
UDDETTS: Unifying Discrete and Dimensional Emotions for Controllable Emotional Text-to-Speech

Jiaxuan Liu, Zhenhua Ling*

NERCSLIP, University of Science and Technology of China, Hefei, China
jxliu@mail.ustc.edu.cn, zhling@ustc.edu.cn

Abstract

Recent neural codec language models have made great progress in the field of text-to-speech (TTS), but controllable emotional TTS still faces many challenges. Traditional methods rely on predefined discrete emotion labels to control emotion categories and intensities, which can't capture the complexity and continuity of human emotional perception and expression. The lack of large-scale emotional speech datasets with balanced emotion distributions and fine-grained emotion annotations often causes overfitting in synthesis models and impedes effective emotion control. To address these issues, we propose UDDETTS, a neural codec language model unifying discrete and dimensional emotions for controllable emotional TTS. This model introduces the interpretable Arousal-Dominance-Valence (ADV) space for dimensional emotion description and supports emotion control driven by either discrete emotion labels or nonlinearly quantified ADV values. Furthermore, a semi-supervised training strategy is designed to comprehensively utilize diverse speech datasets with different types of emotion annotations to train the UDDETTS. Experiments show that UDDETTS achieves linear emotion control along the three dimensions of ADV space, and exhibits superior end-to-end emotional speech synthesis capabilities.

1 Introduction

Recently, a large number of neural codec language models (LMs) [9, 61, 4, 62, 60, 57, 13, 14, 2, 3, 15, 16] with high comprehension have emerged and heralded a new epoch in the field of TTS. These TTS models generate speech semantic tokens from text tokens by predicting the next token in a sequence, and demonstrate significant advantages in synthesizing expressive speech. In the field of human-computer interaction, enhancing speech expressiveness has become increasingly necessary, with emotional TTS as a core element. Currently, emotional LM-based TTS methods [57, 13, 14, 3] primarily rely on emotion prompts for supervised fine-tuning. They simplify emotional expression by mapping emotions into predefined discrete categories such as *happy*, *sad*, *angry*, etc. Although some prompts contain rich information such as emotion, timbre, age, and prosody, emotional control is still fundamentally constrained by the discrete labels in the dataset. Due to the limited variety of labels, this approach generates speech emotions with average expressions per category, failing to capture the inherent complexity and continuity of human emotions. In reality, emotions exist as a highly interconnected continuum in a continuous space rather than isolated categories [22]. Addressing this limitation requires developing continuous emotion modeling mechanisms in LM-based TTS [23, 7] to better capture subtle emotional variations.

With the development of emotion analysis research, dimensional emotion theory [47, 52, 12, 40, 18] provides a more refined and comprehensive framework, enhancing understanding the complexity

*Corresponding author.

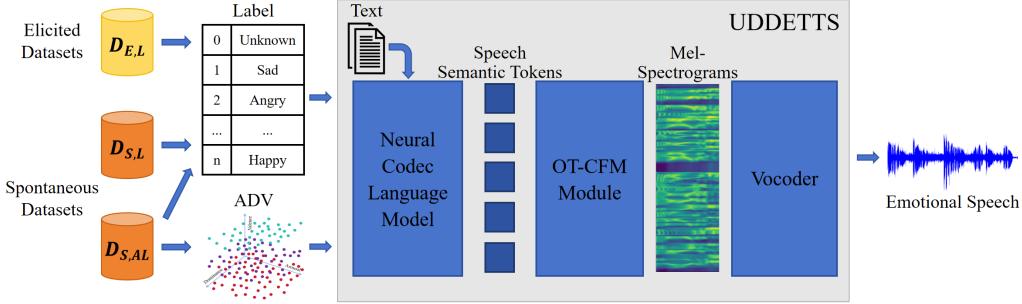


Figure 1: The overview of UDDETTS. UDDETTS integrates both discrete label and dimensional ADV annotations to enable controllable emotional TTS.

and continuity of emotions. Arousal-Dominance-Valence (ADV) space [40] is a commonly three-dimensional framework for describing emotions. Arousal represents psychological alertness levels. Low arousal involves being *sleepy* or *bored*, while high arousal involves being *awake* or *excited*. Dominance measures control over others or being controlled, reflecting emotional expression desires. Low dominance involves being *aggrieved* or *weak*, while high dominance involves being *angry* or *amused*. Valence (also known as pleasure) represents the emotional positivity or negativity, such as being *sad* or *angry* as low valence, while being *happy* or *excited* as high valence. These three dimensions account for most variations in 42 emotion scales and cover almost all speech emotion states [40]. Inspired by the strengths of ADV space in decoupling emotions into interpretable and linearly controllable vectors, how to leverage ADV space and diverse emotional speech datasets in LM-based TTS to enhance emotion controllability remains an open challenge.

One key challenge is that emotion distributions in the ADV space are often imbalanced and limited. On one hand, existing speech datasets tend to overrepresent a few dominant or neutral emotions, leading to overfitting during training. On the other hand, due to the high cost of emotion annotation, most large-scale emotional speech datasets only provide discrete emotion labels, while only a few offer both discrete labels and dimensional ADV values. This scarcity of ADV annotations leads to low controllable coverage rate in the ADV space. Previous studies [37, 56, 32] have addressed the issue of label-based emotional imbalance. However, none of these methods have explored solutions within the ADV space. Meanwhile, other studies [38, 46, 50, 31] have employed semi-supervised training in LMs to tackle the challenges of diverse annotations and limited annotation distributions. In particular, J.Luo et al. [38] demonstrates that semi-supervised training enables interaction across diverse annotation types, and effectively propagates knowledge from labeled data to unlabeled data, offering a promising direction for addressing our challenge.

This paper proposes UDDETTS, a unified framework for controllable emotional TTS, comprising a neural codec language model, an optimal-transport conditional flow matching (OT-CFM) module, and a vocoder, as shown in Figure 1. UDDETTS categorizes all datasets into spontaneous emotion datasets and elicited emotion datasets. To address the low controllable coverage rate of the ADV space, it adopts semi-supervised training to accommodate different types of emotional speech datasets, and fuses ADV and label annotations from these datasets. UDDETTS non-linearly quantizes the ADV space into controllable units as ADV tokens, which are combined with a label token for emotion control. Leveraging the deep understanding of the LM, UDDETTS learns the mapping from text tokens and ADV tokens to the label token and speech semantic tokens using spontaneous emotion datasets, and learns how the label token controls speech semantic tokens using elicited emotion datasets. An ADV predictor is introduced to infer the ADV tokens from text tokens for end-to-end emotional TTS when no explicit emotion conditions are provided during inference. UDDETTS employs an emotional mixture encoder to integrate the masked ADV tokens and label token into emotion conditions. The mel-spectrogram generated by the OT-CFM module is then converted into emotional speech using a HiFi-GAN vocoder [27]. We evaluate UDDETTS using objective and subjective metrics, comparing it with the prompt-based CosyVoice model in terms of label-based emotional naturalness and end-to-end emotional speech synthesis. Experiments demonstrate UDDETTS outperforms CosyVoice across diverse scenarios, and exhibits superior emotion control ability based on ADV or label inputs.

In summary, our contributions to the community include:

1. We propose UDDETTS, a unified emotional TTS framework that unifies discrete and dimensional emotions, featuring the first neural codec language model supporting both ADV and label inputs for emotion control.
2. We use nonlinear binning and semi-supervised training to improve the controllable coverage rate of the ADV space, mitigating the imbalance and scarcity of ADV values in large-scale emotional speech datasets, while capturing the relationships between different emotions.
3. Our proposed UDDETTS achieves linear emotion control along three interpretable dimensions, adapts to diverse speech datasets to improve the naturalness of synthesized speech, and exhibits text-adaptive emotion generation capabilities.

2 Related Work

Current emotional TTS models can be divided into three categories of emotion control methods: label-based control, transfer-based control, and space-based control.

Label-based control models learn from discrete emotion categories or intensity levels, allowing the specification of a target emotion during inference. For example, [20, 13, 3, 57] employ prompt-based LMs to synthesize speech with specified emotion labels. While ZET-Speech [26] uses a diffusion model for zero-shot conversion of neutral speech to a target emotional category. To capture nuanced emotions, some label-based models [25, 33] employ hierarchical control conditions across coarse and fine granularities. Others explore relative ranking matrices [66], interpolation [19], or distance-based quantization [24] methods to derive emotion intensity levels and then control speech emotional intensity. However, these label-based methods struggle to capture the complexity of emotion distributions and typically yield only localized and averaged emotional expressions.

Transfer-based control models learn emotional representations from sources such as audio, texts, or facial expressions and predict corresponding emotional representations for target speech signals. [34] proposes an ECSS model with heterogeneous graph-based context modeling to predict the current emotion category and intensity from the audio, texts, emotion labels and intensities of the dialogue. [28] introduces an end-to-end emotion transfer model with less emotion category confusions. [29] proposes a cross-speaker emotional transfer TTS method by decoupling speaker timbre and emotion. UMETTS [30] proposes a unified TTS framework that transfers emotional representations from multimodal emotion prompts. However, these transferred or extracted emotions are embedded in latent spaces, making them difficult to interpret and limiting the effectiveness of manual control.

Space-based control models aim to construct a continuous embedding space and capture relationships between different emotions. For example, contextual emotion labels can be mapped into hyperbolic space [8] to better capture their hierarchical structure. [54, 65, 43] use interpolation of the embedding space to synthesis speech with a mixture of emotions. AffectEcho [55] uses a vector quantized space to model fine-grained variations within the same emotion. But these models are still fundamentally based on discrete labels and decouple emotions in an interpretable way. Recent studies [21, 53, 10, 11] have explored dimensional emotion spaces for more interpretable control. In particular, EmoSphere-TTS [10] and EmoSphere++ [11] adopt the ADV space and apply a Cartesian-spherical transformation to control emotion categories and intensities, using ADV pseudo-labels. However, these pseudo-labels is not sufficiently accurate, and cannot address the imbalance and sparsity of ADV values — e.g., failing to distinguish differences along the dominance dimension between *angry* and *sad*. This motivates the need to combine a LM with effective strategies to better learn the ADV space.

3 UDDETTS

To build a unified framework for emotional LM-based TTS, UDDETTS needs to learn nuanced emotional representations from large-scale emotional speech datasets. UDDETTS categorizes all datasets into spontaneous emotion datasets D_S and elicited emotion datasets D_E , and further divides them based on annotation types into three types: $D_{S,AL}$ (D_S with label and ADV), $D_{S,L}$ (D_S with label and without ADV), and $D_{E,L}$ (D_E with label and without ADV). D_S are recorded in natural scenarios such as conversations, speeches, or performances. In many samples, the emotional representations in speech align with the textual content, enabling the LM to learn meaningful emotional mappings from a text to speech ADV and label annotations. In contrast, D_E are created by asking speakers to express predefined emotions with varying categories and intensities using the

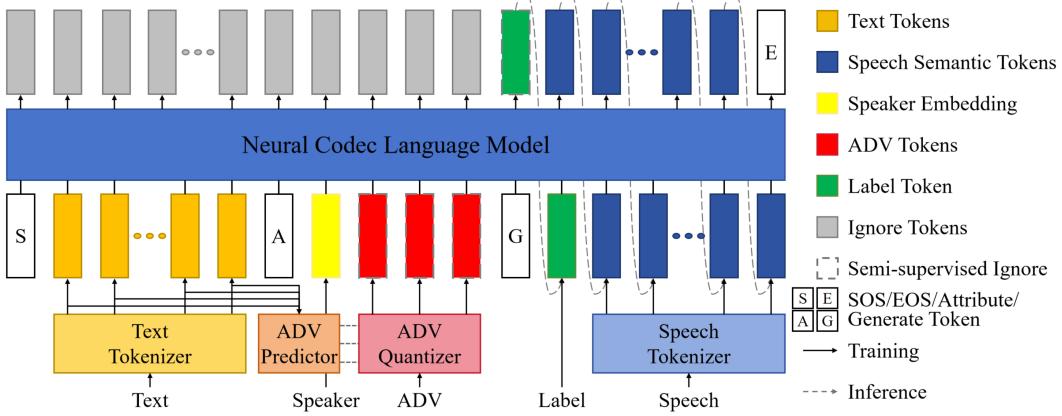


Figure 2: The neural codec language model architecture of UDDETTs operates in an autoregressive manner, predicting next token at a time until the EOS token is generated. The ADV tokens and label token in the input sequence are used for emotion control. During semi-supervised training, they are dynamically masked as ignore tokens depending on the dataset type of each sample.

same text. Here, a single text may correspond to multiple labels that don't match its inherent emotion, making it difficult for the LM to learn emotional mappings from a text to a speech label, and requiring the label to guide the generation of the speech emotional representation.

UDDETTs is designed to control speech emotion using either label or ADV inputs, enabling integration of discrete and dimensional emotion representations. It builds on the CosyVoice model [13] as a scalable baseline, with the core LM and OT-CFM module. Inspired by Spark-TTS [57], UDDETTs separates textual content from speech attribute features, further decoupling speaker timbre from emotional representations within the latter. It quantizes the ADV space using nonlinear binning and employs semi-supervised training in two core architectures to accommodate three dataset types.

3.1 Semi-supervised Neural Codec Language Model

3.1.1 Model Architecture

For the LM as shown in Figure 2, which is based on a Transformer architecture, the construction of input-output sequences is crucial. The LM of UDDETTs adopts semi-supervised training and different input-output sequences for different types of datasets, which are constructed as follows:

$$\begin{aligned} D_{S,AL} : \quad x_{\text{input}} &= [x_{\text{sos}}, x_{\text{text}}, x_{\text{attr}}, x_{\text{spk}}, x_{\text{adv}} \in \mathbb{Z}_{[1,m]}^3, x_{\text{gen}}, x_{\text{lbl}} \in \mathbb{Z}_{[0,n]}^1, x_{\text{sem}}] \\ x_{\text{output}} &= [x_{\text{ign}}, x_{\text{lbl}} \in \mathbb{Z}_{[1,n]}^1, x_{\text{sem}}, x_{\text{eos}}] \end{aligned} \quad (1)$$

$$\begin{aligned} D_{S,L} : \quad x_{\text{input}} &= [x_{\text{sos}}, x_{\text{text}}, x_{\text{attr}}, x_{\text{spk}}, x_{\text{ign}} \in \mathbb{Z}^3, x_{\text{gen}}, x_{\text{lbl}} \in \mathbb{Z}_{[0,n]}^1, x_{\text{sem}}] \\ x_{\text{output}} &= [x_{\text{ign}}, x_{\text{lbl}} \in \mathbb{Z}_{[1,n]}^1, x_{\text{sem}}, x_{\text{eos}}] \end{aligned} \quad (2)$$

$$\begin{aligned} D_{E,L} : \quad x_{\text{input}} &= [x_{\text{sos}}, x_{\text{text}}, x_{\text{attr}}, x_{\text{spk}}, x_{\text{ign}} \in \mathbb{Z}^3, x_{\text{gen}}, x_{\text{lbl}} \in \mathbb{Z}_{[0,n]}^1, x_{\text{sem}}] \\ x_{\text{output}} &= [x_{\text{ign}}, x_{\text{ign}} \in \mathbb{Z}^1, x_{\text{sem}}, x_{\text{eos}}] \end{aligned} \quad (3)$$

where x_{input} and x_{output} are the input sequence and output sequence of the LM. Specifically, x_{sos} , x_{eos} , x_{attr} and x_{gen} represent the start-of-sequence token, end-of-sequence token, attribute-start token, and generation-start token, respectively. All of them are fixed values and belong to \mathbb{Z}^1 . x_{text} is obtained by processing raw text with a Byte Pair Encoding (BPE)-based tokenizer [51]. x_{spk} is the speaker embedding, computed by averaging timbre vectors extracted from all *neutral* emotional speech samples of a speaker using a pre-trained voiceprint model [63]. This embedding captures speaker timbre while excluding emotional representations. x_{adv} is obtained from ADV values $\overrightarrow{\text{adv}}$ using an ADV quantizer based on the nonlinear binning described in Section 3.1.2, and m is the number of bins along each dimension. x_{lbl} is the emotion label token, and n is the number of label token types. In spontaneous emotional datasets D_S , many samples exhibit ambiguous emotional expressions. Therefore, when $x_{\text{lbl}} = 0$ in x_{input} , indicating the label is *Unknown*, the corresponding

x_{lbl} in x_{output} is masked during training. x_{sem} is the speech semantic tokens enriched with emotional representations, extracted with a pre-trained supervised semantic tokenizer [59], which enhances its semantic alignment with both textual and paralinguistic cues. Lastly, x_{ign} is the ignore tokens with a value of -1 , used to mask positions in the x_{output} during training. The LM integrates the input-output sequences defined in Eq. (1), (2), (3) into a unified model. The tokens x_{sos} , x_{atr} , x_{adv} , x_{gen} , x_{lbl} and x_{sem} in x_{input} are fed into the LM through their respective embedding layers. To align speech semantic information, x_{text} is encoded into text embeddings via a Conformer-based text encoder, and x_{spk} is projected to the same hidden dimension as the text embeddings via a linear layer.

3.1.2 Emotion Quantification

In the ADV space, emotions are continuously distributed, and each emotional speech sample can be mapped to a point using the SAM system [41]. For controllability, these continuous representations need to be quantized into tokens $x_{\text{adv}} = [x_a, x_d, x_v] \in \mathbb{Z}_{[1,m]}^3$. However, due to imbalanced emotion distributions and limited ADV values in these speech datasets, the distributions along the three ADV dimensions exhibit approximately normal patterns, and certain regions of the ADV space remain underrepresented, as shown in Figure 6 in Appendix. To mitigate overfitting caused by data imbalance and improve the controllable coverage rate of the ADV space, we design an ADV quantizer by exploring different nonlinear binning algorithms [17] for each of the three dimensions, and finally select the clustering-based binning algorithm to balance uniformity and discriminability. Then, to balance control granularity and coverage rate, the ADV quantizer uses the central limit theorem [49] to determine the number of bins.

We observe that different emotion labels generally form distinct clusters in the ADV space, as shown in Figure 7 in Appendix. However, some labels show substantial overlap, indicating ambiguity in their emotional boundaries. So we unify semantically similar emotion labels in the datasets into a single token. For example, both *happy* and *amused* are grouped under the *happy* category and assigned the same token.

3.1.3 Training and Inference

During training, due to the mixture of datasets, each batch may include samples from multiple sources. For samples with $x_{\text{adv}} \neq [-1, -1, -1]$ in a batch, which belong to $D_{S,AL}$, their corresponding x_{lbl} in x_{output} is not masked. For samples where $x_{\text{adv}} = x_{\text{ign}} = [-1, -1, -1]$, the masking depends on the dataset type: if the sample comes from $D_{S,L}$, x_{lbl} in x_{output} is not masked, but if the sample comes from $D_{E,L}$, x_{lbl} in x_{output} needs to be masked with -1 . We design a label token position-aware smoothing loss function for semi-supervised training, as defined in follow Eq. (4) (5):

$$\mathcal{L}_{LM} = -\frac{1}{L+2} \sum_{l=1}^{L+2} w_{\text{emo}}(l) p(v_l) \log q(v_l), \quad (4)$$

$$\text{where } p(v_l) = \begin{cases} 1 - \epsilon, & \text{if } v_l = \mu_l \\ \frac{\epsilon}{K}, & \text{if } v_l \neq \mu_l \end{cases}, \quad w_{\text{emo}}(l) = \begin{cases} 0, & \text{if } \mu_l = x_{\text{lbl}} = -1 \text{ or } 0 \\ 5.0, & \text{if } \mu_l = x_{\text{lbl}} \neq -1 \text{ or } 0 \\ 1.0, & \text{otherwise} \end{cases}, \quad (5)$$

here, $L + 2$ is the length of $x_{\text{loss}} = [x_{\text{lbl}}, x_{\text{sem}}, x_{\text{eos}}]$ in x_{output} . v_l and μ_l denote the predicted token and the ground-truth token at position l in x_{loss} . $w_{\text{emo}}(l)$ is the position-dependent weighting scale. When the ground-truth value of x_{lbl} in x_{input} is 0 or -1 , indicating that the emotion label is *Unknown* or the sample belongs to $D_{E,L}$ — the loss at x_{lbl} in x_{output} position is masked. Otherwise, the loss at x_{lbl} position is up-weighted to accelerate convergence. $p(v_l)$ is used for label smoothing, where K is the vocabulary size and ϵ is a small smoothing parameter.

During inference, the LM operates in three modes, corresponding to three different tasks:

1. The first task controls emotion categories using a label token: it uses x_{text} and x_{lbl} , with the x_{adv} ignored, to directly generate label-conditioned x_{sem} .
2. The second task controls continuous emotions using ADV tokens: it uses x_{text} and x_{adv} to predict x_{lbl} and then generates x_{sem} autoregressively.
3. The third task predicts text-adaptive emotions directly from texts: it only uses x_{text} to predict x_{sem} , while x_{lbl} and x_{adv} are intermediate tokens generated through self-prediction.

3.1.4 ADV Predictor

We observe that in the third task, directly predicting x_{lbl} and x_{sem} from x_{text} alone performs poorly, often resulting in speech with *neutral* emotion. To address this issue, we introduce an ADV predictor that first estimates $\overrightarrow{\text{adv}}$ from x_{text} . The ADV predictor is inspired by [45, 58] and adopts a pre-trained RoBERTa encoder, followed by a softmax and a sigmoid activation layer over the pooled output. It's trained jointly with the LM and is used to enhance end-to-end emotional speech synthesis. The ADV predictor loss function is defined as:

$$[a_{\text{pred}}, d_{\text{pred}}, v_{\text{pred}}] = \arg \max P(a, d, v | x_{\text{text}}) = \arg \max P(v | x_{\text{text}})P(a | x_{\text{text}})P(d | x_{\text{text}}), \quad (6)$$

$$\mathcal{L}_{\text{ADV}} = \sum_{c \in [a, d, v]} \|c_{\text{pred}} - c_{\text{true}}\|_2^2, \quad (7)$$

where the predicted pseudo-ADV $[a_{\text{pred}}, d_{\text{pred}}, v_{\text{pred}}]$ are then quantized by the ADV quantizer into pseudo-ADV tokens for the input sequences of LM.

3.2 Semi-supervised Conditional Flow Matching

To synthesize emotional speech, UDDETTs reconstructs the speech semantic tokens x_{sem} into mel-spectrograms via an OT-CFM module, which builds upon the CosyVoice's flow module [13]. The OT-CFM module is conditioned on the speaker embedding x_{spk} , the semantic embedding E_{sem} and the emotion conditions E_{emo} . E_{sem} is obtained by encoding the generated x_{sem} via a Conformer-based semantic encoder, while E_{emo} is derived from both x_{adv} and x_{lbl} through an encoder.

To generate E_{emo} , the OT-CFM module employs an emotional mixture encoder, as illustrated in Figure 3. This encoder fuses the masked x_{lbl} and x_{adv} . Specifically, the ADV encoder first encodes x_a , x_d and x_v separately into E_a , E_d and E_v , which are then concatenated and passed through an interaction layer to obtain the ADV embedding E_{adv} . The label encoder directly encodes x_{lbl} into a label embedding E_{lbl} . A multi-head attention layer is applied, using E_{lbl} as the query and E_{adv} as the key and value, resulting in a label-attended emotion embedding $E_{\text{emo}}^{\text{attn}}$. Finally, a gate layer combined with the semi-supervised gating algorithm described in Eq. (8) produces the final emotion conditions E_{emo} .

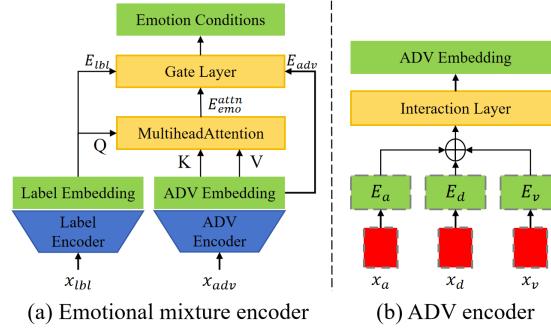


Figure 3: The emotional mixture encoder of OT-CFM module to generate the emotion conditions.

$$E_{\text{emo}} = \begin{cases} E_{\text{adv}} & \text{if } x_{\text{lbl}} = 0 \\ (gate + 1) \cdot E_{\text{lbl}} & \text{if } x_{\text{lbl}} \neq 0 \text{ and } x_{\text{adv}} = [-1, -1, -1] \\ gate \cdot E_{\text{lbl}} + (1 - gate) \cdot E_{\text{emo}}^{\text{attn}} & \text{if } x_{\text{lbl}} \neq 0 \text{ and } x_{\text{adv}} \neq [-1, -1, -1] \end{cases} \quad (8)$$

The OT-CFM module defines a time-dependent vector field $v_t(X) : [0, 1] \times \mathbb{R}^{L \times D} \rightarrow \mathbb{R}^{L \times D}$, and uses an ordinary differential equation [44] to find the optimal-transport (OT) flow ϕ_t^{OT} . All condition, including x_{spk} , E_{sem} and E_{emo} , are fed into a U-net neural network \mathbf{U}_θ to match the vector field $v_t(X)$ to $w_t(X)$ with learnable parameters θ :

$$v_t(\phi_t^{\text{OT}}(X_0, X_1) | \theta) = \mathbf{U}_\theta(\phi_t^{\text{OT}}(X_0, X_1), x_{\text{spk}}, E_{\text{sem}}, E_{\text{emo}}, t), \quad (9)$$

$$w_t(\phi_t^{\text{OT}}(X_0, X_1) | X_1) = X_1 - (1 - \sigma)X_0, \quad (10)$$

where $X_0 \sim \mathcal{N}(0, \tau^{-1}\mathbf{I})$, X_1 is a learned approximation of the mel-spectrogram distributions, t is the timestep using a cosine schedule [42] to prevent rapid noise accumulation from linear addition. The conditional flow matching loss function is shown in Eq. (11):

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{X_0, X_1} \|w_t(\phi_t^{\text{OT}}(X_0, X_1) | X_1) - v_t(\phi_t^{\text{OT}}(X_0, X_1) | \theta)\|_2^2. \quad (11)$$

4 Experiments

4.1 Datasets

To evaluate the UDDETTTS model, we focus on monolingual data and collect a diverse set of English emotional speech datasets, including MSP-PODCAST [36], IEMOCAP [6], MELD [48], ESD [64], EmoV-DB [1], and RAVDESS [35], each annotated with either emotion labels or ADV values. All samples undergo unified preprocessing. We standardize emotion labels, normalize ADV values to [1,7], and remove annotation errors. Speech recordings are resampled to 16 kHz and converted to single-channel format. We remove samples with overlapping speakers, instrumental music, excessive noise, other languages, missing transcriptions, and durations longer than 30 seconds. To reduce speaker timbre confusion, we remove samples from *Unknown* speakers and discard speakers with fewer than four utterances. After cleaning, the final training set contains approximately 300 hours of data. Table 3 in Appendix summarizes the statistics of collected datasets. In total, 13 emotion labels covering five basic emotion categories are used, with corresponding label token types [0, 9] and sample counts listed in Table 4 in Appendix.

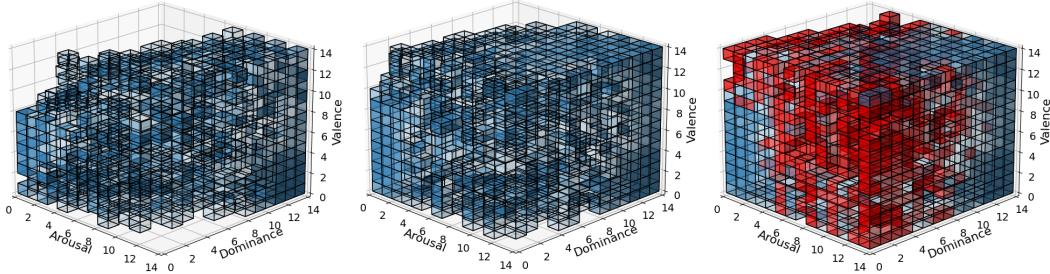


Figure 4: ADV space with $14 \times 14 \times 14$ controllable units: **linear binning** vs. **nonlinear binning** vs. **after semi-supervised training**. Color opacity positively correlates with the number of samples within each controllable unit. Red regions indicate the expanded, reasonably predictable ADV space obtained through semi-supervised training. The coverage rate of ADV space increases from 60.83% to 77.89% and further to 89.35%.

4.2 Implementation Details

We quantize the ADV values $\overrightarrow{adv} \in \mathbb{R}^3$ into controllable units $x_{adv} \in \mathbb{Z}_{[1,14]}^3$ ($m=14$). The statistical results in Figure 4 indicate that nonlinear binning yields more uniformly distributed controllable units compared to linear binning, and increases the ADV space coverage rate from 60.83% to 77.89%.

We adopt CosyVoice-300M as the LM’s backbone and initialize the LM parameters from a pre-trained checkpoint instead of training from scratch. For optimization, we use the Adam optimizer with a learning rate of 1e-4, 2500 warm-up steps, and a gradient accumulation step of 2. The maximum total frame length per batch is set to 5000. We train whole parameters of both the LM and the OT-CFM module for 80 epochs on eight NVIDIA-A800-80GB GPUs paired with 64-core CPUs. The pretrained HiFi-GAN vocoder is fine-tuned for 10 epochs using our datasets. Code and demos are available at <https://anonymous.4open.science/w/UDDETTTS/>.

4.3 Label-Controlled Emotional TTS

We conduct both subjective and objective evaluations to evaluate the performance of speech synthesized via UDDETTTS. As a baseline, we fine-tune the prompt-based CosyVoice-300M using emotion labels as prompt conditions (e.g., "Angry<endofprompt>Content Text"), with the same batch size and epochs as UDDETTTS. We design a corpus of texts including 10 *neutral* texts (see Appendix D). For each *neutral* text, CosyVoice is prompted with 5 emotion labels to synthesize target emotions: *neutral*, *happy*, *angry*, *disgust*, and *sleepiness*. To ensure a fair comparison, UDDETTTS is evaluated under the first task, where emotion control is achieved via the x_{lbl} while ignoring x_{adv} . A total of 10 participants take part in all subjective evaluations. We evaluate the naturalness of the synthesized speech using Mean Opinion Scores (MOS). Emotion control accuracy is evaluated through both

Table 1: Comparison of subjective and objective evaluation results. Acc , P_{macro} , R_{macro} denote Accuracy, macro-Precision, and macro-Recall, respectively. $Subj.$ and $Obj.$ denote subjective and objective evaluations. The w/o EME denotes the OT-CFM module of UDDETTTS without emotional mixture encoder and E_{emo} .

Model	MOS \uparrow	$Acc^{Subj.}$	$P_{macro}^{Subj.}$	$R_{macro}^{Subj.}$	$Acc^{Obj.}$	$P_{macro}^{Obj.}$	$R_{macro}^{Obj.}$
CosyVoice	4.02 ± 0.06	0.79 ± 0.03	0.85 ± 0.03	0.73 ± 0.05	0.60	0.64	0.62
UDDETTTS	4.15 ± 0.05	0.85 ± 0.02	0.90 ± 0.03	0.81 ± 0.02	0.68	0.62	0.73
w/o EME	4.10 ± 0.05	0.86 ± 0.04	0.83 ± 0.04	0.75 ± 0.02	0.65	0.58	0.70

Table 2: Subjective evaluation results of linear emotion control along the three ADV dimensions, measured by Spearman’s Rank Correlation (SRC) and Kendall’s W (KW). The right side *Linear Binning* presents the results of ablation experiments.

Dimension	Range	Nonlinear Binning		Linear Binning	
		SRC	KW	SRC	KW
Arousal	[1-14, 7, 7]	0.85	0.70	0.52	0.48
Dominance	[14, 1-14, 1]	0.78	0.68	0.48	0.50
Valence	[14, 14, 1-14]	0.92	0.83	0.57	0.58

human judgments (Subjective) and automatic classification (Objective) using the emotion2vec [39] model, with classification Accuracy, macro-Precision and macro-Recall computed from the confusion matrices of both evaluations. Table 1 summarizes the results.

The results show that UDDETTTS outperforms CosyVoice in the naturalness of synthesized emotional speech and achieves higher accuracy in label-based emotion control. This indicates that UDDETTTS demonstrates stronger robustness in emotion understanding.

4.4 ADV-Controlled Emotional TTS

We conduct experiments based on the second task by adjusting the values of $x_{adv} \in \mathbb{Z}_{[1,14]}^3$ to control the synthesized emotional speech. To evaluate UDDETTTS’s ability to linearly control emotions along each of three ADV dimensions, we fix two dimensions and vary the third, resulting in three test settings: Arousal test [1-14, 7, 7], Dominance test under strong negative emotions [14, 1-14, 1], Valence test under strong expressiveness [14, 14, 1-14]. Stronger emotions are assumed to exhibit greater perceptual separability during ranking. For each test, we synthesize 14 speech samples from a same *neutral* text and ask participants to rank them according to the SAM system [41]. We use Spearman’s Rank Correlation (SRC) to evaluate the alignment between each participant’s rankings and ground-truth rankings, and report the average score. And Kendall’s W (KW) is used to evaluate inter-rater agreement across 10 participants:

$$SRC = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \quad KW = \frac{12S}{k^2(n^3 - n)}, \quad (12)$$

where d_i is the rank difference between two rankings, n is the number of samples, S is the variance of the rank sums, and k is 10. As shown in Table 2, SRC values near 1.0 indicate that perceived emotions change linearly with the nonlinearly binned x_{adv} . The KW scores above 0.6 reflect strong inter-rater agreement, confirming the reliability of the results. Together, these findings demonstrate that UDDETTTS achieves linear emotion control along all three dimensions.

As shown in Figure 4, We validate that semi-supervised training significantly expands the controllable coverage of the ADV space, increasing it from 77.89% to 89.35%. We highlight in red regions of the ADV space capable of synthesizing emotional speech that aligns with the unseen ADV values. For example, at $x_{adv} = [14, 1, 1]$, where no training samples exist, the model can still synthesize reasonable *sobbing-like* emotional speech during inference. This indicates semi-supervised training promotes the transfer of label knowledge to the ADV space. Additionally, to evaluate the influence of each ADV dimension on emotion expression, we analyze the relationship between x_{adv} and prosodic features of speech, as detailed in the Appendix E.



Figure 5: Percentage results of the subjective preference test on speech-text emotional alignment.

4.5 End-to-End Emotional TTS

For the third task, we supplement the corpus with a set of texts featuring diverse and explicit emotional attributes (see Appendix D). To evaluate the ability of UDDETTS to directly synthesize emotional speech from text, we use only text input to synthesis emotional speech, where the RoBERTa-based ADV predictor predicts pseudo-ADV and then guides the x_{lbl} generation. We compare it with a CosyVoice baseline that uses the same pre-trained RoBERTa encoder but directly predicts the x_{lbl} instead, also using text-only input. A subjective preference test (%) is conducted to evaluate which method generates speech with more appropriate emotion. As shown in Figure 5, participants demonstrated a clear preference for UDDETTS, with an average preference rate of 62.42%, higher than CosyVoice’s 16.70%. We also calculate the p-value of t-test is 0.0053 (< 0.05), indicating that UDDETTS achieves significantly higher emotional consistency between the text and generated speech compared to the baseline. This result shows UDDETTS exhibits superior end-to-end emotional speech synthesis capabilities by integrating pseudo-ADV and label.

4.6 Ablation Studies

We conduct four ablation studies to evaluate the effectiveness of key components in UDDETTS. First, removing the ADV predictor in the third task results in predominantly neutral speech, similar to the CosyVoice baseline in Section 4.5, showing that the ADV predictor is crucial for text-adaptive emotion generation. Second, we remove the emotional mixture encoder and E_{emo} from the OT-CFM module and rely solely on E_{sem} to reconstruct mel-spectrograms. This modification leads to a reduction in emotional expressiveness, as seen in the last row of Table 1. Third, we replace nonlinear binning algorithm with linear binning algorithm in the ADV quantizer. Both SRC and KW scores drop significantly in Table 2, indicating that imbalanced emotion distributions causes the model to overfit dense ADV regions, thereby impeding linear control. Finally, training only on $D_{S,AL}$ without semi-supervised learning reduces the controllable coverage rate of the ADV space to 70%, and fails to synthesize the *sobbing-like* emotion at $x_{\text{adv}} = [14, 1, 1]$ and other unseen emotions. This highlights the pivotal role of unlabeled ADV data in transferring discrete emotion label knowledge into the ADV space and expanding control coverage.

5 Limitations and Future Work

The performance of UDDETTS is limited by the quality of ADV annotations. Subjective variation in emotion perception among annotators can lead to inconsistent ADV labels, which negatively impacts the model’s ability for linear emotional control. Additionally, we observe that for texts with ambiguous emotional attributes, the ADV predictor often struggles to infer appropriate ADV values. Since the same text can express different emotions in different contexts, incorporating multimodal information is necessary for more accurate emotional understanding. In future work, we plan to extract emotional representations from multimodal informations and dialogue context, mapping them into ADV or other dimensional spaces to capture emotions and synthesize more expressive speech.

6 Conclusion

In this paper, we introduce a unified framework named UDDETTS that 1) integrates both ADV and label annotations for the first time, enabling compatibility with diverse types of emotional speech datasets; 2) mitigates the sparsity and imbalance issues in the ADV space; 3) provides a method to linearly control emotional synthesis along three dimensions; and 4) explores the feasibility of ADV in adaptive emotional TTS. Our work can assist developers in building controllable emotional TTS systems based on large-scale emotional datasets, ultimately enhancing the naturalness of emotional expression in human-computer interaction.

References

- [1] A. Adigwe, N. Tits, K. E. Haddad, S. Ostadabbas, and T. Dutoit. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*, 06 2018.
- [2] K. An, Q. Chen, C. Deng, Z. Du, C. Gao, Z. Gao, Y. Gu, T. He, H. Hu, K. Hu, S. Ji, Y. Li, Z. Li, H. Lu, H. Luo, X. Lv, B. Ma, Z. Ma, C. Ni, C. Song, J. Shi, X. Shi, H. Wang, W. Wang, Y. Wang, Z. Xiao, Z. Yan, Y. Yang, B. Zhang, Q. Zhang, S. Zhang, N. Zhao, and S. Zheng. FunAudioLLM: Voice understanding and generation foundation models for natural interaction between humans and llms. *CoRR*, abs/2407.04051, 2024.
- [3] P. Anastassiou, J. Chen, J. Chen, Y. Chen, Z. Chen, Z. Chen, J. Cong, L. Deng, C. Ding, L. Gao, M. Gong, P. Huang, Q. Huang, Z. Huang, Y. Huo, D. Jia, C. Li, F. Li, H. Li, J. Li, X. Li, X. Li, L. Liu, S. Liu, S. Liu, X. Liu, Y. Liu, Z. Liu, L. Lu, J. Pan, X. Wang, Y. Wang, Y. Wang, Z. Wei, J. Wu, C. Yao, Y. Yang, Y. Yi, J. Zhang, Q. Zhang, S. Zhang, W. Zhang, Y. Zhang, Z. Zhao, D. Zhong, and X. Zhuang. Seed-TTS: A family of high-quality versatile speech generation models. *CoRR*, abs/2406.02430, 2024.
- [4] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li, and F. Wei. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738. Association for Computational Linguistics, May 2022.
- [5] M. Borchart and A. Dusterhoff. Emotions in speech - experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. In *2005 International Conference on Natural Language Processing and Knowledge Engineering*, pages 147–151, 2005.
- [6] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4):335–359, December 2008.
- [7] E. Y. Chang. Behavioral emotion analysis model for large language models. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 549–556. IEEE Computer Society, Aug. 2024.
- [8] C. Y. Chen, T. M. Hung, Y.-L. Hsu, and L.-W. Ku. Label-aware hyperbolic embeddings for fine-grained emotion classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10947–10958. Association for Computational Linguistics, July 2023.
- [9] S. Chen, C. Wang, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718, 2025.
- [10] D.-H. Cho, H.-S. Oh, S.-B. Kim, S.-H. Lee, and S.-W. Lee. EmoSphere-TTS: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech. In *Interspeech 2024*, pages 1810–1814, 2024.
- [11] D.-H. Cho, H.-S. Oh, S.-B. Kim, and S.-W. Lee. EmoSphere++: Emotion-controllable zero-shot text-to-speech via emotion-adaptive spherical vector. *IEEE Transactions on Affective Computing*, pages 1–16, 2025.
- [12] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, 2001.
- [13] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma, Z. Gao, and Z. Yan. CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *CoRR*, abs/2407.05407, 2024.

- [14] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang, F. Yu, H. Liu, Z. Sheng, Y. Gu, C. Deng, W. Wang, S. Zhang, Z. Yan, and J.-R. Zhou. CosyVoice 2: Scalable streaming speech synthesis with large language models. *ArXiv*, abs/2412.10117, 2024.
- [15] Y. Fathullah, C. Wu, E. Lakomkin, J. Jia, Y. Shangguan, K. Li, J. Guo, W. Xiong, J. Mahadeokar, O. Kalinli, C. Fuegen, and M. Seltzer. Prompting large language models with speech recognition abilities. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13351–13355, 2024.
- [16] Y. Fathullah, C. Wu, E. Lakomkin, K. Li, J. Jia, Y. Shangguan, J. Mahadeokar, O. Kalinli, C. Fuegen, and M. Seltzer. AudioChatLlama: Towards general-purpose speech abilities for LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5522–5532. Association for Computational Linguistics, June 2024.
- [17] S. Garca, J. Luengo, and F. Herrera. *Data Preprocessing in Data Mining*. Springer Publishing Company, Incorporated, 1st edition, 2016. ISBN 3319377310.
- [18] H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120–136, 2013.
- [19] Y. Guo, C. Du, X. Chen, and K. Yu. EmoDiff: Intensity controllable emotional text-to-speech with soft-label guidance. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.
- [20] Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan. PromptTTS: Controllable text-to-speech with text descriptions. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE, 2023.
- [21] R. Habib, S. Mariooryad, M. Shannon, E. Battenberg, R. J. Skerry-Ryan, D. Stanton, D. Kao, and T. Bagby. Semi-supervised generative modeling for controllable speech synthesis. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, 2020.
- [22] S. Hamann. Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends in Cognitive Sciences*, 16(9):458–466, 2012.
- [23] X. Hong, Y. Gong, V. Sethu, and T. Dang. AER-LLM: Ambiguity-aware emotion recognition leveraging large language models. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025.
- [24] C.-B. Im, S.-H. Lee, S.-B. Kim, and S.-W. Lee. 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, 2022.
- [25] S. Inoue, K. Zhou, S. Wang, and H. Li. Hierarchical emotion prediction and control in text-to-speech synthesis. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10601–10605, 2024.
- [26] M. Kang, W. Han, S. J. Hwang, and E. Yang. ZET-Speech: Zero-shot adaptive emotion-controllable text-to-speech synthesis with diffusion and style-based models. In *Interspeech 2023*, pages 4339–4343, 2023.
- [27] J. Kong, J. Kim, and J. Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Proc. NeurIPS*, 2020.
- [28] T. Li, S. Yang, L. Xue, and L. Xie. Controllable emotion transfer for end-to-end speech synthesis. In *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5, 2021.
- [29] T. Li, X. Wang, Q. Xie, Z. Wang, and L. Xie. Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1448–1460, 2022.

- [30] X. Li, Z.-Q. Cheng, J.-Y. He, J. Chen, X. Fan, X. Peng, and A. G. Hauptmann. UMETTS: A unified framework for emotional text-to-speech synthesis with multimodal prompts. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025.
- [31] Z. Lian, H. Sun, L. Sun, K. Chen, M. Xu, K. Wang, K. Xu, Y. He, Y. Li, J. Zhao, Y. Liu, B. Liu, J. Yi, M. Wang, E. Cambria, G. Zhao, B. W. Schuller, and J. Tao. MER 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9610–9614. Association for Computing Machinery, 2023.
- [32] X. Liang, H. Jiang, W. Xu, and Y. Zhou. Gaussian-smoothed imbalance data improves speech emotion recognition. *CoRR*, abs/2302.08650, 2023.
- [33] J. Liu, Z. Liu, Y. Hu, Y. Gao, S. Zhang, and Z. Ling. DiffStyleTTS: Diffusion-based hierarchical prosody modeling for text-to-speech with diverse and controllable styles. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5265–5272. Association for Computational Linguistics, Jan. 2025.
- [34] R. Liu, Y. Hu, Y. Ren, X. Yin, and H. Li. Emotion rendering for conversational speech synthesis with heterogeneous graph-based context modeling. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*. AAAI Press, 2024.
- [35] S. R. Livingstone and F. A. Russo. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):1–35, 05 2018.
- [36] R. Lotfian and C. Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, 2019.
- [37] M. Lugger and B. Yang. Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4945–4948, 2008.
- [38] J. Luo, X. Luo, X. Chen, Z. Xiao, W. Ju, and M. Zhang. SemiEvol: Semi-supervised fine-tuning for llm adaptation. *CoRR*, abs/2410.14745, 2024.
- [39] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760, Bangkok, Thailand, Aug. 2024.
- [40] A. Mehrabian and J. A. Russell. *An approach to environmental psychology*. The MIT Press, 1974.
- [41] J. D. Morris. Observations: SAM: The self-assessment manikin an efficient cross-cultural measurement of emotional response. *Journal of Advertising Research*, 35(6):63–65, 1995.
- [42] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *Proc. ICML*, volume 139, pages 8162–8171. PMLR, 2021.
- [43] Y. Oh, J. Lee, Y. Han, and K. Lee. Semi-supervised learning for continuous emotional intensity controllable speech synthesis with disentangled representations. In *Interspeech 2023*, pages 4818–4822, 2023.
- [44] D. Onken, S. Wu Fung, X. Li, and L. Ruthotto. OT-Flow: Fast and accurate continuous normalizing flows via optimal transport. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):9223–9232, May 2021.
- [45] S. Park, J. Kim, S. Ye, J. Jeon, H. Y. Park, and A. Oh. Dimensional emotion detection from categorical emotion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4367–4380. Association for Computational Linguistics, Nov. 2021.

- [46] S. Y. Park and C. Caragea. VerifyMatch: A semi-supervised learning paradigm for natural language inference with confidence-aware MixUp. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19319–19335. Association for Computational Linguistics, Nov. 2024.
- [47] R. Plutchik. *Emotion: A Psycho-evolutionary Synthesis*. Harper and Row, 1980.
- [48] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July 2019. Association for Computational Linguistics.
- [49] A. Punhani, N. Faujdar, K. K. Mishra, and M. Subramanian. Binning-based silhouette approach to find the optimal cluster using k-means. *IEEE Access*, 10:115025–115032, 2022.
- [50] L. Qiu, L. Zhong, J. Li, W. Feng, C. Zhou, and J. Pan. SFT-SGAT: A semi-supervised fine-tuning self-supervised graph attention network for emotion recognition and consciousness detection. *Neural Networks*, 180:106643, 2024.
- [51] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. <https://github.com/modelscope/3D-Speaker>, 2023.
- [52] J. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39: 1161–1178, 12 1980.
- [53] S. Sivaprasad, S. Kosgi, and V. Gandhi. Emotional prosody control for speech generation. In *Interspeech 2021*, pages 4653–4657, 2021.
- [54] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao. EmoMix: Emotion mixing via diffusion models for emotional speech synthesis. In *Interspeech 2023*, pages 12–16, 2023.
- [55] H. Viswanath, A. Bhattacharya, P. Jutras-Dubé, P. Gupta, M. Prashanth, Y. Khaitan, and A. Bera. AffectEcho: Speaker independent and language-agnostic emotion and affect transfer for speech synthesis. *CoRR*, abs/2308.08577, 2023.
- [56] S. Wang, J. Guðnason, and D. Borth. Learning emotional representations from imbalanced speech data for speech emotion recognition and emotional text-to-speech. In *Interspeech 2023*, pages 351–355, 2023.
- [57] X. Wang, M. Jiang, Z. Ma, Z. Zhang, S. Liu, L. Li, Z. Liang, Q. Zheng, R. Wang, X. Feng, W. Bian, Z. Ye, S. Cheng, R. Yuan, Z. Zhao, X. Zhu, J. Pan, L. Xue, P. Zhu, Y. Chen, Z. Li, X. Chen, L. Xie, Y. Guo, and W. Xue. Spark-TTS: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *CoRR*, abs/2503.01710, 2025.
- [58] Z. Wen, J. Cao, R. Yang, S. Liu, and J. Shen. Automatically select emotion for response via personality-affected emotion transition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5010–5020. Association for Computational Linguistics, Aug. 2021.
- [59] xingchen Song, D. Zhou, and Y. Yang. Reverse engineering of s3tokenizer proposed in cosyvoice. <https://github.com/xingchensong/S3Tokenizer>, 2024.
- [60] A. Zeng, Z. Du, M. Liu, L. Zhang, shengmin jiang, Y. Dong, and J. Tang. Scaling speech-text pre-training with synthetic interleaved data. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [61] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773. Association for Computational Linguistics, Dec. 2023.
- [62] Z. Zhang, S. Chen, L. Zhou, Y. Wu, S. Ren, S. Liu, Z. Yao, X. Gong, L. Dai, J. Li, and F. Wei. SpeechLM: Enhanced speech pre-training with unpaired textual data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2177–2187, 2024.

- [63] S. Zheng, L. Cheng, Y. Chen, H. Wang, and Q. Chen. 3D-Speaker: A large-scale multi-device, multi-distance, and multi-dialect corpus for speech representation disentanglement. <https://github.com/modelscope/3D-Speaker>, 2023.
- [64] K. Zhou, B. Sisman, R. Liu, and H. Li. Emotional voice conversion: Theory, databases and ESD. *Speech Communication*, 137:1–18, 2022. ISSN 0167-6393.
- [65] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li. Speech synthesis with mixed emotions. *IEEE Transactions on Affective Computing*, 14(4):3120–3134, 2023.
- [66] X. Zhu, S. Yang, G. Yang, and L. Xie. Controlling emotion strength with relative attribute for end-to-end speech synthesis. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 192–199, 2019.

A Dataset Statistics

Table 3: Statistics of cleaned emotional speech datasets used in UDDETTs.

Datasets	Hours	Type	Description
MSP-PODCAST	258.12	$D_{S,AL}$	Large-scale podcast corpus
IEMOCAP	12.28	$D_{S,AL}$	Acted dialogues in lab
MELD	8.86	$D_{S,L}$	TV show dialogues
ESD	29.07	$D_{E,L}$	Emotional voice conversion corpus
EmoV-DB	9.48	$D_{E,L}$	Controlled emotional expressions
RAVDESS	1.47	$D_{E,L}$	Controlled emotional expressions
Total	319.28	-	-

B ADV Statistics in all Datasets

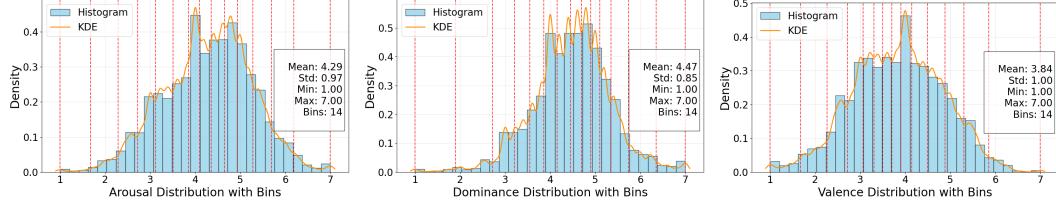


Figure 6: The histograms and kernel density estimations of all training samples along the three dimensions of the ADV space are shown, with the x-axis representing the continuous ADV values. Red dashed lines indicate the division of each dimension into 14 bins.

C Label Statistics in all Datasets

We collect emotion label statistics in all datasets and map them to individual label tokens. Table 4 in Appendix shows the sample count for each label, and Figure 7 shows the distribution of some emotion samples in the ADV space.

Table 4: Emotion labels, corresponding label tokens, and sample counts used in UDDETTs.

Token	Emotion(s)	Samples	Token	Emotion(s)	Samples
0	Unknown	37447	5	Fearful	2189
1	Sad	16287	6	Sleepiness, Bored	1913
2	Angry	24270	7	Neutral	58894
3	Frustrated	1849	8	Surprise	10214
4	Disgust, Contempt	8972	9	Happy, Amused	39433

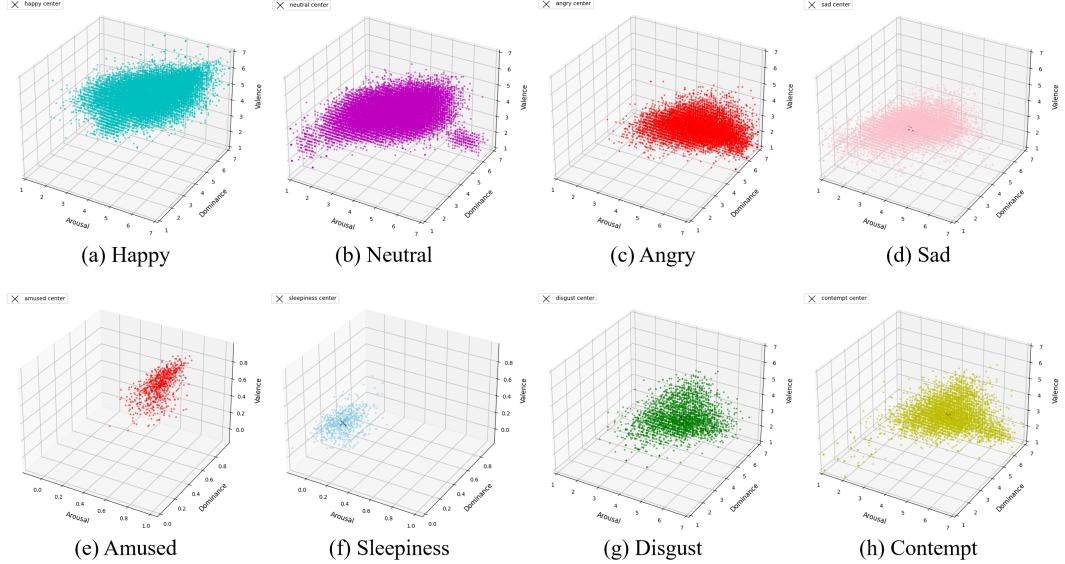


Figure 7: The distribution of some emotional samples in the ADV space. Each emotion tends to form a distinct cluster.

D Text Examples in the Test Set

We construct a test text corpus comprising 10 neutral sentences to evaluate naturalness and 5 sentences with distinct emotional attributes to evaluate end-to-end emotional speech synthesis. All texts are unseen during training. Table 5 shows 3 examples of neutral sentences and 5 examples of sentences with emotional attributes.

Table 5: Some examples of test text corpus with emotional content.

Emotion	Text
Neutral	For the twentieth time that evening the two men shook hands.
Neutral	She open the door and walk into the room.
Neutral	The meeting start promptly at nine in the morning.
Happy	I'm so happy to be friends with you.
Angry	I'm very angry now because you didn't arrive on time!
Sad	Lost wallet, missed last bus, tears drown my voiceless night.
Sleepiness	I'm tired because I had to work overtime until evening.
Mixed	I love you so much, I can't live without you!

E Impact of ADV Control on Prosodic Features

To study the impact of ADV control on emotional representations, we vary all values of $x_{\text{adv}} \in \mathbb{Z}_{[1,14]}^3$ to synthesize emotional speech and extract their prosodic features, including the mean and variance of $\log F0$ and energy, as well as duration and harmonic-to-noise ratio (HNR). We compute the Pearson correlation between each ADV dimension and these prosodic statistics. The results in Figure 8 show that Arousal and Dominance are significantly correlated with $\log F0$ and energy, indicating their role in controlling the excitement and intensity of emotion. Valence is correlated with HNR, which reflects voice quality variations linked to emotional changes [5], and it also affects the shape of the mel-spectrogram in Figure 9, indicating its influence on emotional polarity. Its correlation with duration is likely due to laughter in high-valence speech.

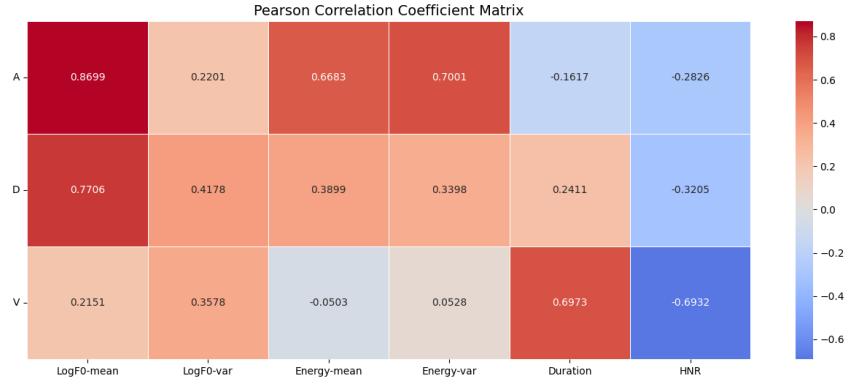


Figure 8: The Pearson correlation coefficient matrix showing the relationship between each ADV dimension and prosodic statistics.

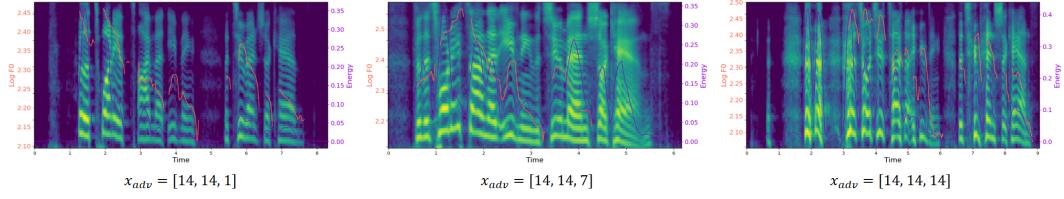


Figure 9: The patterns of F0 contours observed in the mel-spectrogram vary as a function of valence.