

TOWARD NATURAL EMOTIONAL TEXT-TO-SPEECH SYSTEM WITH FINE-GRAINED NON-VERBAL EXPRESSION CONTROL

Wangzixi Zhou, Bagus Tris Atmaja, Sakriani Sakti

Nara Institute of Science and Technology, Ikoma, Japan

ABSTRACT

While current emotional Text-to-Speech (TTS) models have successfully controlled verbal prosody, they often ignore non-verbal vocalizations (NVs), which are essential for authentic human emotion. While some non-verbal datasets have recently emerged, they often suffer from a lack of high-quality, fine-grained annotations. This limitation severely restricts a model's ability to precisely control the generation of NVs. To address this, we propose a novel approach for fine-grained non-verbal expression synthesis. We first reprocess female non-verbal utterances from the EARS corpus, then develop a new annotation scheme using special tags to encode NV types, frequencies, and durations, and finally build a non-verbal emotional TTS to demonstrate its effectiveness. Our subjective evaluation shows that while our fine-grained NV approach leads to a minor trade-off in perceived naturalness, it significantly improves both expressiveness (eMOS 4.20) and emotional recognition accuracy (82.0%). Our emotion-specific analysis further reveals that our non-verbal cues are particularly effective for high-arousal emotions like happy (82.5% accuracy) and fear (82.7% accuracy), and almost perfectly convey sad emotions (98.3% accuracy).

Index Terms— Non-verbal, Speech synthesis, Emotion

1. INTRODUCTION

Emotional Text-to-Speech (TTS) synthesis is a growing research area driven by the increasing integration of conversational AI into daily life. Users now expect emotionally rich and engaging responses from dialogue systems, making the ability to synthesize speech with specific emotional tones crucial for creating more empathetic and human-like AI experiences. While existing emotional TTS systems have made significant strides in generating emotionally speech, they have primarily concentrated on manipulating the verbal content. These models typically operate by using emotional embeddings to control prosodic features—such as pitch, duration, and energy—derived from the linguistic content. While this approach has been successful in many respects, they often ignore a critical component of expressive human communication: non-verbal vocalizations (NVs). As high-

lighted by Mehrabian [1], NVs refer to vocal expressions that carry no linguistic content, such as laughter, screams, and sighs. These vocalizations are pervasive in spoken interaction and are known to play a central role in conveying affective states [2]. They provide a layer of spontaneous, authentic emotional expression that verbal prosody alone cannot capture. Therefore, to achieve truly expressive and human-like emotional TTS, the inclusion and accurate synthesis of NVs are indispensable.

In response to this need, several non-verbal vocalization corpora have emerged. For instance, the AMI Meeting Corpus [3] provides 100 hours of meeting recordings with annotations for laughter and coughs. Similarly, a recent work on NVTTS [4] introduced a 17-hour dataset annotated with 10 types of NVs. However, a key limitation of these datasets is their simplified annotation schemes, which typically rely on coarse-grained tags like <laugh>. Such basic annotations severely restrict the model's ability to achieve fine-grained control over NV generation. For instance, a user might wish to control not just the presence of laughter, but its duration and frequency. The lack of a system for precise control over these NV properties remains a significant obstacle to developing advanced emotional TTS models. The ability to specify a cry's duration or the number of discrete laughs within a single utterance is currently unfeasible with existing annotation methodologies.

In this paper, we address this critical limitation by proposing a Fine-Grained Non-Verbal Expression Dataset designed to support more expressive speech synthesis. Our main contributions are as follows: (1) fine-grained NV expression data construction: We reprocessed non-verbal utterances from the EARS corpus, focusing on female speakers. We then developed a novel annotation scheme using Whisper for transcription and special tags to enable fine-grained control over non-verbal properties, including style, frequency, and duration; (2) fine-grained NV emotional TTS: We established a new non-verbal emotional TTS. This TTS is designed to specifically assess a model's ability to generate speech with fine-grained non-verbal expressions, demonstrating the superiority of our data design in improving both expressiveness and emotional recognition accuracy.

Table 1: The specific transcripts used for each non-verbal category and the total count of each type of utterance within our data.

Category	Cheering	Yelling	Laughter-open	Laughter-closed	Crying	Screaming
Transcript	'Wo ho', 'Yo'	'Hey'	'Ha'	'Ha'	'Whep', 'Wuu', 'Sneeze'	'Ah'
Type Count	2	1	1	1	3	1
Count	262	328	266	220	230	154

2. RELATED WORK

2.1. Non-verbal Dataset

While several non-verbal vocalization (NV) datasets have been proposed, their limitations hinder the development of fine-grained, controllable emotional TTS models.

The NVTTS dataset [4], derived from sources like Vox-Celeb [5] and Espresso [6], suffers from poor acoustic quality. With only 1525 of its 3642 utterances being noise-free, it is challenging for smaller TTS models to learn effectively without reproducing acoustic artifacts. The AMI Meeting Corpus [3] is limited by its narrow scope of NVs (mostly laughter and coughs), its domain-specific context of meetings, and its use of predominantly non-native English speakers. Similarly, the Japanese JNV dataset [7] is unsuitable for English TTS due to its small scale and culturally specific expressions, such as the angry NV "onore."

Therefore, the limited NV scope and cultural specificity of existing datasets necessitate a new, high-fidelity resource for training fine-grained, controllable NV-capable emotional TTS models.

2.2. Emotional TTS with NV

Recent advancements in Emotional TTS have seen the integration of NVs, facilitated by Large Language Models (LLMs) and proprietary large-scale datasets. These models allow for a more natural understanding and generation of NVs. For instance, CosyVoice2[8] is a streaming TTS system that adds coughs and sighs to its repertoire and offers multilingual synthesis with breathing and laughing capabilities. However, while its architecture is open-sourced, its crucial instruction dataset remains proprietary. Building upon this, NVTTS[4] fine-tunes CosyVoice2 on their corpus, enabling the synthesis of a wider variety of non-verbal sounds. This reliance on either proprietary data or fine-tuning large, pre-existing models does not bring substantial technical improvements to conventional TTS systems that lack access to such resources.

3. CONSTRUCT FINE-GRAINED NON-VERBAL EXPRESSION DATA

Our dataset is constructed by annotating and filtering the existing EARS dataset [9]. This section provides an overview

of the original data sources and presents statistics on the non-verbal sounds (NVs) and emotion tags in our final dataset.

3.1. Data Source

We utilized the EARS dataset, a high-quality speech corpus featuring 107 speakers from diverse backgrounds. The dataset encompasses a wide range of speaking styles, including emotional speech, various reading styles, non-verbal sounds, and conversational free-form speech. For our study, we specifically filtered for non-verbal sounds from 60 female speakers. The selected non-verbal types include six distinct categories: **laughter-open**, **laughter-closed**, **cheering**, **yelling**, **crying**, and **screaming**. The duration of the original audio files was approximately 10–12 seconds.

3.2. Dataset Processing

Given that the original audio files contained multiple instances of continuous non-verbal sounds, we first performed preprocessing using the pydub library, a simple yet effective Python tool for audio manipulation. We segmented each audio file based on periods of silence. The segmentation parameters were set as follows: a silence threshold of -40 dBFS, a minimum silence duration of 200 ms, and a silence buffer of 100 ms to be kept around the segmented clips. This process allowed us to extract 739 individual utterances, each lasting approximately 2–6 seconds, from the initial 360 audio recordings.

Table 2: Comparison of the differences between our proposed fine-grained approach and existing coarse-grained methods.

Coarse-grained	Fine-grained
<crying>	<(crying) wuuuuu whep >
• Simple style-level tag	• Style-level tag
-	• Vocalization type and control
-	• Frequency and duration control

Subsequently, we used the Whisper [10] speech recognition model for initial transcription and then performed manual verification to obtain the final non-verbal transcripts. Our transcription strategy, as detailed in Table 2, differs significantly from traditional methods, which often use a single, coarse-grained tag like <laugh> to represent non-verbal

sounds. We designed a more granular transcription system to achieve precise control and representation of different vocalizations. Our standardization process included the following principles:

Discrete Vocalizations: For discrete sounds like ‘ha’ or ‘whep’, we controlled the frequency by repeating specific syllables. For example, we used $\langle (\text{Laughter-open}) \text{ ha ha ha} \rangle$ or $\langle (\text{Laughter-closed}) \text{ ha ha ha} \rangle$ to differentiate between types of laughter and control the number of laughs within an utterance.

Continuous Vocalizations: For continuous sounds such as ‘Wo ho’, ‘Yo’, ‘Hey’, ‘wuu’, or ‘ah’, we controlled the duration by repeating the last character. For instance, a ‘wuu’ for crying might last about 1 second, with each additional ‘u’ adding approximately 0.2 seconds to the duration, allowing for precise control over the length of the cry.

This fine-grained transcription method enabled us to better understand and leverage the underlying structure of non-verbal sounds. The specific transcription types are detailed in Table 1.

4. NON-VERBAL EMOTIONAL TTS

We selected Grad-TTS [11] as our backbone model, which is recognized for its high-quality synthesis of reading-style speech. To enable emotional synthesis, we enhanced the model with an emotion encoder, allowing it to incorporate emotional embeddings. Following Russell’s circumplex model [12] of affect, we utilized arousal and valence as continuous emotional labels, which provides the model with more fine-grained control over emotional expression.

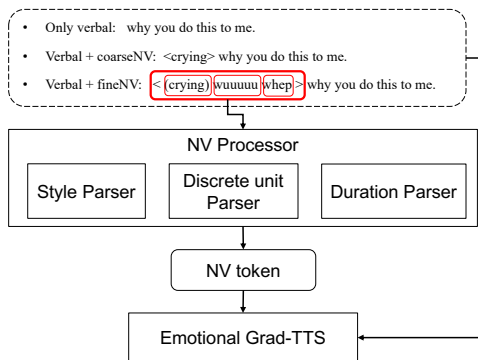


Fig. 1: The pipeline of specialized non-verbal processing.

To ensure the model could effectively learn and utilize our specific non-verbal annotations, we customized its text processing pipeline. As shown in the figure 1, our pipeline is composed of an NV processor that utilizes a style parser, a discrete unit parser, and a duration parser to handle our fine-grained non-verbal expressions. For example, a transcription like $\langle (\text{crying}) \text{ wuuuuu whep} \rangle$ why you do

this to me is first processed to identify the non-verbal segment $\langle (\text{crying}) \text{ wuuuuu whep} \rangle$. This segment is then further analyzed internally: the wyle Parser identifies the overall style *crying*, the discrete unit parser counts the occurrences of discrete vocalizations *whep*, and the duration parser calculates the duration based on continuous vocalizations *wuuuuu*. These components collectively contribute to the creation of NV tokens that structurally encode these non-verbal sounds, allowing them to be distinctly recognized and processed by the emotional Grad-TTS. This hierarchical and structured annotation method was integrated as the initial step in the text cleaning pipeline, ensuring that the non-verbal information is precisely parsed before being fed into the model. This approach allows the model to synthesize not only verbal speech but also highly expressive, emotionally congruent non-verbal sounds.

5. EXPERIMENT

5.1. Training setting

To enable the synthesis of emotional verbal speech, we constructed a comprehensive 9-hour mixed dataset of English female speakers, sampled at 22.05 kHz. It was compiled from: EXPRESSO [6], SEMAINE [13], and ESD[14] datasets. To address missing continuous arousal and valence labels in EXPRESSO and ESD, we used a pre-trained Speech Emotion Recognition (SER) model [15] to predict these values. For the non-verbal component, we utilized two distinct datasets to train our model and compare their synthesis performance. We used our curated fine-grained non-verbal expression data to enable precise control over non-verbal vocalizations. For comparison, we also trained the model using the NVTTS corpus [4] to represent coarse-grained non-verbal annotations. This allowed for a direct evaluation of our fine-grained approach against a conventional method.

For the acoustic features, we extracted 80-dimensional mel-spectrograms from our processed dataset. These spectrograms serve as the primary input for our model, capturing the necessary frequency and energy information of the speech. We then employed Hifi-GAN [16] as the vocoder, which is responsible for converting these mel-spectrograms back into high-fidelity audio waveforms. The model was trained for 400k iterations on a single GPU (NVIDIA RTX A6000 with 48GB of memory).

5.2. Evaluation

5.2.1. Perceptual and Recognition Evaluation

We conducted a subjective evaluation with 15 participants to compare three design: (1) verbal-only, (2) verbal + coarse-grained NV (from the ‘NVTTS corpus’ which only contained ‘happy’ and ‘sad’ emotions due to a lack of fear/anger-related non-verbal cues) (3) verbal + our proposed fine-grained NV.

To ensure emotion was conveyed through vocal cues rather than text, we chose 20 verbally ambiguous sentences as text input(e.g., "what did you do"). Participants rated 60 samples in total for naturalness (nMOS) and emotion expressiveness (eMOS) on a five-point scale. They also performed a four-choice emotion recognition accuracy task (happy, sad, fear, and anger). Due to data limitations, the "coarse-grained NV" was only evaluated for 'happy' and 'sad' NVs.

5.2.2. Non-verbal Expression Preference Evaluation

To further refine our non-verbal design, we conducted a Preference Test specifically for happy- and sad-related expressions, as these emotions were associated with multiple non-verbal cues in our dataset.

For happy emotions, we presented participants with four distinct non-verbal expressions combined with an identical verbal utterance:

- <(cheering) Wo ho>
- <(cheering) Yo>
- <(Laughter-open) ha ha>
- <(Laughter-closed) ha ha>

Similarly, for sad emotions, we provided four variations combined with the same verbal utterance:

- <(crying) whep>
- <(crying) sneeze>
- <(crying) wuuuuuuu whep>
- <(crying) wuuuuuuu>

Participants were asked to rank these combinations based on their preference.

5.3. Result and discussion

The subjective evaluation results are summarized in Figure 2. For Naturalness MOS, the "Only Verbal" achieved the highest score of 3.54, suggesting a minor trade-off in perceived naturalness when non-verbal cues are included. In contrast, for Expressiveness MOS, our "Fine-Grained Non-Verbal" significantly outperformed the others with a score of 4.20, demonstrating its effectiveness in enhancing emotional expressiveness.

For emotional recognition accuracy, our method achieved the highest score of 78.8%, 13.3% improvement over the "Only Verbal" baseline. The "Coarse-grained NV" performed poorly, with low accuracy (especially for 'happy' emotions), indicating its non-verbal cues may have been ambiguous and hindered recognition. A more detailed analysis of this is provided in Section 5.3.1. These results highlight the primary advantage of our approach: our structured and precise non-verbal cues not only increase perceived expressiveness but also significantly improve the listener's ability to correctly identify emotion.

5.3.1. Emotion-Specific Analysis

To provide a more detailed breakdown of each emotion, we conducted an analysis across individual emotion categories.

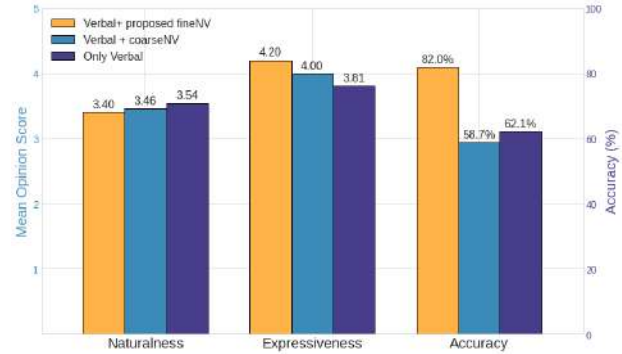


Fig. 2: The figure provides performance result of three designs across three key metrics: naturalness MOS, expressiveness MOS, and emotion recognition accuracy.

Figure3 and table3 present the accuracy, nMOS, and eMOS of each emotion.

Happy: Our proposed design achieved the highest accuracy at 82.5%, significantly outperforming the other two. The particularly low score of the 'Coarse-grained NV' suggests that its happy-related non-verbal cues (laugh) were perceived as ambiguous, its laughter samples were often quiet and subdued, leading many listeners to mistakenly perceive the emotion as 'sad' rather than 'happy'.

Sad: All three designs performed well for sad emotions, with our proposed design achieving the highest accuracy of 98.3%. This indicates that sad emotions are relatively easier to recognize from both verbal and non-verbal cues. However, the subtle differences in scores suggest that our non-verbal cues, which include a variety of crying sounds, still provide a slight edge in conveying sadness compared to the other methods.

Angry: Our method achieved 64.3% accuracy, a modest improvement compared to other emotions. We hypothesize that this is because we lacked a direct non-verbal cue that is uniquely associated with anger. While sad emotions are immediately recognizable through a crying NV, our primary NV for anger was 'yelling'. Yelling, however, is a more general high-arousal expression and does not instantly lead listeners to a specific 'angry' conclusion. This may be why its accuracy did not see a dramatic increase and why its E-MOS score was 4.06, which is lower than the 4.2+ scores of other emotions.

Fear: Our method reached a remarkable 82.7% accuracy, marking an approximately 36% improvement over the 'Only Verbal'. This highlights the crucial role of non-verbal cues (screaming) in expressing fear.

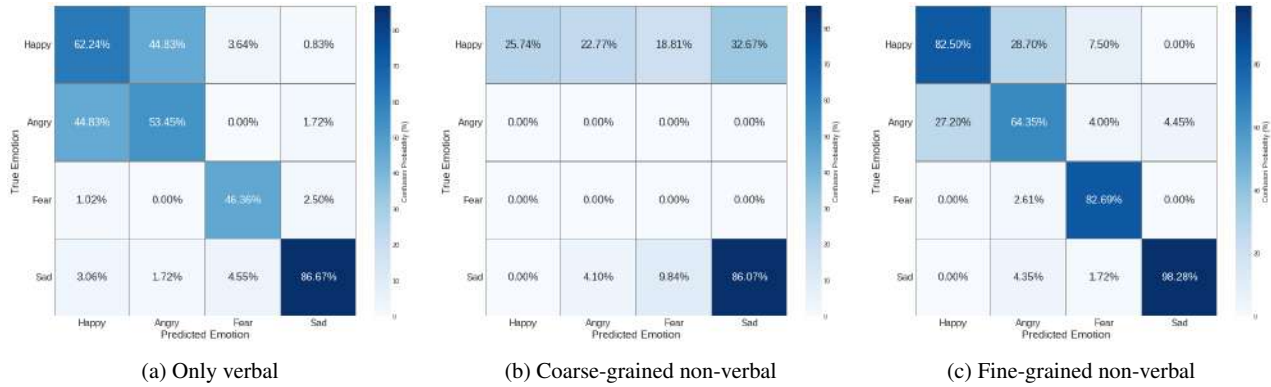
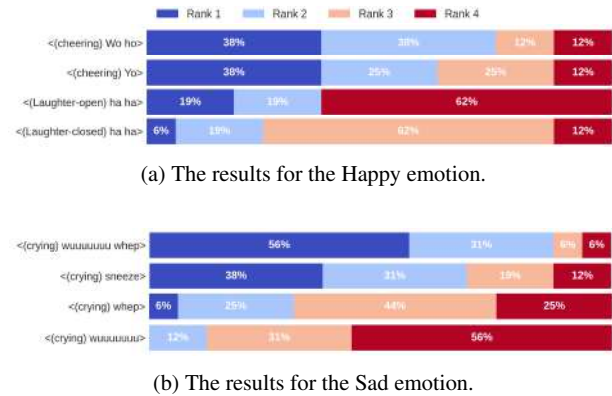
5.3.2. Preference Results

The results of our preference evaluation are visualized in Figures 4a and 4b.

For the Happy emotion, cheering sounds were distinctly

Table 3: Detailed nMOS(naturalness mean opinion score) and eMOS(emotion expressiveness mean opinion score) results for each emotion across the (1) Only Verbal, (2) Verbal + Coarse-grained NV, (3) Verbal + Fine-Grained NV.

	Only Verbal				Coarse-grained NV		Fine-Grained NV			
	Happy	Sad	Anger	Fear	Happy	Sad	Happy	Sad	Anger	Fear
nMOS	3.67	3.69	3.61	3.19	3.19	3.73	3.43	3.67	3.34	3.18
eMOS	3.78	3.83	3.74	3.89	3.85	4.15	4.21	4.25	4.06	4.28

**Fig. 3:** Emotion recognition confusion matrices for (a) Only verbal, (b) Verbal + coarse-grained non-verbal, and (c) Verbal + fine-grained non-verbal.**Fig. 4:** Preference evaluation results for different non-verbal expressions.

preferred over laughter. Both <(cheering) Wo ho> and <(cheering) Yo> were ranked as the top preference by a large portion of participants, indicating that cheering is perceived as a more effective way to convey joy and excitement. In contrast, laughter expressions, especially <(Laughter-closed) ha ha>, were consistently ranked lower, with 62% of participants ranking them third.

For the Sad emotion, a complex, multi-part expression was significantly preferred. The expression <(crying)

wuuuuuuu whep> (combining a prolonged cry with a sob) was ranked first by 56% of participants. This suggests that a nuanced non-verbal expression with multiple elements is perceived as the most authentic representation of sadness. Simple, single-sound expressions like <(crying) whep> and <(crying) wuuuuuuu> received lower preference ratings, with the latter being ranked fourth by 56% of participants.

These results provide valuable guidance for our fine-grained non-verbal design, validating the superiority of specific, structured non-verbal vocalizations in conveying particular emotions.

6. CONCLUSION

In this paper, we constructed a Fine-Grained Non-Verbal Expression Data and built a Fine-Grained Non-Verbal emotional TTS benchmark to enable the use of fine-grained non-verbal cues. Our comprehensive subjective evaluation yielded three key findings.

First, while the inclusion of NVs resulted in a minor trade-off in perceived speech naturalness compared to ‘verbal-only’, this was outweighed by significant gains in other metrics. Second, our “Fine-Grained Non-Verbal” design achieved superior performance in both Expressiveness and Emotional Recognition Accuracy, with an average E-MOS score of 4.20 and average accuracy of 82.0%. This demon-

strated that our designed “Fine-Grained Non-Verbal” is more effective than the coarse-grained NVs found in other corpora. Third, our emotion-specific analysis confirmed that our non-verbal cues were particularly effective in conveying emotions. They were especially potent for high-arousal emotions like happy (82.5% accuracy) and fear (82.7% accuracy). Furthermore, the accuracy for sad emotions was exceptionally high, achieving 98.3%. These results indicate that our non-verbal cues provided a distinct emotional manifestation and burst that was significantly more effective than what could be achieved with verbal-only prosody.

Finally, our preference test provided valuable insights into listener expectations, revealing a strong preference for cheering over laughter for happy emotions and for multi-part crying expressions for sad emotions. These findings validate our approach and highlight the importance of not just including NVs, but also designing them with granularity and user preference in mind to achieve truly expressive and human-like emotional synthesis.

Acknowledgement Part of this work was supported by JSPS KAKENHI Grant Numbers JP21H05054, JP23K21681, JP24K0296, and JP25H01139, as well as JST NEXUS (JP-MJNX25C1).

7. REFERENCES

- [1] Albert Mehrabian, *Nonverbal communication*, Routledge, 2017.
- [2] Klaus R Scherer, “Affect bursts,” *Emotions: Essays on emotion theory*, vol. 161, pp. 196, 1994.
- [3] Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post, “The ami meeting corpus,” in *Proc. International Conference on Methods and Techniques in Behavioral Research*, 2005, pp. 1–4.
- [4] Maksim Borisov, Egor Spirin, and Daria Diatlova, “Nonverbalts: A public english corpus of text-aligned nonverbal vocalizations with emotion annotations for text-to-speech,” *arXiv preprint arXiv:2507.13155*, 2025.
- [5] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Senior, “Voxceleb: Large-scale speaker verification in the wild,” *Computer Speech & Language*, vol. 60, pp. 101027, 2020.
- [6] Tu Anh Nguyen, Wei-Ning Hsu, Antony d’Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Reiz, Jade Copet, Gabriel Synnaeve, Michael Hassid, et al., “Expresso: A benchmark and analysis of discrete expressive speech resynthesis,” *arXiv preprint arXiv:2308.05725*, 2023.
- [7] Detai Xin, Shinnosuke Takamichi, and Hiroshi Saruwatari, “Jnv corpus: A corpus of japanese nonverbal vocalizations with diverse phrases and emotions,” *Speech Communication*, vol. 156, pp. 103004, 2024.
- [8] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al., “Cosyvoice 2: Scalable streaming speech synthesis with large language models,” *arXiv preprint arXiv:2412.10117*, 2024.
- [9] Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinji Watanabe, Alexander Richard, and Timo Gerkmann, “EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation,” in *ISCA Interspeech*, 2024, pp. 4873–4877.
- [10] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*. PMLR, 2023, pp. 28492–28518.
- [11] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov, “Grad-tts: A diffusion probabilistic model for text-to-speech,” in *International conference on machine learning*. PMLR, 2021, pp. 8599–8608.
- [12] James A Russell, “A circumplex model of affect,” *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161, 1980.
- [13] Gary McKeown, Michel F Valstar, Roderick Cowie, and Maja Pantic, “The semaine corpus of emotionally coloured character interactions,” in *2010 IEEE international conference on multimedia and expo*. IEEE, 2010, pp. 1079–1084.
- [14] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li, “Emotional voice conversion: Theory, databases and esd,” *Speech Communication*, vol. 137, pp. 1–18, 2022.
- [15] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller, “Dawn of the transformer era in speech emotion recognition: Closing the valence gap,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, 2023.
- [16] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *Proc. NeurIPS*, Dec. 2020, pp. 17022–17033.