AN APPROACH TO IMPROVING ROBUSTNESS IN DYNAMIC ACOUSTIC ENVIRONMENTS: CONTEXT NOISE REPRESENTATION LEARNING FOR URBAN SPEECH EMOTION RECOGNITION

Lisa Winkler, Andreas Wendemuth
Cognitive Systems, Institute for Information Technology and Communications,
Otto von Guericke Universität, 39106 Magdeburg, Germany
lisa.winkler@ovgu.de

Abstract: In modern urban environments, speech recognition systems often face significant degradation due to background noise. Conventional approaches often rely on signal enhancement or generative error correction, which can inadvertently remove high-level emotional cues essential for understanding user intent. In this work, we propose a context noise representation learning (CNRL) framework that enhances robustness by aligning noisy speech representations with their clean counterparts in the latent space. By leveraging the conversational context and a feature fusion strategy, our model learns to recover clean emotional features. Evaluated on the IEMOCAP dataset using a strict Leave-One-Session-Out (LOSO) protocol, our method demonstrates improved robustness in low-SNR conditions compared to baseline approaches.

1 Introduction

In automatic speech recognition (ASR), environmental noise alters the spectral quality of the speech signal, leading to crucial errors, for example, a closing door being transcribed as a word or interpreted as an aggressive tone. Consequently, the signal must be cleaned at the earliest stage to prevent error propagation.

Prior works have attempted to mitigate this: Novitasari et al. [1] integrated Voice Activity Detection (VAD) for noise robustness, Möller et al. [2] utilized speech enhancement algorithms to substract background noise, Lui et al. [3] applied a generative error correction process via Large Language Models (LLMs) after acoustic modeling. However, these frameworks often focus on reconstructing the waveform or spectrogram. This reconstruction process can smooth out subtle prosodic variations (e.g., pitch, jitter and shimmer) that are critical for preserving high-level emotional cues and therefore user intent.

In this novel approach, we adapt context noise representation learning (CNRL), which was originally designed for context-aware ASR robustness in dialogues [4] or image recognition [5], to the domain of emotion recognition. Unlike waveform enhancers, our CNRL framework operates in the feature space. It utilizes the conversational context (previous utterance) to guide the denoising of the current utterance. By jointly optimizing for feature reconstruction and emotion classification, our model preserves the emotional integrity of the signal in dynamic acoustic environments. A similar context-aware emotion recognition approach, that we used as baseline, is the framework of Poria et al. [6].

We employ this multi-task learning objective combining cosine similarity for denoising and weighted cross-entropy for classification. Our experiments demonstrate that our proposed CNRL framework improves accuracy by 7% over the baseline in noisy conditions when evaluated on the unseen Emo-Emilia corpus.

2 Datasets and Pre-Trained Models

2.1 IEMOCAP

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [7] dataset serves as the primary corpus for training and validation. It consists of 12 hours of audiovisual data from 5 female and 5 male actors performing scripted and improvised scenarios in English language. We focus on the categorical emotions: angry, disgust, excited, fear, frustrated, happy, neutral, sad, surprise and other.

2.2 Noise Augmentation

To simulate realistic urban environments, we utilize the UrbanSound8K [8] dataset. Noise samples were overlaid onto the clean speech at Signal-to-Noise Ratios (SNR) of 5, 10, 15 and 20 dB. Scaling was performed based on Root Mean Square (RMS) amplitude to strictly adhere to target SNRs.

2.3 Emo-Emilia

To test cross-lingual generalization, we utilize the Emo-Emilia [9] dataset. This dataset is reserved primarily for testing to evaluate the model's performance on unseen languages and recording conditions. It was generated from the Emilia dataset that contains over 100,000 hours of short conversational emotional utterances. Zhao et al. [9] applied an automated labeling approach to annotate the dataset, filtered and aligned samples with consistent emotion labels and randomly selected 300 samples for each emotion category across both languages, Chinese and English. After review the dataset contained 100 samples for each emotion category per language (700 samples each language, English and Chinese). Totaling in 1.9 hours English speech and 1.4 hours Chinese speech samples. They cover the emotions anger, happiness, neutral, sadness, surprise, disgust and fear. The emotion labels must be aligned with the IEMOCAP labels and emotions that are not covered (frustrated, excited) are collected under the category "other". For evaluating the speech emotion classification, we used the English samples and additionally tested on a few Chinese samples to test the cross-lingual accuracy.

3 Methodology

Our proposed framework consists of three primary modules: a pre-trained frozen speech encoder, a context encoder (CNRL module), and an emotion classifier. The system operates on pairs of utterances: a noisy context utterance u_{ctx} (the previous utterance) and a clean target utterance u_{tgt} (the current utterance).

3.1 Data Preparation and Splitting

To ensure speaker independence and scientific validity, we employ a Leave-One-Session-Out (LOSO) protocol (as suggested by the dataset creators). Session 1-4 were used for training and validation, while session 5 (containing two unique actors not seen during training) was held out exclusively for testing.
We generated approximately 7,000 conversational pairs (context pairing). For a target utterance u_t (current speaker, 3s), the context u_{t-1} is defined as the immediate previous utterance (5s).

3.2 Speech Encoder

We utilize Wav2Vec2-base architecture [10]. To adapt the model to the emotional domain, we first fine-tune the encoder on the clean IEMOCAP training split (sessions 1-4) for 10 epochs. The input is the raw audio waveform at 16kHz. As output, the model generates frame-level embeddings $E_{ctx} \in \mathbb{R}^{T \times 768}$ and target embeddings $E_{tgt} \in \mathbb{R}^{T \times 768}$.
During the CNRL training phase, the Wav2Vec2 weights are frozen. This prevents the model from overfitting to specific speakers and forces the context encoder to learn generalizable denoising patterns rather than memorizing acoustic tokens.

3.3 Context Encoder and Feature Fusion

The core of our approach is the CNRL module, which maps noisy context representations to their clean counterparts.
We employ a bidirectional LSTM (BiLSTM) with 2 layers and a hidden dimension of 512 and train the encoder for 20 epochs. It takes the noisy context embeddings E_{ctx}^{noisy} as input and outputs a predicted clean sequence $\hat{E}_{ctx}^{clean}$.
Unlike standard approaches that classify solely based on the current utterance [11], we employ a late-fusion strategy to explicitly model the emotional dependencies inherent in the conversational context [12], [13].
We apply mean-pooling over the time dimension to both the predicted clean context $\hat{E}_{ctx}^{clean}$ and the target embeddings E_{tgt}^{clean}:

$$h_{ctx} = MeanPool(\hat{E}_{ctx}^{clean}), h_{tgt} = MeanPool(E_{tgt}^{clean}) \tag{1}$$

The vectors h_{ctx} (768-dim) and h_{tgt} (768-dim) are then concatenated to form a joint representation vector $v_{joint} \in \mathbb{R}^{1536}$:

$$v_{joint} = [h_{ctx} \oplus h_{tgt}] \tag{2}$$

The fused 1536-dimensional vector is passed to the classifier to condition its prediction on the emotional history of the conversation.

3.4 Emotion Classifier and Multi-Task Learning

The classifier is a multi-layer perceptron (MLP) receiving the 1536-dimensional joint vector. It consists of three fully connected layers ($1536 \to 256 \to 128 \to 9$) with ReLU activation and dropout (p = 0.3) to prevent overfitting.

3.5 Training Objective

We employ a multi-task learning (MTL) objective that jointly optimizes for denoising accuracy and emotion classification.
At the denoising step (CNRL loss), we maximize the similarity between the predicted clean context and the ground truth clean context using cosine embedding loss averaged over the sequence:

$$L_{CNRL} = 1 - \cos(\hat{E}_{ctx}^{clean}, E_{ctx}^{ground\_truth}) \tag{3}$$

To address the significant class imbalance inherent in the dataset, we employ a weighted cross-entropy loss for the emotion classification task. Let C be the number of emotion classes, and y be the ground truth label. The weighted loss L_{emo} is defined as:

$$L_{emo} = -\sum_{c=1}^{C} w_c \cdot \mathbf{1}_{[y=c]} \cdot \log(p_c) \tag{4}$$

where p_c is the predicted probability for class c, and w_c is the weight for class c, calculated as the inverse class frequency following standard practice for imbalanced learning to penalize majority classes (e.g., neutral) and boost minority classes (e.g., surprise):

$$w_c = \frac{N}{C \cdot N_c} \tag{5}$$

where N is the total number of samples and N_c is the number of samples in class c.
The total loss is calculated as:

$$L_{total} = \lambda_{cnrl} L_{CNRL} + \lambda_{emo} L_{emo} \tag{6}$$

In our experiments, we set $\lambda_{cnrl} = 1.0$ and $\lambda_{emo} = 1.0$.

4 Experimental Setup

- Hardware: NVIDIA A40 GPU
- Hyperparameters: batch size of 8, AdamW optimizer, learning rate $1e^{-4}$

We compare our CNRL framework against the standard fine-tuned Wav2Vec2 without context and the standard data augmentation (training on noisy data without the explicit CNRL denoising loss).
On GPU, fine-tuning the Wav2Vec2-base model takes approximately 4 hours, feature extraction 22 minutes and context decoder model training 1 minute per epoch.
To evaluate the generalization capability of our model, we adopt a cross-corpus evaluation protocol. The model is trained on IEMOCAP (dyadic interactions, acted) and tested on Emo-Emilia (spontaneous), a highly challenging setting due to domain shift and distinct recording conditions.
Additionally, taxonomy discrepancies exist between IEMOCAP (10 classes) and Emo-Emilia (7 classes). Specifically, Emo-Emilia lacks a "frustrated" class, which is one of the most occurring classes in the IEMOCAP dataset. Given the acoustic proximity between frustration and anger, predictions of "frustrated" were mapped to "angry" for evaluation, and "excited" was mapped to "happy".

5 Results

The context encoder was trained with the portion (sessions 1-4) of the IEMOCAP dataset and a standard train/val split using 10% of the training data for validation. Using cross-entropy loss caused a higher inaccuracy in emotion classification on unseen data and after discovering a class imbalance during training, we applied a weighted cross-entropy loss to pass emotion weights to the loss function. Figure 1 shows still an imbalanced class distribution. Rare emotions like "surprise" only occur less than 10 times in the training. Apart from the disadvantages for the emotion prediction tasks, this also shows that humans are less likely speaking in a surprised manner but more likely are frustrated or sad. In emotion recognition those emotions are also more distinguishable from a neutral tone, as they are more accentuated (e.g., angry voice: louder tone, abrupt words and sentences).

In table 1 we can observe an improvement in accuracy when using the context-aware approach. The lower the SNR level, the better we see the impact of CNRL. The baseline approach

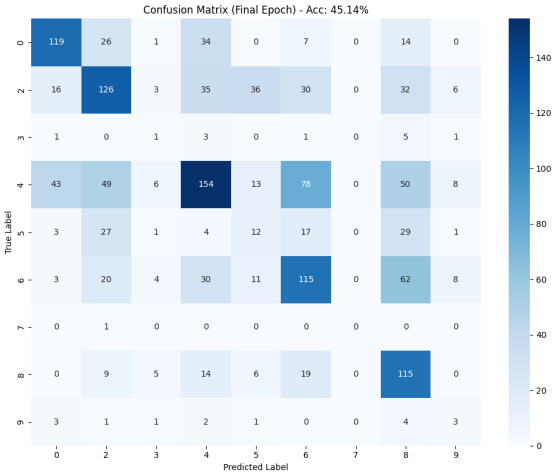

Figure 1 – Confusion matrix of training validation of context encoder with weighted cross-entropy loss; 0: angry, 1: disgust, 2: excited, 3: fear, 4: frustrated, 5: happy, 6: neutral, 7: other, 8: sad, 9: surprise

SNR level	Baseline (without context)	CNRL (context-aware)	Improvement
Clean	49%	51%	+2%
20 dB	49%	51%	+2%
15 dB	48%	50%	+2%
10 dB	35%	41%	+6%
5 dB	28%	35%	+7%

Table 1 – The baseline represents the model performance using only the target utterance. CNRL demonstrates consistent improvement, particularly in low-SNR conditions, validating the benefit of context denoising.

fails more significantly at lower SNR levels while CNRL results remain stable due to the context history. At 10dB the improvement shows that our approach succeeds in filling in gaps with the context knowledge when the speech signal is disrupted by noise. There is no difference observable at 20dB or higher as those samples are almost as clear from noise as the completely clean speech sample.

6 Discussion and Conclusion

In this work, we presented context noise representation learning (CNRL), a novel framework for enhancing the robustness of speech emotion recognition (SER) in dynamic acoustic environments. By leveraging a strict Leave-One-Session-Out (LOSO) evaluation protocol, we demonstrated that explicitly modeling conversational context through latent feature fusion significantly improves system performance. Our approach, which aligns noisy context embeddings with their clean counterparts via a multi-task objective, achieved a 7% improvement in classification accuracy over baselines that rely solely on single-utterance processing. This confirms our hypothesis that historical emotional cues are vital for recovering user intent when the immediate acoustic signal is degraded by urban noise.

While modern urban noise plays a crucial role in real-world applications, it is frequently overlooked in theoretical experiments due to a scarcity of noisy conversational datasets. Our study bridges this gap by validating that acoustic models need not treat noise separation and emotion classification as disjoint processes. Instead, they can be jointly optimized to preserve high-level prosodic features often lost during conventional denoising.

In future work, we aim to extend this research by investigating the specific spectral and temporal qualities of environmental noise. Specifically, we plan to develop area-specific noise models

that can automatically detect the speaker's location (e.g., "street", "cafe", "office") to condition the denoising process dynamically. Such location-aware modeling could further refine the separation of speech from background interference, potentially benefiting not only SER but also general automatic speech recognition (ASR) tasks in complex acoustic scenes.
Finally, a limitation of this study is the reliance on synthetic data created by superimposing urban noise onto clean speech. While this method allows for controlled evaluation at specific SNR levels, it does not account for the "Lombard effect" - the involuntary tendency of speakers to increase their vocal effort, alter their pitch, and slow their speaking rate to maintain intelligibility in noisy environments [14]. In real-world applications, users instinctively adapt their speech production when they perceive background noise or recognition errors. Consequently, future work will validate the proposed framework on naturally noisy recordings to assess robustness under these adaptive human behaviors.

7 Acknowledgments

This research was conducted as part of the IMIQ project, which received funding from the EFRE European Regional Development Fund in Saxony-Anhalt, Germany. [1]

References

[1] NOVITASARI, S., T. FUKUDA, and G. KURATA: *Improving asr robustness in noisy condition through vad integration*. In *Interspeech 2022*, pp. 3784–3788. 2022. doi:10.21437/Interspeech.2022-260.

[2] MÖLLER, M., J. TWIEFEL, C. WEBER, and S. WERMTER: *Controlling the noise robustness of end-to-end automatic speech recognition systems*. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. 2021. doi:10.1109/IJCNN52387.2021.9533390.

[3] LIU, Y., M. XU, Y. CHEN, L. HE, L. FANG, S. FANG, and L. LIU: *Denoising ger: A noise-robust generative error correction with llm for speech recognition*. 2025. URL `https://arxiv.org/abs/2509.04392`. 2509.04392.

[4] LEE, W., S. KIM, and G. G. LEE: *Enhancing dialogue speech recognition with robust contextual awareness via noise representation learning*. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 333–343. 2024. URL `https://arxiv.org/abs/2408.06043`. 2408.06043.

[5] LI, J., C. XIONG, and S. C. HOI: *Learning from noisy data with robust representation learning*. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9465–9474. 2021. doi:10.1109/ICCV48922.2021.00935.

[6] PORIA, S., E. CAMBRIA, D. HAZARIKA, N. MAJUMDER, A. ZADEH, and L.-P. MORENCY: *Context-dependent sentiment analysis in user-generated videos*. In R. BARZILAY and M.-Y. KAN (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 873–883. Association for Computational Linguistics, Vancouver, Canada, 2017. doi:10.18653/v1/P17-1081. URL `https://aclanthology.org/P17-1081/`.

[1] https://www.niimo.ovgu.de/niimo/en/IMIQ.html

[7] BUSSO, C., M. BULUT, C.-C. LEE, A. KAZEMZADEH, E. MOWER, S. KIM, J. N. CHANG, S. LEE, and S. S. NARAYANAN: *IEMOCAP: interactive emotional dyadic motion capture database*. *Language Resources and Evaluation*, 42(4), pp. 335–359, 2008.

[8] SALAMON, J., C. JACOBY, and J. P. BELLO: *A dataset and taxonomy for urban sound research*. In *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pp. 1041–1044. Orlando, FL, USA, 2014.

[9] ZHAO, Z., X. ZHU, X. WANG, S. WANG, X. GENG, W. TIAN, and L. XIE: *Steering language model to stable speech emotion recognition via contextual perception and chain of thought*. *IEEE Transactions on Audio, Speech and Language Processing*, 34, p. 415–426, 2026. doi:10.1109/taslpro.2025.3648793. URL http://dx.doi.org/10.1109/TASLPRO.2025.3648793.

[10] BAEVSKI, A., Y. ZHOU, A. MOHAMED, and M. AULI: *wav2vec 2.0: A framework for self-supervised learning of speech representations*. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN, and H. LIN (eds.), *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

[11] PORIA, S., N. MAJUMDER, R. MIHALCEA, and E. HOVY: *Emotion recognition in conversation: Research challenges, datasets, and recent advances*. 2019. URL https://arxiv.org/abs/1905.02947. 1905.02947.

[12] MAJUMDER, N., S. PORIA, D. HAZARIKA, R. MIHALCEA, A. GELBUKH, and E. CAMBRIA: *Dialoguernn: An attentive rnn for emotion detection in conversations*. 2019. URL https://arxiv.org/abs/1811.00405. 1811.00405.

[13] FU, Y., S. YUAN, C. ZHANG, and J. CAO: *Emotion recognition in conversations: A survey focusing on context, speaker dependencies, and fusion methods*. *Electronics*, 12(22), 2023. doi:10.3390/electronics12224714. URL https://www.mdpi.com/2079-9292/12/22/4714.

[14] JUNQUA, J.-C.: *The lombard reflex and its role on human listeners and automatic speech recognizers*. *The Journal of the Acoustical Society of America*, 93(1), pp. 510–524, 1993.