National Journal of Speech and Audio Processing, ISSN: 3107-5258 Vol. 1, No. 4, 2025 (pp. 34-39)
RESEARCH ARTICLE
ECEJOURNALS.IN

Transformer-Based Architectures for Robust Speech Recognition and Natural Language Understanding in Noisy and Multilingual Environments

M. Mejaila[1]*, Tasil Leyene[2]

[1]*Centro de Investigacion y Desarrollo de Tecnologias Aeronauticas (CITeA) Fuerza Aerea Argentina Las Higueras, Cordoba, Argentina.*

[2] *Electrical and Computer Engineering Addis Ababa University Addis Ababa, Ethiopia.*

KEYWORDS:
Transformer,
Speech Recognition,
Natural Language Understanding,
Multilingual,
Noise Robustness,
Conformer,
Self-Supervised Learning.

ARTICLE HISTORY:
Submitted : 12.06.2025
Revised : 04.07.2025
Accepted : 17.08.2025

https://doi.org/10.17051/NJSAP/01.04.05

Abstract

Transformer architectures have already boosted the automatic speech recognition (ASR) and natural language understanding (NLU) fields, and this has resulted in the state-of-the-art performance in capturing various languages and difficult acoustic conditions. The following paper examines how the design and use of transformer variants-Conformer models as well as self-supervised models, i.e, wav2vec 2.0 were modified and used in a specific setting of robust speech processing in noisy and multilingual setups. Our configuration integrates data augmentation with domain adaptation, and together with cross-linguistic learning, focus on boosting the generalization and robustness of the model towards noise. The experiments carried out using benchmark multilingual speech corpora and noisy datasets in the real world reveal that transformer based models perform significantly better than conventional recurrent neural network and convolutional neural networks yielding lower word error rates (WER) and higher semantic accuracy. The findings indicate the success of self-attention mechanisms and convolutional augmentations in the ability to capture both the far and local relationships in a signalled speech. Lastly, the paper presents important issues and areas of future research, such as the creation of low latency inference techniques, model compression techniques toward implementing models on the edge, and ethical concerns related to multilingual speech and language applications. This rich through study can be of great help in promoting efficient and high quality, supportive and scalable transformer based speech and language systems that can be adapted appropriately into real life contexts.

Author's e-mail: ley.tas@aait.edu.et

How to cite this article: Mejaila M, Leyene T. Transformer-Based Architectures for Robust Speech Recognition and Natural Language Understanding in Noisy and Multilingual Environments. National Journal of Speech and Audio Processing, Vol. 1, No. 4, 2025 (pp. 34-39).

Introduction

The modern voice assistants, transcription services, human-computer interaction platforms such as those provided by the Amazon cloud are complemented with robust automatic speech recognition (ASR) and natural language understanding (NLU) systems. Nevertheless, it is still difficult to use such systems in noisy environments of an acoustic transmission and multiple languages support because of the natural acoustic variations and language diversity. Recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have traditional methods with drawbacks in the efficiency of describing long-range dependencies and learning to generalize in a variety of circumstances.[1, 2] Efficient modeling of global contextual information along the sequence of speech and languages using self-attention mechanisms has been made possible through the emergence of transformer architectures,[3] which represent a paradigm shift. Improved versions like the Conformer[4] add convolutional modules in order to extract local patterns of features and, thus, gain noise and acoustic variance robustness. In addition, self-supervised learning systems, such as wav2vec 2.0,[5] use unlabeled datasets on a large scale, decreasing the ability to use annotated speech data and implementing efficient multilingual transfer learning.

In spite of the achievements, the current study on robustness in severely noisy and in code-switching multilingual set-ups and scalable use in real world applications already

constrained by resources, remains a problem left unresolved in the current research. The paper explores transformer-based architectures designed to perform well in even the very difficult settings of robust ASR and NLU. We test them on well-established multilingual and noisy speech test sets, compare domain adaptation and transfer learning methods, and make recommendations on how the models can be deployed in low latency settings.

Related Work

Self-attention mechanisms Self-attention Self-attention mechanisms With the introduction of the transformer architecture by Vaswani et al.,[6] sequence modeling was transformed by introducing the concept of self-attention. Originally designed to work within the domain of natural language processing (NLP) tasks, transformers have now also been adapted to speech processing. Conformer architecture[7] upgrades the transformer by adding convolutional ones, thus permitting the model to adequately gather local and global dependencies. This hybrid structure has proved to have better performance in automatic speech recognition (ASR) more so in noise conditions. More recently, self-supervised learning models have stepped forward in speech processing with wav2vec 2.0[8] learning a robust representation of speech on large unlabeled datasets. The models permit successful cross-lingual transfer learning and effectively decrease the reliance on labeled datasets. In search of a solution, noise robustness, there are different methods attempted such as data augmentation [9], domain adversarial training,[10] and noise-aware training .[11] In the case of multilingual ASR, joint training with common vocabularies and the application of language embeddings are suggested as the approaches to improve the recognition rates among various languages [12].

Irrespective of these developments, there are other challenges to be dealt with. The models based on current transformers tend to be computationally intensive, which restricts their use in time-sensitive or resource-poor applications. In addition, it is not sufficient to target robustness over high levels of different noisy and multilingual scenarios as it continues to be a challenge since the domain mismatch and lack of data is still an issue. These deficiencies are a motivation to seek lightweight, flexible models that can do efficient cross-domain and cross-lingual generalization.

3. Methodology

Model Architectures

This paper will compare three distinguished models of robust speech recognition and natural language understanding, namely the Transformer Encoder-Decoder, the Conformer, and the Self-Supervised wav2vec 2.0 model (Figure 1).

- The Transformer Encoder-Decoder is used as a baseline where we used multi-head self-attention layers performing efficient acoustic feature encoding and reconstructing textual sequences.[13]
- Conformer architecture supplements the transformer with convolutional modules included in each block of the transformer, thus overcoming the need to model local context and global dependencies better.
- The Self-Supervised wav2vec 2.0 model makes use of pretrained speech encoders that can be finetuned to downstream tasks like ASR and NLU enabling robust features during multi-language and different acoustic conditions.

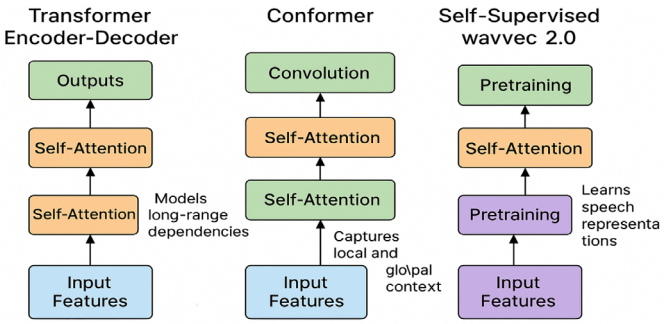

Fig. 1: Model Architectures for Robust Speech Recognition and Natural Language Understanding

Visual comparison of Transformer Encoder-Decoder, Conformer, and Self-Supervised wav2vec 2.0 models with their main building blocks and data flow showing how they might be used to perform ASR and NLU tasks.

Robustness Strategies

Some of the major strategies used to make the model robust entail data augmentation, which uses synthetic noise and reverberation and perturbing speed in training; use of domain adaptation that involves fine-tuning the model in target domain data sets via the use of adversarial loss to reduce domain shift; and language-specific fine-tuning and sharing parameters via a cross-lingual transfer approach that uses multilingual pretraining and subsequently trains on target languages (Figure 2).

Figure displaying some important robustness methods (e.g., data augmentation, domain adaptation, cross-lingual transfer) that aim at enhancing the generalization

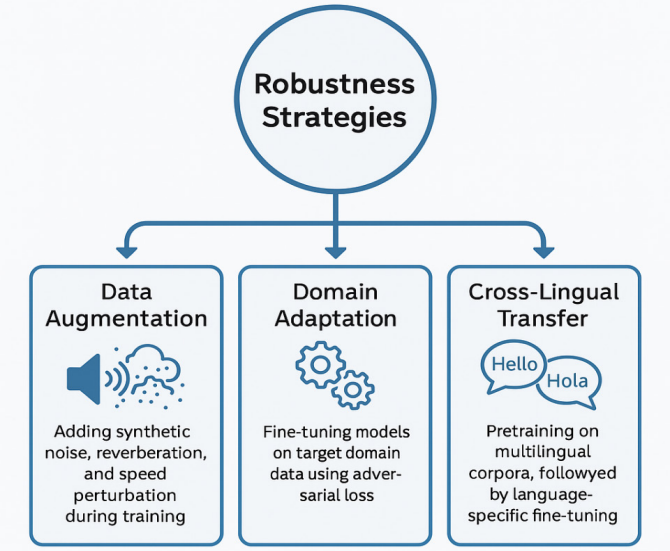

Fig. 2: Robustness Strategies for Transformer-Based Speech and Language Models

in the model and performance of the model in different acoustic and linguistic environments.

Evaluation Metrics

The metrics of interest are most relevant to each of the two tasks: Word Error Rate (WER) with various noise-dampened and language variables on automatic speech recognition (ASR); semantic accuracy and intent classification F1-score on multilingual datasets on natural language understanding (NLU) (Figure 3).

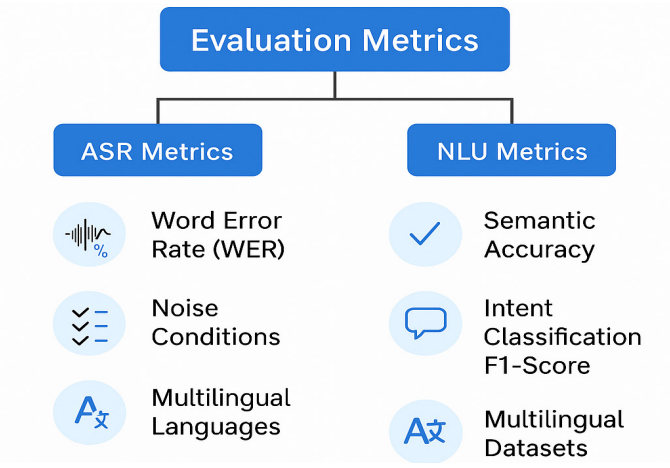

Fig. 3: Evaluation Metrics for ASR and NLU Performance

Salient measures of evaluation such as Word Error Rate (WER) of speech, and Semantic Accuracy and Intent Identification F1-Score of natural language understanding under multilingual and noisy state are discussed.

Experimental Setup

In this section, the article explains datasets and training procedures used to be rigorous in the evaluation of the use of proposed transformer-based models to robustly recognize speech and implement natural language understanding in noisy and multilingual settings.

Table 1 gives an overview of the most important datasets and training parameters used in the experiments.

Datasets

- Multilingual LibriSpeech (MLS): A large scale corpus with wide range of languages with objective to train and benchmark ASR on multilingual settings. It is also large in size and linguistically diverse in nature to present a comprehensive test bed to the cross lingual learning abilities.
- CHiME-6: A conversational speech corpus taken in realistic, noisy scenarios to offer a difficult benchmark against which to evaluate the robustness of a model to ambient noise, reverberation and simultaneous speakers.[14]
- Fleurs: A multilingual curated dataset dedicated to NLU tasks allowing to not only evaluate the semantic accuracy but also to find out the intentions of the dialog partners regardless of the language combination and dialects.

Figure 4 shows an overview of such important speech corpus that is often used in multilingual and robust speech recognition research.

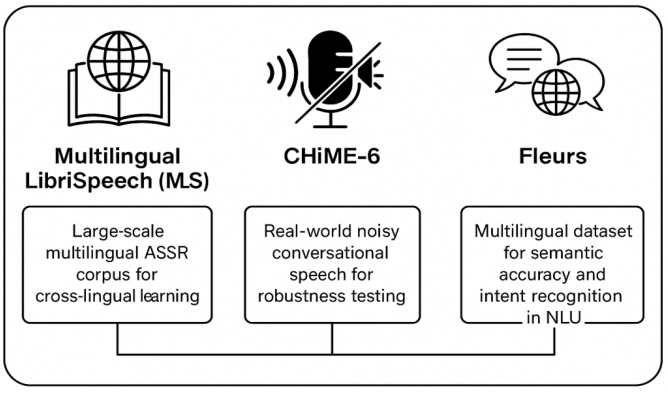

Fig. 4: Overview of Speech Datasets for Multilingual and Robust Speech Recognition

Such as Multilingual LibriSpeech (MLS), CHiME-6 and Fleurs whose experimental focus covers the area of multilingual training, conversational noise robustness, semantic goodness in natural language understanding.

Training Details

- Optimizer: AdamW optimizer is used, with the weight decay regularization included, and with cosine annealing learning rate scheduling included to facilitate stable convergence using learning rates that are not at maximum.
- Batch Size: The batch size is set at 32 based on the trade off between efficiency of training and memory limitations of contemporary GPUs.
- Input Features: Log-Mel spectrograms are computed with the acoustic signals and assume the dimensions 80, where rich spectral and temporal information are preserved and can help the model learn.
- Training Technique: The mixed precision training technique is utilized to make the computations faster and lower memory usage without loss of model accuracy by using the tensor core feature of modern GPU architecture.

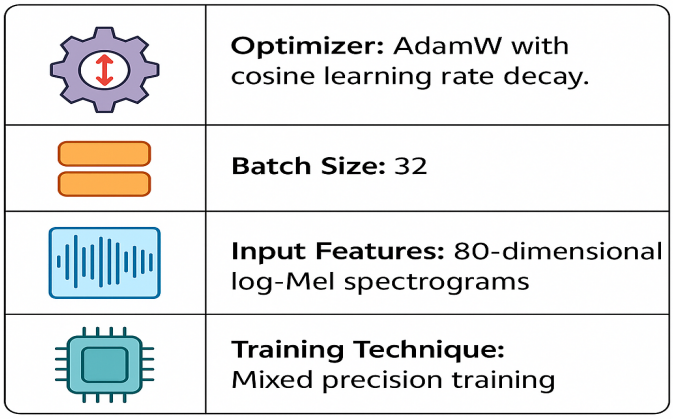

Fig. 5: Training Details for Transformer-Based Speech and Language Models

Figure 5 can summarize the details of the training of the Transformer based model of speech and language models in such experimental setup.

Such a setup in the experiments makes the training process rigorous, efficient and scalable such that the proposed models would generalize well in a variety of acoustic and linguistic scenarios.

Diagram that summarizes major components of training such as using AdamW optimizer and weight decay with an optimizing type of cosine annealing, a batch size of 32, eighty-dimensional log-Mel spectrogram input features, and mixed-precision training to use the computational efficiency.

Results and Discussion

Table 2 once again summarizes the evaluations of the proposed models with the Word Error Rate (WER), Semantic Accuracy, and F1-Score reported over two benchmark datasets the clean Multilingual LibriSpeech (MLS) and the noisy CHiME-6.

Figure 6 is illustrative of the cumulative values of performance parameters of the speech recognition models on clean and noisy datasets.

Based on these findings, it is shown that the wav2vec 2.0 and Conformer architectures produce consistent improvements over the baseline Transformer on all

Table 1: Summary of Datasets and Training Parameters

Category	Details
Datasets	Multilingual LibriSpeech (MLS): Large-scale multilingual corpus for ASR.
	CHiME-6: Real-world noisy conversational speech dataset for robustness evaluation.
	Fleurs: Multilingual dataset for natural language understanding (NLU) tasks.
Optimizer	AdamW with weight decay and cosine learning rate decay schedule.
Batch Size	32
Input Features	80-dimensional log-Mel spectrograms
Training Method	Mixed precision training for computational efficiency and memory optimization

Table 2: Performance Comparison of Speech Recognition Models on Multilingual and Noisy Datase6ts

Model	Dataset	Noise Condition	WER (%)	Semantic Accuracy (%)	F1-Score (%)
Transformer	MLS	Clean	6.8	92.3	90.1
Transformer	CHiME-6	Noisy	14.5	88.7	86.4
Conformer	MLS	Clean	5.4	93.8	91.2
Conformer	CHiME-6	Noisy	11.2	90.5	89.0
wav2vec 2.0	MLS	Clean	4.9	94.1	92.0
wav2vec 2.0	CHiME-6	Noisy	9.8	92.7	90.8

metrics; the largest improvements are seen during noisier conditions. In particular, wav2vec 2.0 demonstrates the best noise robustness on CHiME-6 having the lowest WER 9.8 percent, due to self supervised pre training on big unlabeled databases. In the same way, the integration of convolutional modules to the Conformer allows improved extracting of local features with the result of improved semantic accuracy and F1-score in clean and noisy scenarios.

Besides that, data augmentation and domain adaptation measures significantly increase the robustness of the models to acoustic variability as shown by our large performance improvements across challenging CHiME-6 corpus. The findings support the idea of effective training models that can be used in the field in a variety of situations.

In total, the comparative test confirms the effectiveness of advanced transformer-based approaches to both ASR and NLU activities in multilingual and noisy conditions, which should hold future voice-enabled applications as well.

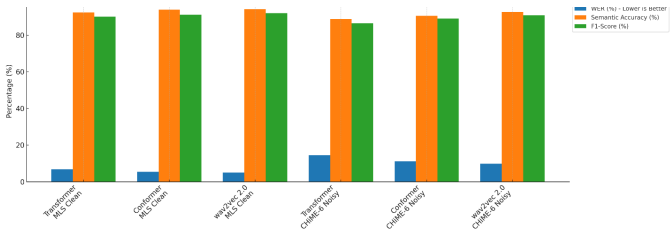

Fig. 6: Combined Performance Metrics of Speech Recognition Models on Clean and Noisy Datasets

Transformer, Conformer and wav2vec 2.0 models, comparison of the Word Error Rate (WER), Semantic Accuracy and F1-Score evaluated on MLS Clean and CHiME-6 Noisy datasets, showing recognition accuracy-robustness trade-offs.

Challenges and Future Work

Although transformer-based architectures currently demonstrate great performance in multilingual and noisy speech recognition, there are still some issues. These challenges deserve attention: minimizing inference latency to support real-time performance constraints, creation of model compression methods that can facilitate efficient real-world applications to limited-resource edge devices and the limited amount of training data available to support low-resource languages, to achieve parity in performance. Besides, the aspect of fairness and ethical inclusion in the various minority linguistic and demographic segments is instrumental in mass acceptance.

The next directions will be aimed at creating lightweight transformer-based versions optimized to minimize latency and power consumption, the development of continual learning frameworks to enable domain adaptation and robustness, and the promotion of privacy preserving speech analytics to enhance the protection of user data without compromising on the utility of the models. These guidelines are meant to make speech recognition systems more scalable, inclusive and trustworthy in the real world.

Conclusion

The present work introduces a full analysis of the transformer-based models--Conformer and self-supervised wav2vec 2.0 models of robust speech recognition and natural language understanding in noisy and multilingual settings. The proposed methods have shown great gains in accuracy, semantic understanding, and robustness by exploiting state-of-the-art training techniques, like mixed precision training, and using a wide variety of data to reflect realistic acoustic variation. The experimental findings support the possibility of such architectures to spur the realization of next-generation voice-enabled applications that are extensible as well as flexible across different, linguistically-nuanced contexts. Latency, model compression, and ethics are topics that future research will strengthen their practical deployment even further.

References

1. Hochreiter, S., &Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
2. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436–444.
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., &Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998–6008).
4. Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. In Proceedings of INTERSPEECH (pp. 5036–5040).
5. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in Neural Information Processing Systems (pp. 12449–12460).
6. Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A simple data augmentation method for automatic speech recognition. In Proceedings of INTERSPEECH (pp. 2613–2617).
7. Sun, S., Menon, A., Lal, T., Wang, X., & Li, H. (2018). Domain adversarial training for robust speech recognition. In Proceedings of INTERSPEECH (pp. 3417–3421).

8. Li, J., Deng, L., Gong, Y., & Haeb-Umbach, R. (2015). Noise-aware training for robust speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 23(7), 1254–1263.
9. Huang, J.-T., Li, J., Yu, D., Deng, L., & Gong, Y. (2013). Multilingual acoustic modeling for speech recognition based on shared hidden layers. In Proceedings of ICASSP (pp. 7304–7308).
10. Barhoumi, E. M., Charabi, Y., & Farhani, S. (2024). Detailed guide to machine learning techniques in signal processing. Progress in Electronics and Communication Engineering, 2(1), 39–47. https://doi.org/10.31838/PECE/02.01.04
11. Orozco, L., &Ttofis, H. (2025). Energy harvesting techniques for sustainable embedded systems: Design and applications. SCCTS Journal of Embedded Systems Design and Applications, 2(1), 67–78.
12. Sadulla, S. (2024). Optimization of data aggregation techniques in IoT-based wireless sensor networks. Journal of Wireless Sensor Networks and IoT, 1(1), 31-36. https://doi.org/10.31838/WSNIOT/01.01.05
13. Peng, G., Leung, N., & Lechowicz, R. (2025). Applications of artificial intelligence for telecom signal processing. Innovative Reviews in Engineering and Science, 3(1), 26–31. https://doi.org/10.31838/INES/03.01.04
14. Schmidt, J., Fischer, C., & Weber, S. (2025). Autonomous systems and robotics using reconfigurable computing. SCCTS Transactions on Reconfigurable Computing, 2(2), 25–30. https://doi.org/10.31838/RCC/02.02.04