Performance Optimization in the Cascade of VAD and ASR Systems: A Study on Evaluation and Alignment Strategies

Zhentao Lin*, Zihao Chen*§, Bi Zeng*, Leqi Chen* and Jia Cai†‡§
* Computer Science and Engineering, Guangdong University of Technology School
† China Electronic Product Reliability and Environmental Testing Research Institute
‡ Key Laboratory of MIIT for Intelligent Products Testing and Reliability
§ Corresponding Author

***Abstract*—Voice Activity Detection (VAD) is a critical pre-processing step in Automatic Speech Recognition (ASR) systems, tasked with distinguishing speech from non-speech segments in audio signals. The accuracy of this segmentation is essential, as errors can significantly impair ASR performance. This study integrates multiple VAD algorithms with ASR systems and analyzes their interaction by focusing on the Word Error Rate (WER) of ASR in relation to VAD's recall and precision. The research reveals a strong correlation between ASR's insertion and deletion errors and VAD's performance metrics. Insertions are mainly due to VAD misclassifying non-speech as speech, while deletions stem from misclassifying speech as non-speech. To address these issues, we introduce the Aligned Token Word Error Rate (atWER), a novel metric based on forced alignment from the Connectionist Temporal Classification (CTC) framework. Our experiments show that atWER accurately reflects the impact of VAD on ASR performance, effectively reducing discrepancies between VAD annotations and classification outcomes.**

I. Introduction

Voice Activity Detection (VAD) is a crucial initial step in Automatic Speech Recognition (ASR) systems, designed to accurately distinguish speech from non-speech elements within audio signals. The precision of this segmentation is vital, as non-speech interference—such as ambient noise or silence—can significantly impair ASR performance.

The accuracy of VAD is critical; incorrect classification of speech and non-speech can lead to a cascade of errors in ASR processing. Misclassifications may result in the loss of essential speech data or the inclusion of irrelevant noise, undermining the system's ability to accurately interpret and respond to spoken commands.

The correlation between VAD effectiveness and ASR error types is complex and significant [1]. The precision of the VAD mechanism significantly impacts the two predominant error categories in ASR systems, namely the occurrences of insertion and deletion errors.

- **Insertion Errors in ASR** occur when the system misinterprets non-speech sounds as speech, a failure of the VAD to filter out background noise. This leads to the ASR system processing unnecessary acoustic data, contaminating the output with false linguistic content and reducing reliability.
- **Deletion Errors in ASR** happen when the system misses actual speech segments due to the VAD misclassifying them as non-speech. This oversight can be due to stringent VAD detection criteria or inadequate representation of atypical speech in the training data, resulting in incomplete ASR outputs.

This study investigates the integration of VAD and ASR systems to enhance their performance. The methodology includes:

- A comparative temporal analysis of System VAD versus Oracle VAD using case studies, revealing a significant correlation between their metrics and temporal alignment.
- Experimental scrutiny was employed to examine the relationship between VAD and ASR metrics. The results indicated that when VAD and ASR systems are integrated sequentially, the composite performance metrics are predominantly determined by the VAD system's characteristics.
- Introduction of an algorithm for calculating Aligned Token Word Error Rate (atWER) to address error propagation challenges in the sequential deployment of VAD and ASR systems, mitigating misalignment between VAD annotations and classification outcomes.

II. Related Work

VAD, serving as a pivotal component in the pre-processing phase of ASR systems, has garnered escalating significance. The primary goal of VAD is to identify vocal segments within a continuous audio signal while reducing the impact of background noise and non-vocal sounds. This segmentation is vital for enhancing ASR system performance by minimizing incorrect interpretations and improving efficiency through reduced data processing.

Advancements in deep learning have propelled the development of VAD methods. Recurrent Neural Networks (RNNs) have been instrumental in capturing the temporal dynamics of audio signals [2]. Convolutional Neural Networks (CNNs) have also demonstrated effectiveness in feature extraction [3]. The Convolutional Recurrent Deep Neural Network (CRDNN) combines CNN and RNN features, utilizing convolutional layers for feature extraction and recurrent layers for time

series processing. This hybrid approach effectively captures relevant audio features to distinguish between speech and non-speech segments. The Feedback Sequential Memory Network (FSMN) leverages a feedforward architecture with memory units capable of storing and updating information, optimizing the handling of time series data [4].

Despite these advancements, challenges remain, particularly in the robustness of VAD systems when dealing with diverse accents, dialects, and noisy environments. The performance of VAD directly influences ASR outcomes. A recent study [5] introduces a multi-task learning framework that integrates VAD into the ASR system. This framework aims to improve ASR performance by using VAD alignment information during the training phase. [6] primarily introduces a novel method that applies Minimum Word Error training to a RNN to optimize VAD for improving the accuracy of speech recognition, especially on noisy data.

However, [7] point out that ASR systems are typically deployed in conjunction with a VAD system to operate ASR only on the voiced acoustic signals, thereby maintaining ASR performance by removing unnecessary non-speech parts from input audio signals during inference. However, if VAD fails to correctly split speech and non-speech segments, errors can propagate. Particularly in noisy environments or unknown acoustic domains, VAD is more prone to failure, which can trigger more significant insertion errors in ASR. [8] experimented with the hyperparameters of multi-channel VAD to mitigate the problems of insertion and deletion errors in ASR, but the approach continues to depend on the precision of the front-end System's VAD.

The problem of ASR accuracy being affected by VAD metrics such as Recall and accuracy is still resolved. Insertion errors in ASR can occur when VAD incorrectly classifies non-speech as speech, while deletion errors arise when speech is mistakenly categorized as non-speech.

III. Methods

A. System VAD and Oracle VAD

This section aims to delve into the intricacies of two significant approaches within the VAD domain: System VAD and Oracle VAD.

In System VAD, the decision-making process can be modeled as a binary classification problem, where the system must classify each frame of the audio signal as either speech or non-speech. A common approach is to define a decision function f_{vad} that takes a feature vector x_{vad} as input and outputs a binary decision y_{vad}:

$$y_{vad} = f_{vad}(x_{vad}, \theta) \tag{1}$$

Here, x_{vad} represents the feature vector extracted from the audio signal, which could include spectral, temporal, and statistical features. θ denotes the parameters of the VAD algorithm, which are typically learned from a training dataset.

Oracle VAD is an idealized model that represents the perfect detector with complete knowledge of the true speech segments. It serves as a theoretical benchmark to evaluate the performance of System VAD algorithms. The oracle detector can be mathematically represented as:

$$y_{oracle\_vad} = 1[x_{vad} \in Speech] \tag{2}$$

Here, 1 is the indicator function that returns 1 if the condition is true (i.e., the feature vector x_{vad} belongs to a speech segment) and 0 otherwise.

B. VAD Metrics

The objective of this research is to delve into the effectiveness of VAD systems by examining four critical evaluation metrics: Accuracy (Acc), Precision (Pre), Recall, and F1-Score (F1). These metrics are essential for assessing the performance of VAD algorithms, providing a comprehensive view of their reliability and efficiency in detecting speech segments.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Pre = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = 2 * \frac{Pre * Recall}{Pre + Recall} \tag{6}$$

where true positives (TP) is the number of segments correctly identified as speech, true negatives (TN) is the number of segments correctly identified as non-speech, false positives (FP) is the number of non-speech segments incorrectly identified as speech, and false negatives (FN) is the number of speech segments incorrectly identified as non-speech. Precision measures the proportion of true speech segments among all segments identified as speech by the system. Recall measures the proportion of true speech segments correctly identified out of all actual speech segments. The F1-Score provides a balanced measure, taking into account both precision and recall.

C. CTC-based and non-CTC-based ASR models

This section delves into the heart of ASR technology, examining two distinct approaches: CTC (Connectionist Temporal Classification)-based models and their non-CTC counterparts.

CTC-based ASR Models: The Connectionist Temporal Classification framework, introduced in [9], has been a pivotal breakthrough in the field of ASR. CTC-based models excel in handling variable-length input and output sequences, making them ideal for tasks where the alignment between the speech signal and the corresponding text is not predefined. This approach has been instrumental in the development of deep learning-based ASR systems, particularly those utilizing Transformer, RNN, and CNN. The CTC loss function, which allows for the direct alignment of input sequences with output labels, has been a key factor in the success of these models.

Non-CTC-based ASR Models: While CTC has set the standard for many ASR systems, alternative approaches have

emerged that challenge the status quo. Non-CTC-based models, such as those utilizing Attention Mechanisms and Transformer architectures, have gained prominence due to their ability to capture long-range dependencies and contextual information more effectively. These models often employ encoder-decoder frameworks that can directly predict the output sequence from the input sequence, offering a more interpretable and flexible approach to speech recognition.

D. ASR Metric

The WER (Word Error Rate) is an important metric for evaluating the performance of ASR systems. It measures the accuracy of recognition by quantifying the differences between the predicted text and the reference text. The formula for calculating WER is as follows:

$$WER = \frac{S + D + I}{N} \tag{7}$$

where S represents the number of substitution (SUB) errors, D represents the number of deletion (DEL) errors, I represents the number of insertion (INS) errors, and N represents the total number of words or characters in the reference text. The lower the WER value, the better the performance of the ASR system.

E. atWER

In this section, we will focus on introducing the atWER method, which mainly includes two steps: the alignment of the hypothesis and label, and the calculation of atWER.

We utilized the U2 Attention rescoring method [10] to recognize speech after VAD segmentation. Meanwhile, the CTC algorithm effectively addressed the alignment issue in sequence labeling by introducing a blank symbol and dynamic programming techniques.

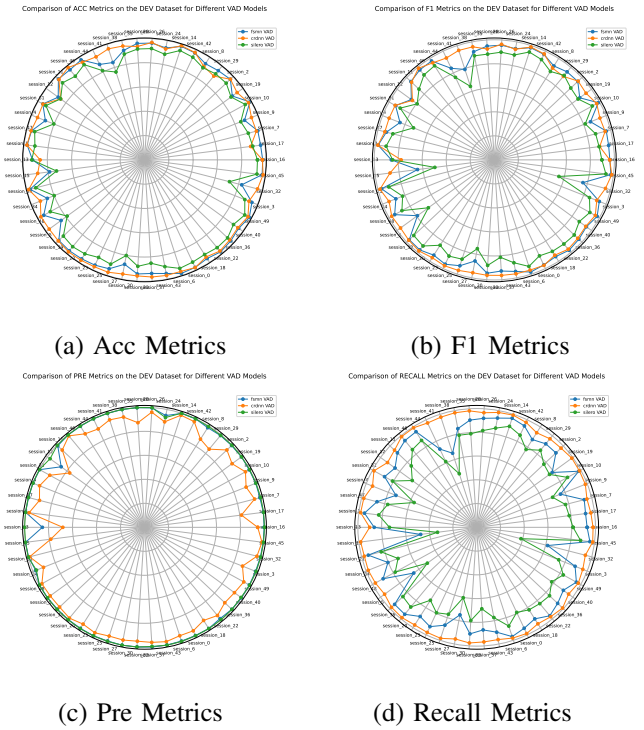
Specifically, the CTC algorithm initially procures the n-best candidate ensemble in a sequential stream, which is then subjected to the attention-based rescoring mechanism. This mechanism refines the n-best candidates through an apportioned scoring methodology, culminating in the derivation of a hypothesis that encapsulates sequence alignment data. Ultimately, the alignment outcomes at the character tier, predicated on the aforementioned hypothesis, are harmonized with the reference label of the Oracle VAD. The token is deemed credible when both the System VAD and the Oracle VAD are concurrently operational.The formula for calculating atWER is as follows:

$$atWER = \frac{S_c + D_c + I_c}{N_c} \tag{8}$$

where S_c represents the number of SUB errors among the credible tokens, D_c represents the number of DEL errors among the credible tokens, I_c represents the number of INS errors among the credible tokens, and N_c represents the total number of words or characters in the reference text within the credible tokens. The lower the atWER value, the better the performance of the ASR system.

VAD Model	Acc	F1	Recall	Pre
FSMN VAD	0.92	0.9	0.84	0.99
CRDNN VAD	0.73	0.71	0.73	0.7
Silero VAD	0.89	0.85	0.75	1.0

TABLE I
COMPARISON OF DIFFERENT SYSTEM VADS.



(a) Acc Metrics (b) F1 Metrics

(c) Pre Metrics (d) Recall Metrics

Fig. 1. Radar chart of the performance of different systems in different sessions on the DEV dataset.

IV. EXPERIMENT

This section delineates the specific contents of the experiments conducted in this study.

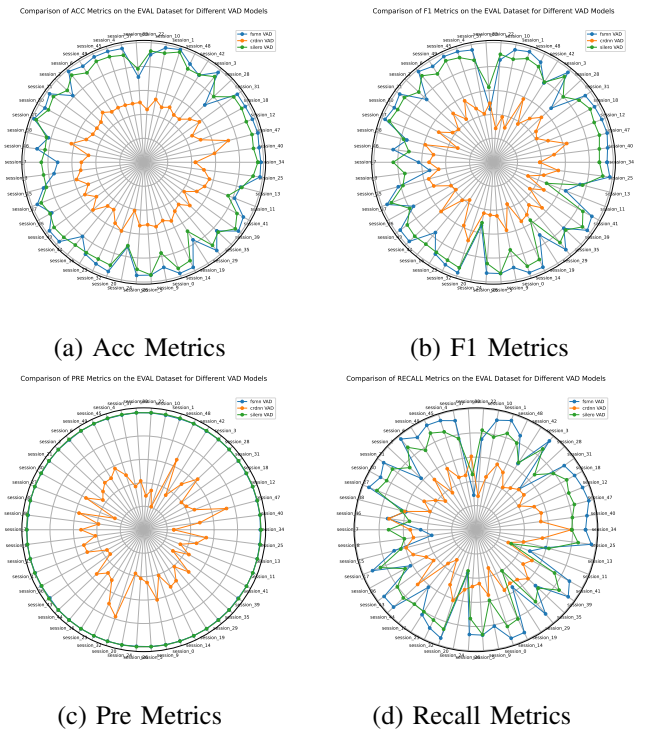
A. VAD System Evaluation

This section assesses the performance of three VAD models—the FSMN [4], CRDNN, and Silero [11] VAD models—on the LibriParty [1] dataset using the metrics introduced in Section III-E.

Initially, the experiment evaluated the performance of three VAD models using the Development (DEV) dataset, calculating metrics such as Accuracy, F1-score, Precision, and Recall. The results were compared using radar charts, as depicted in Figure 1. Figures 1.a and 1.b illustrate that the CRDNN VAD model outperforms the other two models in overall performance. However, Figure 1.c reveals a higher false positive rate for the CRDNN VAD model, indicating a propensity to misclassify non-speech segments as speech.

Subsequently, the experiment assessed the models' scores on the Evaluation (EVAL) dataset, as shown in Figure 2.

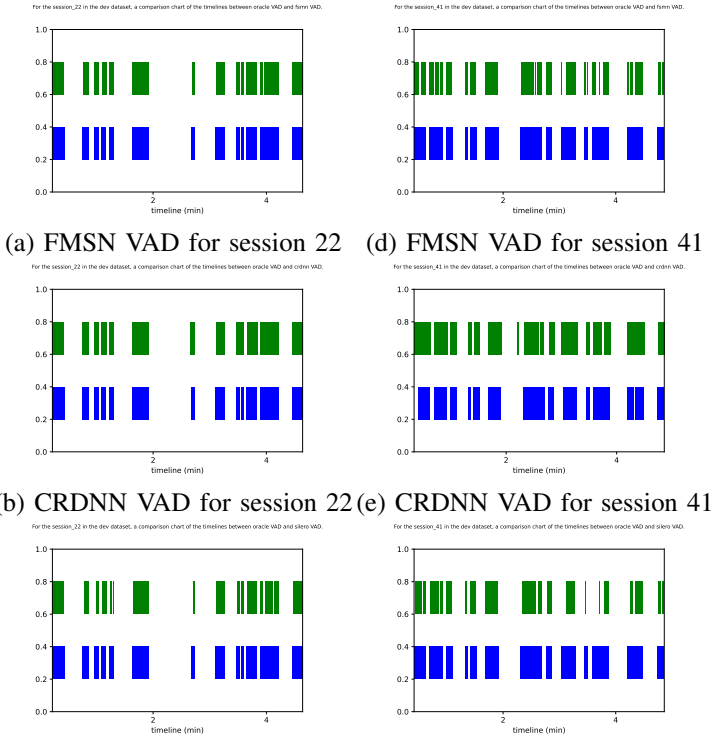
[1]https://github.com/speechbrain/speechbrain/tree/develop/recipes/LibriParty



(a) Acc Metrics (b) F1 Metrics

(c) Pre Metrics (d) Recall Metrics

Fig. 2. Radar chart of the performance of different systems in different sessions on the EVAL dataset.



(a) FMSN VAD for session 22 (d) FMSN VAD for session 41

(b) CRDNN VAD for session 22 (e) CRDNN VAD for session 41

(c) Silero VAD for session 22 (f) Silero VAD for session 41

Fig. 3. Sampled dev_session_22 and dev_session_41 from the DEV dataset, the alignment results of the Oracle VAD (blue) and the System VAD (green) on the timeline.

The findings diverged from the previous experiment, with the CRDNN VAD model significantly underperforming compared to the other models. Figures 2.a and 2.b demonstrate that the FSMN VAD model achieved the best performance, closely followed by the Silero VAD model. The latter's slightly lower performance is primarily attributed to its tendency to misclassify speech segments as non-speech, as illustrated in Figure 2.d.

The paper further illustrates the performance of the VAD models through visual representations of sample results in Figures 3 and 4. These figures delineate the manual segmentation outcomes in blue and the model predictions in green. Notably, in Figure 3.b and 3.e, the CRDNN VAD model is observed to misclassify non-speech frames as speech. Conversely, Figures 3.a, 3.c, 3.d, and 3.f indicate a tendency for both the FSMN and Silero VAD models to classify speech frames as non-speech, a behavior that is also evident in Figure 4. The sample analyses corroborate the findings from the radar chart assessments, reinforcing the consistency of the models' performance across different evaluation methods.

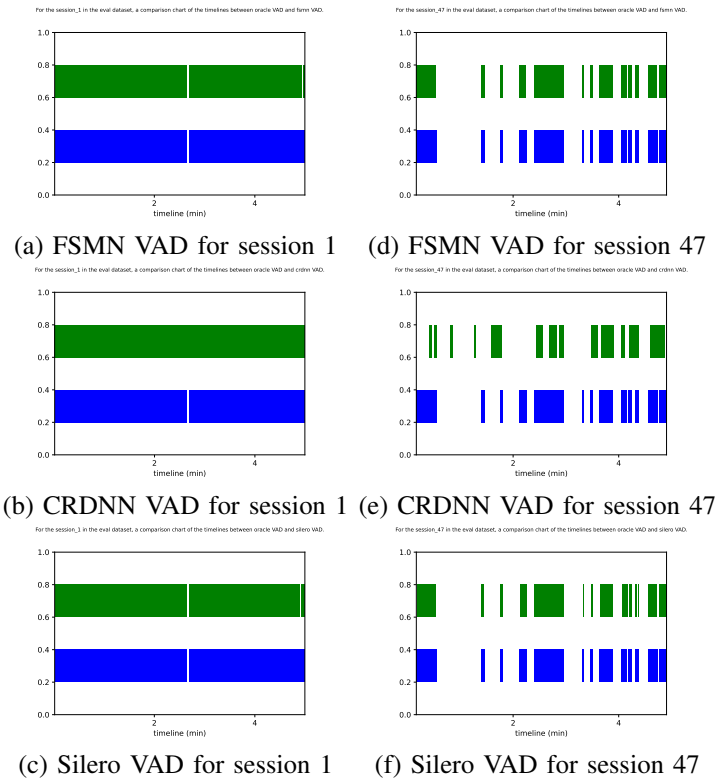
This experiment also systematically documents the comparative performance of the three VAD models across two datasets in Table I. The FSMN VAD model leads in performance, closely followed by the Silero VAD model, with the CRDNN VAD model ranking third. A minor performance discrepancy exists between the FSMN and Silero models, particularly in the Recall metric. This discrepancy aligns with our observations that the Silero VAD model is more prone to false negatives.

B. VAD cascading ASR System Evaluation

To assess the impact of cascading VAD systems on ASR systems, we integrated the three VAD systems from Section IV-A with the Paraformer [12] model and the SenseVoice [13] model, respectively, and measured the evaluation outcomes based on WER and atWER. The results are presented in Table II and III. In accordance with the principles of atWER, this experiment aligned the System VAD outcomes with the VAD labels, evaluating ASR results only under true positive conditions. It was observed that all three types of ASR errors were reduced, with the most significant decrease observed in deletion errors.

We argue that the atWER metric excludes false positive cases, where System VAD misclassifies non-speech as speech, from the evaluation. Consequently, the content erroneously identified as speech does not influence the atWER assessment. Additionally, atWER does not consider false negatives, where speech is misclassified as non-speech, resulting in omitted content in ASR recognition. We posit that these misclassifications are beyond the purview of ASR system enhancements and should not be included in ASR evaluation metrics. As a result, in our statistical analysis, we have excluded insertion and deletion errors associated with these scenarios.

Furthermore, we observed an unexpected increase in substitution errors, which seemingly conflicts with the anticipated effects of atWER. This anomaly can be explained by the



(a) FSMN VAD for session 1 (d) FSMN VAD for session 47

(b) CRDNN VAD for session 1 (e) CRDNN VAD for session 47

(c) Silero VAD for session 1 (f) Silero VAD for session 47

Fig. 4. Sampled eval_session_1 and eval_session_47 from the EVAL dataset, the alignment results of the Oracle VAD (blue) and the System VAD (green) on the timeline.

Exp	Evaluation Method	WER	SUB	INS	DEL
E1: FSMN VAD	WER	40.0	2121	198	6212
	atWER	38.88	2122	198	5821
E2: CRDNN VAD	WER	31.75	2250	304	4727
	atWER	30.24	2240	208	4379
E3: Silero VAD	WER	40.02	2695	202	6281
	atWER	38.83	2685	191	5861

TABLE II
WER RESULTS AND ATWER RESULTS OF THE VAD-CASCADED ASR SYSTEM EVALUATION ON THE DEV DATASET.

fact that atWER, by eliminating insertion errors, paradoxically generates new deletion errors. These deletions, in turn, can merge with existing errors, transforming them into a single substitution error.

V. CONCLUSIONS

This research presented highlights the significant correlation between the effectiveness of VAD and the types of errors encountered in ASR systems, specifically insertion and deletion errors. To address the challenges posed by insertion and deletion errors, this paper introduces the atWER metric calculation algorithm. This novel approach aims to resolve the alignment discrepancies between VAD labels and classification results, thereby enhancing the reliability and accuracy of both VAD and ASR systems. By improving the accuracy of speech detection and addressing the alignment issues between VAD and ASR, we can significantly enhance the user experience

Exp	Evaluation Method	WER	SUB	INS	DEL
E1: FSMN VAD	WER	34.5	2462	285	5165
	atWER	33.93	2444	215	5092
E2: CRDNN VAD	WER	67.45	1482	76	12828
	atWER	57.53	1359	56	7990
E3: Silero VAD	WER	42.38	2678	196	6164
	atWER	39.67	2681	196	5201

TABLE III
WER RESULTS AND ATWER RESULTS OF THE VAD-CASCADED ASR SYSTEM EVALUATION ON THE EVAL DATASET.

and the practical utility of speech recognition technologies in various applications.

ACKNOWLEDGMENT

This work was supported by Key Laboratory of MIIT for Intelligent Products Testing and Reliability 2023 Key Laboratory Open Project Fund(No.CEPREI2023-01).

REFERENCES

[1] S. Tong, N. Chen, Y. Qian, and K. Yu, "Evaluating vad for automatic speech recognition," in *2014 12th International Conference on Signal Processing (ICSP)*, IEEE, 2014, pp. 2308–2314.

[2] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132 306, 2020.

[3] S. Tong, H. Gu, and K. Yu, "A comparative study of robustness of deep learning approaches for vad," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, pp. 5695–5699.

[4] S. Zhang, M. Lei, Z. Yan, and L. Dai, "Deep-fsmn for large vocabulary continuous speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 5869–5873.

[5] M. Li, Y. Xia, and F. Lin, "Incorporating vad into asr system by multi-task learning," in *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, 2022, pp. 160–164.

[6] G. Gelly and J.-L. Gauvain, "Minimum word error training of rnn-based voice activity detection.," in *INTERSPEECH*, 2015, pp. 2650–2654.

[7] S. Novitasari, T. Fukuda, and G. Kurata, "Improving asr robustness in noisy condition through vad integration.," in *INTERSPEECH*, 2022, pp. 3784–3788.

[8] O. Ichikawa, K. Nakano, T. Nakayama, and H. Shirouzu, "Multi-channel vad for transcription of group discussion.," in *Interspeech*, 2021, pp. 336–340.

[9] P. Paterlini-Brechot and N. L. Benali, "Circulating tumor cells (ctc) detection: Clinical impact and future directions," *Cancer letters*, vol. 253, no. 2, pp. 180–204, 2007.

[10] Z. Yao, D. Wu, X. Wang, *et al.*, "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Proc. Interspeech*, IEEE, Brno, Czech Republic, 2021.

[11] S. Team, "Silero vad: Pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," *Retrieved March*, vol. 31, p. 2023, 2021.

[12] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," in *INTERSPEECH*, 2022.

[13] T. SpeechTeam, "Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms," *arXiv preprint arXiv:2407.04051*, 2024.