 1D CNN, its performance has further improved due to removal of bias due to higher frequency noise. However its impact on 2D CNN is inversely reported due to the reduction of high amplitude frequency signatures at lower frequency levels that are likely to be contributing more in developing a robust model.

IV. CASE STUDY: TESTING THE MODELS ON NON-NATIVE MALAYALAM SPEAKERS

The final developed CNN models are expected to be more generalised than any other model as it includes all four important pre-processing steps to reduce the bias in the dataset. To test the general transfer ability of the model that has been solely trained on the data that has been acquired from native-Malayalam speakers, the data from two non-native Malayalam speakers (one from India while the other from Poland) has been collected for testing purposes. In test data, the speakers tried their best to record the week's names without associating any emotions and with natural accent, hence giving more diversity to the case study data compared to the trained one. Table IV presents the week's name that are either correctly or incorrectly predicted by the final 1D and 2D CNNs models. The prediction made by 1D CNN model is comparatively lesser than 2D CNN, and in addition the overall accuracy of correct prediction is far worse than reported in Table II. The poor performance of 1D CNN is likely due to the fact that the case study data includes speech samples from non-native Malayalam speakers which produces very different pattern of speech compared to training dataset. On the other hand, 2D CNN has produced 100% results for Indian accent and 42.85% results for Polish accent. The poor performance of Polish accent is due to fact that the duration of speech data was very slow and hence produces a very different spectrogram of spoken language. Prediction accuracy of such type of data can be improved by incorporating additional pre-processing steps such as data augmentation, audio segmentation, additive feature extraction etc.

TABLE IV
NUMBER OF DAYS CORRECTLY/INCORRECTLY PREDICTED BY FINAL 1D AND 2D CNNS MODELS DEVELOPED AFTER UNDERGOING ALL THE PREPROCESSING STEPS NAMELY $S_iR + OE + AN + S_pR$.

Days in Malayalam	Target prediction	Prediction by 1D CNN		Prediction by 2D CNN	
		Indian accent	Polish accent	Indian accent	Polish accent
Thinkal	Monday	Correct	Incorrect	Correct	Correct
Chovva	Tuesday	Incorrect	Incorrect	Correct	Incorrect
Budhan	Wednesday	Incorrect	Incorrect	Correct	Correct
Vyaazham	Thursday	Incorrect	Incorrect	Correct	Incorrect
Velli	Friday	Incorrect	Correct	Correct	Incorrect
Shani	Saturday	Incorrect	Incorrect	Correct	Correct
Njaayar	Sunday	Incorrect	Incorrect	Correct	Incorrect

V. CONCLUSION

Interlingual transcription in real-time faces several hurdles such as noise, emotion, accent, etc. which can impact the overall transcription accuracy of the system. However, the evolution of end-to-end speech-to-text translation systems has been significantly impacted by the advancement of neural networks and feature engineering. Directly using speech data without any pre-processing steps has many disadvantages, such as, it reduces the reliability, and scalability of machine learning models hence constraining the consistency, stability, and generalisation of the transcription models. To investigate the role of using various pre-processing methodologies in dealing with more realistic real-time speech-to-text interlingual transcription, two types of CNN architecture have been adopted. The proposed study has revealed that higher dimension of feature space such as spectrograms has a more robust tendency to deal with emotions and accents in the dataset. Further preprocessing of the dataset may not have any additional advantage for distinctive words but may be helpful for less distinguishable words.

REFERENCES

[1] I. Sutskever, O. Vinyals, Q.V. Le, "Sequence to sequence learning with neural networks", In Proceedings of the Annual Conference on Neural Information Processing Systems, pp. 3104-3112, 2014.

[2] P. Bahar, A. Zeyer, R. Schlter, H. Ney, "On using 2D sequence-to-sequence models for speech recognition", In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 5671-5675, 2019.

[3] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, Y. Bengio, "End-to-end attention-based large vocabulary speech recognition", In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4945-4949, 2016.

[4] A. Zeyer, K. Irie, R. Schluter, H. Ney, "Improved training of end-to-end attention models for speech recognition", In Proceedings of the19th Annual Conference of the International Speech Communication Association, pp. 7-11, 2018.

[5] M.A. Di Gangi, M. Negri, R. Cattoni, R. Dessi, M. Turchi, "Enhancing transformer for end-to-end speech-to-text translation", In Proceedings of the Machine Translation Summit XVII, pp. 21-31, 2019.

[6] R.J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, Z. Chen, "Sequence-to-sequence models can directly translate foreign speech", In Proceedings of the 18th Annual Conference of the International Speech Communication Association, pp. 2625-2629, 2017.

[7] K. Hara S.T. Iqbal, "Effect of machine translation in interlingual conversation: Lessons from a formative study", In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 3473-3482, 2015.

[8] A. Tjandra, S. Sakti, S. Nakamura, "Machine speech chain", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 976-989, 2020.

[9] A. Berard, O. Pietquin, L. Besacier, C. Servan, "Listen and translate: A proof of concept for end-to-end speech-to-text translation", NIPS Workshop on End-to-End Learning for Speech and Audio Processing, Article ID: hal-01408086f, 2016.

[10] S. Bansal, H. Kamper, A. Lopez, S. Goldwater, "Towards speech-to-text translation without speech recognition", In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pp. 474-479, 2017.

[11] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, T. Cohn, "An attentional model for speech translation without transcription", In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 949-959, 2016.

[12] Y.H. Ghadage, S.D. Shelke, "Speech to text conversion for multilingual languages." In Proceedings of the IEEE International Conference on Communication and Signal Processing, pp. 236-240, 2016.

[13] Abdusalomov, Akmalbek Bobomirzaevich, Furkat Safarov, Mekhriddin Rakhimov, Boburkhon Turaev, and Taeg Keun Whangbo. "Improved Feature Parameter Extraction from Speech Signals Using Machine Learning Algorithm." Sensors 22, no. 21 (2022): 8122.

[14] Shah, Nirmesh J., and Hemant A. Patil. "A novel approach to remove outliers for parallel voice conversion." Computer Speech and Language 58 (2019): 127-152.

[15] Dawson, Robert. "How significant is a boxplot outlier?." Journal of Statistics Education 19, no. 2 (2011).

[16] A. Dutt, P. Gader, "Wavelet multiresolution analysis based speech emotion recognition system using 1D CNN LSTM networks"m IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2043-2054, 2023.

[17] S.P. Joysingh, P. Vijayalakshmi, T. Nagarajan, "Quartered spectral envelope and 1D-CNN-based classification of normally phonated and whispered speech", Circuits, Systems, and Signal Processing, vol. 42, no. 5, pp. 3038-3053, 2023.

[18] Y.B. Singh, S. Goel, "A lightweight 2D CNN based approach for speaker-independent emotion recognition from speech with new Indian emotional speech corpora", Multimedia Tools and Applications, vol. 82, pp. 23055-23073, 2023.

[19] K. Zvarevashe, O.O. Olugbara, "Recognition of speech emotion using custom 2D-convolution neural network deep learning algorithm", Intelligent Data Analysis, vol. 24, no. 5, pp. 1065-1086, 2020.

[20] P. Gambhir, A. Dev, P. Bansal, D.K. Sharma, "End-to-end multi-modal low-resourced speech keywords recognition using sequential Conv2D nets", ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 23, no. 1, Article no. 7, 2024.