Affectron: Emotional Speech Synthesis with Affective and Contextually Aligned Nonverbal Vocalizations

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, Seong-Whan Lee*
Department of Artificial Intelligence, Korea University, Seoul, Korea
{dh_cho,hs_oh,sb-kim,sw.lee}@korea.ac.kr

Abstract

Nonverbal vocalizations (NVs), such as laughter and sighs, are central to the expression of affective cues in emotional speech synthesis. However, learning diverse and contextually aligned NVs remains challenging in open settings due to limited NV data and the lack of explicit supervision. Motivated by this challenge, we propose Affectron as a framework for affective and contextually aligned NV generation. Built on a small-scale open and decoupled corpus, Affectron introduces an NV-augmented training strategy that expands the distribution of NV types and insertion locations. We further incorporate NV structural masking into a speech backbone pre-trained on purely verbal speech to enable diverse and natural NV synthesis. Experimental results demonstrate that Affectron produces more expressive and diverse NVs than baseline systems while preserving the naturalness of the verbal speech stream.

1 Introduction

Nonverbal vocalizations (NVs), including laughter, sighs, and cries, are essential for conveying affect in human communication, complementing prosodic modulation (Cortes et al., 2021; Kanda et al., 2024; Wu et al., 2024; Ye et al., 2025). However, most expressive text-to-speech (TTS) systems (Lee et al., 2022; Im et al., 2022; Zhou et al., 2023a,b; Inoue et al., 2024; Cho et al., 2025a; Wang et al., 2025b; Gao et al., 2025; Gudmalwar et al., 2025) remain limited in their ability to generate expressive speech that incorporates natural NVs.

Existing studies on speech synthesis incorporating NVs can be broadly categorized into two main approaches. The first is tag-controlled TTS (Zhang et al., 2023; Kanda et al., 2024; Wu et al., 2024; Borisov et al., 2025; Ye et al., 2025), which manually inserts explicit tags (e.g., ⟨laughing⟩, ⟨filler⟩) to specify NV type and location. Although this approach enables fine-grained control, it relies on aligned annotations or NV detection models (Omine et al., 2024; Schmid et al., 2025), whose biases and error propagation often result in temporal inconsistencies in NV locations. The second is spontaneous-style TTS (Li et al., 2024a,b), which predicts NVs from contextual cues without explicit alignment. Reproducibility in this approach is constrained by reliance on proprietary datasets and by limitations in the scale and quality of publicly available corpora. Several NV-integrated datasets have been introduced in recent work to mitigate these issues. However, richly annotated corpora (Zhang et al., 2023; Li et al., 2024a) are generally not publicly available. In contrast, publicly available corpora (Xin et al., 2024b,a; Wang et al., 2025a; Liao et al., 2025; Borisov et al., 2025) are generally skewed toward basic NVs, such as breathing and laughter, and frequently exhibit acoustic artifacts. Consequently, the absence of large, diverse, and high-quality public NV corpora continues to hinder accurate modeling of fine-grained NV variations, such as subtle chuckles, giggles, and snickers.

Recent developments in generative speech modeling have significantly enhanced the naturalness and flexibility of TTS systems. Neural codec language model (NCLM)-based zero-shot TTS systems (Peng et al., 2024; Chen et al., 2025; Du et al., 2024; Huang et al., 2025) can be trained on diverse and even low-quality speech corpora while still producing highly natural synthesized speech. Nevertheless, current NCLM-based TTS systems primarily focus on voice cloning, and the generation of human-like expressive speech with integrated NVs remains insufficiently explored. As a result, the ability to control NVs with fine-grained prosodic variation remains limited.

To address these limitations, we introduce Affectron, a method for generating affectively and contextually aligned NVs. During training, we mit-

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 27502–27525

igate data scarcity and bias by leveraging a small-scale open and decoupled corpus in which verbal speech and NVs are recorded separately (Richter et al., 2024). Based on this setting, we propose an NV-augmented training strategy that enhances the capability of a verbal-only pre-trained NCLM (Peng et al., 2024) to model diverse NV types and insertion locations. This augmentation comprises two components: (i) we introduce emotion-driven top-K NV matching, which selects emotionally aligned NVs for each verbal utterance to enhance affective consistency and increase NV diversity. (ii) we propose emotion-aware top-K routing, which locates the selected NVs at contextually appropriate locations, thereby eliminating dependence on alignment annotations or NV detectors. These two modules are used only during training to construct NV-augmented samples. Additionally, we incorporate NV structural masking into the NCLM to condition generation on the affective context of the surrounding verbal speech. During inference, the model generates speech from an NV-tagged text and an emotional reference utterance, without requiring any matching or routing procedure. Experimental results demonstrate that Affectron produces more expressive and diverse NV synthesis than previous NCLM-based TTS systems, while maintaining the naturalness of the verbal stream. Moreover, the proposed augmentation yields NV type-location distributions that more closely align with empirical data compared to competing approaches. Our audio samples and implementation code are publicly available at `https://choddeok.github.io/Affectron/`.

2 Related Work

2.1 Speech Synthesis with NVs

Recent studies have explored the generation of NVs in emotional TTS systems. Laughter Synthesis (Xin et al., 2023) feeds pseudo-phonetic tokens into the TTS to produce laughter. However, this method determines the location and variation of laughter in a largely stochastic manner, which limits controllability and constrains the expressive range. ELaTE (Kanda et al., 2024) and EmoCtrl-TTS (Wu et al., 2024) condition flow-matching-based TTS systems on NV embeddings. However, these works attempt to reduce labeling cost using NV detectors (Omine et al., 2024), but these methods are generally applicable only to laughter and crying, which restricts their generalizability. Spontaneous-TTS (Li et al., 2024b) supplies behavior labels and syntactic cues derived from linguistic features to model NVs explicitly. However, this approach depends on proprietary labeled datasets and necessitates explicit supervised annotations for training. Meanwhile, CosyVoice (Du et al., 2024, 2025) successfully synthesizes natural speech with NVs and supports fine-grained control. These methods require extensive, high-quality annotated corpora, and the resulting synthesized speech often lacks naturalness when multiple NV types are generated concurrently.

2.2 Neural Codec Language Models

NCLMs have recently emerged as a prominent paradigm for speech generation. They discretize audio into codec tokens, model sequential dependencies via next-token prediction, and decode the predicted tokens back into high-quality audio. Recent zero-shot TTS approaches frequently employ NCLMs as the foundational architecture. A representative example is VALL-E (Chen et al., 2025), which predicts part of the EnCodec codebooks (Défossez et al., 2022) using an autoregressive (AR) codec language model, while generating the remaining codebooks with a non-AR model. This design demonstrates that zero-shot TTS can be achieved by conditioning a codec language model on a brief reference prompt. VoiceCraft (Peng et al., 2024) is a Transformer-based NCLM that performs AR token infilling under a bidirectional context using a two-stage token rearrangement. This architecture facilitates efficient multi-codebook modeling and exhibits strong performance in both speech editing and zero-shot TTS. However, current NCLM-based TTS systems are primarily designed for voice cloning and have not been extensively investigated for generating human-like expressive speech. These systems frequently fail to reliably generate NVs with fine-grained prosodic variations, even when such cues are explicitly provided in the prompt.

3 Background

3.1 Affective Dynamics for NV Type and Location

NVs convey affective states more directly and effectively than verbal prosody, serving as subtle emotional cues in speech. Incorporating NVs into a speech synthesis system requires careful selection of NV types that align with the surrounding affective context (Gupta et al., 2012; Bänninger-Huber

and Salvenauer, 2023). In addition to selecting appropriate NV types, the location of NV occurrences significantly influences the perception and integration of emotional cues with verbal speech. Previous research (Ephratt, 2008; Hoey, 2014) demonstrates that the location of NVs within a sentence substantially affects the expressiveness and immersive quality of emotional communication.

Furthermore, studies on continuous emotion modeling (Wöllmer et al., 2013; Huang and Epps, 2018; Liu et al., 2025) indicate that emotional states typically evolve gradually rather than shift abruptly over time. Motivated by these characteristics of emotional transitions, we analyzed temporal patterns of emotional attribute changes across consecutive word segments in real data (Richter et al., 2024; Borisov et al., 2025). To compute these patterns, emotional attribute pseudo-labels (Wagner et al., 2023) were transformed into spherical coordinates (Cho et al., 2024, 2025b,c), and angular distances were measured on the unit sphere. This representation highlights directional changes in affective dynamics, enabling angular distance to more reliably capture relative emotional transitions between adjacent segments. As shown in Figure 1, shorter temporal intervals between words are associated with smaller angular distances. These findings indicate that affective states change gradually and exhibit local stability over brief temporal spans. Based on this observation, locations exhibiting minimal emotional attribute change are identified as emotionally stable locations and are assumed to serve as natural anchor points for NV event insertion. In our framework, inserting NVs at such locations is expected to preserve affective coherence while enhancing expressiveness. Appendix A provides a more detailed analysis of these temporal patterns of emotional attributes.

3.2 VoiceCraft Overview

VoiceCraft (Peng et al., 2024) serves as the backbone of Affectron, leveraging causal masking and delayed stacking to enable context-aware editing and infilling. Causal masking (Aghajanyan et al., 2022; Donahue et al., 2020; Bavarian et al., 2022) moves a masked span to the end of the sequence, allowing the model to condition on both past and future context. This bidirectional conditioning improves boundary naturalness and contextual coherence. Based on this, we hypothesize that bidirectional conditioning can benefit NV synthesis by leveraging the surrounding affective context. De-

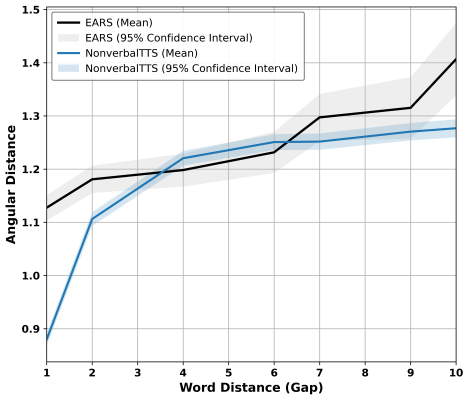


Figure 1: Analysis of emotional attribute change patterns across temporal gaps (up to a gap of 10) for the EARS (Richter et al., 2024) dataset (verbal only) and the NonverbalTTS (Borisov et al., 2025) dataset (verbal-nonverbal combined speech).

layed stacking (Copet et al., 2023) enables efficient multi-codebook AR modeling by adding a cumulative delay across EnCodec codebook streams (Défossez et al., 2022). This mechanism facilitates high-fidelity token prediction across parallel acoustic channels and supports the generation of fine-grained, high-quality audio signals. Therefore, high-fidelity token modeling is expected to be particularly advantageous for NV synthesis, which poses greater modeling challenges than verbal speech.

A decoder-only Transformer is trained to predict speech tokens, including the masked spans, conditioned on a transcript and optimized with a cross-entropy loss (Aghajanyan et al., 2022). Training on both observed and masked regions provides supervision at every timestep, stabilizing optimization and accelerating convergence.

4 Affectron

In this section, we introduce Affectron, which fine-tunes a speech backbone pre-trained on purely verbal speech (Peng et al., 2024) using affectively aligned NV augmentation constructed from an open-source decoupled corpus (Richter et al., 2024). Figure 2 illustrates the training-time augmentation and fine-tuning framework of Affectron, while the detailed components are described in the following subsections.

4.1 Emotion-Driven Top-K NV Matching

We design the emotion-driven top-K NV matching module to ensure affective consistency while preserving diversity among NV inputs. Given a

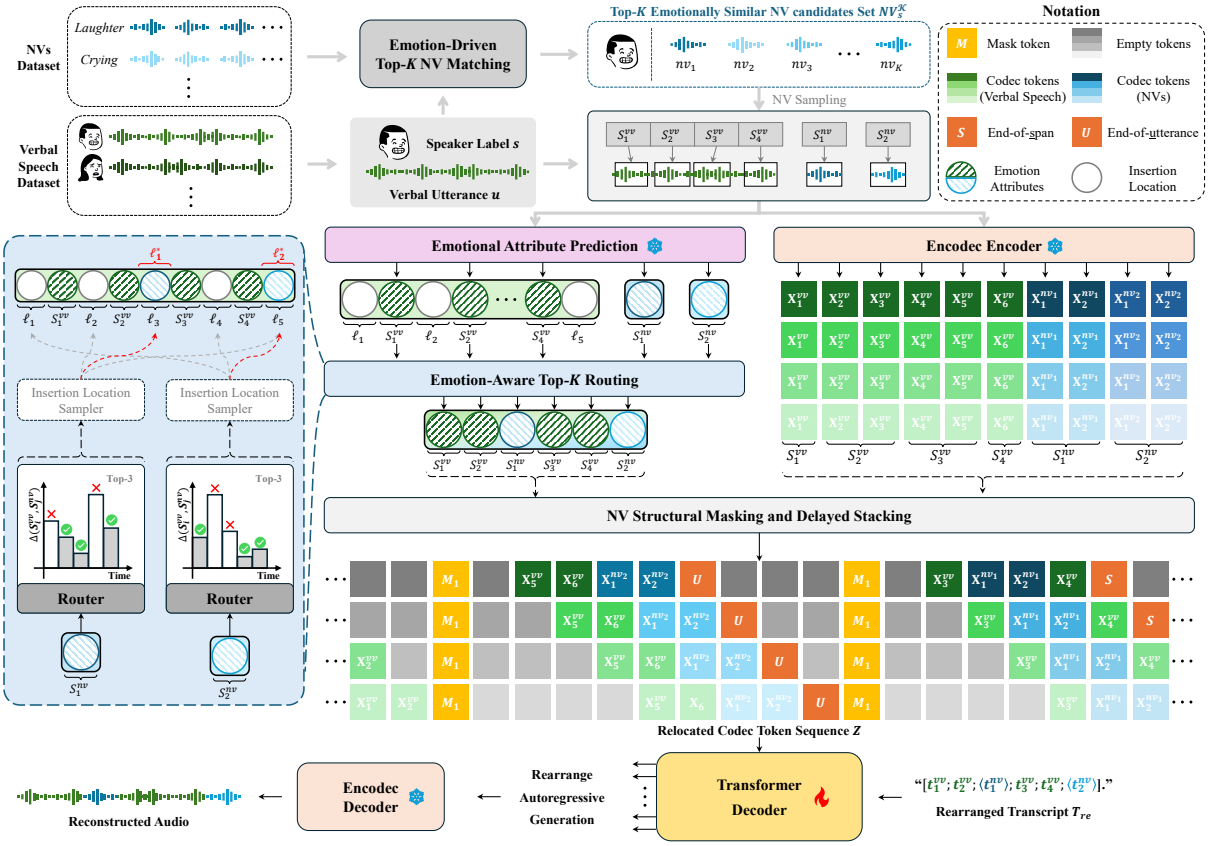


Figure 2: Overview of the Affectron training framework. NV candidates are selected and routed to contextually appropriate locations to construct NV-augmented training samples, which are then used to fine-tune the VoiceCraft backbone for affect-aware NV synthesis.

verbal utterance u and its corresponding speaker label s, all NV candidates NV_s^{all} associated with the speaker are retrieved. The emotional similarity between each candidate $nv_n \in NV_s^{\text{all}}$ and the utterance u is calculated using Emotion2Vec embeddings (Ma et al., 2024). The top-K candidates form an index set $\mathcal{K}$, and the corresponding NV subset is denoted as $NV_s^{\mathcal{K}}$. To facilitate probability-based selection among emotionally similar candidates, the top-K similarity scores are normalized using a temperature-scaled softmax distribution (Shazeer et al., 2017; Fan et al., 2018; Fedus et al., 2022):

$$p(nv_k \mid u) = \frac{\exp(\mathrm{CS}(\mathbf{e}^u, \mathbf{e}_k^{nv})/\tau)}{\sum_{m \in \mathcal{K}} \exp(\mathrm{CS}(\mathbf{e}^u, \mathbf{e}_m^{nv})/\tau)}, \quad (1)$$

where $\mathbf{e}^u$ and $\mathbf{e}_k^{nv}$ denote the emotion embeddings of u and the k-th NV candidate $nv_k \in NV_s^{\mathcal{K}}$, $\mathrm{CS}(\cdot,\cdot)$ denotes the cosine similarity, and τ is the temperature parameter. To handle cases where NVs occur multiple times within an utterance, we sample up to two NVs from the selection distribution:

$$NV_s^* = \{S_j^{nv} \mid S_j^{nv} \sim p(nv_k \mid u),\ j \le 2\}, \quad (2)$$

where NV_s^* denotes the set of NV candidates S_j^{nv} sampled from the probability distribution $p(nv_k \mid u)$ over the top-K candidates. Additional analyses of alternative embedding choices for NV matching and the validity of pseudo-label-based NV–emotion alignment are provided in Appendices B and C, respectively.

4.2 Emotion-Aware Top-K Routing

The emotion-aware top-K routing module determines contextually appropriate locations for NV insertion. Initially, word-level segments are extracted from each verbal utterance u using the Montreal Forced Aligner (McAuliffe et al., 2017). Emotional attribute pseudo-labels are assigned to each verbal segment and each NV candidate using a pre-trained emotional attribute predictor (Wagner et al., 2023). The extracted emotional attributes are transformed into a spherical coordinate space to quantify subtle local affective dynamics (Cho et al., 2024, 2025b,c). For each NV candidate S_j^{nv}, emotional attribute changes relative to each verbal segment S_i^{vv} are measured using angular distance

on the unit sphere:

$$\Delta(S_j^{nv}, S_i^{vv}) = \arccos(\sin\theta_j^{nv}\sin\theta_i^{vv} + \cos\theta_j^{nv}\cos\theta_i^{vv}\cos(\phi_j^{nv} - \phi_i^{vv})). \quad (3)$$

Here, θ and ϕ represent the elevation and azimuth angles. The affective distance $\mathrm{d}(\cdot,\cdot)$ for the t-th potential insertion location ℓ_t is computed as:

$$\mathrm{d}(S_j^{nv}, \ell_t) = \begin{cases} \Delta(S_j^{nv}, S_t^{vv}), & \text{if } t = 1, \\ \Delta(S_j^{nv}, S_{t-1}^{vv}), & \text{if } t = \mathrm{I}+1, \\ \frac{\Delta(S_j^{nv}, S_{t-1}^{vv}) + \Delta(S_j^{nv}, S_t^{vv})}{2}, & \text{otherwise.} \end{cases} \quad (4)$$

Here, I denotes the total number of verbal segments in the utterance. After computing distances for all candidate insertion locations, the top-K locations with the smallest distances are selected to define the index set $\mathcal{K}$. Subsequently, the negative distances are transformed into a temperature-scaled softmax distribution (Shazeer et al., 2017; Fan et al., 2018; Fedus et al., 2022):

$$p(\ell_k | S_j^{nv}) = \frac{\exp(-\mathrm{d}(S_j^{nv}, \ell_k)/\tau)}{\sum_{m\in\mathcal{K}} \exp(-\mathrm{d}(S_j^{nv}, \ell_m)/\tau)}, \quad (5)$$

where ℓ_k denotes the k-th candidate insertion location and τ is the temperature parameter. The insertion location ℓ_j^* for the NV candidate S_j^{nv} is determined by sampling from the selection distribution:

$$\ell_j^* \sim p(\ell_k | S_j^{nv}), \quad k \in \mathcal{K}. \quad (6)$$

Detailed formulations of the spherical coordinate transformation and the routing computation procedure are provided in Appendix D.

4.3 NV Structural Masking and Delayed Stacking

Building on the causal masking strategy (Peng et al., 2024), we propose an NV structural masking scheme that applies this mechanism to NV codec tokens. The codec sequence is initially rearranged according to the selected insertion locations determined by emotion-aware routing. For clarity, consider a token sequence composed of a verbal span $X^{vv} = (x_1^{vv}, \ldots, x_6^{vv})$ and two NV spans $X^{nv_1} = (x_1^{nv_1}, x_2^{nv_1})$ and $X^{nv_2} = (x_1^{nv_2}, x_2^{nv_2})$. This sequence is rearranged into $X^{\mathrm{re}} = (X_1^{vv}; X^{nv_1}; X_2^{vv}; X^{nv_2})$, where ";" denotes concatenation, $X_1^{vv} = (x_1^{vv}, \ldots, x_3^{vv})$, and $X_2^{vv} = (x_4^{vv}, \ldots, x_6^{vv})$. To enable emotion-conditioned infilling based on verbal context, one NV span is randomly sampled, and a masked span is constructed around it. The masked span length is sampled as $l \sim \mathrm{Uniform}(1, L)$, which allows the span to optionally include adjacent verbal tokens. For instance, if the masked span length is 4 and the masking is applied around X^{nv_1}, a masked span such as $\langle MASK_1 \rangle = (x_3^{vv}, x_1^{nv_1}, x_2^{nv_1}, x_4^{vv})$ can be selected. For each masked span $\langle MASK_n \rangle$, the original tokens are relocated to the end of X^{re}, with a mask token M_n inserted immediately in front of the span. Finally, delayed stacking (Copet et al., 2023) is applied to facilitate efficient AR modeling across parallel codebook streams, resulting in the final relocated token sequence Z.

4.4 Modeling with Aligned NV-Augmented Tokens

The Transformer decoder employs a GPT-style architecture to model the relocated codec token sequence Z, in which nonverbal and verbal segments are jointly aligned and augmented with NV tokens. To enable AR generation, the model is conditioned on the rearranged speech transcription $T_{re} = [t_1^{vv}; t_2^{vv}; \langle t_1^{nv} \rangle; t_3^{vv}; t_4^{vv}; \langle t_2^{nv} \rangle]$, constructed using the optimal insertion location ℓ_j^* identified by the emotion-aware routing module. The model is optimized with the standard language modeling objective, applying cross-entropy loss across all tokens (Peng et al., 2024).

4.5 Training and Inference Workflow

During training, Affectron constructs NV-augmented utterances by applying emotion-driven top-K NV matching and emotion-aware top-K routing to verbal utterances and candidate NV segments. The resulting NV-augmented transcripts and rearranged codec token sequences are then used to fine-tune the VoiceCraft backbone (Peng et al., 2024) with an AR codec language modeling objective. During inference, in contrast, the model directly generates speech from an NV-tagged transcript and a reference utterance that provides the target speaker identity and emotional condition, without applying matching or routing.

5 Experiments

5.1 Dataset

The EARS dataset (Richter et al., 2024) was utilized for fine-tuning. It consists of approximately 100 hours of clean speech recorded under anechoic conditions from 107 speakers, encompassing di-

verse reading styles and emotions. In addition to verbal speech, the corpus contains separately recorded NVs across 15 types, totaling approximately four hours. Each audio file contains either multiple sentences of verbal speech or multiple stylistic variations of a single NV type, such as fillers (e.g., um, oh). Whisper (Radford et al., 2023) was employed to segment multi-sentence verbal recordings into individual utterances. Silence detection based on simple acoustic cues was applied to divide NV files containing multiple stylistic variations into individual NV events. The seen-speaker test set was constructed by sampling one random utterance for each combination of speaker and emotion, resulting in approximately 2,200 samples. To evaluate zero-shot generalization, four speakers (p001, p004 for males; p002, p003 for females) were held out as unseen speakers. An unseen-speaker test set was additionally constructed by sampling multiple random utterances for each combination of speaker and emotion, yielding approximately 400 evaluation samples.

NonverbalTTS (Borisov et al., 2025) is a 17-hour open-access English speech corpus with aligned text annotations for 10 NV types. This dataset was used to evaluate NV locations and types. The test set includes approximately 1,000 sentences, each containing a single NV. The onset locations of NV events were labeled using FlexSED (Hai et al., 2025), following prior work (Borisov et al., 2025).

5.2 Implementation Details

We initialized our model using the 330M-parameter VoiceCraft checkpoint[1] (Peng et al., 2024), which was pre-trained on purely verbal speech (Chen et al., 2021), and employed EnCodec (Défossez et al., 2022) as the speech tokenizer. For fine-tuning, we used the AdamW optimizer with a learning rate of 1×10^{-5}, a batch size of 100 via gradient accumulation, trained for 50,000 steps. Training was conducted on four NVIDIA RTX A6000 GPUs over a period of five days. During training, the number of masked spans was sampled from a truncated $\mathrm{Poisson}(1)$ distribution in $[1, 3]$, with span lengths drawn from $\mathrm{Uniform}(1, 600)$. For both the emotion-driven NV matching and emotion-aware routing modules, we set the top-K values to 10 and 5, and applied a temperature parameter of $\tau = 0.7$.

[1] https://huggingface.co/pyp1/VoiceCraft/tree/main

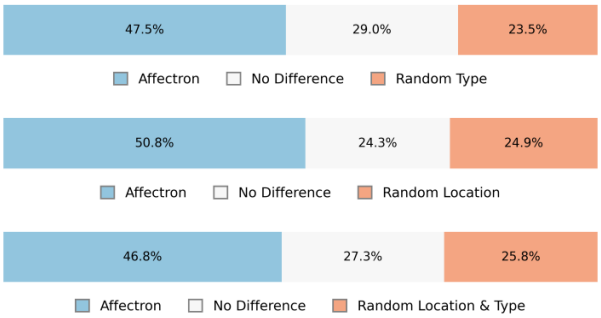


Figure 3: AB preference test comparing the proposed NV augmentation with rule-guided randomized strategies following CapSpeech (Wang et al., 2025a).

5.3 Evaluation Metrics

We evaluated our model using both subjective and objective metrics to assess the quality of synthesized speech. For subjective evaluation, we conducted AB preference tests and mean opinion score (MOS) assessments for NV-type naturalness (NTN-MOS) and NV-context emotional congruence (NEC-MOS). For objective evaluation, we measured verbal and nonverbal speaker embedding cosine similarity (V/NV-SECS), verbal and nonverbal emotion embedding cosine similarity (V/NV-EECS), nonverbal classification accuracy (NV-Acc), nonverbal similarity (NV-Sim), and word error rate (WER). Additionally, we evaluated NV-type and location prediction using top-K accuracy (Acc@K), Jensen-Shannon Divergence (JSD), and Hellinger Distance (HD). Detailed definitions for metrics are provided in Appendix E.

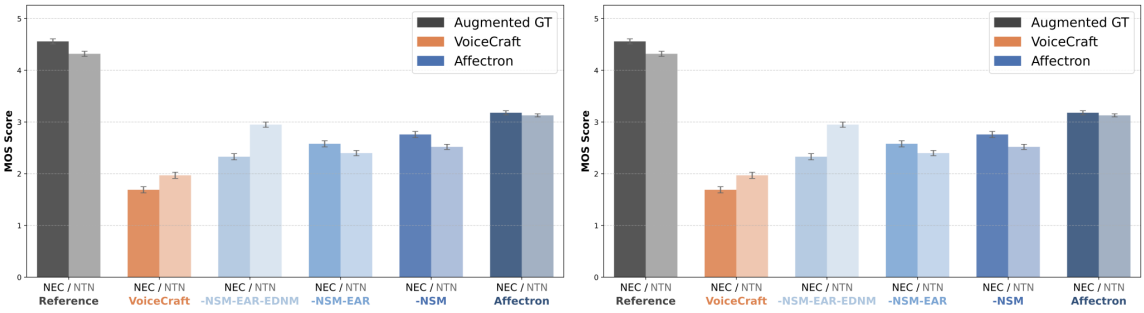
6 Results

6.1 Preference Evaluation of Augmentation

We evaluated whether different NV augmentation methods yield perceptually natural and contextually appropriate insertions using AB preference tests on ground-truth (GT) utterances. For a fair comparison, the proposed augmentation method was replaced with three rule-guided randomized strategies, following CapSpeech (Wang et al., 2025a). These strategies restrict NV insertion to silent regions and randomly select NV locations and types within the permitted set. Figure 3 indicates that the proposed affect-aware NV augmentation is consistently preferred over the three rule-guided randomized strategies. These results suggest that selecting NV types and locations based on affective context leads to more natural and better-integrated NV expressions.

Method	DA	EDNM	EAR	NSM	Nonverbal Metrics NV-Acc (↑)	NV-Sim (↑)	EECS (↑)	SECS (↑)	Verbal Metrics WER (↓)	EECS (↑)	SECS (↑)
SEEN SPEAKERS											
Augmented GT	-	-	-	-	85.96	-	0.5796	0.9231	1.13	0.6186	0.9163
VoiceCraft (Peng et al., 2024)	✗	✗	✗	✗	10.49	0.5898	**0.6149**	0.8950	9.05	0.6212	0.8927
Affectron (Proposed)	✓	✗	✗	✗	**58.78**	0.5988	0.5455	**0.8927**	6.06	0.6190	**0.8950**
	✓	✓	✗	✗	35.83	0.6085	0.5648	0.8897	**6.02**	0.6126	0.8892
	✓	✓	✓	✗	32.93	0.6090	0.5707	0.8915	7.52	0.6211	0.8889
	✓	✓	✓	✓	37.75	**0.6118**	0.5748	0.8906	6.59	**0.6216**	0.8886
UNSEEN SPEAKERS											
Augmented GT	-	-	-	-	89.29	-	0.5487	0.9092	2.93	0.6733	0.8995
VoiceCraft (Peng et al., 2024)	✗	✗	✗	✗	11.90	0.4766	0.5479	0.8755	10.50	0.5582	**0.8690**
Affectron (Proposed)	✓	✗	✗	✗	**52.38**	0.5186	0.5012	**0.8757**	9.48	0.5472	0.8657
	✓	✓	✗	✗	26.19	0.5127	0.5252	0.8694	8.82	0.5580	0.8687
	✓	✓	✓	✗	33.33	0.5230	0.5281	0.8676	10.21	0.5547	0.8644
	✓	✓	✓	✓	36.90	**0.5427**	**0.5506**	0.8686	**8.31**	**0.5591**	0.8630

Table 1: Experimental results of the proposed method for both seen and unseen speakers. DA, EDNM, EAR, and NSM denote data augmentation, emotion-driven top-K NV matching, emotion-aware top-K routing, and NV structural masking, respectively. Augmented GT applies our NV augmentation to the ground truth.



(a) Seen Speaker (b) Unseen Speaker

Figure 4: Comparison of our proposed method in terms of NTN-MOS and NEC-MOS. Augmented GT applies our NV augmentation to the ground truth. Vertical lines illustrate the 95% confidence intervals.

6.2 Model Performance

We performed both subjective and objective evaluations to assess the impact of each proposed module within Affectron. Experiments were conducted under two conditions: synthesis of verbal-only speech and synthesis of speech with NVs.

As shown in Figure 4 and Table 1, we conducted ablation studies by removing each proposed module from the Affectron model. 1) "w/o NSM" discarded the NV structural masking module and reverted to baseline causal masking (Peng et al., 2024). Unlike random masking, the proposed structure leverages both past and future affective context from verbal speech during generation, enhancing the naturalness and expressiveness of the NV synthesis. 2) "w/o EAR" removed the emotion-aware top-K routing module, inserting NVs at rule-guided random locations (Wang et al., 2025a). Aligning insertion locations with the progression of affective change enables the full model to achieve a more emotionally expressive integration of NV and verbal components. 3) "w/o EDNM" removed the emotion-driven top-K NV matching module, resulting in random pairing of NVs with verbal segments from the same speaker (Wang et al., 2025a). Random matching increased NV diversity, resulting in higher NV-Acc and NTN-MOS, but a lack of emotional congruence with verbal speech significantly reduced EECS. 4) "w/o DA" removed the NV-augmented training process, resulting in separate training of verbal and NV data. Without augmentation, the model overfitted to the emotional information present in each input, resulting in higher EECS while degrading NV-related performance.

Overall, these results demonstrate that Affectron enhances NV expressiveness and diversity while maintaining natural speech continuity. Additional comparisons with NV-capable zero-shot TTS models, analyses of speaker entanglement under cross-speaker NV mixing, and top-K ablation results are provided in Appendices F, G, and H, respectively.

Method	Acc@1 (↑)		Acc@3 (↑)		Acc@5 (↑)		JSD (↓)		HD (↓)	
	Location	Type	Location	Type	Location	Type	Location	Type	Location	Type
Qwen 2.5-7B (Hui et al., 2024)	7.48	25.65	14.13	47.03	25.30	69.39	0.1425	0.2139	0.3970	0.5027
LLaMA 3.1-8B (Grattafiori et al., 2024)	5.46	16.03	18.76	49.88	21.85	77.32	0.4399	0.3020	0.3884	0.5741
GPT-oss-20B (Agarwal et al., 2025)	**26.84**	16.98	**50.24**	51.19	**62.47**	72.68	0.1278	0.1130	0.4328	0.3523
Vicuna-7B (Chiang et al., 2023)	13.78	11.28	23.63	41.57	39.07	69.48	0.1535	0.2431	0.3847	0.5176
Affectron-330M (Proposed)	12.59	**75.77**	29.69	**85.99**	44.06	**91.69**	**0.0523**	**0.0051**	**0.2414**	**0.0723**

Table 2: Comparison of zero-shot NV type and location prediction across LLMs and the Affectron model. Prediction accuracy is measured using top-K accuracy (Acc@1/3/5) and distributional alignment is assessed using Jensen-Shannon Divergence (JSD) and Hellinger Distance (HD).

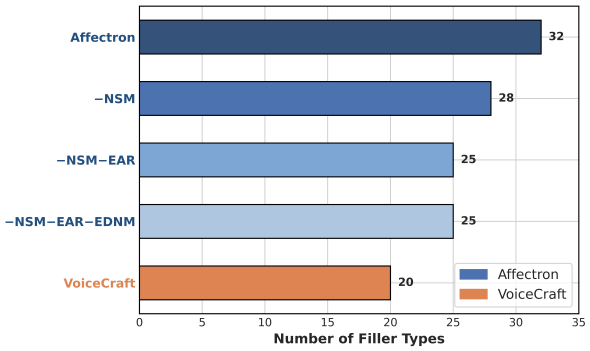


Figure 5: Comparison of the diversity of generated fine-grained filler variations across models.

6.3 Analysis of Generated Filler Diversity

To further investigate the expressive range of NVs, the number of distinct filler variations synthesized by each model when provided with the same NV tag was evaluated. A filler tag ⟨filler⟩ was inserted at affectively appropriate locations in the test set, and the realized filler type was extracted from each generated utterance, following (Kim et al., 2025). As shown in Figure 5, the number of distinct filler variations produced by the baseline model was compared with those produced by the ablated variants of Affectron. VoiceCraft (Peng et al., 2024) generated only a limited subset of filler variations, and removing any of the proposed modules progressively reduced filler diversity. In contrast, Affectron produced a wider variety of fine-grained filler NV realizations, suggesting that the NV-augmented training process promotes richer NV expression. A detailed analysis of NV categories and their distributional patterns is provided in Appendix I.

6.4 Comparison with LLM Baselines on NV Type and Location Prediction

Recent studies on NV-aligned emotional corpora (Xin et al., 2024a) and filler placement (Kim et al., 2025) have explored large language model (LLM)-based approaches. LLMs are capable of partially capturing discourse-level regularities, including those associated with NV expression patterns. Building on these findings, the proposed type and location prediction methods were evaluated against several LLMs without task-specific training. The evaluated LLMs include Qwen 2.5-7B (Hui et al., 2024), LLaMA 3.1-8B (Grattafiori et al., 2024), Vicuna-7B (Chiang et al., 2023), and GPT-oss-20B (Agarwal et al., 2025), which represent a range of model sizes and conversational capabilities. Both NV type and location prediction were evaluated using the NonverbalTTS dataset (Borisov et al., 2025), which offers high-quality NV annotations suitable for model comparison. Additional details regarding the LLM baselines and prompting procedures are available in Appendix J.

As shown in Table 2, Affectron achieved the highest type prediction accuracy as well as the lowest JSD and HD. These results indicate that emotion-driven top-K NV matching effectively captures affective category priors that align with actual NV distributions. In terms of location prediction, GPT-oss-20B attained the highest top-K accuracy. Conversely, Affectron achieved the lowest JSD and HD, demonstrating closer alignment with empirical positional distributions through its emotion-aware top-K routing. Although text-only LLMs can implicitly model discourse structure, the absence of prosodic and emotional cues limits their ability to distinguish fine-grained affective nuances in NV usage. Overall, these results suggest that while text-only LLMs can capture coarse discourse structure, explicit emotion-aware modeling is necessary to achieve well-aligned NV type and location prediction. Since this comparison focuses on text-only LLM baselines motivated by prior text-driven augmentation studies, we additionally report comparisons with multimodal audio-capable LLMs in Appendix K.

7 Conclusion

We introduce Affectron, an NCLM-based framework for expressive NV generation with emotion-aware control mechanisms. By leveraging a small-scale open and decoupled corpus, the framework augments data through emotion-driven top-K matching and emotion-aware top-K routing. This approach expands the distribution of NV types and locations while preserving the naturalness of verbal content. In contrast to conventional NCLM-based TTS systems, Affectron incorporates NV structural masking, enabling smoother transitions and preserving contextual coherence. Across subjective and objective evaluations, the proposed approach improves NV realism and diversity without degrading the naturalness of verbal speech. Overall, Affectron offers a practical and scalable solution for affect-aware NV synthesis using small open corpora, thereby reducing reliance on costly alignment processes and large proprietary datasets.

8 Limitations

Our methodology utilizes a relatively small-scale open corpus, such as the EARS dataset (Richter et al., 2024), where verbal speech and NVs are recorded independently. This limitation hinders direct comparison with models trained on large-scale or proprietary in-the-wild corpora. Additionally, the separation of verbal and nonverbal data in training restricts the model's capacity to represent overlapping verbal and nonverbal speech segments. However, accurately modeling the overlap between verbal speech and NVs remains an open challenge (Ludusan and Wagner, 2020). Human annotators frequently disagree regarding the precise boundaries and timing of these events (Truong et al., 2019). In future research, we aim to expand the NV inventory using semi-supervised mining and to enhance the modeling of overlapping verbal and nonverbal segments.

9 Acknowledgments

This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) under the artificial intelligence graduate school program (Korea University) (No. RS-2019-II190079) and artificial intelligence star fellowship support program to nurture the best talents (IITP-2026-RS-2025-02304828) grant funded by the Korea government (MSIT).

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. Gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and 1 others. 2022. Cm3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 33:12449–12460.

Eva Bänninger-Huber and Stefan Salvenauer. 2023. Different types of laughter and their function for emotion regulation in dyadic interactions. *Current Psychology*, 42(28):24249–24259.

Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. 2022. Efficient training of language models to fill in the middle. *arXiv preprint arXiv:2207.14255*.

Rudolf Beran. 1977. Minimum hellinger distance estimates for parametric models. *The annals of Statistics*, pages 445–463.

Maksim Borisov, Egor Spirin, and Daria Diatlova. 2025. Nonverbaltts: A public english corpus of text-aligned nonverbal vocalizations with emotion annotations for text-to-speech. *arXiv preprint arXiv:2507.13155*.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Interspeech 2021*, pages 3670–3674.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 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.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2025. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and 1 others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, Sang-Hoon Lee, and Seong-Whan Lee. 2024. Emosphere-tts: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech. In *Proceedings of the Interspeech*, pages 1810–1814.

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, and Seong-Whan Lee. 2025a. DiEmo-TTS: Disentangled Emotion Representations via Self-Supervised Distillation for Cross-Speaker Emotion Transfer in Text-to-Speech. In *Interspeech 2025*, pages 4373–4377.

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, and Seong-Whan Lee. 2025b. Emosphere++: Emotion-controllable zero-shot text-to-speech via emotion-adaptive spherical vector. *IEEE Transactions on Affective Computing*, 16(3):2365–2380.

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, and Seong-Whan Lee. 2025c. EmoSphere-SER: Enhancing Speech Emotion Recognition Through Spherical Representation with Auxiliary Classification. In *Interspeech 2025*, pages 4653–4657.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generationn. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 47704–47720.

Diana S Cortes, Christina Tornberg, Tanja Bänziger, Hillary Anger Elfenbein, Håkan Fischer, and Petri Laukka. 2021. Effects of aging on emotion recognition from dynamic multimodal expressions and vocalizations. *Scientific reports*, 11(1):2647.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online. Association for Computational Linguistics.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Xian Shi, Keyu An, and 1 others. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Michal Ephratt. 2008. The functions of silence. *Journal of pragmatics*, 40(11):1909–1938.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F Chen. 2025. Emo-dpo: Controllable emotional speech synthesis through direct preference optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Lucas Goncalves, Ali N Salman, Abinay R Naini, Laureano Moro Velazquez, Thomas Thebaud, Leibny Paola Garcia, Najim Dehak, Berrak Sisman, and Carlos Busso. 2024. Odyssey 2024-speech emotion recognition challenge: Dataset, baseline framework, and results. *Development*, 10(9,290):4–54.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ashishkumar Prabhakar Gudmalwar, Ishan Darshan Biyani, Nirmesh J Shah, Pankaj Wasnik, and Rajiv Ratn Shah. 2025. Emoreg: Directional latent vector modeling for emotional intensity regularization in diffusion-based voice conversion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23960–23968.

Rahul Gupta, Chi-Chun Lee, and Shrikanth Narayanan. 2012. Classification of emotional content of sighs in dyadic human interactions. In *ICASSP 2012-2012 IEEE International Conference on Acoustics, Speech*

and Signal Processing (ICASSP), pages 2265–2268. IEEE.

Jiarui Hai, Helin Wang, Weizhe Guo, and Mounya Elhilali. 2025. Flexsed: Towards open-vocabulary sound event detection. *arXiv preprint arXiv:2509.18606*.

Elliott M Hoey. 2014. Sighing in interaction: Somatic, semiotic, and social. *Research on Language and Social Interaction*, 47(2):175–200.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Sung-Feng Huang, Heng-Cheng Kuo, Zhehuai Chen, Xuesong Yang, Pin-Jui Ku, Ante Jukic, Huck Yang, Yu Tsao, Yu-Chiang Frank Wang, Hung yi Lee, and Szu-Wei Fu. 2025. VoiceNoNG: Robust High-Quality Speech Editing Model without Hallucinations. In *Interspeech 2025*, pages 3469–3473.

Zhaocheng Huang and Julien Epps. 2018. Prediction of emotion change from speech. *Frontiers in ICT*, 5:11.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2.5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Chae-Bin Im, Sang-Hoon Lee, Seung-Bin Kim, and Seong-Whan Lee. 2022. Emoq-tts: Emotion intensity quantization for fine-grained controllable emotional text-to-speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6317–6321. IEEE.

Sho Inoue, Kun Zhou, Shuai Wang, and Haizhou Li. 2024. Hierarchical emotion prediction and control in text-to-speech synthesis. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10601–10605.

Naoyuki Kanda, Xiaofei Wang, Sefik Emre Eskimez, Manthan Thakker, Hemin Yang, Zirun Zhu, Min Tang, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, and 1 others. 2024. Making flow-matching-based zero-shot text-to-speech laugh as you like. *arXiv preprint arXiv:2402.07383*.

Seung-Bin Kim, Jun-Hyeok Cha, Hyung-Seok Oh, Heejin Choi, and Seong-Whan Lee. 2025. FillerSpeech: Towards human-like text-to-speech synthesis with filler insertion and filler style control. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34096–34113, Suzhou, China. Association for Computational Linguistics.

Sang-Hoon Lee, Seung-Bin Kim, Ji-Hyun Lee, Eunwoo Song, Min-Jae Hwang, and Seong-Whan Lee. 2022. Hierspeech: Bridging the gap between text and speech by hierarchical variational inference using self-supervised representations for speech synthesis. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 35:16624–16636.

Hanzhao Li, Xinfa Zhu, Liumeng Xue, Yang Song, Yunlin Chen, and Lei Xie. 2024a. Spontts: Modeling and transferring spontaneous style for tts. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12171–12175.

Weiqin Li, Peiji Yang, Yicheng Zhong, Yixuan Zhou, Zhisheng Wang, Zhiyong Wu, Xixin Wu, and Helen Meng. 2024b. Spontaneous Style Text-to-Speech Synthesis with Controllable Spontaneous Behaviors Based on Language Models. In *Interspeech 2024*, pages 1785–1789.

Huan Liao, Qinke Ni, Yuancheng Wang, Yiheng Lu, Haoyue Zhan, Pengyuan Xie, Qiang Zhang, and Zhizheng Wu. 2025. Nvspeech: An integrated and scalable pipeline for human-like speech modeling with paralinguistic vocalizations. *arXiv preprint arXiv:2508.04195*.

Jianhua Lin. 2002. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Jingwen Liu, Kan Jen Cheng, Jiachen Lian, Akshay Anand, Rishi Jain, Faith Qiao, Robin Netzorg, Huang-Cheng Chou, Tingle Li, Guan-Ting Lin, and 1 others. 2025. Emo-reasoning: Benchmarking emotional reasoning capabilities in spoken dialogue systems. *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Bogdan Ludusan and Petra Wagner. 2020. Speech, laughter and everything in between: A modulation spectrum-based analysis. In *Speech Prosody 2020*, pages 995–999.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760, Bangkok, Thailand. Association for Computational Linguistics.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech 2017*, pages 498–502.

Nari-labs. 2025. Dia. https://github.com/nari-labs/dia.

Hyung-Seok Oh, Sang-Hoon Lee, Deok-Hyeon Cho, and Seong-Whan Lee. 2025. Durflex-evc: Duration-flexible emotional voice conversion with parallel generation. *IEEE Transactions on Affective Computing*.

Taisei Omine, Kenta Akita, and Reiji Tsuruno. 2024. Robust Laughter Segmentation with Automatic Diverse Data Synthesis. In *Interspeech 2024*, pages 4748–4752.

Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. 2024. VoiceCraft: Zero-shot speech editing and text-to-speech in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12442–12462, Bangkok, Thailand. Association for Computational Linguistics.

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

Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinji Watanabe, Alexander Richard, and Timo Gerkmann. 2024. EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation. In *Interspeech 2024*, pages 4873–4877.

Florian Schmid, Tobias Morocutti, Francesco Foscarin, Jan Schlüter, Paul Primus, and Gerhard Widmer. 2025. Effective pre-training of audio transformers for sound event detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *Proceedings of the International Conference on Learning Representations*.

Khiet P. Truong, Jürgen Trouvain, and Michel-Pierre Jansen. 2019. Towards an annotation scheme for complex laughter in speech corpora. In *Interspeech 2019*, pages 529–533.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. 2023. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Helin Wang, Jiarui Hai, Dading Chong, Karan Thakkar, Tiantian Feng, Dongchao Yang, Junhyeok Lee, Laureano Moro Velazquez, Jesus Villalba, Zengyi Qin, and 1 others. 2025a. Capspeech: Enabling downstream applications in style-captioned text-to-speech. *arXiv preprint arXiv:2506.02863*.

Tianrui Wang, Haoyu Wang, Meng Ge, Cheng Gong, Chunyu Qiang, Ziyang Ma, Zikang Huang, Guanrou Yang, Xiaobao Wang, Eng Siong Chng, and 1 others. 2025b. Word-level emotional expression control in zero-shot text-to-speech synthesis. In *Proceedings of the 39th International Conference on Neural Information Processing Systems*.

Martin Wöllmer, Moritz Kaiser, Florian Eyben, Björn Schuller, and Gerhard Rigoll. 2013. Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163.

Haibin Wu, Xiaofei Wang, Sefik Emre Eskimez, Manthan Thakker, Daniel Tompkins, Chung-Hsien Tsai, Canrun Li, Zhen Xiao, Sheng Zhao, Jinyu Li, and 1 others. 2024. Laugh now cry later: Controlling time-varying emotional states of flow-matching-based zero-shot text-to-speech. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 690–697. IEEE.

Detai Xin, Junfeng Jiang, Shinnosuke Takamichi, Yuki Saito, Akiko Aizawa, and Hiroshi Saruwatari. 2024a. Jvnv: A corpus of japanese emotional speech with verbal content and nonverbal expressions. *IEEE Access*, 12:19752–19764.

Detai Xin, Shinnosuke Takamichi, Ai Morimatsu, and Hiroshi Saruwatari. 2023. Laughter synthesis using pseudo phonetic tokens with a large-scale in-the-wild laughter corpus. In *Interspeech 2023*, pages 17–21.

Detai Xin, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2024b. Jnv corpus: A corpus of japanese nonverbal vocalizations with diverse phrases and emotions. *Speech Communication*, 156:103004.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Runchuan Ye, Yixuan Zhou, Renjie Yu, Zijian Lin, Kehan Li, Xiang Li, Xin Liu, Guoyang Zeng, and Zhiyong Wu. 2025. A scalable pipeline for enabling nonverbal speech generation and understanding. *arXiv preprint arXiv:2508.05385*.

Haitong Zhang, Xinyuan Yu, and Yue Lin. 2023. Nsvtts: Non-speech vocalization modeling and transfer in emotional text-to-speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Kun Zhou, Berrak Sisman, Rajib Rana, Björn W. Schuller, and Haizhou Li. 2023a. Emotion intensity and its control for emotional voice conversion. *IEEE Transactions on Affective Computing*, 14(1):31–48.

Kun Zhou, Berrak Sisman, Rajib Rana, Björn W. Schuller, and Haizhou Li. 2023b. Speech synthesis with mixed emotions. *IEEE Transactions on Affective Computing*, 14(4):3120–3134.

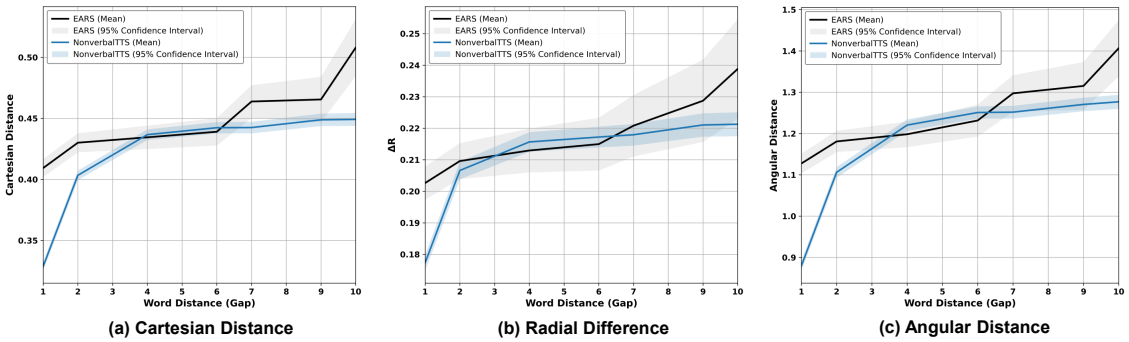


Figure 6: Comparison of emotional attribute change patterns across temporal gaps using angular distance, Cartesian distance, and spherical radial differences.

Method	Acc@1 (↑)	Acc@3 (↑)	Acc@5 (↑)	JSD (↓)	HD (↓)
Cartesian Distance	11.05	28.86	41.69	0.0568	0.2517
Radial Difference	11.88	28.98	41.45	0.0544	0.2495
Angular Distance	**12.59**	**29.69**	**44.06**	**0.0523**	**0.2414**

Table 3: Comparison of zero-shot NV type and location prediction across distance metrics. Prediction accuracy is measured using top-K accuracy (Acc@1/3/5) and distributional alignment is assessed using Jensen-Shannon Divergence (JSD) and Hellinger Distance (HD).

A Analysis of Emotional Attribute Dynamics Across Distance Metrics

Emotional attribute changes were analyzed across multiple distance metrics to assess the robustness of the observed affective dynamics. Specifically, three distance metrics were compared to measure emotional attribute changes between pairs of speech segments separated by varying temporal gaps: (i) Cartesian distance, (ii) radial distance, and (iii) angular distance on the unit sphere, following (Cho et al., 2024, 2025b,c). Further details regarding emotional attribute extraction and coordinate transformation are provided in Appendix D.

Cartesian Distance. Affective changes were measured directly in Cartesian coordinates. For two speech segments S_i and S_{i+t} in the sequence, separated by a word-level gap t, the Cartesian distance is defined as:

$$\Delta(S_i, S_{i+t}) = \|\mathbf{C}_i - \mathbf{C}_{i+t}\|_2, \tag{7}$$

where $\mathbf{C}_i = (A_i, V_i, D_i)$ denotes the emotional attribute vector of segment S_i. Each segment S_i represents either a verbal or nonverbal segment.

Radial Difference. After transforming emotional attributes into spherical coordinates (r, θ, ϕ), affective changes were measured based solely on radial magnitude. Given two speech segments S_i and S_{i+t} separated by a word-level gap t, the spherical radial difference is defined as:

$$\Delta(S_i, S_{i+t}) = |r_i - r_{i+t}|, \tag{8}$$

where r_i denotes the radial magnitude of segment S_i.

Angular Distance. To capture directional affective changes, angular distance was computed in spherical coordinate space. Given two speech segments S_i and S_{i+t} separated by a word-level gap t, the angular distance is defined as:

$$\Delta(S_i, S_{i+t}) = \arccos(\sin\theta_i \sin\theta_{i+t} + \cos\theta_i \cos\theta_{i+t} \cos(\phi_i - \phi_{i+t})), \tag{9}$$

where θ_i and ϕ_i denote the elevation and azimuth angles of segment S_i.

Figure 6 compares the emotional attribute change curves obtained using the three distance metrics for temporal gaps up to 10. For all metrics, affective differences increased as temporal gaps widened. This observation indicates that local affective states exhibit relative stability over short timescales. The influence of these differences on NV location prediction was further evaluated using each distance measure. Table 3 presents the top-K accuracy results (Acc@K) and distributional alignment as measured by Jensen-Shannon Divergence (JSD) and Hellinger Distance (HD). Angular distance achieved higher top-K accuracy and lower values for JSD and HD compared to the other metrics. Overall, these results confirmed that emotional attribute change patterns are consistent across distance formulations, with angular distance showing slightly more stable alignment characteristics for NV location modeling.

Method	Acc@1 (↑)	Acc@3 (↑)	Acc@5 (↑)	JSD (↓)	HD (↓)
Wav2Vec2-Base	75.15	85.04	86.94	0.0074	0.0881
HuBERT-Base	74.60	85.15	88.72	0.0104	0.1041
WavLM-Base	75.10	85.27	87.29	0.0071	0.0869
CLAP	40.93	58.84	73.43	0.1288	0.3870
Emotion2Vec	**75.77**	**85.99**	**91.69**	**0.0051**	**0.0723**

Table 4: Comparison of alternative embedding choices for NV type prediction within the top-K NV matching framework.

B Comparison with Alternative Embedding Choices for NV Matching

To broaden the discussion of embedding choices for NV modeling, we conducted additional experiments under the NV type prediction setting described in Section 6.4 by replacing Emotion2Vec in the top-K NV matching module while keeping all other components unchanged. We considered the following alternative embedding models:

- **Wav2Vec2-Base**[2] (Baevski et al., 2020): a self-supervised speech representation model trained to capture general acoustic and phonetic structure from raw audio.
- **HuBERT-Base**[3] (Hsu et al., 2021): a self-supervised speech model that learns hidden-unit representations from masked prediction over clustered acoustic targets.
- **WavLM-Base**[4] (Chen et al., 2022): a speech representation model designed to encode both content and speaker-related characteristics for speech processing tasks.
- **CLAP**[5] (Elizalde et al., 2023): a joint audio–text embedding model trained for cross-modal alignment between audio signals and natural language descriptions.

For the self-supervised speech models, NV matching was performed based on similarity between utterance and NV audio embeddings. For CLAP, NV type prediction was performed by comparing audio embeddings with textual NV label embeddings in the shared cross-modal space.

Table 4 summarizes the results. Prosody-oriented speech representations achieved competitive performance, indicating that general speech features are helpful for NV matching. However, Emotion2Vec achieved the best overall results, including the highest Acc@1/3/5 and the lowest JSD and HD, suggesting more stable and distributionally consistent NV matching. CLAP showed substantially lower performance, possibly because its representation space is optimized for cross-modal retrieval rather than fine-grained affective alignment between speech and NV categories. Overall, these findings support the use of affect-oriented embeddings as a practical choice for emotion-aware NV matching.

[2] https://huggingface.co/facebook/wav2vec2-base

[3] https://huggingface.co/facebook/hubert-base-ls960

[4] https://huggingface.co/microsoft/wavlm-base

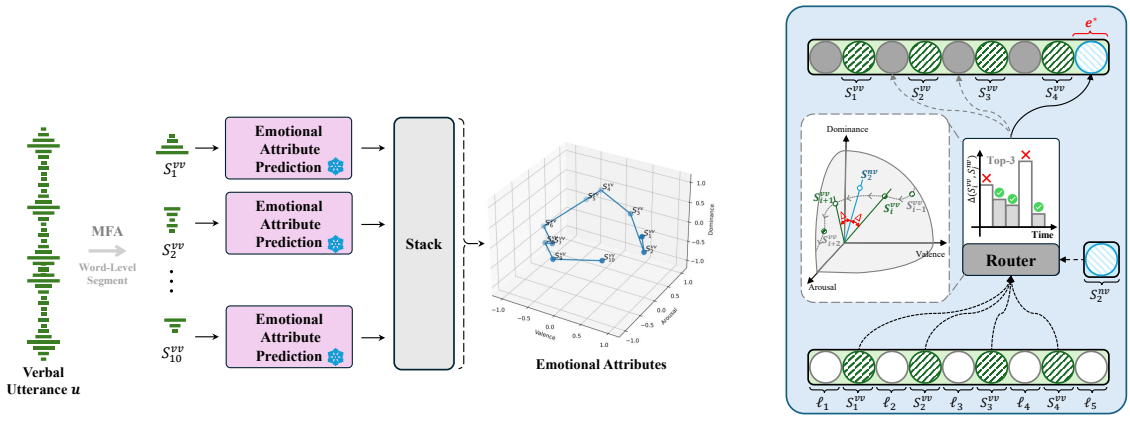
[5] https://huggingface.co/laion/clap-htsat-unfused

NV Category	Angry	Disgusted	Fearful	Happy	Neutral	Sad	Surprised
Agreement	236	30	3	**297**	45	20	148
Anger	**571**	98	4	115	44	36	80
Congratulations	339	150	8	429	61	50	**470**
Filler	72	89	–	91	**185**	56	144
Greetings	151	30	2	254	27	15	**313**
Cheering	32	16	11	**181**	7	10	83
Crying	–	26	3	122	3	**138**	54
Laughter	22	32	11	**549**	22	110	65
Screaming	22	9	**115**	38	2	5	31
Yelling	113	16	13	**166**	4	2	111
Coughing	1	**253**	3	39	5	14	21
Eating	–	**18**	–	5	2	1	5
Sneezing	61	**163**	10	18	–	8	161
Throat	16	**267**	3	39	5	14	21
Yawning	4	15	–	10	1	9	**43**

Table 5: Distribution of Emotion2Vec-based pseudo-label predictions across NV categories in the EARS dataset.

C Analysis of Pseudo-Label Distributions and NV–Emotion Pairing Validity

To further examine the reliability of pseudo-labels used in Affectron, we analyzed the distribution of Emotion2Vec pseudo-label across NV categories in the EARS dataset. Table 5 shows that most NV categories are not associated with a single discrete emotion, but instead exhibit broad affective distributions. For example, laughter is most frequently mapped to happy, but also appears with sad and surprised predictions, which is consistent with context-dependent forms such as nervous or hollow laughter. These observations help explain why some augmented NV–emotion pairings may appear non-intuitive when viewed only at the categorical level. Affectron does not construct augmented samples based on fixed emotion–NV rules, but instead relies on embedding-level affective alignment that captures more nuanced contextual compatibility. The validity of this strategy is further supported by the strong NV type prediction and diversity results reported in Sections 6.3 and 6.4, suggesting that the proposed augmentation preserves meaningful NV variation beyond dominant emotion categories.



(a) Word-level emotional attribute extraction pipeline

(b) Affective distance computation using spherical coordinates

Figure 7: Overview of the emotion-aware top-K routing process. (a) Word-level emotional attributes are extracted from verbal speech using MFA and an emotional attribute predictor. (b) Affective distance computation between an NV candidate and verbal segments via angular distance in spherical coordinates.

D Detailed Description of Emotion-Aware Top-K Routing

We provided a detailed description of the emotion-aware top-K routing procedure introduced in Section 4.2. As illustrated in Figure 7 (a), word-level boundaries were first obtained using the Montreal Forced Aligner[6] (McAuliffe et al., 2017). Word-level emotional attribute pseudo-labels for each verbal segment and each NV candidate were extracted using a pre-trained emotional attribute predictor[7] (Wagner et al., 2023). Pseudo-labels were generated in the arousal–valence–dominance space, with each dimension ranging approximately from 0 to 1 in Cartesian coordinates. First, the Cartesian coordinates were translated so that the neutral emotion center M was positioned at the origin. Specifically, each emotional attribute pseudo-label e was translated as follows:

$$e' = e - M, \quad \text{where} \quad M = \frac{1}{N_n} \sum_{i=1}^{N_n} e_i^n, \quad (10)$$

where N_n denotes the total number of neutral emotion pseudo-labels e_i^n. The translated pseudo-labels $e' = (A, V, D)$, where A denotes arousal, V valence, and D dominance, were then transformed into spherical coordinates (r, θ, ϕ) (Cho et al., 2024, 2025b,c) as follows:

$$r = \sqrt{A^2 + V^2 + D^2}, \quad (11)$$

$$\theta = \arccos\left(\frac{D}{r}\right), \quad \phi = \arctan\left(\frac{V}{A}\right), \quad (12)$$

where θ and ϕ correspond to the elevation and azimuth angles, which are used to compute affective distances during the emotion-aware top-K routing procedure.

As shown in Figure 7 (b), the affective difference between an NV candidate S_j^{nv} and each verbal segment S_i^{vv} was computed using the angular distance on the unit sphere. These angular distances were then aggregated to obtain an affective distance $\mathrm{d}(S_j^{nv}, \ell_t)$ for each candidate insertion index ℓ_t. Based on these distances, the top-K locations with the smallest values were selected. Finally, a temperature-scaled softmax (Shazeer et al., 2017; Fan et al., 2018; Fedus et al., 2022) over the negative distances defined the routing distribution $p(\ell_k | S_j^{nv})$, from which the insertion index ℓ_j^* was sampled for each NV candidate.

E Detailed Description of Evaluation Metrics

E.1 Subjective Evaluation

For subjective evaluation, we conducted AB preference tests and mean opinion score (MOS) evaluations. In AB preference tests, listeners compared pairs of utterances in which different augmentation strategies were applied to the same ground-truth (GT) speech and selected the sample exhibiting more natural and contextually appropriate NVs. For evaluation data construction, 100 utterances were randomly sampled from the test set. For MOS evaluation, we focused on two perceptual aspects of NV expressiveness: naturalness and correctness of the realized NV type (NTN-MOS), and emotional congruence of the NV with the surrounding

[6] https://montreal-forced-aligner.readthedocs.io

[7] https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim

verbal context (NEC-MOS). For evaluation data construction, a total of 150 utterances were randomly sampled from the test set, comprising 100 samples from seen speakers and 50 from unseen speakers. All MOS scores were collected using a five-point scale and were reported with 95% confidence intervals.

We utilized crowdsourcing for these evaluations via Amazon Mechanical Turk[8]. A total of 20 native English speakers residing in the United States participated in each evaluation. AB preference tests and MOS-based evaluations were conducted with 20 participants each, at total costs of 60 USD and 180 USD, respectively. To ensure evaluator reliability, only workers with an approval rate of at least 98% and more than 90 approved HITs were permitted to participate. Attention checks were implemented by inserting fake samples and verifying whether participants rated them appropriately. Evaluations from listeners who failed the attention check or spent less than half of the audio duration on the task were excluded from analysis. These measures were implemented to filter out inattentive listeners and ensure the integrity of the collected ratings. Further details on listener requirements and the evaluation interfaces for AB preference tests, NEC-MOS, and NTN-MOS are provided in Figures 11, 12, and 13.

E.2 Objective Evaluation

For objective evaluation, we employed several metrics to quantify the similarity and consistency between reference and generated speech in both verbal and nonverbal domains. Experiments were conducted under two synthesis conditions: verbal-only speech synthesis and speech synthesis with NVs. Verbal metrics were computed using samples generated under the verbal-only condition, whereas nonverbal metrics were evaluated using samples synthesized with NVs.

Speaker similarity was measured using verbal and nonverbal speaker embedding cosine similarity (V/NV-SECS) using WavLM-base for speaker verification[9], where cosine similarity is computed between the reference verbal speech and the generated verbal speech or NVs. Emotional expressiveness was assessed using verbal and nonverbal emotion embedding cosine similarity (V/NV-EECS) with Emotion2Vec[10] (Ma et al., 2024), which computed emotion embeddings between the reference and generated speech, following (Oh et al., 2025). Nonverbal classification accuracy (NV-Acc) was further evaluated using an internally trained classifier on the EARS dataset (Richter et al., 2024), following the architecture of a categorical emotion recognition model (Goncalves et al., 2024). Nonverbal similarity (NV-Sim) was computed as the cosine similarity between the generated NV and the GT NV selected via emotion-driven matching, using Emotion2Vec embeddings. Linguistic consistency was assessed by calculating the word error rate (WER) with the Whisper large model (Radford et al., 2023).

Finally, we employed two complementary metrics to evaluate the accuracy of NV type and location prediction. Top-K accuracy (Acc@K) was reported, measuring whether the GT NV type or location is included among the top-K predictions. JSD (Lin, 2002), and HD (Beran, 1977) were also computed between the GT and predicted distributions of NV types and locations to quantify how closely the model reproduces the occurrence patterns observed in real data.

F Comparison with Zero-Shot TTS Models Trained on NV-Labeled Data

Recent zero-shot TTS models have demonstrated the capability to generate NVs. However, direct comparison with existing NV-capable systems remains challenging due to differences in data availability and annotation protocols. To provide a supplementary point of reference, we additionally compared our method with publicly released NV-capable zero-shot TTS models trained on large-scale NV-labeled corpora. Specifically, we considered the following NV-capable systems:

- **CosyVoice2-0.5B**[11] (Du et al., 2024): an improved streaming neural codec language model (NCLM) for zero-shot TTS, trained on large-scale multilingual speech data with NV annotations. The model supports 12 predefined NV categories.
- **Fun-CosyVoice3-0.5B**[12] (Du et al., 2025): an improved NCLM for zero-shot multilingual

[8] https://www.mturk.com/

[9] https://huggingface.co/microsoft/wavlm-base-sv

[10] https://github.com/ddlBoJack/emotion2vec

[11] https://huggingface.co/FunAudioLLM/CosyVoice2-0.5B

[12] https://huggingface.co/FunAudioLLM/Fun-CosyVoice3-0.5B-2512

Method	Nonverbal Metrics				Verbal Metrics		
	NV-Acc (↑)	NV-Sim (↑)	EECS (↑)	SECS (↑)	WER (↓)	EECS (↑)	SECS (↑)
Augmented GT	89.29	-	0.5487	0.9092	2.93	0.6733	0.8995
VoiceCraft-330M♠ (Peng et al., 2024)	-	-	-	-	8.89	0.5410	0.8483
CosyVoice2-0.5B♠ (Du et al., 2024)	25.00	0.4097	0.4991	0.8751	1.97	0.4876	0.8739
Fun-CosyVoice3-0.5B♠ (Du et al., 2025)	27.38	0.4222	0.4792	**0.8952**	**1.65**	0.4904	**0.8916**
Dia-1.6B♠ (Nari-labs, 2025)	13.10	0.4894	0.4549	0.8484	11.5	0.5204	0.8690
VoiceCraft-330M♣ (Peng et al., 2024)	11.90	0.4766	0.5479	0.8755	10.5	0.5582	0.8690
Affectron-330M♣ (Proposed)	**36.90**	**0.5427**	**0.5506**	0.8686	8.31	**0.5591**	0.8630

Table 6: Comparison of NV-capable zero-shot TTS models on unseen speakers in the EARS dataset (Richter et al., 2024) zero-shot evaluation setting. ♠ denotes a pre-trained model using the official implementation without any training or fine-tuning on the EARS dataset. ♣ denotes a model fine-tuned on the EARS dataset based on the official pre-trained checkpoint.

Method	Nonverbal Metrics				Verbal Metrics		
	NV-Acc (↑)	NV-Sim (↑)	EECS (↑)	SECS (↑)	WER (↓)	EECS (↑)	SECS (↑)
Augmented GT	89.29	-	0.5487	0.9092	2.93	0.6733	0.8995
Cross-speaker NV mixing	14.29	**0.4939**	0.4731	0.8571	9.40	**0.5520**	0.8604
Same-speaker NV mixing	**16.67**	0.4819	**0.4979**	**0.8675**	**8.89**	0.5518	**0.8660**

Table 7: Ablation study on cross-speaker and same-speaker NV mixing in unseen-speaker zero-shot TTS synthesis.

TTS, extending CosyVoice2 with enhanced data scale and modeling capacity. The model supports 12 predefined NV categories.

- **Dia-1.6B**[13] (Nari-labs, 2025): a zero-shot TTS system trained on a large NV-labeled corpus that directly generates expressive speech from transcripts. Dia can produce a broad inventory of 21 predefined NV categories.

All compared models were evaluated in a zero-shot setting using their publicly available pre-trained checkpoints. All generated audio was resampled to a uniform 16 kHz sampling rate prior to evaluation to ensure a fair comparison. Because the supported NV tag categories did not exactly match those used in this evaluation, semantically compatible NV tags were mapped to a shared representation (e.g., ⟨coughing⟩ to ⟨cough⟩, ⟨screaming⟩ to ⟨noise⟩). Accordingly, the category set used for NV classification was adapted for each model based on its supported NV inventory.

As shown in Table 6, existing NV-capable zero-shot TTS models demonstrated complementary strengths across both verbal and nonverbal dimensions. CosyVoice2 and CosyVoice3 achieved strong performance on linguistic and speaker-related metrics, reflecting the advantages of large-scale supervised training and robust acoustic modeling. However, their performance on NV-related metrics was comparatively limited, suggesting that NV generation was not explicitly optimized, despite NV annotations being available during training. Dia trained on a broader inventory of NV categories, demonstrated relatively higher nonverbal similarity compared to CosyVoice-based models. Nevertheless, its NV accuracy and emotion consistency scores were comparatively lower. In contrast, Affectron outperformed baseline systems on NV-related metrics while maintaining competitive verbal quality. These findings indicate that, beyond data scale, explicitly modeling affective alignment for NV selection and placement plays a critical role in achieving coherent and expressive NV synthesis.

G Analysis of Speaker Entanglement and Cross-Speaker NV Mixing

To further examine the concern regarding speaker entanglement, we analyzed the effect of relaxing the same-speaker constraint used in NV augmentation. In Affectron, the same-speaker constraint was introduced to preserve speaker voice coherence and stabilize acoustic transitions between verbal speech and inserted NV segments, rather than to enforce speaker-specific NV patterns. Importantly, the candidate selection itself is still guided by affective similarity in a shared embedding space, indicating that emotional alignment remains the primary criterion during NV matching. This interpretation is also consistent with the results in Table 1, where

[13] https://huggingface.co/nari-labs/Dia-1.6B-0626

Method	K_{EDNM}	K_{EAR}	Nonverbal Metrics				Verbal Metrics		
			NV-Acc (↑)	NV-Sim (↑)	EECS (↑)	SECS (↑)	WER (↓)	EECS (↑)	SECS (↑)
SEEN SPEAKERS									
Augmented GT	-	-	85.96	-	0.5796	0.9231	1.13	0.6186	0.9163
Affectron (Proposed)	1	1	27.92	0.5965	0.5684	0.8887	8.35	0.6148	0.8885
	3	3	32.18	0.5961	0.5673	0.8891	6.59	0.6216	0.8886
	5	5	36.16	0.5970	0.5656	0.8905	6.19	0.6155	0.8878
	10	5	37.75	0.6118	0.5748	0.8906	6.59	0.6216	0.8886
	10	10	42.48	0.6084	0.5620	0.8900	6.19	0.6153	0.8899
	15	10	45.95	0.6183	0.5611	0.8936	5.99	0.6117	0.8936
	15	15	42.80	0.6106	0.5521	0.8921	7.06	0.6073	0.8898
UNSEEN SPEAKERS									
Augmented GT	-	-	89.29	-	0.5487	0.9092	2.93	0.6733	0.8995
Affectron (Proposed)	1	1	27.38	0.5067	0.5261	0.8662	11.52	0.5683	0.8651
	3	3	23.81	0.5041	0.5182	0.8699	9.41	0.5669	0.8690
	5	5	33.33	0.5132	0.5168	0.8635	9.30	0.5449	0.8688
	10	5	36.90	0.5427	0.5506	0.8686	8.31	0.5591	0.8630
	10	10	34.52	0.5075	0.5342	0.8689	9.97	0.5670	0.8665
	15	10	43.33	0.5224	0.5306	0.8728	13.12	0.5416	0.8665
	15	15	50.00	0.5211	0.5269	0.8729	9.06	0.5443	0.8671

Table 8: Grid search over the top-K hyperparameters for emotion-driven NV matching (EDNM) and emotion-aware routing (EAR). Augmented GT applies our NV augmentation to the ground truth.

Affectron maintains competitive NV-related performance under the unseen-speaker zero-shot setting.

To further validate this point, we conducted an additional ablation study with cross-speaker NV mixing under the same training setup for 10,000 steps and evaluated the model on unseen-speaker zero-shot TTS synthesis. As shown in Table 7, cross-speaker mixing slightly improved NV-Sim, but did not improve affective consistency and resulted in lower speaker similarity and weaker acoustic coherence overall. In contrast, same-speaker NV mixing achieved more balanced performance across NV-related and verbal metrics. These findings suggest that the same-speaker constraint mainly serves as an acoustic stabilization strategy, while affect-driven matching remains the key mechanism enabling robust NV generalization beyond seen speakers.

H Ablation Analysis of Top-K Hyperparameters for NV Matching and Routing

Top-K selection was implemented to balance affective consistency and diversity by enabling controlled probabilistic sampling. We analyzed the sensitivity of Affectron to the top-K hyperparameters used in emotion-driven NV matching (EDNM) and emotion-aware routing (EAR). Specifically, K_{EDNM} and K_{EAR} were varied from 1 to 15, with each configuration trained and evaluated under otherwise identical experimental settings.

As shown in Table 8, setting K to a small value resulted in both NV matching and routing becoming largely deterministic. Such restrictive configurations limited probabilistic sampling, which reduced NV diversity and decreased the overall naturalness of generated speech. Conversely, excessively large K_{EDNM} values expanded the candidate set, increasing NV diversity but reducing nonverbal similarity. Similarly, overly large K_{EAR} values weakened the emotional coherence of NV placement, resulting in degraded emotional expressiveness. Overall, the results indicate that intermediate top-K settings offer a beneficial trade-off between NV diversity and affective coherence. Specifically, the configuration with $K_{\text{EDNM}} = 10$ and $K_{\text{EAR}} = 5$ consistently achieved balanced performance across both nonverbal and verbal metrics. Accordingly, this setting was adopted as a practical operating point for Affectron in all subsequent experiments.

I Filler NV Variations Analysis

To supplement the analysis of filler diversity, a visualization of the distribution of generated fine-grained filler NV variations was provided. As shown in Figure 8, word-cloud representations that visualize the realized filler variants produced by both the VoiceCraft (Peng et al., 2024) and the proposed model under identical filler tag ⟨filler⟩

(a) VoiceCraft **(b) Affectron**

Figure 8: Word-cloud visualization of generated fine-grained filler variations for VoiceCraft (Peng et al., 2024) and Affectron. The size of each keyword corresponds to its relative frequency among the realized filler variants.

conditions. Within the word-cloud visualization, the size of each keyword corresponds to its relative frequency among the generated filler realizations. This representation highlights both the number of distinct filler variations and their distributional balance and usage patterns.

VoiceCraft demonstrates a strong concentration on a limited set of common filler types. This pattern indicates repetitive generation behavior and limited expressive diversity. In contrast, Affectron generates a broader range of filler variants with a more balanced frequency distribution, including less frequent yet contextually appropriate realizations. Collectively, these results suggest that the proposed NV-augmented training framework encourages richer and more diverse NV expression beyond dominant filler variants.

J LLM Baselines and Prompting Strategy

Building on previous research regarding NV-aligned emotional corpora (Xin et al., 2024a) and filler placement (Kim et al., 2025), we followed large language model (LLM) baselines and a prompting strategy. LLMs can partially capture discourse-level regularities associated with NV events, even in the absence of explicit acoustic supervision. For baseline comparisons with the proposed method, four widely adopted LLMs were selected to represent a range of sizes, architectures, and conversational capabilities:

- **Qwen 2.5-7B**[14] (Hui et al., 2024): an instruction-tuned multilingual model with robust instruction-following capabilities.
- **LLaMA 3.1-8B**[15] (Grattafiori et al., 2024): a Transformer-based autoregressive language model designed for general-purpose text generation.
- **Vicuna-7B**[16] (Chiang et al., 2023): a dialogue-centric model fine-tuned using large-scale human conversational data.
- **GPT-oss-20B**[17] (Agarwal et al., 2025): a larger open-weight model with advanced long-context reasoning capabilities.

In all cases, a unified prompting scheme was employed for zero-shot NV type and location prediction, as illustrated in Figure 9. For type predic-

Prompt Template for NV Type Prediction

Target Sentence: *<TGT_SEN>*
Candidate Labels: {snoring, groaning, sighing, coughing, laughter, grunting, sniffing, sneezing}
Prediction count (K): *<K_Num>*
USER:
Select the TOP-*<K_Num>* most natural nonverbal labels that would plausibly co-occur with the target sentence *<TGT_SEN>* . You MUST choose only from Candidate labels. Return ONLY valid JSON: {"predicted_types": ["label1", ...]}.
ASSISTANT:

Prompt Template for NV Position Prediction

Target Sentence: *<TGT_SEN>*
Nonverbal Label: *<NV_LABEL>*
Prediction count (K): *<K_Num>*
USER:
Choose the TOP-*<K_Num>* most natural gap positions for inserting *<NV_LABEL>* into the target sentence *<TGT_SEN>*. Return ONLY valid JSON: {"predicted_positions": [int, ...]} with unique integers.
ASSISTANT:

Figure 9: Sample prompt templates for NV type and location prediction.

[14] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[15] https://huggingface.co/meta-llama/Llama-3.1-8B
[16] https://huggingface.co/lmsys/vicuna-7b-v1.5
[17] https://huggingface.co/openai/gpt-oss-20b

Method	Acc@1 (↑)		Acc@3 (↑)		Acc@5 (↑)		JSD (↓)		HD (↓)	
	Location	Type	Location	Type	Location	Type	Location	Type	Location	Type
Qwen2-Audio-7B (Text only) (Chu et al., 2024)	6.68	12.00	16.39	53.56	18.68	76.96	0.2862	0.4346	0.5704	0.7150
Qwen2-Audio-7B (Text and Speech) (Chu et al., 2024)	7.36	12.83	18.05	57.60	18.41	80.17	0.2353	0.4338	0.5188	0.7023
Qwen2.5-Omni-7B (Text only) (Xu et al., 2025)	6.53	12.47	12.11	43.94	27.32	69.90	0.2213	0.3334	0.4967	0.6138
Qwen2.5-Omni-7B (Text and Speech) (Xu et al., 2025)	7.36	20.07	12.23	56.41	28.86	75.30	0.2155	0.3307	0.4961	0.6096
Affectron-330M (Proposed)	**12.59**	**75.77**	**29.69**	**85.99**	**44.06**	**91.69**	**0.0523**	**0.0051**	**0.2414**	**0.0723**

Table 9: Comparison of zero-shot NV type and location prediction across LLMs and the Affectron model. Prediction accuracy is measured using top-K accuracy (Acc@1/3/5) and distributional alignment is assessed using Jensen-Shannon Divergence (JSD) and Hellinger Distance (HD).

tion, each model received the transcript of a target utterance and a fixed list of NV categories, and was instructed to select the most appropriate NV type based solely on textual context. For location prediction, the utterance was presented as a sequence of word tokens indexed by their locations, and the model was prompted to select the most suitable insertion index for a single NV event. This experimental setup enabled a direct and controlled comparison between text-only LLM baselines and the proposed framework.

K Comparison with Multimodal Audio-LLM Baselines on NV Type and Location Prediction

The comparison in Section 6.4 was originally designed to evaluate Affectron against prior LLM-based augmentation strategies, which are predominantly text-driven (Xin et al., 2024a; Kim et al., 2025). In that setup, text-only LLMs were prompted to predict NV type and insertion location from transcripts alone, whereas Affectron used speech-conditioned affective representations derived from the utterance and NV candidates. To address modality fairness, we additionally evaluated multimodal audio-capable LLMs under both text-only and text-and-speech input settings. For the multimodal fairness comparison, we included two open-source audio-capable LLMs:

- **Qwen2-Audio-7B**[18] (Chu et al., 2024): an instruction-tuned multimodal audio language model that supports both spoken audio and text inputs for speech understanding and generation tasks.
- **Qwen2.5-Omni-7B**[19] (Xu et al., 2025): an open-source omni-modal language model designed to process text and audio inputs in a unified framework, with improved multimodal reasoning and speech interaction capabilities.

Table 9 summarize the results for NV location and NV type prediction, respectively. Incorporating audio inputs yielded only modest improvements over text-only settings for both multimodal LLMs. In contrast, Affectron consistently outperformed all multimodal baselines by a large margin across both top-K accuracy and distributional alignment metrics. These results suggest that the gains of Affectron cannot be explained solely by access to acoustic cues, but instead arise from the explicit affect-aware modeling of NV selection and placement. Even when provided with audio input, general-purpose multimodal LLMs showed limited ability to capture fine-grained NV type and location distributions, highlighting the benefit of a dedicated affect-driven framework.

L Potential Risks

Advances in speech synthesis have substantially expanded the capabilities of expressive and natural speech generation. At the same time, such progress introduces the possibility of misuse, particularly in scenarios where highly realistic synthesized speech could be leveraged to create deceptive or misleading audio content. These risks include the generation of fabricated speech for impersonation or misinformation, with potential societal consequences. Addressing such concerns requires complementary safeguards, including synthesized speech detection and watermarking, to support authentication, traceability, and responsible deployment.

M AI Assistance

GPT-5.2 assisted with language polishing, including proofreading and grammatical corrections. The model was not involved in generating technical content, experimental design, or scientific claims.

[18] https://huggingface.co/Qwen/Qwen2-Audio-7B-Instruct

[19] https://huggingface.co/Qwen/Qwen2.5-Omni-7B

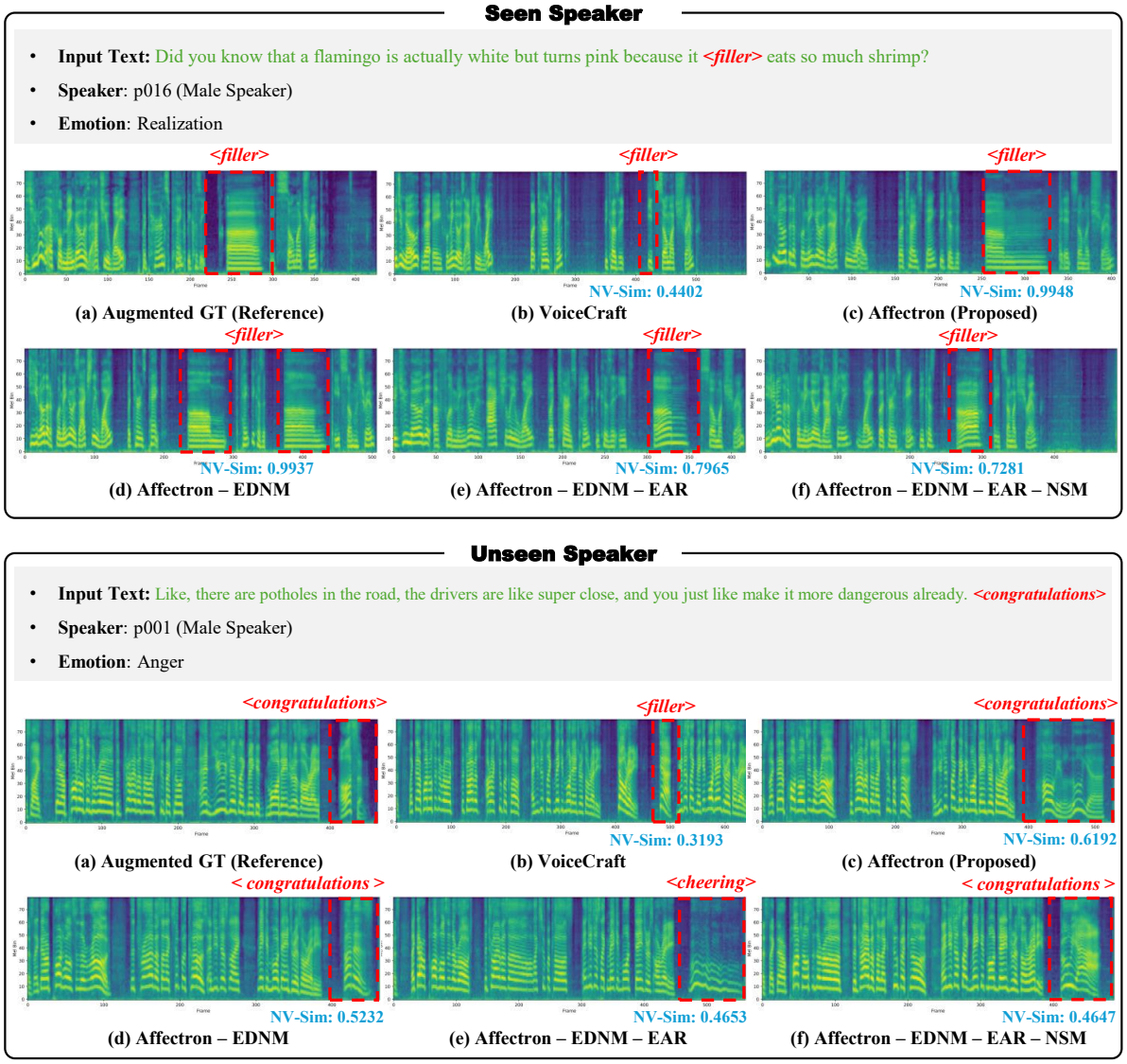


Figure 10: Comparison of Mel-spectrograms generated by different models for the same utterance. Results are shown for a seen speaker and an unseen speaker. NV segments are highlighted in red, and NV-Sim scores are annotated in blue. EDNM, EAR, and NSM denote emotion-driven top-K NV matching, emotion-aware top-K routing, and NV structural masking, respectively. Augmented GT applies our affectively aligned NV augmentation to the ground truth.

N Visualization Analysis of NV Expressive Generation

Figure 10 compares the Mel-spectrograms generated by Affectron, baseline models, and ablated variants under identical text and NV tag conditions. From this comparison, we made the following observations. First, Affectron consistently generated Mel-spectrograms that are well aligned with the given NV tags, accurately realizing NV events at the specified locations. In contrast, several baseline and ablated models either produced NVs that do not correspond to the provided tags or generated NVs at incorrect locations. Second, Affectron generated NV expressions that are emotionally aligned with adjacent verbal segments, resulting in smoother emotional transitions. This behavior was reflected in the higher NV-Sim scores achieved by Affectron. By comparison, most baseline and ablated models failed to maintain a consistent affective context between verbal speech and NVs, resulting in fragmented or incongruent expressive patterns. Overall, these visualizations demonstrated that Affectron generates NVs that follow the given tags and integrate them coherently with the surrounding verbal speech. For a broader set of qualitative examples and audio demonstrations, readers are referred to https://choddeok.github.io/Affectron/.

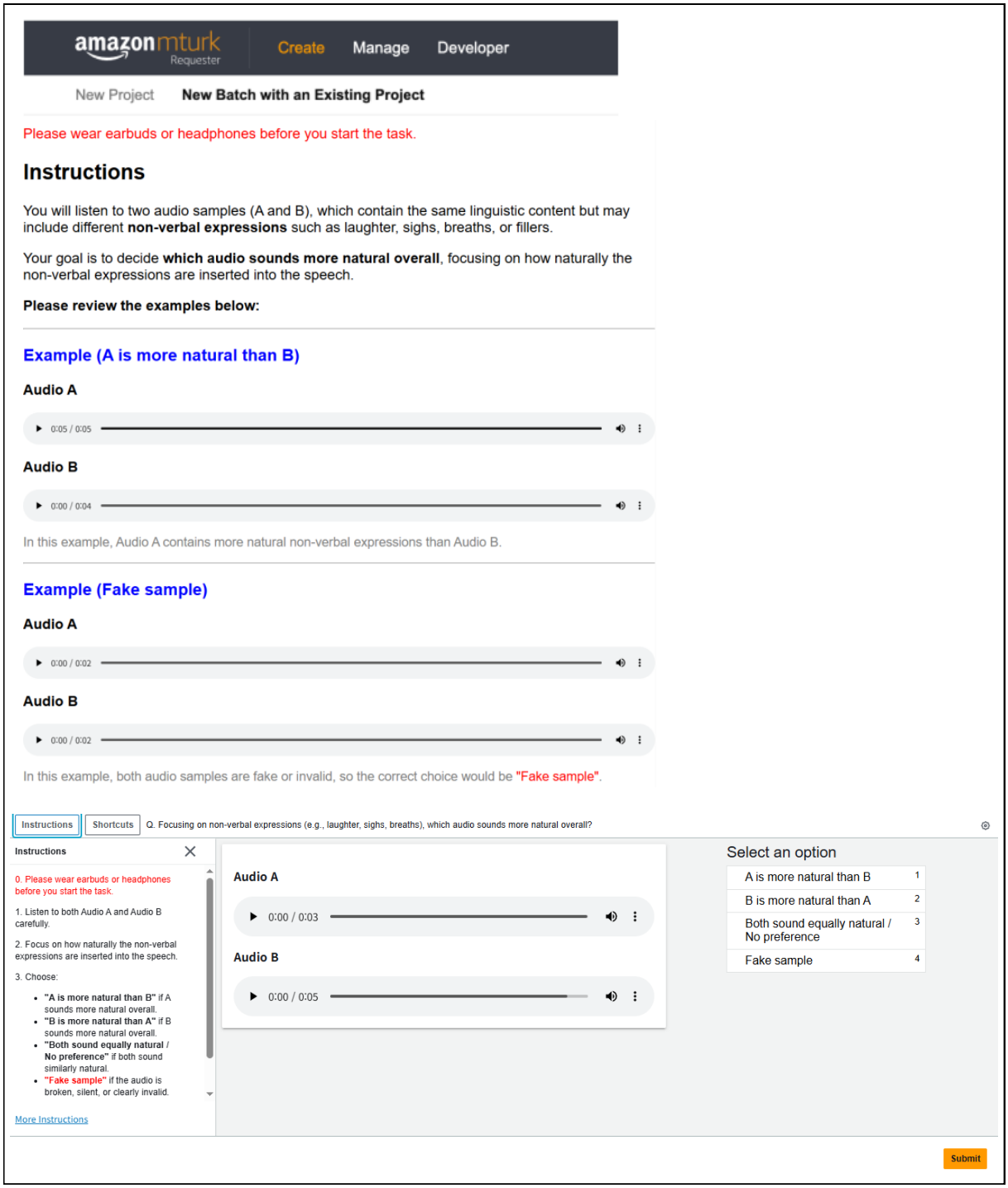


Figure 11: Detailed information on listener requirements and the AB preference test interfaces.

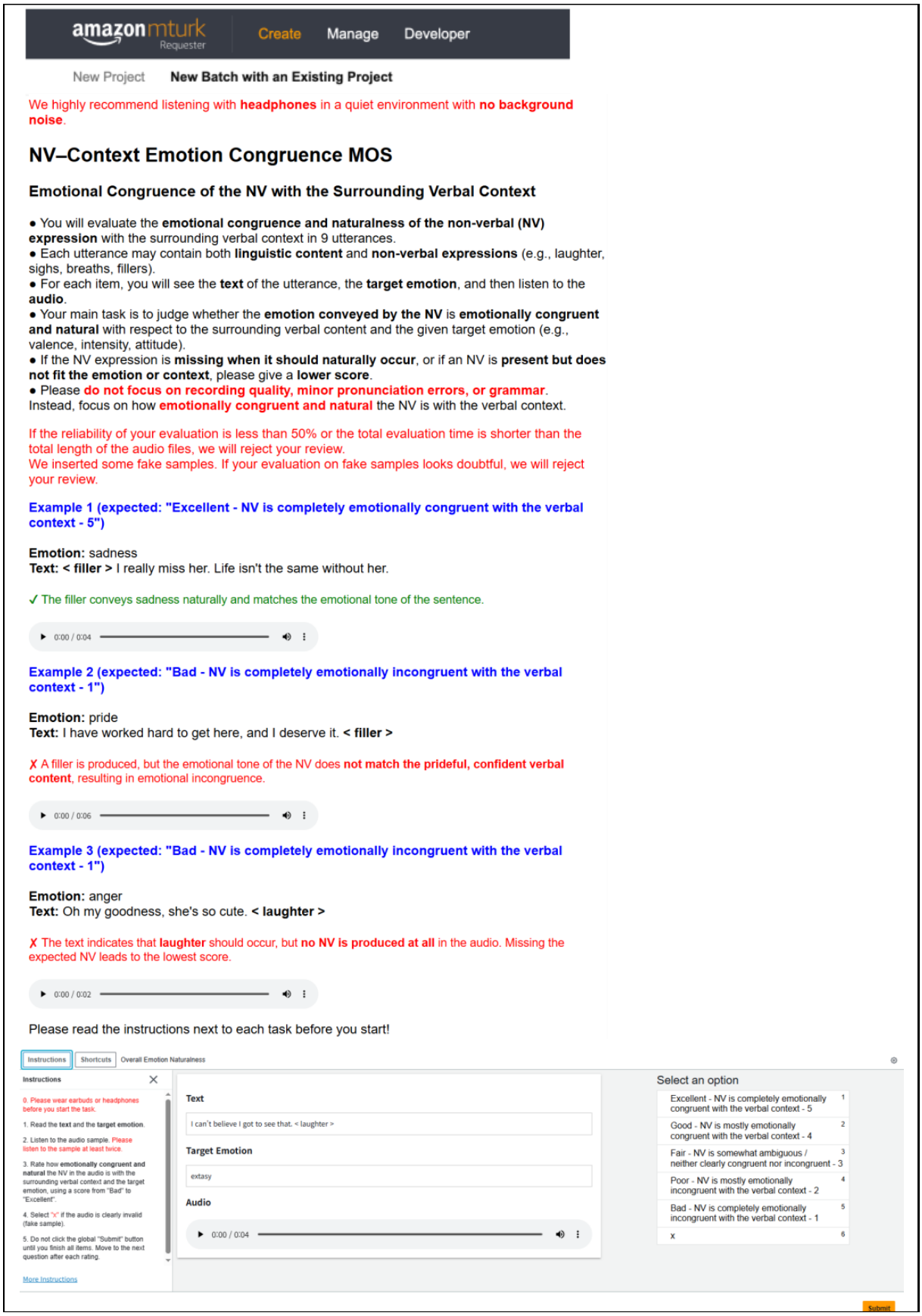


Figure 12: Detailed information on listener requirements and NV-context emotional congruence (NEC-MOS) evaluation interfaces.

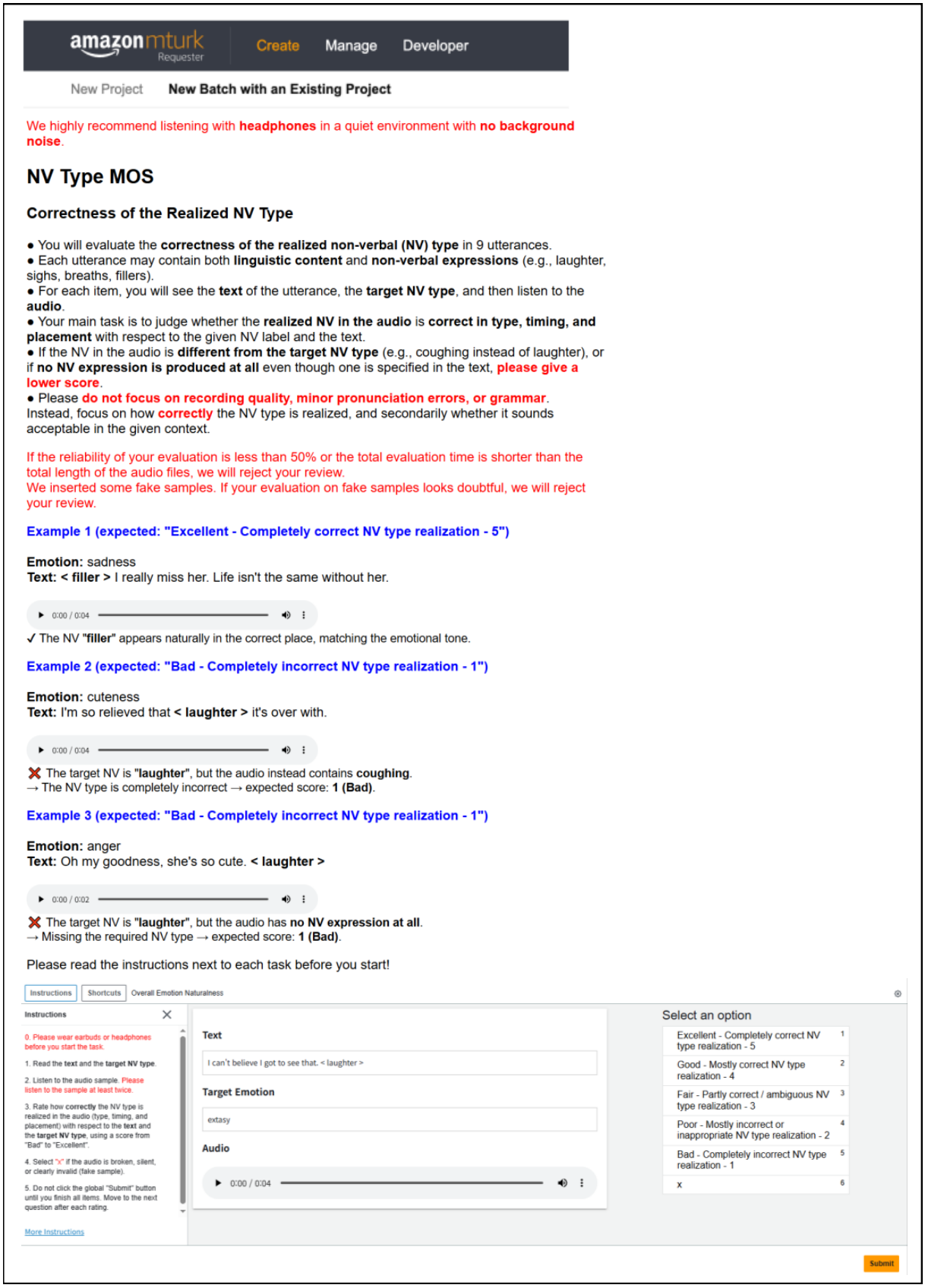


Figure 13: Detailed information on listener requirements and NV-type naturalness (NTN-MOS) evaluation interfaces.