Conversation Context-aware Direct Preference Optimization for Style-Controlled Speech Synthesis

Atsushi Kojima* Yusuke Fujita* Hao Shi* Tomoya Mizumoto* Mengjie Zhao* Yui Sudo*
* SB Intuitions, Japan
E-mail: atsushi.kojima@sbintuitions.co.jp Tel/Fax: +81-80-7685-5125

***Abstract*—Conversational text-to-speech (TTS) generates context-appropriate expressive speech by modeling dialogue context, without requiring manual paralinguistic labels such as emotion or intent. However, acoustically similar emotions (e.g., happy vs. angry) often cause inappropriate renderings. To address this, we propose a preference-based optimization approach that reinforces the model toward selecting contextually appropriate emotional expressions. Specifically, we adopt direct preference optimization (DPO) in a unit-based TTS architecture, enabling fine-grained supervision of emotional expression by comparing discrete unit sequences for more and less suitable speech samples. Our method employs a text-to-unit model predicting discrete acoustic units from phonemes and context, followed by a unit-based HiFi-GAN vocoder. Training proceeds in two stages: first, the text-to-unit model is optimized with cross-entropy (CE) loss on ground-truth units; second, DPO is applied using unit pairs reflecting appropriate versus inappropriate emotions within the same context. Experiments on the STUDIES dataset show that preference-based optimization substantially improves emotion recognition accuracy (67.14% with CE+DPO vs. 48.45% with CE only). These results demonstrate that our approach enables fine-grained control of emotional expression and better captures subtle distinctions in conversational speech.**

I. Introduction

Conversation text-to-speech (TTS) aims to generate expressive and contextually appropriate speech in conversational systems [1], [2], [3], [4]. Unlike traditional emotional TTS systems that rely on explicit paralinguistic cues—such as emotion labels (e.g., happy, sad) or intent types (e.g., questions, greetings) [5], [6], [7], conversation TTS generates speech with the appropriate emotional tone by implicitly leveraging the preceding conversation context. Recent studies have demonstrated that incorporating conversation history enables effective style control, showing that context can guide emotional expression in synthesized speech [4], [8].

However, even when conversation context is taken into account, a fundamental challenge remains: some emotions, such as happy and angry, can be difficult to distinguish because they share similar prosodic features, like high pitch or strong energy. This overlap leads to confusion in both human perception and automatic recognition tasks [9], [10], [11]. It is essential to train models not only to recognize subtle acoustic differences, but also to favor emotional expressions that are appropriate for the given conversational context. To overcome this ambiguity, it is crucial not only to enhance the model's ability to differentiate between acoustically similar emotions, but also to introduce a mechanism that explicitly favors emotional expressions appropriate to the conversational context.

In this paper, we propose to adopt Direct Preference Optimization (DPO) [12] as a learning strategy, allowing the model to learn preference signals between emotional expressions that are more or less suitable for the given context. We adopt a unit-based architecture, as representing speech with discrete units enables the application of DPO, which operates by comparing discrete outputs during preference learning. The text-to-unit model takes phoneme sequences and preceding conversation context as input, and predicts discrete acoustic units. These units are then converted into speech using a unit-based HiFi-GAN vocoder [13], [14]. To train the text-to-unit model, a two-step training strategy is adopted. In the first step, the model is trained by minimizing the cross-entropy (CE) loss between the ground-truth discrete units—extracted from speech spoken with a contextually appropriate emotion given the conversation context—and the predicted units. In the second step, the model is further trained by minimizing the DPO loss. In this stage, discrete units from speech with an appropriate emotion for the given context are used as preferred samples, while units from speech with an inappropriate emotion for the same context serve as dispreferred samples. Our experiments show that incorporating preference-based optimization improves the ability to generate emotionally and contextually appropriate speech responses.

II. Related Work

A. Context-Aware Speech Synthesis with Joint Training on Tacotron2 Loss

As a context-aware TTS model, a method based on gated recurrent units (GRU) has been proposed [4]. In this approach, the hidden vector obtained from the conversation context text via a GRU is fed as an auxiliary feature into the encoder of Tacotron2 [15], allowing the model to control the style of the synthesized speech. Evaluations have shown that this method outperforms the original Tacotron2 in both naturalness and the similarity of style between real and synthesized speech.

However, in this method, the encoder that extracts the context representation is trained solely with the Tacotron2 loss, typically a reconstruction loss such as mean squared error (MSE). This type of loss encourages the model to reproduce observed acoustic features as closely as possible, but does not directly optimize for the appropriateness or preference of emotional expression within each conversational context. As

979-8-3315-7206-8/25/$31.00

a result, the model may generate acoustically accurate speech that is not always the most contextually suitable or preferred in terms of subtle emotional cues.

In contrast, our preference optimization method with DPO enables the model to directly learn context- and emotion-appropriate speech by optimizing for human or system-level preferences, rather than only reconstructing acoustic features as in standard Tacotron2 loss. As a result, our approach produces more expressive and contextually suitable synthesized speech, overcoming the limitations of prior methods.

B. Emotion Embedding Switching with DPO

A method for controlling speech style using DPO loss has been proposed, in which DPO is used to learn emotion embeddings for flexible emotion switching [16]. This approach enables preference-based optimization of emotion representations, allowing the system to generate speech with various emotional expressions. However, this approach requires that emotion embeddings be manually selected or switched during inference, preventing automatic and context-sensitive style control based on the actual conversation context. As a result, such methods are unable to dynamically adapt speaking style to the flow of conversation, limiting their ability to generate truly natural and contextually appropriate expressive speech.

In contrast, our approach employs DPO within a unit-based architecture, enabling direct supervision of subtle emotional preferences driven by conversation context. This allows for more natural and automatic style adaptation, eliminating the need for manual intervention at inference time.

C. Contrastive Learning for Context Representation

A recent approach, Contrastive-based conversation speech synthesis (CONCSS) [8], tackles the challenge of generating contextually appropriate prosody by applying contrastive learning to conversation TTS. In this method, the context encoder is trained with a pretext task that encourages it to produce context-sensitive latent vectors: it minimizes the distance between vectors from similar conversation contexts while maximizing the distance for dissimilar ones. To further enhance the discriminability of these context vectors, the framework incorporates a hard negative sampling strategy and triplet loss. Experimental results show that this approach increases the sensitivity of synthesized prosody to different conversation contexts.

However, CONCSS focuses exclusively on optimizing the context representation space and does not directly account for user or system-level preferences regarding the emotional appropriateness of the generated speech. Moreover, while the model is optimized for contrastive objectives at the representation level, it lacks explicit supervision over the actual prosodic realization. In contrast, our proposed method applies DPO at the generation level, directly guiding the model to produce emotionally preferred speech samples for each context. This enables more fine-grained and human-aligned control over style in conversation TTS.

III. Proposed Method

A. Model Architecture

Our proposed method integrates DPO within a text-to-unit architecture conditioned on conversation context. By explicitly modeling conversation context, our approach enables context-aware, automatic, and fine-grained control of emotional style in speech synthesis. This allows the system to generate responses with appropriate emotional nuance for each conversational turn, without any manual intervention.

Our proposed model consists of the text-to-unit model and the unit-based HiFi-GAN vocoder. The text-to-unit model consists of the conversation context encoder and phoneme embedding layer. The conversation context encoder converts conversation context into a hidden vector that captures emotional context. Phoneme embedding layer converts phoneme symbols into phoneme embeddings. Then, we concatenate the hidden vector and phoneme embeddings along the time axis. The text-to-unit model predicts discrete units using the concatenated vectors. The unit-based HiFi-GAN vocoder generates speech waveforms using discrete units predicted by the text-to-unit model. For details of the overall architecture, please refer to Section III-B.

TABLE I
Examples of conversation context and corresponding target utterances to be synthesized.

Conversation context	Target utterance for synthesis	Emotion
You look happy. Something good?	Nothing special. Just a sunny day.	Happy
I expected 90%, but got 60%.	I can't accept such a bad score.	Angry
How was the test? Not great.	Yeah, I made some mistakes.	Sad

TABLE I shows examples of conversation context and corresponding target utterances used for synthesis. Each context consists of K turns of conversation between two speakers, followed by a target utterance that reflects an appropriate emotional response. For instance, in the first row, the context is "You look happy." and "Something good?", and the target utterance is "Nothing special. Just a sunny day.", labeled as a happy emotion.

The text-to-unit model is attention-based encoder-decoder. The encoder converts phoneme to hidden vectors. The decoder outputs discrete units at each decoding step based on attention calculated between encoder hidden vector and decoder hidden vector. The decoding process terminates when the decoder generates a special token that indicates the end of the sequence.

The conversation context encoder converts previous conversation context into a hidden vector that captures emotional context. This context vector is concatenated with phoneme embeddings and fed into the text-to-unit model. Detailed implementation of the context encoder and input representation is described in Section III-B. For the first utterance in the conversation, the conversation context encoder obtains a special tag, which represents that the conversation context to refer to does not exist.

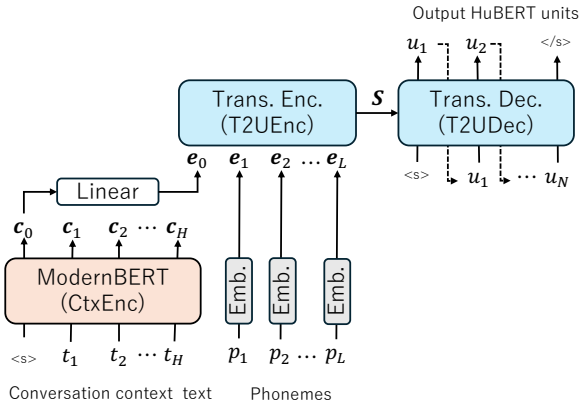

Fig. 1. Overall architecture of the proposed text-to-unit model. The model predicts HuBERT units from input phonemes and preceding conversation context using a Transformer encoder-decoder architecture. The context encoder (ModernBERT) processes conversation context, which is combined with phoneme embeddings and passed to the encoder. The decoder generates unit sequences autoregressively.

B. Two-Step Training with CE and DPO

For training the proposed text-to-unit model, we adopt a two-step strategy combining CE and DPO losses. Fig. 1 illustrates the overall architecture of the text-to-unit model and notation used in the formulation. In the first step, the model is trained with CE loss to predict HuBERT [17] unit sequences from phonemes and conversation context. Let $\mathbf{h} = \{t_h \in \mathcal{V}^{\text{text}} \mid h = 1, ..., H\}$ denote the conversation context, and $\mathbf{p} = \{p_l \in \mathcal{V}^{\text{phoneme}} \mid l = 1, ..., L\}$ the input phoneme sequence. The goal is to estimate the conditional probability $p(u_n \mid \mathbf{u}_{<n}, \mathbf{h}, \mathbf{p})$ for each unit u_n, where $\mathbf{u}_{<n}$ is the sequence of previously generated units.

The encoder processes the input as follows: the conversation context $\mathbf{h}$ is encoded by a pretrained ModernBERT model (CtxEnc) [18], yielding contextual embeddings $\mathbf{c}_0, ..., \mathbf{c}_H$. The first embedding $\mathbf{c}_0$, which encodes global information of the conversation context, is linearly projected to form $\mathbf{e}_0 = \text{Linear}(\mathbf{c}_0)$. Each phoneme p_l is embedded as $e_l = \text{Emb}(p_l)$ for $1 \leq l \leq L$. The input sequence $[e_0, e_1, ..., e_L]$ is passed through the Transformer encoder (T2UEnc) to obtain source representation $\mathbf{S} \in \mathbb{R}^{d \times (L+1)}$. The decoder (T2UDec) autoregressively predicts the HuBERT units $\{u_1, u_2, ..., u_N\}$, with

$$p(u_n \mid \mathbf{u}_{<n}, \mathbf{h}, \mathbf{p}) = \text{T2UDec}(\mathbf{u}_{<n}, \mathbf{S}), \quad 1 \leq n \leq N. \quad (1)$$

The CE loss for preferred unit sequences $\mathbf{u}^{(\text{w})}$ is defined as

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{n=1}^{N} \log p(u_n^{(\text{w})} \mid \mathbf{u}_{<n}^{(\text{w})}, \mathbf{h}, \mathbf{p}). \quad (2)$$

In the second step, the model is fine-tuned using the DPO loss. Our method applies preference-based supervision at the generation level, allowing the model to learn acoustically discriminative latent features for more accurate and contextually appropriate emotional expression. Dispreferred units $\mathbf{u}^{(\text{l})}$ (e.g., generated with mismatched emotion) are contrasted with the preferred units using

$$\mathcal{L}'_{\text{CE}} = -\frac{1}{M} \sum_{n=1}^{M} \log p(u_n^{(\text{l})} \mid \mathbf{u}_{<n}^{(\text{l})}, \mathbf{h}, \mathbf{p}), \quad (3)$$

and

$$\mathcal{L}_{\text{DPO}} = -\log \sigma \left(\beta (\mathcal{L}'_{\text{CE}} - \mathcal{L}_{\text{CE}}) \right), \quad (4)$$

TABLE II
EMOTION CATEGORIES IN THE TRAINING DATASETS.

Emotion	STUDIES	JVNV
Angry	1511	209
Disgust	0	218
Fear	0	225
Happy	3990	240
Normal	4245	0
Sad	2109	217
Surprise	0	266

where β is a temperature parameter controlling the strength of preference, and $\sigma(\cdot)$ denotes the sigmoid function.

This objective, shown in (4), encourages the model to assign higher likelihood to emotionally appropriate responses for the same conversation context. The CE losses for the preferred and dispreferred sequences are computed as in (2) and (3), respectively. The autoregressive decoding process is given in (1).

C. Preference Data Construction based on emotion-aware TTS

To construct preference data for DPO training, we need pairs of discrete unit sequences for the same text and conversation context that differ only in emotional appropriateness. This requires an emotion-aware text-to-unit model capable of generating discrete units conditioned on both phoneme and emotion labels, since standard text-to-unit models cannot control for emotion.

We implement the emotion-aware text-to-unit model as an attention-based encoder-decoder that takes both phoneme and emotion inputs, each represented as a one-hot vector [5]. These inputs are embedded by separate layers, concatenated along the time axis, and used to predict discrete unit sequences. For each text with ground-truth emotion annotation, we extract units from speech uttered with the correct emotion as the preferred sample, and units from speech with an incorrect emotion as the dispreferred sample.

IV. Experiments and Results

A. Experimental Setup

1) Dataset: We used the STUDIES dataset [19] for training both the text-to-unit model and the unit-based HiFi-GAN vocoder. In addition, we used JVNV dataset [20], which

includes a broader set of emotional categories including surprise, fear, and disgust in both verbal and nonverbal vocal expressions, to compensate for the limited emotional diversity. TABLE II shows number of emotion in these datasets. The speech in these datasets was downsampled to 16 kHz during training. We treated the JVNV dataset as utterances without conversation context ($K = 0$) because JVNV dataset does not have conversation context.

2) Model Architecture: For the text-to-unit model, we trained the text-to-unit model based on CE loss using JVNV and STUDIES datasets in step1. In step2, we fine-tune the text-to-unit model trained in step1 based on DPO loss using STUDIES. During DPO-based fine-tuning, the number of conversation turns fed into the conversation context encoder is randomly varied between 0 and 5 in order to augment the conversation context. We limit the maximum context length to 5 turns because excessively long conversation context may introduce irrelevant or noisy information, and in real dialogue scenarios, long conversation contexts are not always available particularly at the beginning of an interaction. Randomly varying the number of turns also helps the model adapt to situations with both short and long available context, thereby improving robustness across different conversation stages. For training the unit-based HiFi-GAN vocoder, we used STUDIES and JVNV for pre-training, then fine-tuned with female voice (LD06) in the JVNV dataset.

3) Hyperparameters: The text-to-unit model is an attention-based encoder-decoder, which consists of a 6 Transformer encoder layers and 6 Transformer decoder layers. The hidden_dimension and position-wise feed-forward were set to 1024 and 4096 respectively. Also, the number of heads was set to 16. For the conversation context encoder, we used ModernBERT [1] with 130M parameters that outputs a 128-dimensional hidden vector. The hidden vector is converted by linear layer. The phoneme embedding layer converts 41 types of phonemes into phoneme embeddings. The dimension of the phoneme embeddings was set to 128. The text-to-unit model predicts discrete units with 64 classes from the phoneme embeddings and hidden vector representing conversation context. The maximum number of conversation turns K input to the conversation context encoder was set to 5.

We describe the architecture of the unit-based HiFi-GAN vocoder. The output dimension of the generator's unit embedding layer was set to 128, and a transposed convolution-based method was used for the upsampling layers. For extracting discrete units, a 64-class k-means clustering model was trained on hidden vectors (extracted using a HuBERT [2] with 99.4M parameters), using speech data from STUDIES, JVNV and JSUT [21].

4) Preference Data Construction: We describe the architecture and training process of the emotion-aware TTS model presented in Section III-C. For architecture, we use the same architecture as the text-to-unit model described in the Section IV-A3, except that it receives an emotion label as a one-hot vector as additional input. As training data, we used STUDIES and JVNV. Emotion-aware TTS model gets text and emotion label, and outputs discrete units. Preferred samples were obtained by inputting the correct emotion labels assigned to STUDIES and the utterance content. Dispreferred samples were obtained by randomly selecting labels other than the correct emotion label and inputting them along with the utterance content. We selected 300 utterances from STUDIES dataset, and made 300 preferred and dispreferred pairs of units. For sampling discrete units from emotion-aware TTS, we used temperature sampling to ensure sufficient diversity in the generated samples, which helps the model learn more robust emotional representations. We set temperature to 0.98 and top-p to 0.8.

5) Training Configurations: We describe training conditions for the text-to-unit model and unit-based HiFi-GAN vocoder. For training the text-to-unit model in step 1, we set the learning rate to $5e{-}6$ and the batch size to 16. Furthermore, we applied label smoothing [22] of 0.2 to improve robustness. For training the text-to-unit model in step 2, we set the learning rate to 0.0001, the batch size to 4, and the DPO margin β to 2. For training the unit-based HiFi-GAN vocoder, the log mel spectrograms of the audio were calculated with an FFT size of 1024 samples, a frame shift of 160 samples, and 128 frequency bins for training the unit-based HiFi-GAN. Additionally, the discrete units were extracted from real speech using HuBERT and k-means clustering models. During training, targets were prepared by dividing the speech waveforms into chunks with a window width of 16000 samples and a frame shift of 8000 samples.

6) Evaluation Protocol: As the baseline system, we used the emotion-aware text-to-unit model, which predicts discrete units based on phonemes and emotion. For predicting using the emotion-aware text-to-unit model, we give ground-truth emotion label given dataset. In addition, we compare the text-to-unit model trained using only CE loss.

For the evaluation, we evaluated emotion recognition accuracy using an SVM-based classifier with HuBERT hidden representations [23]. The SVM was trained on a subset of the STUDIES dataset. We used 20 test samples from the STUDIES dataset. During synthesis using text-to-unit models trained using CE and DPO, the maximum number of conversation turns fed into the conversation context encoder was set to 5. Moreover, we evaluated 5-scale naturalness MOS test and speaker style similarity between synthesized speech and real speech. These scores were recorded with 95% confidence intervals (CI). 33 native Japanese raters listened to the test samples randomly, where they were allowed to evaluate each audio sample once.

B. Results

TABLE III shows comparison results for the text-to-unit model trained using DPO and the text-to-unit model trained using CE.

[1]https://huggingface.co/sbintuitions/modernbert-ja-130m
[2]https://huggingface.co/rinna/japanese-hubert-base

TABLE III
COMPARISON RESULTS BETWEEN MODELS TRAINED WITH ONLY CE AND CE+DPO.

ID	Loss	MOS (↑)	Similarity (↑)	Happy (%)	Angry (%)	Sad (%)	Mean accuracy (%)
exp1	CE	2.50 ± 0.09	3.05 ± 0.10	65.00	42.86	37.50	48.45
exp2	CE+DPO	$\mathbf{2.56 \pm 0.07}$	$\mathbf{3.06 \pm 0.08}$	**80.00**	**71.43**	**50.00**	**67.14**

TABLE IV
COMPARISON RESULTS BETWEEN PROPOSED MODEL AND COMPARISON MODEL.

ID	Input	MOS (↑)	Similarity (↑)
exp2	Conversation context (proposed)	$\mathbf{2.56 \pm 0.07}$	$\mathbf{3.06 \pm 0.08}$
exp3	Emotion (one-hot vector)	2.39 ± 0.07	2.79 ± 0.09

In terms of emotion recognition accuracy, the DPO-trained model demonstrates a clear improvement over the CE baseline. As shown in the table, DPO increases the recognition rates for all emotion classes, especially for "happy" and "angry", resulting in a substantial gain in overall mean accuracy (67.14% vs. 48.45%). This indicates that preference-based optimization enables the model to more reliably capture subtle emotional differences in conversational speech.

For MOS, we could observe that the DPO-trained model (exp2) achieves a slightly higher score (2.56 ± 0.07) than the CE-trained counterpart (2.5 ± 0.09) (exp1), suggesting improved naturalness in the synthesized speech [3]. Similarly, the similarity score, which evaluates how appropriately the generated speech reflects the intended emotion in context, also shows a marginal improvement with DPO (3.06 ± 0.08 vs. 3.05 ± 0.10).

TABLE IV shows comparison results between the proposed text-to-unit model utilizing conversation context and an emotion-aware text-to-unit model that uses one-hot emotion labels as auxiliary input. The MOS results indicate that our proposed method (exp2) outperforms the one-hot model (exp3) (2.56 ± 0.07 vs. 2.39 ± 0.07), suggesting that conversation context provides richer and more effective guidance for generating natural-sounding speech. Furthermore, the similarity scores also demonstrate a clear advantage for the proposed model (3.06 ± 0.08) (exp2) compared to the one-hot emotion model (2.79 ± 0.09) (exp3).

C. Analyzing the Effectiveness of DPO Using Emotion Recognition for Similar Emotions

To further analyze the effectiveness of DPO in style-controlled speech synthesis, We analyzed the effect of DPO in two ways: preference margin analysis and emotion recognition experiments results analysis to evaluate discrimination of similar emotions as highlighted in Section I.

Unlike typical TTS training method using MSE, our approach incorporates DPO-based preference optimization to explicitly reinforce contextually suitable emotional realizations. We evaluated preference margins—log-probability gaps between preferred and dispreferred outputs on 30 test samples. As shown in Fig. 2, the DPO-trained model exhibited significantly larger margins than the CE-trained baseline, indicating stronger alignment with preference signals.

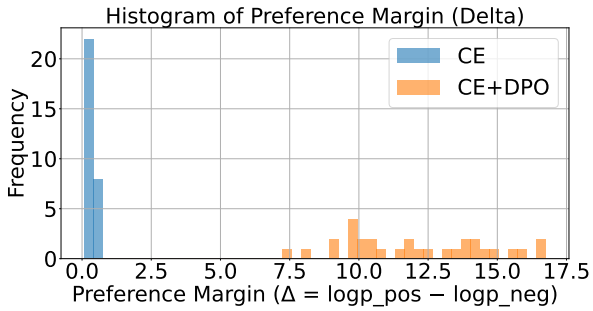

Fig. 2. Histogram of preference margins for 30 test samples.

TABLE V
CONFUSION MATRIX COMPARISON: ONLY CE VS CE+DPO — RECOGNITION ACCURACY FOR SIMILAR EMOTIONS.

CE				**CE+DPO**			
GT \ Pred	Happy	Angry	Sad	GT \ Pred	Happy	Angry	Sad
Happy	13	3	4	Happy	16	4	0
Angry	3	3	1	Angry	1	5	1
Sad	7	3	6	Sad	6	2	8

Moreover, we analyzed confusion matrix for emotion recognition experiments. TABLE V shows that DPO improved recognition accuracy, increasing correct "happy" samples from 13 to 16 and "angry" from 3 to 5. Similarly, the number of "sad" samples accurately classified rose from 6 to 8.

These improvements suggest that DPO facilitates a more effective separation of emotion-related features in the model's latent space, particularly for emotions like happy and angry, which often present overlapping acoustic characteristics in spontaneous speech. The enhanced recognition of sad samples further indicates that DPO fine-tuning produces more discriminative embedding representations. Overall, these results demonstrate that DPO contributes to more robust and reliable emotion recognition, especially for challenging or easily confusable categories.

D. Ablation Study: Effect of Context Length on Emotion Prediction Confidence

In dialogue systems, understanding emotional intent often requires contextual information from previous turns. However, the impact of context length on model confidence has not been systematically explored. To address this, we varied the number of preceding utterances provided to the model and assessed their influence on the accuracy and confidence of emotion

[3]The relatively low MOS values in this work are primarily due to the use of 64-class HuBERT discrete units for waveform synthesis, which inherently limits the upper bound of naturalness compared to conventional mel-spectrogram-based TTS systems. This trade-off is typical in unit-based neural vocoder approaches and is consistent with previous work.

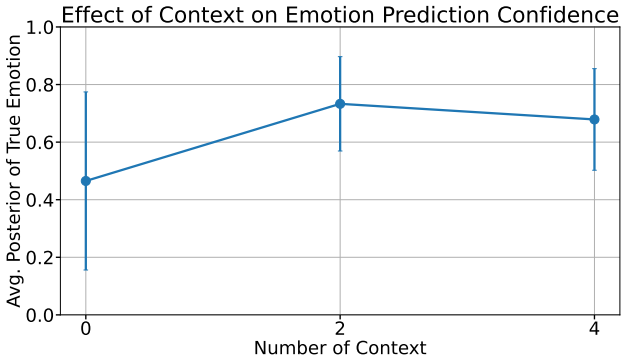

Fig. 3. Effect of context length on emotion recognition confidence.

recognition. This analysis was conducted on 20 dialogue samples. As shown in Fig. 3, the average and standard deviation of the posterior probability for the ground-truth emotion label were calculated using the SVM described in Section IV-C, across different context lengths.

The results demonstrate that incorporating even a small amount of context such as two preceding utterances—significantly increases the model's confidence in its predictions. In contrast, when no context is available, the model's confidence is not only lower on average but also more variable, reflecting greater uncertainty in its decisions.

These findings highlight that emotional intent in dialogue can be recognized more reliably when the model has access to relevant conversational cues, emphasizing the crucial role of context in emotion recognition.

V. Conclusion

In this paper, we proposed a conversation context-aware DPO method for style-controlled speech synthesis. By integrating conversation context into a text-to-unit model and fine-tuning with preference-based optimization, our method generates speech that better matches the emotional context of dialogues. Experimental results showed that our approach substantially improved emotion recognition accuracy compared to the model trained with CE, while MOS and style similarity also showed slight improvements. Future work will explore larger datasets, additional context factors, and direct experimental comparisons with recent methods such as CONCSS to further enhance the system.

References

[1] Y. Saito, T. Moriyama, K. Kubo, Y. Akita, and S. Takaki, "CALLS: Japanese empathetic dialogue speech corpus of complaint handling and attentive listening in customer center," in *Proc. Interspeech*, 2023, pp. 5561–5565.

[2] Y. Saito, S. Takamichi, E. Iimori, K. Tachibana, H. Saruwatari, "ChatGPT-EDSS: Empathetic Dialogue Speech Synthesis Trained from ChatGPT-Derived Context Word Embeddings," in *Proc. Interspeech*, 2023.

[3] G. Bruce, B. Granström, M. Filipsson, K. Gustafson, M. Horne, D. House, B. Lastow, P. Touati, "Speech Synthesis in Spoken Dialogue Research," in *Proc. Eurospeech*, 1995.

[4] H. Guo, S. Zhang, F. K. Soong, L. He, L. Xi, "Conversational End-to-End TTS for Voice Agent," in *Proc. SLT*, 2021.

[5] Y. Lee, A. Rabiee, and S. Y. Lee, "Emotional End-to-End Neural Speech Synthesizer," arXiv preprint arXiv:1711.05447, 2017.

[6] Y. Lee and T. Kim, "Robust and Fine-grained Prosody Control of End-to-End Speech Synthesis," in *Proc. ICASSP*, 2019, pp. 5911–5915.

[7] D.-H. Cho, H.-S. Oh, S.-B. Kim, S.-H. Lee, and S.-W. Lee, "EmoSphere-TTS: Emotional Style and Intensity Modeling via Spherical Emotion Vector for Controllable Emotional Text-to-Speech," in *Proc. Interspeech*, 2024, pp. 1810–1814.

[8] Y. Deng, J. Xue, Y. Jia, Q. Li, Y. Han, F. Wang, Y. Gao, D. Ke, and Y. Li, "CONCSS: Contrastive-based Context Comprehension for Dialogue-Appropriate Prosody in Conversational Speech Synthesis," in *Proc. ICASSP*, 2024, pp. 10706–10710.

[9] H. Feng, S. Ueno, and T. Kawahara, "End-to-End Speech Emotion Recognition Combined with Acoustic-to-Word ASR Model," in *Proc. Interspeech*, 2020, pp. 501–505.

[10] Y. Chiba, T. Nose, and A. Ito, "Multi-condition training for noise-robust speech emotion recognition," *Acoustical Science and Technology*, vol. 40, no. 6, pp. 406–409, 2019.

[11] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *Proc. ICASSP*, 2015, pp. 4749–4753.

[12] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," in *Proc. NeurIPS*, 2023.

[13] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, E. Dupoux, "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations," in *Proc. Interspeech*, 2021.

[14] J. Kong, J. Kim, J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Proc. NeurIPS*, 2020.

[15] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, Y. Wu, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," in *Proc. ICASSP*, 2018.

[16] X. Gao, C. Zhang, Y. Chen, H. Zhang, and N. F. Chen, "Emo-DPO: Controllable Emotional Speech Synthesis through Direct Preference Optimization," in *Proc. ICASSP*, 2025, pp. 1–5.

[17] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, Oct. 2021.

[18] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, and I. Poli, "Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference," arXiv preprint arXiv:2412.13663, 2024.

[19] Y. Saito, Y. Nishimura, S. Takamichi, K. Tachibana, and H. Saruwatari, "STUDIES: Corpus of Japanese Empathetic Dialogue Speech Towards Friendly Voice Agent," in *Proc. Interspeech*, 2022, pp. 5155–5159.

[20] D. Xin, J. Jiang, S. Takamichi, Y. Saito, A. Aizawa, and H. Saruwatari, "JVNV: A Corpus of Japanese Emotional Speech with Verbal Content and Nonverbal Expressions," *IEEE Access*, vol. 12, pp. 19752–19764, 2024.

[21] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," arXiv preprint arXiv:1711.00354, 2017.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. CVPR*, 2016, pp. 2818–2826.

[23] A. Chakhtouna, S. Sekkate, and A. Adib, "Unveiling embedded features in Wav2vec2 and HuBERT models for Speech Emotion Recognition," *Procedia Computer Science*, vol. 232, pp. 2560–2569, 2024.