Enhancing Dialogue Speech Recognition with Robust Contextual Awareness via Noise Representation Learning

Wonjun Lee [*1], **San Kim** [*2] and **Gary Geunbae Lee** [1,2]
[1] Department of Computer Science and Engineering, POSTECH, Republic of Korea
[2] Graduate School of Artificial Intelligence, POSTECH, Republic of Korea
{lee1jun, sankm, gblee}@postech.ac.kr

Abstract

Recent dialogue systems rely on turn-based spoken interactions, requiring accurate Automatic Speech Recognition (ASR). Errors in ASR can significantly impact downstream dialogue tasks. To address this, using dialogue context from user and agent interactions for transcribing subsequent utterances has been proposed. This method incorporates the transcription of the user's speech and the agent's response as model input, using the accumulated context generated by each turn. However, this context is susceptible to ASR errors because it is generated by the ASR model in an auto-regressive fashion. Such noisy context can further degrade the benefits of context input, resulting in suboptimal ASR performance. In this paper, we introduce Context Noise Representation Learning (CNRL) to enhance robustness against noisy context, ultimately improving dialogue speech recognition accuracy. To maximize the advantage of context awareness, our approach includes decoder pre-training using text-based dialogue data and noise representation learning for a context encoder. Based on the evaluation of speech dialogues, our method shows superior results compared to baselines. Furthermore, the strength of our approach is highlighted in noisy environments where user speech is barely audible due to real-world noise, relying on contextual information to transcribe the input accurately.

1 Introduction

Automatic Speech Recognition (ASR) is central in accurately interpreting human speech, serving as a fundamental resource for numerous subsequent downstream tasks. The advent of robust ASR modules, such as wav2vec2.0 (Baevski et al., 2020) and Whisper (Radford et al., 2023), has significantly enhanced the capabilities of ASR systems, facilitating their integration into a wide array of research and application domains. The integration of ASR modules into various works highlights the pivotal role of ASR in enhancing human-computer interaction, signifying a notable development in interactive technologies.

Despite the successful advancement of the ASR system, its inaccuracy poses significant risks to the efficacy of downstream tasks, such as speech-to-text translation (Liu et al., 2020; Le et al., 2024; Tang et al., 2021) and spoken language understanding (Serdyuk et al., 2018; Arora et al., 2022; Huang and Chen, 2020). These tasks predominantly rely on the textual output generated by ASR systems, highlighting the importance of accuracy in the initial speech recognition process. Especially for the dialogue system, the quality of the ASR system is paramount to ensure seamless interaction between user and dialogue agent, as models trained on written conversations perform poorly on spoken data (Kim et al., 2021). To minimize the impact of ASR error on the dialogue model, various endeavors have been made. Jiang et al. (2023) used an ASR correction module which employs multiple ASR models, while others focused on augmenting data with plausible ASR errors (Park et al., 2023; Wang et al., 2020; Tian et al., 2021). However the limitation is evident as they primarily focus on the robustness of dialogue models, which may not address the core issue compared to directly rectifying ASR models.

Conversely, incorporating a context encoder for dialogue history to improve the ASR model has been proposed, resulting in notable performance enhancements (Ortiz and Burud, 2021; Shenoy et al., 2021; Hou et al., 2022; Hori et al., 2020). Nevertheless, since the context is transcribed at each turn by the ASR model, it may contain errors, potentially disrupting the use of contextual information.

In this work, we present a novel Context Noise Representation Learning (CNRL) method to encode accurate contextual information, even from

[*] Equally contributed

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 333–343
September 18–20, 2024.

noisy ASR transcriptions. This approach aims to improve the performance of speech recognition in Task Oriented Dialogue (TOD) by minimizing the impact of ASR errors in dialogue history as context. Furthermore, we explore the advantages of decoder pre-training in context-aware ASR systems, emphasizing their improved robustness in noisy environments. The overall training pipeline can be decomposed by three steps: 1) Decoder pre-training on text-based dialogue data between user and agent. 2) ASR fine-tuning with speech encoder and context encoder jointly. 3) CNRL on context encoder to minimize the impact of ASR-noise context. Our contributions are as follows:

- We propose a novel training pipeline for dialogue speech recognition that leverages the dialogue history between user and agent.
- We demonstrate the effectiveness of CNRL by comparing it to various baseline models, showing a relative 13% reduction in Word Error Rate (WER) compared to the current state-of-the-art ASR model (Radford et al., 2023).
- In evaluations conducted in highly noisy environments, our model exhibits robust transcription accuracy, achieving up to a 31.4% reduction in WER compared to the baseline.

2 Related work

2.1 Context-aware speech recognition

Several studies have shown that leveraging contextual information in dialogue scenarios can enhance ASR performance. Shenoy et al. (2021) used a context carry-over mechanism to enhance the recurrent model's accuracy. Hou et al. (2022) proposed utilizing a context encoder in RNN-T architecture, adopting the semantic embedding of dialogue context from BERT (Devlin et al., 2019). Hori et al. (2020) targeted considering long-context by sliding-window fashion. Wang et al. (2023) and Wang et al. (2024) proposed an audio-augmented retriever to directly transcribe and track the dialogue state. These Context-Aware ASR(CA-ASR) models have a potential drawback: the context generated for each turn is based on ASR transcriptions, which inevitably contain errors, potentially degrading context-awareness. In this paper, we introduce the CNRL method, which trains only the context encoder independently. The goal is to enable the context encoder to produce similar encoding for noisy (ASR output) contexts to match clean context.

2.2 Decoder pre-training

Compared to pre-training encoder layers (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022), pre-training the decoder for ASR has received comparatively less attention. Notably, in scenarios where input speech is flawed or incomplete, the decoder can still play a crucial role in transcribing user utterances by leveraging contextual language modeling. To harness the decoder's capabilities, the use of external datasets like phoneme-to-grapheme paired data (Masumura et al., 2020) or text data (Gao et al., 2021) has been suggested. This approach enables the model to benefit from numerous external, non-paired data sources. Tsunoo et al. (2023) trained decoder for both ASR task and language modeling task, enabling improved linguistic understanding and leading to better ASR performance. Following these works, we pre-trained the decoder for a context-aware ASR model using voluminous text-only data. Specifically, we focus on turn-based dialogue data between user and agent, where each utterance is highly related to each other.

2.3 Noise Representation Learning

Noise in input data is inevitable in various forms across many datasets. Training models with such data negatively impacts their generalization performance. To address this challenge, numerous studies have adopted contrastive learning to enhance model robustness. Ma et al. (2023) improved named entity recognition performance by employing a token-level dynamic loss function and contrastive learning, leveraging noisy data and accounting for noise-distribution changes during training. Xu et al. (2023) enhanced contrastive learning through a dimension-wise method to mitigate feature corruption in sentence embeddings. Sun et al. (2023) used a K-NN graph to identify confident samples and applied mixup supervised contrastive learning to create robust representations, leading to improved relation extraction performance. Zheng et al. (2023) utilized both class-wise and instance-wise contrastive learning in their novel representation learning module. In this work, we adopt representation learning to enhance context awareness when noisy ASR transcriptions are used for context. The proposed CNRL is integrated solely with the context encoder in the CA-ASR model to

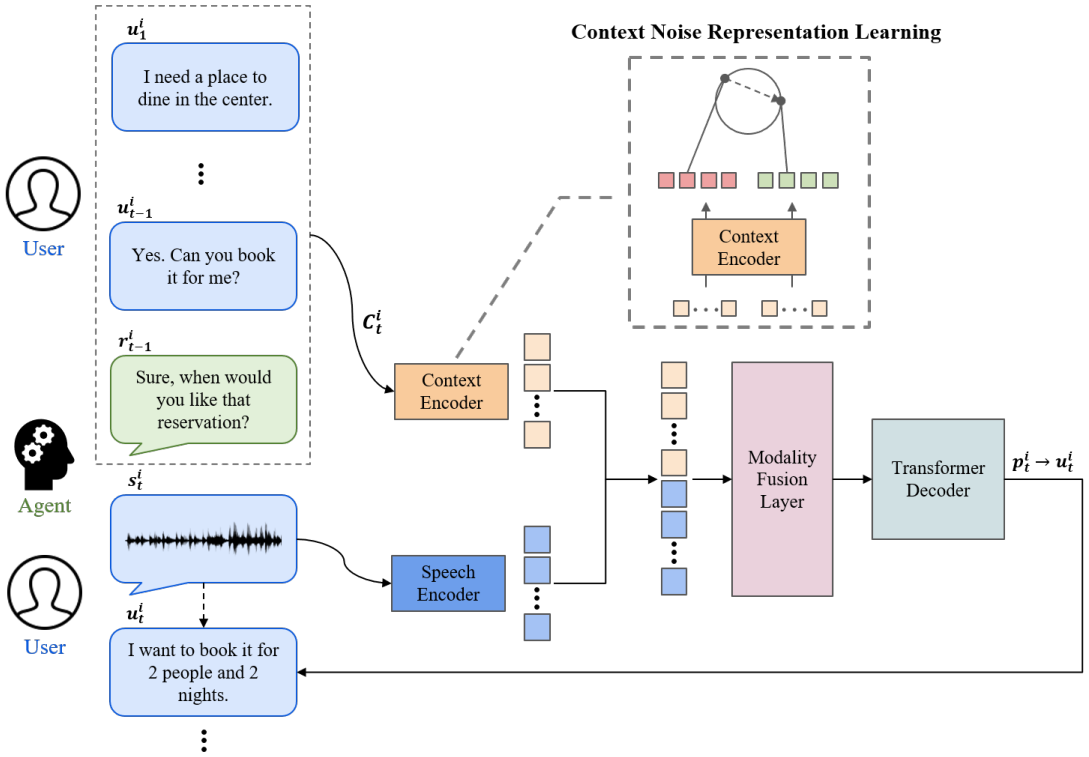

Figure 1: The architecture of a Context-Aware ASR(CA-ASR), featuring separate speech and context encoders to process the user's current speech s_t^i and dialogue history C_t^i, respectively. These representations are concatenated and fused using a modality fusion layer and transcribed to the predicted user utterance p_t^i by the transformer decoder. The predicted user utterance will be added to context ($p_t^i \rightarrow u_t^i$) for the next turn ($t+1$). After the training, the context encoder can improve itself by our CNRL method, detailed in Figure 2 and Section 3.3.

minimize training costs.

3 Methodology

3.1 Preliminary

We define D_t^i as the turn-based dialogue dataset for turn t in the i-th dialogue, which includes the speech input s_t^i, the corresponding text labels u_t^i (transcriptions) of user utterances, and the turn-based dialogue history $C_t^i = (u_1^i, r_1^i, ..., u_{t-1}^i, r_{t-1}^i)$, accumulating up to turn $t-1$, where r_t^i represents the agent's response at turn t. Each dialogue instance at the k-th turn, denoted as (u_k^i, r_k^i), comprises a single-turn conversation consisting of both a user utterance and an agent response. During inference, the predicted utterance (transcription) from model p_t^i is used instead of u_t^i for user utterance to form context C_t^i.

The CA-ASR model integrates the user's speech and dialogue history. For each turn t, the model predicts the current user utterance u_t^i from the speech input s_t^i and the context C_t^i. The dialogue history comprises text logs from both the user and the agent, where the user's speech is transcribed in real-time, while the agent's responses are given in text format. To transcribe the user's speech at turn t, the model draws upon past conversations from turn 1 to $t-1$. Utilizing an encoder-decoder architecture for the CA-ASR model, dedicated encoders initially process each input type—speech and text. These encodings are then concatenated and fused through a modality fusion layer, yielding a fused representation. Subsequently, the fused representation is passed through a decoder layer to transcribe the user utterance. Figure 1 illustrates the CA-ASR architecture, highlighting the interaction between user utterances and agent responses.

3.2 Decoder pre-training for Dialogue

We adopt a pre-training method specifically targeting decoders in the CA-ASR model. This method employs an encoder-decoder architecture, where the model takes the text-form dialogue history C_t^i as input. For the output, since the decoder is eventually used for transcribing user utterances, it aims to predict the next user utterance u_t^i. Additionally, the utilization of text data as input enables the training process to use external text datasets, further enhancing the decoder's performance. We demonstrate

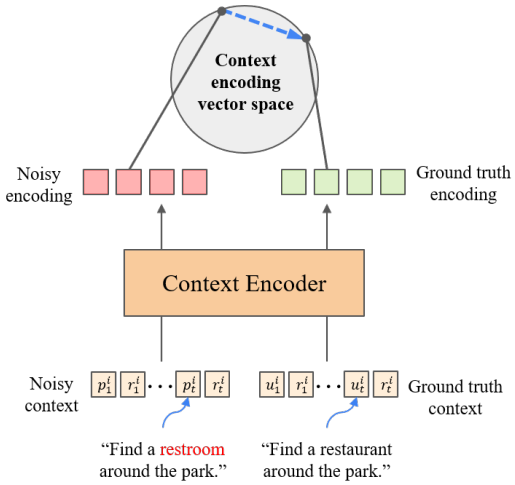

Figure 2: **Context Noise Representation Learning:** The noisy context including user utterances generated by the CA-ASR model during inference (p_t^i), and the ground truth context with clean user utterances (u_t^i), are encoded by the context encoder. The noisy encoding is adjusted to closely match the ground truth encoding in the context encoding vector space.

this efficacy in Section 5.2 . This approach enables the decoder to anticipate the subsequent user utterance based on contextual information derived from the dialogue history. This training method is particularly effective because dialogues in TOD are more predictable from the dialogue history than other types of conversations. In typical user-agent interactions, the agent often asks specific questions, and the user responds with relevant answers, making the dialogue structure more consistent and easier to predict.

When integrated into the CA-ASR model and fine-tuned for ASR tasks, the pre-trained decoder can significantly enhance transcription performance. By leveraging its ability to anticipate user responses from the agent's response (or the entire dialogue history), the decoder contributes to more accurate and robust transcription results, even with imperfect input speech, such as noisy audio signals.

3.3 Context Noise Representation Learning

During inference, the CA-ASR model uses context from previous transcriptions of user utterances and agent responses. However, inaccuracies in the ASR-generated transcriptions can degrade the advantage of using context, as training typically uses only ground truth context for each turn. To address this, we introduce CNRL. This method involves an additional training step where the model transcribes and utilizes noisy transcriptions to train the context encoder in a representation learning manner, as illustrated in Figure 2. The context encoder is fine-tuned to generate similar encoding for noisy input context as it does for the ground truth context. This method focuses solely on enhancing the context encoder, maintaining training efficiency.

To create the training set for CNRL, we first generate noisy transcriptions using the CA-ASR model with the ASR training set (See Section 4.1) divided into 10 folds. In each fold, 90% of the training set is used to train the CA-ASR model, and the remaining 10% is used to generate noisy ASR transcriptions. By iterating through all 10 folds, we obtain a complete noisy context training set. The dataset for CNRL comprises pairs of noisy and ground truth contexts, each containing multiple conversation turns. Each turn pairs a user utterance with an agent response, except for the initial turn, which consists only of the user's utterance.

We trained context encoder with cosine embedding loss:

$$\text{loss}(x, y) = \begin{cases} 1 - \cos(x_1, x_2), & \text{if } y = 1 \\ \max(0, \cos(x_1, x_2) - \text{margin}), & \text{if } y = -1 \end{cases} \quad (1)$$

Where x_1 is the encoding vector from the context encoder within the ASR-generated context and x_2 is the encoding from ground truth context. y is the label that indicates these two (x_1 and x_2) are of the same class($y = 1$) or not($y = -1$). Since we trained the context encoder to generate a similar output encoding for the noisy input (x_1) to match the clean ground truth (x_2), we set $y = 1$ for training. During training, x_1 gets close to x_2 on context encoding vector space, ensuring the context encoder produces similar encoding for a given noisy context. By using CNRL, the context encoder can maintain accurate context information, leading to improved speech recognition accuracy.

4 Experimental setup

4.1 Datasets

The DSTC11 Challenge Dataset The DSTC11 (Soltau et al., 2022) dataset is derived from the MultiWoZ 2.1 (Eric et al., 2020) by adding speech recordings and synthesized voices generated by a TTS model. The training set is built using the TTS model, while the evaluation sets are recorded by human volunteers. Each dialogue consists of audio files of user utterances paired with corre-

sponding agent responses. In every dialogue, the user initiates the conversation, making the first user utterance has no preceding context.

Since the official transcription for the DSTC11 test split (test-dstc11.human-verbatim) is not publicly available, we evaluate our experiments on the DSTC11 development split with human recording (dev-dstc11.human-verbatim)[1] as test set. Additionally we randomly sampled 3000 audios from the training set and used them as our development set during training.

The DSTC11 training set consists of 8,434 dialogues comprising 56,750 user utterances synthesized by four TTS voices, generating a total of 227,000 audio files. Our development set, randomly sampled from the training set, contains 3000 user utterances and is excluded from the training data. The test set includes human recordings of 7,374 user utterances from 1,000 dialogues. The average audio duration is 3.31 seconds for the training and development sets and 5.35 seconds for the test set.

Evaluation in Noisy Environments Environmental noise is a significant challenge for ASR systems in real-world scenarios. However, contextual information can mitigate this issue. To test our ASR system's resilience to real-world noises, we use the ESC-50 dataset (Piczak, 2015), which includes 50 classes of common urban noises, such as drilling and sirens. Noise samples are randomly selected from 2000 audio files and injected into our test set at Signal-to-Noise Ratios (SNR) of 20dB and 0dB, representing soft and hard noise conditions, respectively. This evaluation replicates challenging acoustic environments to test the ASR system's robustness rigorously. Note that the noisy audio is used exclusively for evaluation, not training. Our goal is to show that contextual information can be helpful in noisy environments where the audio signal is significantly degraded.

Decoder pre-training To facilitate the use of context information, we first trained CA-ASR's decoder using exclusively text-based data before ASR fine-tuning. For this purpose, we employ large datasets of turn-based dialogue text, combining the Schema-Guided Dialogue (SGD) (Rastogi et al., 2020) dataset with the DSTC11 text dataset to pre-train the decoder. SGD consists of over 20,000 task-oriented conversations between human and virtual assistant. From 8434 English dialogues from DSTC11 and approximately 16,000 English dialogues from the SGD training dataset, we use about 260,000 turn conversations. To evaluate the effect of decoder pre-training, we varied the volume of text data used for this process. The effects of these variations are detailed in Table 2.

4.2 Model configuration

Baselines We compare our CA-ASR model against several baselines, including those reported in DSTC11 (Soltau et al., 2022) and the current state-of-the-art ASR model Whisper (Radford et al., 2023). Additionally, we present a model that uses wav2vec2.0 (Baevski et al., 2020) as the encoder and BART (Lewis et al., 2019) as the decoder. This model shares the same architecture as the CA-ASR model, except for removing the context encoder and modality fusion. For transcription post-processing, we normalize common English patterns (e.g., "I've" to "I have"), remove punctuation, and normalize digits to ensure a fair comparison between models.

Context-Aware ASR Compared to the baselines, the CA-ASR model leverages previous user utterances and agent responses as textual input to enhance transcription accuracy. To encode this contextual information, CA-ASR uses the BART encoder as the context encoder. The speech encoder is wav2vec2.0 with the checkpoint *wav2vec2-large-960h*[2], and the pretrained BART encoder and decoder with the checkpoint *bart-large*[3] are utilized as the context encoder and the CA-ASR decoder, respectively. Given that the maximum token length for BART-large is limited to 1024, we truncate the context to the last 1024 tokens if necessary.

For modality fusion, the wav2vec2.0 speech encoder and the BART context encoder each produce hidden representations with dimensions of token $\times$ 1024. Since the BART decoder requires an encoder hidden state with a dimension of 1024, we concatenate these hidden representations along the 1024 dimension. This concatenated representation is then passed through a linear layer (1024, 1024) with ReLU activation to create a fused representation. This fused representation is subsequently fed into the BART decoder to transcribe the user utterance.

Total parameter size of our model is 774M, consisting of 315M for the speech encoder, 203M for

[1] https://storage.googleapis.com/gresearch/dstc11/dstc11_20221102a.html

[2] https://huggingface.co/facebook/wav2vec2-large-960h

[3] https://huggingface.co/facebook/bart-large

the BART context encoder, 254M for the BART decoder, and 1M for the linear fusion layer.

4.3 Training configuration

Our training pipeline consists of three sequential steps: decoder pretraining, ASR fine-tuning with audio masking, and CNRL. We evaluate the effect of each step in the subsequent Result & Analysis section.

Decoder pre-training We initially adopt the BART encoder-decoder model to pre-train the decoder, which is subsequently used for ASR fine-tuning. The optimization is performed using the AdamW algorithm (Loshchilov and Hutter, 2017) with $(\beta_1, \beta_2) = (0.9, 0.999)$, learning rate of 5e-5, weight decay of 1e-5, and a batch size of 32. We select the best model based on the lowest validation loss over 10 epochs of training, spanning 50 hours. The encoder functions as the context encoder, while the decoder serves as the transformer decoder in the CA-ASR model. Utilizing Cross-Entropy loss, we aim to input the dialogue history with the agent's response, which is the last turn, into the encoder and generate the user's response as the output from the decoder.

ASR fine-tuning In ASR fine-tuning stage, a speech encoder (wav2vec2.0) is attached to the pre-trained BART decoder from decoder pre-training. We adopt a batch size of 64 and an Adam optimizer with a learning rate of 2e-5. Across 10 epochs of training for 20 hours, the model with the lowest WER on development set at the end of each epoch was chosen as the best model for the speech encoder.

Audio masking Motivated by other multi-modal ASR study (Shi et al., 2022), a small portion of the speech data is obscured by masking to reduce the model's reliance on speech input. Specifically, 10% of speech data are randomly chosen for masking, and each selected data is masked for 20% of its total duration. Note that this configuration of masking probability and duration was empirically determined to yield optimal results in our experiments, with the proportion of masked data and masking length varied between 10% to 30% and 10% to 50%, respectively. To implement the masking process, we segment each audio into discrete chunks of 1-second duration. These chunks serve as the minimum unit for the masking, e.g. in an audio input with a duration of 10 seconds, two randomly chosen chunks would be masked. Unless otherwise specified, all results of the CA-ASR model include audio masking during training.

CNRL Setup We utilized the noisy context training set from the 10-folds described in Section 3.3. The average WER for the noisy context was 6.53% across the 10 folds. We filtered out transcriptions with a WER exceeding 20% to prevent interference with CNRL, resulting in the exclusion of 8.2% of the noisy context training set. We evaluated the effect of CNRL noisy context data by modifying the dialogue turns and introducing arbitrary word drops. For arbitrary word drop, we remove words for user utterances from golden context by 10% of change for each word and iterate it until we match the WER for each dialogue up to 6.5%, which is similar to WER with 10-folds. The training data setups are listed below:

- **S1**: Arbitrarily remove words from the golden context (user utterance only) to match an average WER of 6.5%.
- **S2**: Using the 10-fold training set, only the last user utterance contains noisy text.
- **S3**: Using the 10-fold training set, all user utterances may contain noisy text.
- **S4**: Combining S1 with S3. If a user utterance for each turn does not contain noisy text, arbitrary word drops are applied to increase the noise.

Unless otherwise specified, subsequent experimental results with CNRL use the **S4** setup. We use a batch size of 128 and the Adam optimizer with a learning rate of 5e-4. Training is conducted for up to 5 epochs, selecting the epoch with the lowest cosine embedding loss on our development set.

All experiments are conducted using 4 NVIDIA A6000 GPUs.

5 Result & Analysis

5.1 Context Aware-ASR

Table 1 illustrates the WER across various models and noise levels. The CA-ASR model significantly improves performance on our test set, reducing relative WER by **33.4%** compared to the RNN-T (Soltau et al., 2022) baseline **(7.92% vs. 11.90%)** and by **14.2%** compared to the wav2vec2.0 with BART baseline, even without additional methods like CNRL or decoder pre-training. This highlights the advantage of using multi-modality with a context encoder for dialogue speech recognition.

Configurations			Audio Noise Level		
Model	**Modality**	**Parameter size**	**No Noise**	**SNR:20dB**	**SNR:0dB**
DSTC11 RNN-T (Soltau et al., 2022)	Speech	220M	11.90%	-	-
DSTC11 Whisper (Soltau et al., 2022)*	Speech	1550M	8.50%	-	-
Whisper-large-v2 (Radford et al., 2023)**	Speech	1550M	8.10%	8.45%	14.82%
Wav2Vec2.0+BART (baseline)	Speech	569M	9.23%	11.89%	18.45%
CA-ASR (Ours)	Speech+Text	774M	7.92%	8.23%	15.65%
+CNRL	Speech+Text	774M	7.66%	8.10%	15.03%
+Decoder Ptr.	Speech+Text	774M	7.39%	7.51%	13.33%
+Decoder Ptr. & CNRL	Speech+Text	774M	**7.04%**	**7.24%**	**12.65%**

Table 1: WER comparison of various baselines and CA-ASR settings under different noise conditions. Our proposed CA-ASR model is evaluated with and without Context Noise Representation Learning (CNRL) and Decoder Pretraining (Decoder Ptr.) enhancements. * : reported. **: re-evaluated with our post-processing.

Decoder pre-training further enhances the performance of the CA-ASR model, significantly reducing relative WER by **6.7%**, especially under severe noise conditions (SNR:0dB) where the voice is barely audible. This is expected since the decoder is initially tuned to the dialogue domain, enabling it to predict the user's subsequent probable response from the context even with incomplete speech input.

The benefits are maximized when CNRL is applied, resulting in a relative WER reduction of **11.1%** in clean conditions and **19.1%** in noisy environments compared to the basic CA-ASR model. Since CNRL is designed to make the context encoder resilient to context errors, it significantly enhances the model's robustness against strong noise.

Under the noisy audio test set (refer to Section 4.1), each model's performance declines as the noise level increases (SNR:20dB to SNR:0dB). However, incorporating decoder pre-training and CNRL significantly mitigates this performance drop compared to the basic CA-ASR model (**12.65% vs. 15.65%**).

While the Whisper model shows robust performance under severe noise conditions (SNR:0dB), our CA-ASR model with CNRL and decoder pre-training demonstrates even greater robustness (**12.65% vs. 14.82%**).

5.2 Decoder Pre-training for Dialogue

Table 2 demonstrates the effectiveness of pre-training the decoder with varying the number of turns and pre-training dataset sizes. Note that the baseline model is the same as in Table 1, consisting only of a speech encoder (wav2vec2.0) and a BART decoder. As illustrated, pre-training the decoder on the dialogue domain benefits both the speech-only model (baseline) and the speech-text multimodal model (CA-ASR). Compared to the best result of baseline, the inclusion of the context encoder leads to significant improvements, resulting in a relative WER reduction of approximately **16.7%** at best in CA-ASR with multi-turn (**8.88% vs. 7.39%**). This finding suggests that the efficacy of pre-training the decoder is maximized when the model incorporates information from previous dialogues. Additionally, the WER of CA-ASR with multi-turn improves relatively by up to 6.7% as the dataset size increases (adding SGD), indicating the utility of incorporating external datasets as long as they involve user-agent conversations. Moreover, models considering multiple turns of dialogue exhibit a relatively 3.2% better WER compared to those considering a single turn, as shown in the comparison of best results (**7.64%** vs **7.39%**). This highlights the importance of considering a longer context.

Model	**Input Dialogue**	**Decoder Pre-traing**	**WER**
baseline	-	BART(Lewis et al., 2019)	9.23%
baseline	-	+ MultiWoZ 2.1	8.95%
baseline	-	+ SGD	8.88%
CA-ASR	single-turn	BART	8.14%
CA-ASR	single-turn	+ MultiWoZ 2.1	7.98%
CA-ASR	single-turn	+ SGD	7.64%
CA-ASR	multi-turn	BART	7.92%
CA-ASR	multi-turn	+ MultiWoZ 2.1	7.45%
CA-ASR	multi-turn	+ SGD	**7.39%**

Table 2: WER across various accumulated datasets and a number of turn-takings. Note that CNRL and noise evaluation are not applied in this result to focus on the efficacy of decoder pre-training.

Model (Modality)	Audio Masking	No Noise	SNR:20db	SNR:0db
baseline (Speech)	No	8.94%	11.20%	18.02%
baseline (Speech)	Yes	8.88%	10.58%	17.61%
CA-ASR (Speech + Text)	No	7.45%	7.88%	14.28%
CA-ASR (Speech + Text)	Yes	**7.04%**	**7.24%**	**12.65%**

Table 3: WER comparison between madality and audio masking in clean and noisy samples. Each model's decoder is pre-trained with Multio-WoZ 2.1 and SGD, and CNRL is additionally applied to CA-ASR.

5.3 Effect of Audio masking

Since audio masking can serve as data augmentation, we conducted additional experiments to compare the performance improvement between the baseline (speech-only) model and the CA-ASR (multimodal) model. As shown in Table 3, audio masking enhances ASR performance in both the baseline and CA-ASR models. While the baseline models exhibit marginal performance improvements of about 0.6% in clean sample evaluations, CA-ASR benefits from audio masking with a **5.5%** relative WER reduction. The improvement in CA-ASR becomes more pronounced in noisy environments as noise levels increase. Although the WER is highest at SNR:0dB, indicating the strongest noise, the relative WER reduction is **11.4%**, compared to 8.12% at SNR:20dB. These results suggest that while audio masking is beneficial in both clean and noisy environments, its effect is maximized when the model can utilize contextual information.

5.4 Context Noise Representation Learning

To investigate the impact of noise data on CNRL, we conducted experiments using different types of noise (S1-S4) as described in the CNRL setup in Section 4.3. In Table 4, compared to the model without CNRL, S1 (which arbitrarily removes words) degraded performance, indicating that using only artificial noise is not beneficial for CNRL. S2 and S3, which use real ASR noise from 10-fold data generation, showed better performance, with multi-turn noise (S3) outperforming single-turn noise (S2).

In our evaluation, we found that S4, which combines S1 with S3, performed the best, with WERs of 7.04%, 7.24%, and 12.65% for No-Noise, SNR:20dB, and SNR:0dB conditions, respectively. For comparison, we evaluated our model with ground truth context during inference without CNRL, serving as the upper bound of our experiment. As expected, using ground truth context showed robust results across noise levels, while

CNRL.	No Noise	SNR:20db	SNR:0db
No	7.39%	7.51%	13.33%
S1	7.53%	7.45%	13.45%
S2	7.30%	7.41%	12.94%
S3	7.22%	7.29%	12.83%
S4	7.04%	**7.24%**	12.65%
Ground Truth Context*	**7.01%**	7.25%	**12.28%**
full fine-tune w/ S4	7.24%	7.63%	13.50%

Table 4: CNRL result on different training data settings (S1, S2, S3 and S4) including evaluation result with ground truth context (*) and full fine-tuning result.

CNRL with S4 produced similar results with a small margin. This demonstrates that CNRL enables the context encoder to handle noisy contexts effectively, generating representations close to the ground truth.

We also experimented with training the full CA-ASR model, not just the context encoder, using S4 with corresponding audio for ASR fine-tuning. Training the full model showed lower performance gains than CNRL (**7.04% vs. 7.24%**) and required much larger training costs. We believe this is because training all components with noisy data can disrupt optimization. CNRL allows us to maintain ASR performance against noisy contexts while keeping training efficient.

6 Conclusion

This study introduced Context Noise Representation Learning (CNRL) to improve context-aware ASR systems, especially in noisy environments. By integrating decoder pre-training with dialogue data, ASR fine-tuning, and CNRL, we significantly reduced transcription errors. Our training pipeline demonstrated significant improvements in dialogue speech recognition, even in noisy environments where speech input is defective. Experiments showed CNRL's efficacy, reducing Word Error Rate (WER) by up to 11.1% in clean conditions and 19.1% in noisy settings. By making the model more robust against noisy context, our approach consistently outperformed baselines in various settings.

Limitations

Due to the scarcity of spoken turn-based dialogue datasets, we could only validate our method on a single dataset DSTC11. However validating on the various test datasets would improve its credibility if applicable.

Our primary goal is to enhance ASR performance. However, these enhancements could be even more valuable for downstream Dialogue State Tracking (DST) tasks. Future work could explore optimizing ASR specifically for DST applications to further increase the impact and value of our contributions.

Acknowledgements

This work was supported by the Technology Innovation Program(20015007, Development of Digital Therapeutics of Cognitive Behavioral Therapy for treating Panic Disorder) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea).
This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-2020-0-01789) supervised by the IITP(Institute for Information Communications Technology Planning Evaluation)

References

Siddhant Arora, Siddharth Dalmia, Pavel Denisov, Xuankai Chang, Yushi Ueda, Yifan Peng, Yuekai Zhang, Sujay Kumar, Karthik Ganesan, Brian Yan, et al. 2022. Espnet-slu: Advancing spoken language understanding through espnet. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7167–7171. IEEE.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Changfeng Gao, Gaofeng Cheng, Runyan Yang, Han Zhu, Pengyuan Zhang, and Yonghong Yan. 2021. Pre-training transformer decoder for end-to-end asr model with unpaired text data. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6543–6547. IEEE.

Takaaki Hori, Niko Moritz, Chiori Hori, and Jonathan Le Roux. 2020. Transformer-based long-context end-to-end speech recognition. In *Interspeech*, pages 5011–5015.

Junfeng Hou, Jinkun Chen, Wanyu Li, Yufeng Tang, Jun Zhang, and Zejun Ma. 2022. Bring dialogue-context into rnn-t for streaming asr. In *INTERSPEECH*, pages 2048–2052.

Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: How much can a bad teacher benefit asr pre-training? In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6533–6537. IEEE.

Chao-Wei Huang and Yun-Nung Chen. 2020. Learning asr-robust contextualized embeddings for spoken language understanding. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8009–8013. IEEE.

Ridong Jiang, Wei Shi, Bin Wang, Chen Zhang, Yan Zhang, Chunlei Pan, Jung Jae Kim, and Haizhou Li. 2023. Speech-aware multi-domain dialogue state generation with ASR error correction modules. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 105–112, Prague, Czech Republic. Association for Computational Linguistics.

Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Hakkani-Tür. 2021. "how robust ru?": Evaluating task-oriented dialogue systems on spoken conversations. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1147–1154. IEEE.

Chenyang Le, Yao Qian, Long Zhou, Shujie Liu, Yanmin Qian, Michael Zeng, and Xuedong Huang. 2024.

Comsl: A composite speech-language model for end-to-end speech-to-text translation. *Advances in Neural Information Processing Systems*, 36.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Yuchen Liu, Jiajun Zhang, Hao Xiong, Long Zhou, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2020. Synchronous speech recognition and speech-to-text translation with interactive decoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8417–8424.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Zhiyuan Ma, Jintao Du, and Shuheng Zhou. 2023. Noise-robust training with dynamic loss and contrastive learning for distantly-supervised named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10119–10128.

Ryo Masumura, Naoki Makishima, Mana Ihori, Akihiko Takashima, Tomohiro Tanaka, and Shota Orihashi. 2020. Phoneme-to-grapheme conversion based large-scale pre-training for end-to-end automatic speech recognition. In *INTERSPEECH*, pages 2822–2826.

Pablo Ortiz and Simen Burud. 2021. Bert attends the conversation: Improving low-resource conversational asr. *arXiv preprint arXiv:2110.02267*.

Cheonyoung Park, Eunji Ha, Yewon Jeong, Chi-young Kim, Haeun Yu, and Joo-won Sung. 2023. Copyt5: Copy mechanism and post-trained t5 for speech-aware dialogue state tracking system. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 89–94.

Karol J. Piczak. 2015. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, page 1015–1018, New York, NY, USA. Association for Computing Machinery.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. 2018. Towards end-to-end spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5754–5758. IEEE.

Ashish Shenoy, Sravan Bodapati, Monica Sunkara, Srikanth Ronanki, and Katrin Kirchhoff. 2021. Adapting Long Context NLM for ASR Rescoring in Conversational Agents. In *Proc. Interspeech 2021*, pages 3246–3250.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*.

Hagen Soltau, Izhak Shafran, Mingqiu Wang, Abhinav Rastogi, Jeffrey Zhao, Ye Jia, Wei Han, Yuan Cao, and Aramys Miranda. 2022. Speech aware dialog system technology challenge (dstc11). *arXiv preprint arXiv:2212.08704*.

Xin Sun, Qiang Liu, Shu Wu, Zilei Wang, and Liang Wang. 2023. Noise-robust semi-supervised learning for distantly supervised relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13145–13157.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. 2021. Improving speech translation by understanding and learning from the auxiliary text translation task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4252–4261.

Xin Tian, Xinxian Huang, Dongfeng He, Yingzhan Lin, Siqi Bao, Huang He, Liankai Huang, Qiang Ju, Xiyuan Zhang, Jian Xie, Shuqi Sun, Fan Wang, Hua Wu, and Haifeng Wang. 2021. Tod-da: Towards boosting the robustness of task-oriented dialogue modeling on spoken conversations. *Preprint*, arXiv:2112.12441.

Emiru Tsunoo, Hayato Futami, Yosuke Kashiwagi, Siddhant Arora, and Shinji Watanabe. 2023. Decoder-only architecture for speech recognition with ctc prompts and text data augmentation. *arXiv preprint arXiv:2309.08876*.

Longshaokan Wang, Maryam Fazel-Zarandi, Aditya Tiwari, Spyros Matsoukas, and Lazaros Polymenakos. 2020. Data augmentation for training dialog models robust to speech recognition errors. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 63–70.

Mingqiu Wang, Izhak Shafran, Hagen Soltau, Wei Han, Yuan Cao, Dian Yu, and Laurent El Shafey. 2024. Retrieval augmented end-to-end spoken dialog models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12056–12060. IEEE.

Mingqiu Wang, Izhak Shafran, Hagen Soltau, Wei Han, Yuan Cao, Dian Yu, and Laurent El Shafey. 2023. Speech-to-text adapter and speech-to-entity retriever augmented llms for speech understanding. *arXiv preprint arXiv:2306.07944*.

Jiahao Xu, Wei Shao, Lihui Chen, and Lemao Liu. 2023. Simcse++: Improving contrastive learning for sentence embeddings from two perspectives. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Xiaolin Zheng, Mengling Hu, Weiming Liu, Chaochao Chen, and Xinting Liao. 2023. Robust representation learning with reliable pseudo-labels generation via self-adaptive optimal transport for short text clustering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10493–10507.