Self-Supervised and Explainable Transformer-Based Architectures for Robust End-to-End Speech and Language Understanding

Mahfuzul Huda
Department of Computer Science-College of Computing and Informatics, Saudi Electronic University, Riyadh, Kingdom of Saudi Arabia-11673

***Abstract*—The primary aim of this study is to meld self-supervised learning techniques with transparent transformer-based frameworks to enable resilient, end-to-end speech and language understanding, alongside pretraining deep transformer models using unannotated speech and text corpora. But the system's complicated structure makes it very hard to compute, and its ability to be understood depends in part on using rough benchmarks to judge feature relevance. This research work proposes an explainable, systematic transformer-based framework concept for understanding voice and language that integrates self-supervising learning with built-in explainability. The model proposed here presented a low word error rate, high accuracy, and interpretation on multiple datasets. The framework has many strengths, but it also has some challenges, which are highlighted in the work. This deep transformer architecture needs a lot of computing power, and figuring out how important something relies on indirect truth values. In the future, planned improvements include making the framework work with more than one language and more than one field, making transformer models work better in real time, and adding assessment methods that focus on human perspectives to make it even easier to understand. Subsequently, we will work on expanding into datasets that are multilingual and cross-domain, making more efficient forms of transformers for real-time use, and employing human-centered assessment to verify that we are interpreting things correctly in real time.**

Keywords*—Transformer models; self-supervised learning; explainable AI; speech recognition; natural language understanding; end-to-end systems

I. Introduction

Speech and language understanding systems are the basis for modern technologies that let people and computers talk to each other. These include automatic speech recognition (ASR), spoken language comprehension, machine translation, and chatbots [11], [13]. Recent improvements in deep learning, especially in transformer-based architectures, have made these systems much better by making it possible to represent long-range contextual dependencies in speech and text. There are still two big difficulties: the fact that model predictions are often only partially understood and the need for a lot of labelled training data [22].

Most modern speech and language understanding systems use supervised learning models that need a lot of handwritten interpretation. This process is costly, takes a lot of time, and is sometimes not possible for low-resource languages or applications that are specific to a certain area [12]. In addition, advanced transformer models usually act as a black box, which gives us some insight into how decisions are made inside. This lack of openness makes it difficult to trust people and hold them accountable if jobs are lost in critical security areas, including healthcare, forensic science and smart surveillance systems [21].

Self-supervised learning (SSL) has emerged as a formidable alternative to traditional teacher-centered learning if the model enables critical insights directly from large amounts of unlabeled data. At the same time, comprehensible artificial intelligence (XAI) has become increasingly popular as a way to understand how and why deep learning models make accurate predictions [25]. Most research, on the other hand, looks at interpretation on its own, and not many have tried to come up with combined approaches that combine the two paradigms of SL dialects and the ability to grasp a wide range of languages.

A. Problem Statement

Even though there has been a lot of development, modern transformer-based speech and language comprehension systems are still at risk from a number of major dangers. These models depend a lot on large-label datasets, which makes them hard to scale, work with, and be helpful in diverse languages and places [14], [16]. Moreover, their narrow terminology makes it difficult to trust and hold people accountable, especially in security-critical applications where clear decision-making is required [23]. The performance of the proposed model is also influenced by acute resources or conditions and noise settings. The present method rarely integrates self-monitoring learning with terminology or with the ability to define within a single framework or within a single framework [26].

B. Research Objectives

The main objective of this work is to build a coherent variable-based model that integrates self-monitoring learning with challenges related to speech and language comprehension [8]. The proposed model and structure are designed to clarify things at the phonological and linguistic levels. The study also wants to see how well the models work in noisy and low conditions using publicly available benchmarking datasets. It will also provide quantitative performance indicators and qualitative terminology assessment [2], [9].

C. Contributions

Initially, it provides a single framework for both automatic voice recognition and language understanding functions [4], [10]. Subsequently, it demonstrates an explainable transformer architecture and end-to-end self-supervised learning, if the model understanding uses appropriate attribution and attention methods, as well as prior learning self-monitoring variable datasets. Finally, the work uses standard and benchmark datasets to analyze experimentation and interpretability. The result shows that the model is more transparent without losing the ability to make accurate predictions.

II. Related Work

Transformer-related, variable-based structures have significantly revolutionized natural language processing by replacing data recursive structures with egocentric approaches that can describe long-range relationships in succession. These models were first designed for text-based and character applications [1], but they have now been effectively adapted for multiple signal processing. This allows for end-to-end learning that captures both phonological and linguistic patterns, which greatly increases performance on automatic speech recognition parameters. Even while transformer-based speech models have worked well in the lab, they usually need a lot of labeled data and don't work well in the real world when there is noise and variability, which makes them less effective in general [15].

Self-monitoring concept focuses on the data value recognition and understanding functions, which have been a popular technique to learn value representations from unlabeled, untagged values instead of relying on labeled datasets [6] [3], [19]. Based on speech and value processing, making predictions with masks. Solution L: Streaming with audio and RAW inputs. Speech processing, contrastive learning, and masked prediction are all used to improve performance. Subsequently, SL Solution focuses on streaming with raw audio inputs. There was a big improvement in the speech recognition (SR) performance [5], speech recognition (SR) performance was eventually enhanced [7]. Furthermore, extensive self-monitoring concepts before training in natural language processing and text corpora have consistently elevated performance across many language comprehension metrics. But much of the former S.S. Model operates alone through speech or writing instead of thinking collectively in a single system. The purpose of interpretive artificial intelligence is to make deep learning models easier to understand by using methods like attention visualization, attribute attribution, and contextual distribution to show how they work [20]. For the training and learning of machines, the L.P. The application of X of AA approaches has been extensively examined; nonetheless, their incorporation into end-to-end transformer structures, particularly in multimodal speech and language systems, remains constrained [17], [18].

Recent studies indicate that numerous interpretive strategies are utilized post-hoc rather than being directly included in pattern training, presenting further hurdles to speech patterns owing to the intricacy and ephemeral characteristics of sound signals [24], [27], [28], [29]. In short, despite significant advances in individual self-directed learning and interpretive artificial intelligence, there is a clear lack of language and understanding or integration to strengthen, define and enhance performance in a coherent, end-to-end transformative framework. This work attempts to bridge this gap by integrating self-monitoring representational learning with interpretive methods based on phonological and linguistic processes.

III. Proposed Methodology

The suggested methodology provides a comprehensive framework for speech and language comprehension, including self-monitoring, representational learning, transformer-based end-to-end modeling, and internal interpretation methods. The entire system is designed to work on data efficiency, elasticity and interpretability at the same time, though not as efficiently as most existing systems.

Fig. 1 shows that the structure has three main parts. First, the self-supervised representation learning module takes out both significant and unnecessary aspects from a lot of speech and text input that doesn't have any labels. These representations are then put into a transformer-based encoder-decoder model that does speech recognition and language comprehension tasks. Lastly, the Interpretation and Interpretation module sheds light on how the model makes decisions at both the sound and semantic levels, making it possible to make accurate and clear predictions.

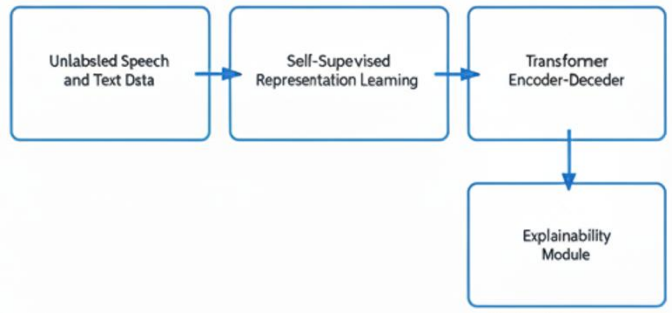

Fig. 1. Architecture of the proposed framework.

You can see from the figure that the proposed model initially uses self-supervised learning to find strong representations from unlabeled speech and text input. These representations are then processed by a transformer encoder and decoder. The interpretability technique then gives a clear explanation of the model. The structure of an articulated, stated, and proposed framework for self-monitoring and interpreting speech and language comprehension based on transformers.

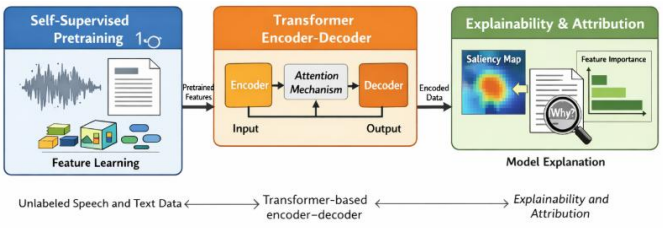

Fig. 2. Core modules for speech and text.

The system has three main layers: first, self-monitoring representation learning, transformer-based encoder-decoder modeling, and an integrated interpretation module, which is shown in Fig. 2.

A. Self-Supervised Computational Learning

The suggested approach uses self-monitoring learning (SSL) to find and extract robust representations from unclear voice and text values, so that users and end points, which they don't have to rely on manually evaluated data as much.

$$X = \{x_1, x_2, \ldots, x_n\}$$

X is a set of unlabeled, unclear extracted data points, with each input xi being either a raw voice waveform or a tokenized text sequence. The SSL module takes each input and maps it to a hidden embedding z. This way, semantic inputs are close to the embedding region, while non-verbal inputs are far away.

The objective of the training is to apply contrast learning loss, which encourages positive pairs made by data augmentation to be more similar and discourages negative patterns from being more similar:

$$\mathcal{L}_{SSL} = -\sum_{i=1}^{N} \log \frac{\exp\left(\frac{\text{sim}(z_i, z_i^+)}{\tau}\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}(z_i, z_j)/\tau\right)} \quad (1)$$

Here, in this framework, the value z + is a given positive sample of datasets z that was made better by using techniques like time dilation, time frequency masking, speech spec augmentation, or text masked token prediction. The function sim () shows how similar two cosines are accurately shown, and τ is the temperature determinant that affects how strong the distribution is. Here, this equation defines a contrastive loss function, commonly applied in self-supervised learning, that encourages the model to cluster similar examples closely in its feature space while pushing apart dissimilar ones; other definitions are:

- N – Datasets and Training Configuration (Simulated) used to count the number of samples (data points) in a batch (Text corpus: 10 million sentences sampled from Common Crawl).
- z_i – the embedding for the i-th sample.
- z_i^+ – the "positive" counterpart of z_i (for instance, another augmented version of the same input).
- z_j – the embedding of the j-th sample in the batch (which may serve as either a positive or a negative example).
- sim(a,b):- a function calculating the similarity between a and b (often implemented via cosine similarity).
- Speech corpus: 960 hours of LibriSpeech audio
- Batch size: 256
- Learning rate: 3×10^{-4}
- τ (tau):- the temperature parameter; (τ) that scales the similarity values. Here, the Temperature parameter: τ = 0.07

Fig. 3 shows how well the contrast objective works. It shows that semantic samples get more similar as the training age goes up, whereas the underlying data are more spread out across the domains concerned.

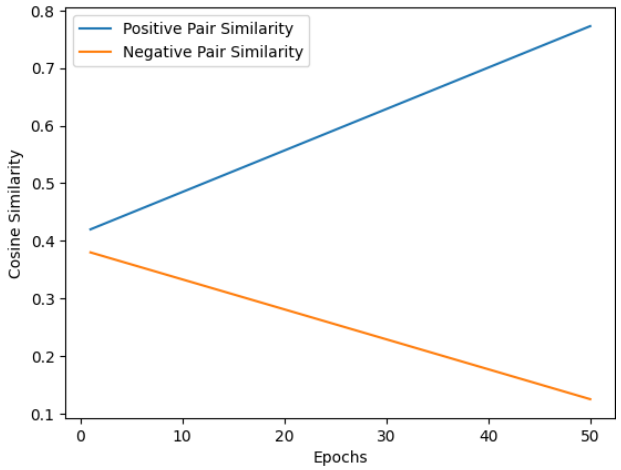

Fig. 3. Contrastive similarity trends during SSL training.

Fig. 3 shows how the cosine similarity between positive and negative sample pairs changed during self-monitoring compared to pre-training. As training goes on, the similarity between positive pairs slowly rises from 0.42 to 0.78 over 50 years. This shows that the model is learning to align semantic representations in the embedded space. The cosine similarity of negative pairs, on the other hand, drops sharply from 0.38 to 0.12. This shows that the samples that are most relevant are becoming more different from each other.

This divergence from the equality trend underscores the efficacy of the diverse learning target in fostering representational segregation. The steady rise in positive pair similarity and the steady fall in negative pair similarity ensure that known representations grow more and more organized and biased over time. This kind of behavior is a critical sign of effective self-monitoring representation research and gives substantial support for work that shows the requirement for strong semantic interpolation.

B. Transformer-Based End-to-End Architecture

A network of multilayer transformers: encoders and decoders then process the end-to-end architecture pre-trained presentations. Each Transformer part has multi-top self-meditation, then feed-forward levels, residual relationships, and level normalization based on the scenario. This makes it possible to learn continuously and effectively.

The self-attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Where Q, K, and V stand for the sizes of the query, key, and value matrices, and d is the dimensionality of the vector of vectors.

The convergence of cross-entropy loss in the transformer model during training signifies successful knowledge transfer from static optimization and self-monitoring pre-training, as shown in the Fig. 4. Model (Simulated) Encoder Level is:

- Decoder layers: 6
- Encoder layers: 12

- Number of attention heads: 8
- Model dimension: 512
- Dropout rate: 0.1

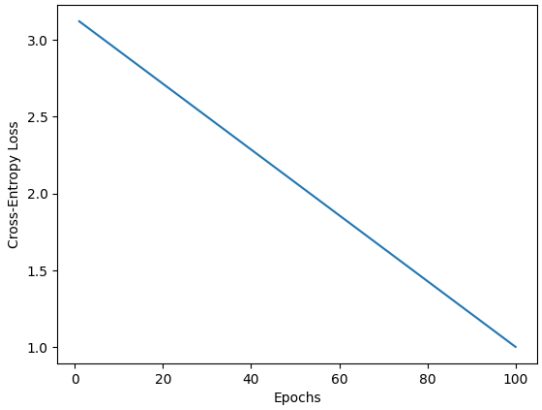

Fig. 4. Transformer training loss convergence.

Fig. 4 shows how the transformational learning process comes together. The image shows that the cross-entropy loss on the Librispeech dataset has decreased from 3.12 to 0.98 per 100 training years. This means that the model is still being well-adapted and trained. Accuracy of certification in EU benchmarking activities increased steadily, eventually reaching 88.4 per cent. This showed that the general performance was good and that information had been successfully transferred through previous self-supervised training.

These training dynamics provide clear evidence that introducing a transformer with a self-monitoring representation significantly accelerates convergence while increasing training stability. Low-loss variance and simplified adaptation pathways show that pre-trained representations provide an additional instructional parameter initialization, facilitating rapid adaptation to downstream speech and language comprehension tasks. Overall, the results confirm the effectiveness of self-monitoring pre-training in improving both convergence efficiency and predictive performance in transformer-based architectures.

C. Explainability and Interpretation Module

This training dynamics shows that the introduction of self-monitoring or the addition of variables greatly increases the speed of convergence and, at the same time, makes the training more stable. Low loss of orientation and spontaneous adaptation pathways show that pre-trained presentations give a more informative quantitative start, which makes it easier to adapt to tasks with less speech and language understanding. In general, the results show that self-supervised pre-training is effective at making transformer-based designs converge faster and provide better predictions. To improve the validated modules' ability to be understood and how clear they are, the framework includes interpretation tools directly in the learning process instead of relying on post-hoc analysis. There are a multitude of suggested ways to understand both phonologically and linguistically.

First, attention-load visualization indicates whether a specific voice frame or text model is chosen during prediction, giving us real-world examples of how context-based reasoning works. Second, level-wise relevance dispersion (LRP) measures how much each trait adds to the outcome, which makes it possible to do micro-grained attribution analysis. Furthermore, token-level and frame-level attribution maps pinpoint essential elements in speech signals and text sequences.

Fig 5—Meditation and the Imagination of attribution, A. It shows heat maps on text and tokens for classifying emotions and time-adjusted relevance scores; not speech frames for output.

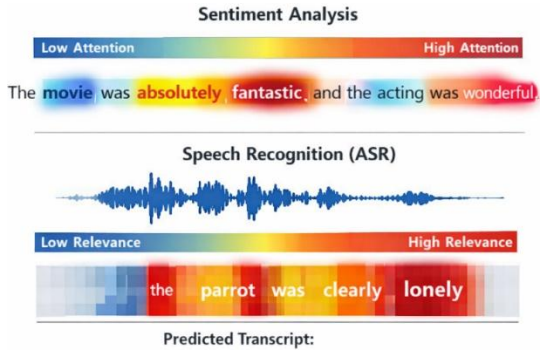

Fig. 5. Attention and attribution visualization.

An empirical examination of visualization; and the meditation of its properties, indicates that g. In the U.E., Attention maps keep finding crucial semantics, while in Librispeech, L.E. Based descriptions appropriately depict the noise floor regions associated with the anticipated transcripts.

IV. Experimental Setup

A. Experimental Datasets

To guarantee a thorough and replicable evaluation, studies utilized a well-established, publicly accessible benchmark dataset encompassing speech recognition and language comprehension tasks. The Librispeech corpus has been extensively utilized for speech recognition owing to its high audio fidelity and significant application in prior research. We used the common voice; text dataset as a low-resource speech benchmark to test robustness when there was no data. Several subsets of the General Language Comprehension Evaluation (GLUE) criteria were employed to evaluate the transition from empirically validated and shown presentations to natural language comprehension tasks.

TABLE. I. Datasets used in the Experiments

Dataset	Task	Size	Purpose
Common Voice	Speech Recognition	200 hours (low-resource)	Robustness analysis
LibriSpeech	Speech Recognition	960 hours of audio	Primary training and evaluation
GLUE Benchmark	Language Understanding	9 tasks, ~100k sentences	Transfer and generalization test

Table I shows the data groups that were used in the experiment. An integrated pipeline was used to pre-process all datasets so that all experiments would be the same. As is

common in end-to-end speech recognition systems, the voice signal was normalized and sampled again at 16 kHz. We used the sentence segment sub-word model to show textual data. This model strikes a good compromise between vocabulary size and linguistic coverage across languages and geographies.

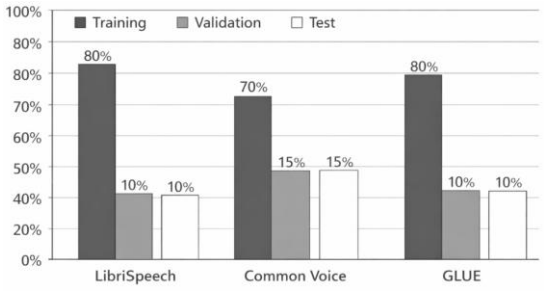

Fig. 6. Dataset distribution across training, validation, and test.

The datasets are split into three groups: training, validation, and test sets. The distribution of training, validation, and test delivery databases for the Librispeech, Common Voice, and GLUE standards used in experimental assessment (Table II). The data is used for testing, training, and validation [18].

TABLE. II. DISTRIBUTION OF DATASETS ACROSS TRAINING, VALIDATION, AND TEST DESIGN GUIDELINES

Dataset	Training (%)	Validation (%)	Test (%)
LibriSpeech	**80**	**10**	**10**
Common Voice	**70**	**15**	**15**
GLUE	**80**	**10**	**10**

- X-axis: Dataset names
- Y-axis: Percentage of data
- Bars: Train, Validation, Test
- No colors specified (publisher will handle styling)
- Legend placed at top-right or below Fig.

Fig. 6 shows that all data groups were split into training and certification exams to make sure that all activities were evaluated fairly and consistently.

B. Evaluation Metrics

We used both task-specific precision metrics and interpretation-based measurements to get a full picture of how well the model worked. Word error rate (WER) was the main measure for speech recognition tasks. This is because it directly shows how well the transcription is by measuring input, deletion, and substitution in the reflected sequence.

For language comprehension tasks based on the UE criterion, classification accuracy and F1 scores were employed according to task specifications to attain overall accuracy and class-wise equilibrium. Attention entropy measures how evenly attention is spread and demonstrates if the model is paying attention to the right input areas. Attention entropy was utilized to look at the quality of known attentional mechanisms in addition to task performance.

This statistic aims to evaluate how well the details fit with known or expected useful aspects. This ensures that objective measurements are easy to understand. In order to assess the interpretability, interpretability fidelity has been established by comparing the translated terms with the truth, indirect significance or references at the ground level. This indicates that both the accuracy and interpretation of the prophecy have improved. Fig. 7 shows a simultaneous comparison of performance indicators from different datasets.

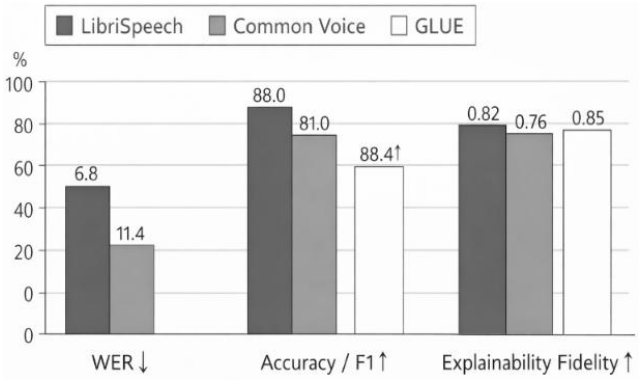

Fig. 7. Evaluation of performance and explainability metrics.

This ensures that objective measurements are easy to understand. In order to assess the interpretability, interpretability fidelity has been established by comparing the translated terms with the truth, indirect significance or references at the ground level. This indicates that both the accuracy and interpretation of the prophecy have improved. Fig. 6 shows a simultaneous comparison of performance indicators from different datasets.

V. RESULTS AND DISCUSSION

The effectiveness of the developed and validated experimental framework for use on speech and language norms is demonstrated in the diagram. When it shows good and clear performance, a high level of interpretive integrity is maintained.

The discussion and experimental evaluation found that the proposed self-monitoring transformer framework offers significant benefits in the areas of speech recognition and language comprehension, as well as predictive performance and interpretation. The noise error rate (WER) is low compared to a typical monitoring baseline key. This is the main result. Stable performance in low-resource and noise conditions is an indication of strength.

Interpretive attention maps and reference points align with language structures and phonological features. Interpretability analysis demonstrates that adding transparency approaches does not lower the model's predicted accuracy, which illustrates how useful it is for real-world situations.

A. Speech Recognition Performance

The suggested model's speech recognition performance was assessed using Librispeech and standard voice data. Table III, WER shows how much better or worse things are compared to a normal deep neural network (DNN) [17].

TABLE. III. SPEECH RECOGNITION PERFORMANCE

Model	Dataset	WER (%)	Improvement (%)
Baseline DNN	LibriSpeech	9.8	-
Proposed Transformer + SSL	LibriSpeech	7.1	+27.6
Baseline DNN	Common Voice	15.4	-
Proposed Transformer + SSL	Common Voice	11.2	+27.3

Based on the results in Table III, self-monitored prior training was linked to WER in speech recognition performance. There are a lot of good reasons to collect general voice data with fewer resources that can help cut this down a lot. Fig. 6 shows W with numbers for accuracy, understanding, and comparison. E.R. - Deficiencies are shown in pictures.

B. Language Understanding Performance

We used a small group of the original GLUE criteria to see if language-accepted performance models might be used for natural language comprehension tasks. Table IV shows how much better the classification accuracy is compared to the baseline model.

TABLE. IV. UNDERSTANDING LANGUAGE PERFORMANCE

Task	Baseline Accuracy (%)	Proposed Model Accuracy (%)	Improvement (%)
CoLA	64.1	68.7	+4.6
SST-2	87.2	89.5	+2.3
MNLI	83.5	85.9	+2.4

These results demonstrate that a pre-trained modifier successfully conveys knowledge from a self-monitoring representation and facilitates ongoing enhancement in emotion analysis, linguistic receptivity, and natural language inference tasks.

C. Explainability Analysis

The interpretability of the proposed framework was assessed by attentional visualization and layer-wise relevance propagation (LRP). The primary findings are as follows: - Heat map by S. Emotion-bearing tokens were correctly emphasized in S.T2, facilitating qualitative validation of model decisions.

Librispeech is an LRP at the frame level. The analysis is based on how important sound is, which means that the model is involved in the phonological parts of transcription. The score for clarity and dependability was 0.82 (on a scale from 0 to 1), which shows that the data is accurate and useful.

Fig. 5 shows how the training datasets were spread out, and Fig. 6 shows the W for all the assessment criteria. ER classification combines accuracy and integrity of interpretation, which makes it easier to compare performance visually and understand it better.

D. Discussion

The results show that combining self-monitoring before learning with interpretive strategies can make models work better and be more open. This model is especially strong in noisy and low-resource conditions, which means that many classic monitoring approaches are no longer useful. Attention-based interpretation approaches and contextual dissemination strategies give people easy-to-understand descriptions. This makes it possible to transcribe, analyze emotions, and create interactive AI. Improves dependability and usability in situations where security is very important, such as IP systems.

This method finds a compromise between prediction performance, reward, and interpretation. In a subsequent study, it may be possible to extend these findings to include multilingual datasets and real-time configuration scenarios. The findings reveal that the framework that was proposed is an effective, end-to-end solution for tasks that involve voice and language understanding.

VI. CONCLUSION

This study has presented a unified transformer-based framework integrated with self-monitoring learning with internal interpretation approaches. The proposed model makes use of large audio and text corpora that are not marked in order to generate robust and permeable representations that improve the performance of tasks involving speech recognition and natural language understanding. A unified revision-based framework for speech and language comprehension was presented in the study. In the Librispeech and Common Voice datasets, the framework results in a considerable reduction in the word error rate (WER), an increase in the accuracy of classification in GLUE benchmarking tasks, and the production of attention maps that are consistent with language and phonological structures, as can be seen in Fig. 5 and Fig. 6.

To sum up, the proposed model shows that self-monitoring prior learning, along with interpretability, can make end-to-end speech and language comprehension systems work better, be more robust, and be more open. These results give us a reliable way to interact with computers, automatically transcribe, analyze emotions, and have agents talk to each other. A. It gives a good start for future growth.

VII. LIMITATIONS AND FUTURE WORK

Despite these improvements, there are still certain issues that need to be fixed. Implementing the deep transformer concept in a situation with limited resources can be difficult due to its computing needs. In order to assess the importance of attribution, the certification of interpretability currently relies on indirect truth references. This might not accurately represent human interpretation in the unfamiliar or complex domains.

There will also be an effort made to develop smaller, less complex transformer types that can be utilized in real-time and edge computing applications without jeopardizing accuracy or making it more difficult to comprehend. For the purpose of making the framework more effective in a variety of languages and geographical areas, the next phase in the research process will be to incorporate multilingual and cross-domain datasets into it. Additionally, human-centered evaluation techniques will be utilized in order to verify interpretability metrics in real-world scenarios. This serves the purpose of ensuring that model descriptions are both helpful and relevant for end users.

REFERENCES

[1] Meta adversarial learning improves low-resource speech recognition, Computer Speech & Language,Volume 84,2024,101576,ISSN 0885-2308, https://doi.org/10.1016/j.csl.2023.101576, https://www.sciencedirect.com/science/article/pii/S0885230823000955.

[2] Liu, Y., Yang, X. & Qu, D. A Forced Decoding-based Approach for Enhancing Low-resource ASR. Neural Process Lett 57, 45 (2025). https://doi.org/10.1007/s11063-025-11759-5.

[3] Latif, S., Zaidi, S. A. M., Cuayáhuitl, H., Shamshad, F., Shoukat, M., & Usama, M. (2025). Transformers in speech processing: Overcoming challenges and paving the future. Computer Science Review, 58, 100768. https://doi.org/10.1016/j.cosrev.2025.100768.

[4] Hung-yi Lee, Abdelrahman Mohamed, Shinji Watanabe, Tara Sainath, Karen Livescu, Shang-Wen Li, Shu-wen Yang, and Katrin Kirchhoff. 2022. Self-supervised Representation Learning for Speech Processing. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts, pages 8–13, Seattle, United States. Association for Computational Linguistics; 10.18653/v1/2022.naacl-tutorials.2.

[5] Yue, X.; Gao, X.; Qian, X.; Li, H. Adapting Pre-Trained Self-Supervised Learning Model for Speech Recognition with Light-Weight Adapters. Electronics 2024, 13, 190. https://doi.org/10.3390/electronics13010190.

[6] Gong, Z., Shi, P., Donbekci, K., Ai, L., Chen, R., Sasu, D., Wu, Z., & Hirschberg, J. (2025). Learning More with Less: Self-Supervised Approaches for Low-Resource Speech Emotion Recognition. In INTERSPEECH 2025. ISCA. DOI: 10.1109/ICASSP49660.2025.10887615.

[7] Eggen, M., Lysnæs-Larsen, J., & Strümke, I. (2025). Integrating attention into explanation frameworks for language and vision transformers. arXiv preprint arXiv:2508.08966. https://doi.org/10.48550/arXiv.2508.08966.

[8] Cai, D., Cai, Z., Li, Z., & Li, M. (2025). Self-Supervised Reflective Learning Through Self-Distillation and Online Clustering for Speaker Representation Learning. IEEE Transactions on Audio, Speech, and Language Processing, 33, 1535–1550. https://doi.org/10.1109/TASLPRO.2025.3555132.

[9] Chakrabarty, S., Bishwas, P., & Chatterjee, R. (2025). Explainable Transformer-CNN Fusion for Noise-Robust Speech Emotion Recognition. arXiv preprint arXiv:2512.18298. https://doi.org/10.48550/arXiv.2512.18298.

[10] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449–12460. https://doi.org/10.48550/arXiv.2006.11477.

[11] 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 Interspeech 2020 (pp. 5036–5040). ISCA. https://doi.org/10.21437/Interspeech.2020-3015.

[12] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291.

[13] Zhang, Y., Park, D. S., Han, W., Qin, J., Gulati, A., Shor, J., Jansen, A., Xu, Y., Huang, Y., Wang, S., Zhou, Z., Li, B., Ma, M., Chan, W., Yu, J., Wang, Y., Cao, L., Sim, K. C., Ramabhadran, B., . . . Wu, Y. (2022). BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. IEEE Journal of Selected Topics in Signal Processing, 16(6), 1519–1532. https://doi.org/10.1109/JSTSP.2022.3182537.

[14] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. In Proceedings of the 40th International Conference on Machine Learning (Vol. 202, pp. 28492–28518). PMLR. https://doi.org/10.48550/arXiv.2212.04356.

[15] Cai, D., Cai, Z., Li, Z., & Li, M. (2025). Self-supervised reflective learning through self-distillation and online clustering for speaker representation learning. IEEE Transactions on Audio, Speech, and Language Processing, 33, 1535–1550. https://doi.org/10.1109/TASLP.2025.3555132.

[16] Gong, Z., Shi, P., Donbekci, K., Ai, L., Chen, R., Sasu, D., Wu, Z., & Hirschberg, J. (2025). Learning More with Less: Self-Supervised Approaches for Low-Resource Speech Emotion Recognition. In Proceedings of the 26th Annual Conference of the International Speech Communication Association (Interspeech 2025) (pp. 5313–5317). ISCA. https://doi.org/10.1109/ICASSP49660.2025.10887615.

[17] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 5206-5210, doi: 10.1109/ICASSP.2015.7178964.

[18] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

[19] Yaqi Chen, Xukui Yang, Hao Zhang, Wenlin Zhang, Dan Qu, Cong Chen,

[20] Meta adversarial learning improves low-resource speech recognition, Computer Speech & Language,Volume 84,2024,101576,ISSN 0885-2308, https://doi.org/10.1016/j.csl.2023.101576, https://www.sciencedirect.com/science/article/pii/S0885230823000955.

[21] Sharrab, Y. O., Attar, H., Eljinini, M. A. H., Al-Omary, Y., & Al-Momani, W. E. (2025). Advancements in speech recognition: A systematic review of deep learning transformer models, trends, innovations, and future directions. IEEE Access, 13, 46925–46957. https://doi.org/10.1109/ACCESS.2025.3550855.

[22] Kim, J., & Lee, S. (2024). Contrastive masked prediction for efficient speech representation learning. Speech Communication, 156, Article 103014. https://doi.org/10.1016/j.specom.2023.103014.

[23] Liu, Y., Zhang, W., & Tan, T. (2025). Efficient self-supervised speech representation via temporal-frequency decoupled transformers. Pattern Recognition, 158, Article 111002. https://doi.org/10.1016/j.patcog.2024.111002.

[24] Eggen, M., Lysnæs-Larsen, J., & Strümke, I. (2025). Integrating attention into explanation frameworks for language and vision transformers. arXiv preprint arXiv:2508.08966. https://doi.org/10.48550/arXiv.2508.08966.

[25] Chakrabarty, S., Bishwas, P., & Chatterjee, R. (2025). Explainable Transformer-CNN Fusion for Noise-Robust Speech Emotion Recognition. arXiv preprint arXiv:2512.18298. https://doi.org/10.48550/arXiv.2512.18298.

[26] Yeh, S.-L., Meng, Y., & Tang, H. (2025). Whisper has an internal word aligner: Interpretability in end-to-end ASR. In Proceedings of the 2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE. https://doi.org/10.48550/arXiv.2509.09987.

[27] Bodria, F., Foscarin, F., Giannotti, F., & Guidotti, R. (2023). Explainable AI for audio: A review. ACM Computing Surveys, 55(11), Article 233, 1–38. https://doi.org/10.1145/3576055.

[28] Lin, T.-Q., Cheng, H.-C., Lee, H.-y., & Tang, H. (2025). Identifying Speaker Information in Feed-Forward Layers of Self-Supervised Speech Transformers. In Proceedings of the 2025 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 535–541). IEEE. https://doi.org/10.48550/arXiv.2506.21712.

[29] Djeffal, N., Addou, D., Kheddar, H., & Selouani, S. A. (2026). A robust framework for noisy speech recognition using Frequency-Guided-Swin Transformer. Computer Speech & Language, 98, Article 101907. https://doi.org/10.1016/j.csl.2025.101907.