applied sciences

MDPI

Article

ECE-TTS: A Zero-Shot Emotion Text-to-Speech Model with Simplified and Precise Control

Shixiong Liang [1], Ruohua Zhou [1,*] and Qingsheng Yuan [2]

[1] School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 102627, China
[2] National Computer Network Emergency Response Technical Team Coordination Center of China, Beijing 100029, China
* Correspondence: zhouruohua@bucea.edu.cn

Abstract: Significant advances have been made in emotional speech synthesis technology; however, existing models still face challenges in achieving fine-grained emotion style control and simple yet precise emotion intensity regulation. To address these issues, we propose Easy-Control Emotion Text-to-Speech (ECE-TTS), a zero-shot TTS model built upon the F5-TTS architecture, simplifying emotion modeling while maintaining accurate control. ECE-TTS leverages pretrained emotion recognizers to extract Valence, Arousal, and Dominance (VAD) values, transforming them into Emotion-Adaptive Spherical Vectors (EASV) for precise emotion style representation. Emotion intensity modulation is efficiently realized via simple arithmetic operations on emotion vectors without introducing additional complex modules or training extra regression networks. Emotion style control experiments demonstrate that ECE-TTS achieves a Word Error Rate (WER) of 13.91%, an Aro-Val-Domin SIM of 0.679, and an Emo SIM of 0.594, surpassing GenerSpeech (WER = 16.34%, Aro-Val-Domin SIM = 0.627, Emo SIM = 0.563) and EmoSphere++ (WER = 15.08%, Aro-Val-Domin SIM = 0.656, Emo SIM = 0.578). Subjective Mean Opinion Score (MOS) evaluations (1–5 scale) further confirm improvements in speaker similarity (3.93), naturalness (3.98), and emotional expressiveness (3.94). Additionally, emotion intensity control experiments demonstrate smooth and precise modulation across varying emotional strengths. These results validate ECE-TTS as a highly effective and practical solution for high-quality, emotion-controllable speech synthesis.

Keywords: ECE-TTS; emotional speech synthesis; zero-shot text-to-speech

Academic Editor: Douglas O'Shaughnessy

Received: 17 March 2025
Revised: 27 April 2025
Accepted: 2 May 2025
Published: 4 May 2025

Citation: Liang, S.; Zhou, R.; Yuan, Q. ECE-TTS: A Zero-Shot Emotion Text-to-Speech Model with Simplified and Precise Control. *Appl. Sci.* **2025**, *15*, 5108. https://doi.org/10.3390/app15095108

1. Introduction

With the rapid advancement of Text-to-Speech (TTS) technology, modern speech synthesis models have achieved the capability to generate speech that closely resembles human voices and accurately convey semantic information [1–7]. However, human speech not only carries semantic content but also expresses rich emotional information through variations in intonation, making emotional speech synthesis a topic of significant interest in recent years. Emotion TTS has made significant progress in recent years, but still faces many challenges in emotion control, mainly due to the limitations of existing models, the inherent complexity of human emotions (including emotional style and intensity) [8], and the lack of adequate emotional datasets.

Existing emotional speech synthesis models, such as Emo-VITS [9], E3-VITS [10], and YourTTS [11], which utilize the VITS model [12], as well as GenerSpeech [13] and EmoSphere-TTS [14], which are based on the FastSpeech 2 model [15], have been widely

adopted in the field of speech synthesis. However, these models generally produce audio with limited quality and restricted emotional expressiveness. In contrast, the recently proposed F5-TTS model, a non-autoregressive TTS framework based on Conditional Flow Matching and Diffusion Transformers (DiT) [16], offers a more streamlined and efficient approach. Recent research [17–19] indicates that removing phoneme-level duration predictors improves speech naturalness in zero-shot generation while eliminating the need for explicit phoneme alignment, thereby enhancing both naturalness and emotional diversity. Unlike conventional models such as VITS and FastSpeech 2, which rely on phoneme alignment and duration predictors, F5-TTS simplifies the synthesis pipeline by eliminating phoneme alignment, duration predictors, and text encoders, reducing architectural complexity. Additionally, it integrates ConvNeXt V2 [20] to enhance text representations and improve text–speech alignment during in-context learning. This design removes the constraints of traditional alignment mechanisms, leading to a more natural and expressive synthesis. Furthermore, F5-TTS employs the Sway Sampling strategy, which optimizes the sampling process in flow-based steps, further improving speech quality, naturalness, and speaker similarity in synthesized outputs. In addition to TTS models, effective emotion control is crucial for emotional speech synthesis. It can be broadly divided into emotional style control and emotional intensity control, each addressing different aspects of expressive speech generation.

Emotional style control determines the type of emotion expressed in speech, shaping how an utterance conveys emotional intent. In this aspect, existing emotional TTS systems have primarily adopted two approaches. The first approach relies on supervised training using discrete labeled emotional datasets [21–23], where explicit labels are used to control emotional style. However, due to the complexity of human emotions, manually labeling emotional data is costly and time-consuming. As a result, existing datasets are typically short and lack diversity, limiting the effectiveness of this method and leading to synthesized speech with monotonous emotional styles. The second approach extracts emotional features from short reference audio that matches specific style characteristics through conditional formatting [24–26], a method also known as reference encoding [27], or style transfer [28]. This approach uses a reference encoder to extract style information from a reference audio sample and transfer it to the generated speech. While this method does not rely on explicit emotion labels and enables the synthesis of diverse emotional expressions, a major challenge lies in effectively disentangling emotional features from other attributes (e.g., speaker identity, age) present in the reference audio and transferring only the desired emotional style to the main synthesis network [29]. Recent research has explored speech disentanglement techniques to tackle this issue. For instance, the FACodec framework [2] decomposes speech into three components: content, prosody, and acoustic details. However, it has been observed that the prosody encoder encodes phonetic details, leading the synthesized speech to reflect the content of the emotion prompt audio instead of the intended text prompt [30].

The above two kinds of methods for controlling emotional style are typically limited by the finite types of emotions available in emotional datasets. This limitation often leads to synthesized audio with a lack of stylistic variety or stereotypical emotional expressions, making it difficult to achieve fine-grained emotional control [31,32]. To address this issue, continuous emotional representations, such as the Russell emotional model [33], are utilized to achieve more nuanced control over variations in emotional style. Based on Russell's Circumplex Model of Affect, emotional attributes can be represented in a continuous three-dimensional space defined by valence, arousal, and dominance. Valence reflects the degree of positivity or negativity of an emotion, arousal indicates the level of physiological or emotional activation, and dominance measures the extent to which an individual feels

in control or subordinate in a given emotional state. While this representation enables continuous modeling of affective states, further clarification of how fundamental emotions are distributed within this space enhances interpretability and control in emotional speech synthesis. While this continuous model supports nuanced emotional control, it is important to understand how canonical emotion types are positioned within the VAD space to further improve interpretability and synthesis precision.

To this end, a structured emotional spectrum is typically formed using seven representative categories that are widely recognized in emotional speech datasets: neutral, happy, angry, sad, disgusted, excited, and fear. These categories not only represent distinct affective states, but also exhibit characteristic positions in the VAD space, as illustrated in Figure 1. Neutral is typically located near the center of the space, representing emotional balance, low activation, and the absence of strong valence or dominance. In contrast, happy occupies the region of high valence, high arousal, and strong dominance, conveying a positive and energetic affective state. Angry also exhibits high arousal and dominance but it is characterized by low valence, reflecting confrontation and tension. Sad lies in the low-valence, low-arousal, and low-dominance region, indicating a withdrawn and passive emotional state. Fear shares low valence and high arousal with anger but shows significantly lower dominance, denoting helplessness. Excited is close to happy in valence and arousal but tends to express greater outward intensity. Disgusted generally appears in the low-valence, moderate-arousal range, associated with rejection and aversion. This structured spectrum not only enhances understanding of the relationships among emotion categories but also provides a psychologically grounded reference framework for designing emotion control mechanisms. It supports interpretable emotion encoding, enables interpolation in the affective space, and offers a solid foundation for both data annotation and conditional generation in TTS. Moreover, the corresponding values of arousal, valence, and dominance can be automatically extracted using a pretrained emotion recognition model [34], eliminating the need for manual emotion annotation. Building on this foundation, a recent study has applied an emotion-adaptive coordinate transformation to convert VAD into an Emotion-Adaptive Spherical Vector (EASV) [35]. Experimental results indicate that using EASV as the emotion style representation enables more precise control over emotional styles.

Emotion intensity, which reflects the degree of emotional expression along specific dimensions, plays a crucial role in emotional modeling. For example, happiness can range from pleasure to exuberance, while sadness can manifest as mild sorrow or deep grief [36]. However, most emotional speech datasets lack explicit emotion intensity labels, and compared to discrete emotion category annotations, labeling emotion intensity is a more subjective and complex process, posing additional challenges for emotion modeling. To address emotion intensity control, researchers have made numerous efforts to define and compute emotion intensity values effectively for model training. One of the most commonly used methods is Relative Attributes Ranking (RAR) [37], which has been widely applied in studies [38–42]. RAR constructs a ranking matrix through maximum margin optimization and solves it using a Support Vector Machine (SVM). The resulting rankings are then used for model training. Despite its popularity, RAR is a manually designed and independent stage, which may lead to suboptimal results and introduce biases during training. Beyond RAR, researchers have further explored emotion embedding space manipulations. For instance, ref. [43] proposed an algorithm that maximizes the distance between emotional embeddings and interpolates within the embedding space to control emotion intensity. Similarly, ref. [44] quantified the distances within the emotion embedding space to determine emotion intensity. However, the structure of the embedding space significantly impacts the performance of these models, necessitating careful constraint application to

ensure stability and generalization [45]. Recently, Task Arithmetic has been introduced for emotion intensity control [46], extending its application from Natural Language Processing (NLP) and Computer Vision (CV) to emotional speech synthesis. Unlike manual ranking or embedding manipulations, this method leverages weight-space interpolation, enabling direct and efficient intensity modulation. By computing emotion vectors as the difference between pretrained and fine-tuned model weights, Task Arithmetic provides precise, scalable, and smooth emotion intensity control without additional ranking-based processes.

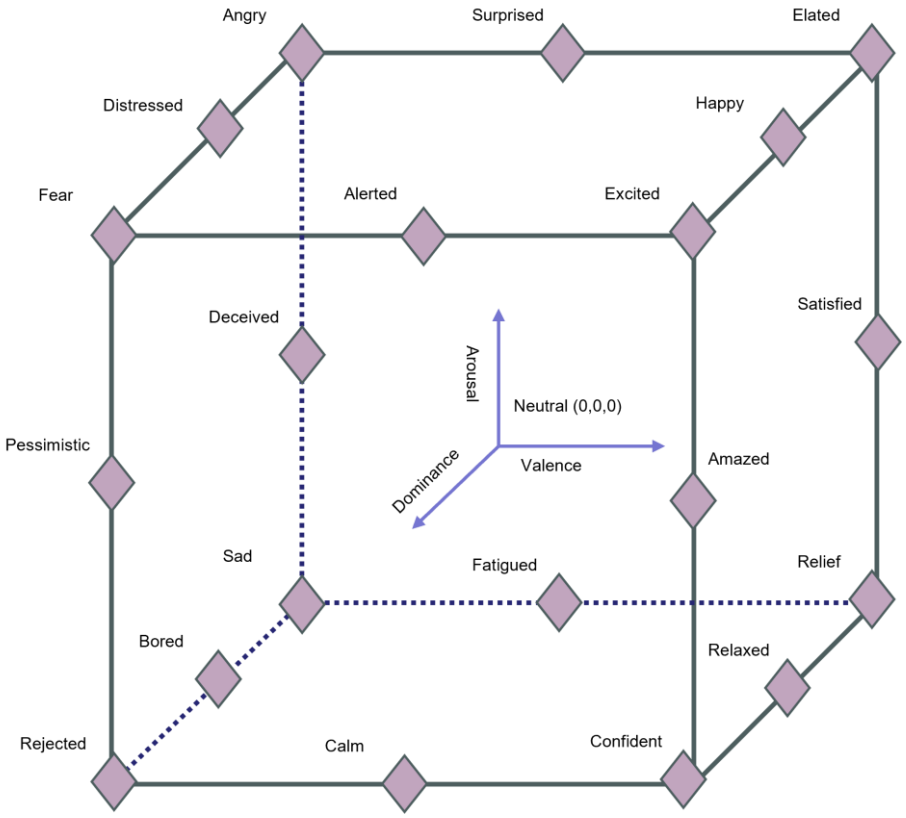

Figure 1. Three-dimensional valence–arousal–dominance (VAD) emotional space representation.

In addition to model limitations and challenges in emotion control, the limited scale of existing emotional speech datasets and the high cost of manual emotion annotation also restrict the performance of emotional speech synthesis. While early studies, Ekman et al. [47] identified six basic emotions—happiness, anger, sadness, disgust, surprise, and fear—through cross-cultural facial expression studies. These emotions exhibit high consistency and distinguishability across different cultures. Most existing emotional speech datasets are built upon these fundamental emotions, ensuring high consistency and distinguishability across different cultures. Mainstream emotional datasets, such as ESD [48], IEMOCAP [49], and EXPRESSO [50], predominantly use discrete labels for emotion annotation, limiting expressiveness and diversity. Among them, IEMOCAP dataset includes continuous emotion annotations, offering a more refined representation of emotional variations. Specifically, the ESD dataset contains approximately 30 h of speech and covers five emotional categories; IEMOCAP spans 12 h and includes seven emotion categories, while EXPRESSO contains 47 h of speech and supports 26 different expressive styles, providing a broader range of emotional diversity. Despite progress in the construction of emotional speech datasets, most existing datasets remain relatively small in scale, with many containing less than 100 h of recorded speech. This limitation constrains the model's ability to learn diverse emotional patterns and compromises the quality of synthesized speech. Moreover, studies suggest that humans can perceive up to 34,000 distinct emotions [51],

far exceeding the number of emotions covered by current TTS systems. Due to the high complexity of human emotions, manually annotating emotional data is extremely costly, making it difficult to scale dataset size beyond 100 h. To address this challenge, adopting an automatic annotation pipeline for constructing an emotional speech dataset is essential. This approach not only removes the dependency on manual labeling, significantly reducing costs, but also contributes to expanding the dataset scale, ensuring a more extensive and scalable data foundation for emotional speech synthesis.

In this paper, we propose Easy-Control Emotion TTS (ECE-TTS), a zero-shot emotion-controllable TTS system that simplifies emotion control while enhancing both the naturalness and expressiveness of synthesized speech. Built upon a flow-matching-based zero-shot TTS architecture, ECE-TTS simplifies model design by removing the need for phoneme alignment, duration predictors, and text encoders. By leveraging ConvNeXt V2 modules, it refines text representations and enhances text–speech alignment during in-context learning, leading to improved prosody and rhythm. To achieve fine-grained emotional style control, we employ the Emotion-Adaptive Coordinate Transformation, which converts VAD values into EASV representations. Compared to conventional approaches that rely on discrete emotion labels or reference-based style transfer, this method allows for more precise and adaptable emotional expression. For emotion intensity modulation, we adopt an emotion arithmetic vector method, leveraging weight-space interpolation techniques to adjust emotion intensity in a smooth and controllable manner. By applying a scaling factor α, this approach enables continuous emotion intensity transitions, overcoming the limitations of conventional ranking-based or embedding-space manipulation methods. To further strengthen the model's performance, we construct a 200 h high-quality emotional speech dataset through an automatic annotation pipeline, eliminating the need for manual labeling while scaling up the dataset. Extensive experiments demonstrate that ECE-TTS effectively generates highly natural and expressive speech while providing precise and flexible control over both emotional style and intensity, making it a practical and efficient solution for zero-shot emotional speech synthesis.

2. ECE-TTS

This study introduces ECE-TTS, a streamlined emotion-controllable TTS model. Based on F5-TTS, it employs text-guided speech infilling for training, eliminating reliance on phoneme-level duration aligners and thereby simplifying the model architecture. To further improve text–speech alignment, it integrates ConvNeXt V2 for enhanced text representation and Classifier-Free Guidance (CFG) [52] to refine alignment effectively balancing conditional and unconditional score estimations. For emotion control, it utilizes EASV for precise style modulation and adopts the emotion arithmetic vector approach to ensure accurate and flexible intensity control while reducing model complexity.

2.1. Modeling Principles

2.1.1. Flow Matching-Based TTS

In recent years, Flow Matching (FM) [53] has been widely adopted in image generation tasks. Its core idea is to guide samples from a prior distribution to a target data distribution through time-dependent vector fields in the latent space, thereby enabling efficient generative modeling. Compared to traditional diffusion-based approaches, FM eliminates the need for score function estimation or complex numerical solvers, significantly improving generation efficiency while maintaining comparable synthesis quality.

This modeling paradigm has also been extended to speech generation. Voicebox [54], proposed by Meta, is the first to apply Conditional Flow Matching (CFM) to speech synthesis, particularly targeting speech-infilling tasks. In Voicebox, the model is trained

to reconstruct missing audio segments given partial speech context and a frame-level phoneme sequence as conditions. By combining masking strategies with conditional vector field supervision, Voicebox achieves natural and fluent speech generation in multilingual, cross-speaker, and even zero-shot settings. These results demonstrate the effectiveness and generalizability of CFM in sequential and condition-aware generation scenarios, offering a novel perspective for continuous modeling in TTS.

Although speech and image data differ significantly in modality and structure, their generative processes can both be abstracted as continuous mappings from a noise distribution to a target distribution. As a result, the unified generation trajectory modeled by FM exhibits strong cross-modal transferability. In the speech domain, acoustic features such as Mel-spectrograms can be treated as continuous points in latent space, where their evolution is guided by time-dependent vector fields. Unlike image generation, which emphasizes local spatial structure, speech synthesis demands strict temporal coherence and semantic alignment with textual inputs.

Inspired by the modeling principles of Voicebox, this work adopts Flow Matching as the core training objective, incorporating both textual and emotional representations as conditional inputs. Furthermore, we employ the Optimal Transport variant of CFM (OT-CFM) to constrain the generative trajectory between the prior and target distributions, enabling stable training while supporting fine-grained control over speaking style and emotional expression. This framework captures temporal dependencies along the entire generative path under conditional guidance, effectively enhancing the model's ability to produce natural and expressive speech.

The flow matching loss function is defined in Equation (1), where the objective is to predict the expected behavior of the flow transformation given an input x_0. The term $u_t(x)$ represents the expected flow direction vector:

$$L_{FM}(\theta) = E_{t,p_t(x)} \| v_t(x) - u_t(x) \|^2, \tag{1}$$

in this equation, θ represents the neural network parameters, $t \sim \mathcal{U}[0,1]$, and $x \sim p_t(x)$. Since no prior knowledge is available to approximate the explicit form of $p_t(x)$, the training process follows a conditional probability path:

$$p_t(x \mid x_1) = N\Big(x \mid \mu_t(x_1), \sigma_t(x_1)^2 I\Big), \tag{2}$$

where x_1 is a random variable corresponding to the training data, and μ and σ represent the time-dependent mean and scalar standard deviation of a Gaussian distribution. It has been proven in [53] that the loss function of conditional flow matching is equivalent to the original flow matching loss, providing the theoretical foundation for our modeling approach.

The goal of conditional flow matching is to model the transition between two distributions, such as transforming a random normal noise distribution $p(x_0)$ into a target data distribution $p(x_1)$. This transformation relies on the flow path $\psi_t(x_0) = x_t$ and a time-dependent vector field $v_t(x_t)$, which determines the movement direction and speed of data points at each time step. The flow path $\psi_t(x_0) = x_t$ describes the complete dynamic process of transitioning from the initial distribution $p(x_0)$ to the target distribution $p(x_1)$, and can be formulated as a mapping between two probability density functions:

$$\frac{d}{dt}\psi_t(x) = v_t(\psi_t(x)), \quad \psi_0(x) = x, \tag{3}$$

Our objective is to generate data samples that conform to the target distribution while starting from a simple distribution such as Gaussian noise. In this process, flow matching provides a vector field:

$$\frac{d}{dt}\psi_t(x_0) = u_t(\psi_t(x_0) \mid x_1), \tag{4}$$

By applying reparameterization, we derive the conditional flow matching loss function:

$$L_{CFM}(\theta) = E_{t,q(x_1),p(x_0)} \left\| v_t(\psi_t(x_0)) - \frac{d}{dt}\psi_t(x_0) \right\|^2, \tag{5}$$

Utilizing the Optimal Transport (OT) formulation $\psi_t(x) = (1-t)x + tx_1$, we can further express the OT-CFM loss function as follows:

$$L_{CFM}(\theta) = E_{t,q(x_1),p(x_0)} \|v_t((1-t)x_0 + tx_1) - (x_1 - x_0)\|^2, \tag{6}$$

During the inference stage, given a sample point x_0 from the initial distribution, we solve for the flow's time step t and generation constraints using an Ordinary Differential Equation (ODE) solver [55]. The integral result of $\psi_t(x_0)$ is then evaluated as a function of time, with the initial condition set as $\psi_0(x_0) = x_0$. This approach ensures mathematical rigor in the generation process and maintains dynamic controllability over the data distribution.

2.1.2. Classifier-Guided and Classifier-Free Guidance

Classifier Guidance (CG) [56] is a conditioning technique introduced in diffusion models to guide conditional generation. Its primary goal is to explicitly enhance the alignment between generated samples and the target conditions by incorporating gradient information from a classifier, thereby significantly improving the quality of synthesized samples. Compared to unconditional generation, CG effectively steers the generated samples towards the target distribution, ensuring that the results adhere more closely to user-specified conditions. However, this explicit conditional generation approach also comes with several notable limitations.

First, CG requires training an additional classifier, which must be capable of efficiently handling noisy input data, increasing the overall implementation complexity. Moreover, the quality of generated samples is directly dependent on the classifier's accuracy; if the classifier is unreliable or susceptible to external noise, the alignment between the generated samples and the intended conditions may deteriorate. Additionally, since CG relies on classifier gradients for guidance, it is inherently vulnerable to adversarial attacks, potentially leading to the generation of misleading samples. For instance, these adversarial samples may contain subtle imperceptible details that appear to meet the given conditions but deviate from the true target distribution. Consequently, while CG offers an explicit conditional control mechanism, its dependence on a classifier and the associated security concerns limit its practical applicability.

To address these challenges, Classifier-Free Guidance (CFG) was introduced as an alternative guidance approach that eliminates the dependence on an explicit classifier. The core idea of CFG is to leverage both conditional and unconditional score estimations during the generation process, thereby enhancing sample quality while avoiding the complexity and risks associated with classifier-based guidance. This method randomly drops conditioning information with a certain probability during training, allowing the model to learn both conditional and unconditional score estimates, ultimately improving generation flexibility.

Specifically, CFG utilizes a single model to simultaneously generate conditional scores ϵ_{cond} and unconditional scores ϵ_{uncond}. The final guided score is computed using the following linear combination formula:

$$\widetilde{\epsilon}_\theta = \epsilon_{uncond} + s \cdot (\epsilon_{cond} - \epsilon_{uncond}), \tag{7}$$

where s is a scaling parameter that controls the strength of generation guidance. By adjusting the value of s, CFG can dynamically balance sample fidelity and diversity during generation.

Compared to CG, CFG offers several significant advantages. First, CFG does not require an additional classifier, greatly simplifying implementation and reducing training complexity in diffusion models. Second, CFG avoids vulnerabilities to adversarial attacks that could arise from classifier guidance, inherently enhancing the security and reliability of generated samples. Lastly, through the random dropping of conditions during training, CFG enables seamless switching between conditional and unconditional generation modes, providing a more flexible and adaptive generation process.

By employing this technique, CFG not only simplifies conditional generation but also enhances the model's control capabilities. Experimental results demonstrate that CFG achieves an effective balance between sample quality and diversity, as expressed in Equation (8), where α represents the strength of CFG:

$$v_{t,\text{CFG}} = v_t(\psi_t(x_0), c) + \alpha(v_t(\psi_t(x_0), c) - v_t(\psi_t(x_0))). \tag{8}$$

Beyond improving general conditional generation, text–speech alignment is a critical factor in ensuring the semantic accuracy and emotional consistency of synthesized speech. CFG enhances text–speech alignment by randomly dropping text conditions during training, allowing the model to learn both text-conditioned and unconditioned score distributions. In our training process, the masked speech input is randomly dropped with a probability of 30%, and in addition, there is a 20% probability of simultaneously dropping both the masked speech and the corresponding text input. By exposing the model to training samples with partially or entirely missing conditions, this strategy enables it to robustly model score distributions under varying conditioning scenarios. This enhances the association between textual content and generated speech, and improves the model's robustness and flexibility when handling incomplete or uncertain inputs. Therefore, we adopt CFG in our approach, leveraging its advantages to simplify model architecture while significantly improving the semantic coherence and emotional fidelity of synthesized speech.

2.2. Model Architecture

2.2.1. Model Training

Our model is trained using an infilling task, where it predicts a speech segment by leveraging surrounding audio information and complete text input (including both contextual transcripts and the segment to be generated). We denote x as the audio sample, y as the corresponding transcript, and E as the emotion feature sequence. The input data format is represented as $(x,\ y,\ E)$.

As illustrated in Figure 2, during training, the input audio data x is transformed into a Mel-spectrogram feature matrix $x_1 \in \mathbb{R}^{F \times N}$, where F represents the Mel frequency dimension, and N denotes the sequence length.

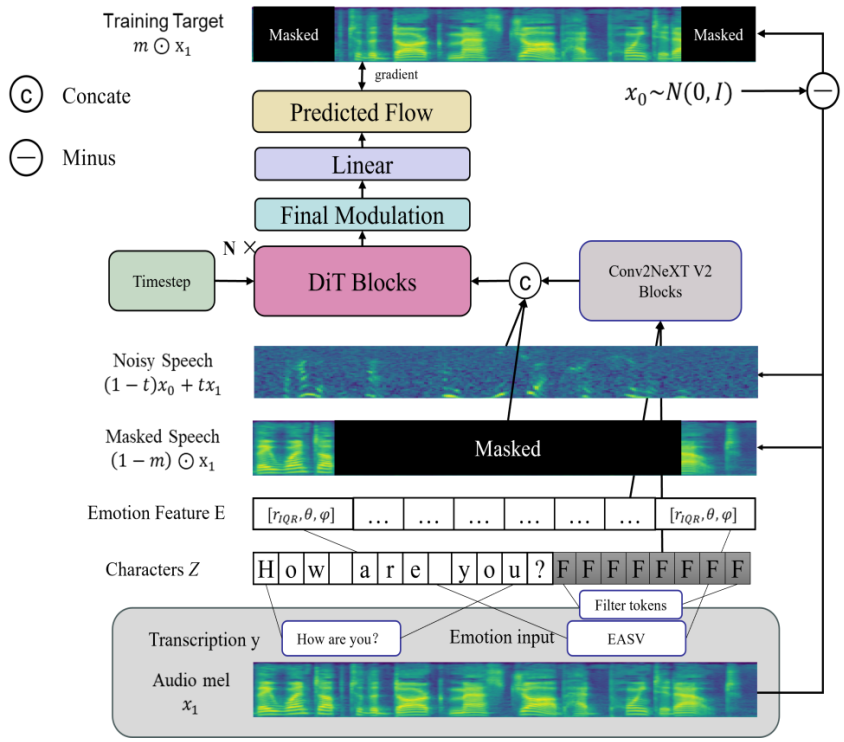

Figure 2. An overview of ECE-TTS training.

Unlike phoneme-based models that require phoneme-level text alignment, ECE-TTS only requires character-level text sequences. The input text is first converted into a corresponding character sequence, which is then padded to match the length N of the Mel-spectrogram. The resulting expanded character sequence Z is formulated as follows:

$$z = (c_1, c_2, \ldots, c_M, \langle F \rangle, \ldots, \langle F \rangle), \tag{9}$$

where c_i represents the i-th character, and F denotes padding characters, ensuring that the sequence length matches N, with M being the original text length.

Since we directly use character sequences Z instead of phoneme sequences, this may affect the model's alignment capability. To address this, before concatenating features, we process the text sequence using ConvNeXt V2 blocks, which enhances the model's ability to align speech and text representations. Experimental results demonstrate that this approach provides a dedicated modeling space, significantly improving context learning ability. Similarly, we apply ConvNeXt V2 blocks to the emotion feature sequence E to further improve its alignment with the text representation.

Within the CFM framework, we introduce noisy audio and masked audio into the model for training. Specifically, we apply Gaussian noise perturbation using $(1-t)x_0 + tx_1$, where x_0 represents sampled Gaussian noise, and t denotes the flow step size. Additionally, we introduce masked audio as $(1-m) \odot x_1$, where $m \in [0,1]^{F \times N}$ is a binary temporal mask. The training objective is to reconstruct the masked speech segment $m \odot x_1$ based on the unmasked speech segment $(1-m) \odot x_1$, the character sequence Z, and the emotion feature sequence E. This process is equivalent to learning the target distribution p_1, formulated as: $P(m \odot x_1 | (1-m) \odot x_1, Z, E)$ which aims to approximate the true speech data distribution q.

2.2.2. Model Inference

To achieve content and emotion-controllable speech synthesis, our model integrates the following key inputs: reference audio spectrogram x_{ref}, corresponding transcript y_{ref},

target transcript y_{gen}, and emotion prompt E. The reference audio provides crucial information about the target speaker's timbre and prosody, the text prompt guides the content of the generated speech, and the emotion prompt enables precise control over emotional expression.

In Text-to-Speech models, sequence length N (i.e., speech duration) is a key factor that directly affects length control in speech synthesis. Unlike other models that require a separately trained duration predictor, we directly predict the duration based on the character ratio between y_{gen} and y_{ref}. This approach simplifies the model architecture, introduces greater randomness, and enhances the naturalness of the generated speech. We assume that the total character length does not exceed the Mel-spectrogram length. Therefore, similar to the training process, we concatenate y_{gen} and y_{ref}, then pad the character sequence to match the Mel-spectrogram length. Additionally, following the emotion-adaptive coordinate transformation method introduced in Section 2.3.1, we extract emotion-adaptive spherical coordinates from the emotion reference audio e_{ref} and use them as the emotion feature input E for the model.

To sample from the learned distribution, we use the converted Mel features x_{ref}, the concatenated extended character sequence $z_{ref \cdot gen}$, and the emotion dimension sequence E as conditional inputs in Equation (8), obtaining

$$v_t(\psi_t(x_0), c) = v_t\big((1-t)x_0 + tx_1 \mid x_{ref}, z_{ref \cdot gen}, E\big), \tag{10}$$

As illustrated in Figure 3, the inference process begins with sampling from simple noise x_0, with the goal of gradually generating the final output x_1. Given the differential equation condition $d\psi_t(x_0)/dt = v_t\Big(\psi_t(x_0), x_{ref}, z_{ref \cdot gen}, E\Big)$, we employ an ODE solver to iteratively recover the target state from the initial condition $\psi_0(x_0) = x_0$ to $\psi_1(x_0) = x_1$. During inference, flow transformation steps are executed in a structured manner; for example, by uniformly sampling a predefined number of time steps from 0 to 1, based on the Number of Function Evaluations (NFE) setting.

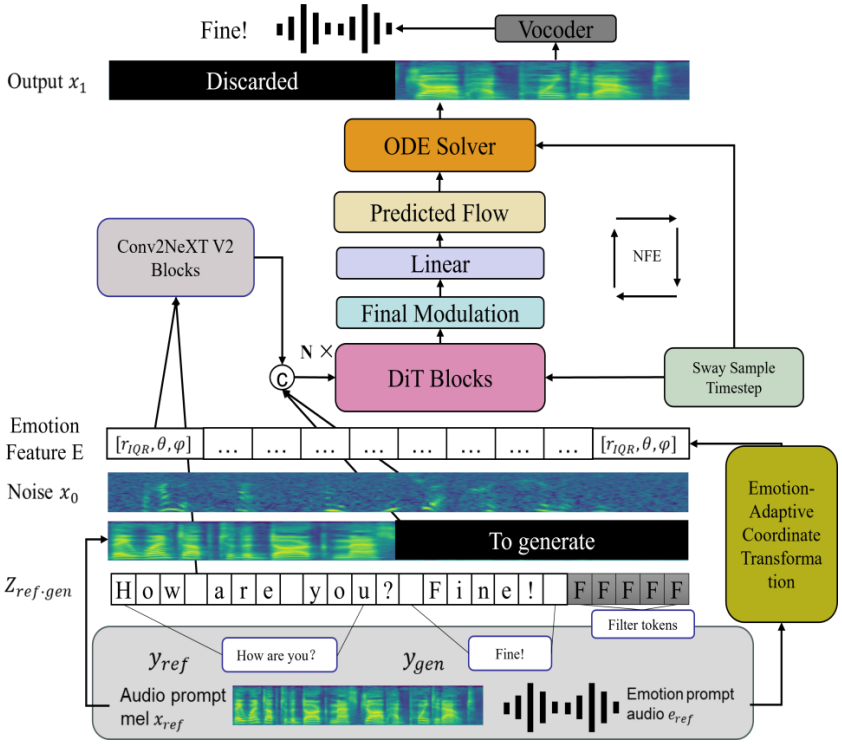

Figure 3. An overview of ECE-TTS inference.

After generating the Mel-spectrogram using the model v_t and ODE solver, we discard the reference audio component x_{ref}, retain only the generated speech component x_{gen}, and finally convert the Mel-spectrogram into waveform audio using a vocoder.

2.3. Emotion Control Strategy

2.3.1. Emotion Style Adjustment Through Adaptive Coordinate Transformation

To achieve accurate modeling of emotional features, we adopt the Emotion-Adaptive Coordinate Transformation method [35], which transforms the extracted VAD values into EASV. The transformation process consists of the following steps:

1. Emotion Dimension Extraction. First, we use an emotion recognition model ψ to predict VAD values for each reference speech sample x_i in the speech dataset X, obtaining

$$e_i^k = \psi(x_i), \tag{11}$$

where e_i^k is a three-dimensional vector (d_v, d_a, d_d), representing V, A, and D, respectively.

2. Calculation of the Neutral Emotion Centroid. For neutral emotion data, we compute the neutral emotion centroid as a reference baseline for the coordinate transformation:

$$M = \frac{1}{N_n} \sum_{i=1}^{N_n} e_i^n, \tag{12}$$

where N_n is the total number of neutral samples, and e_i^n denotes the VAD values of the i-th neutral emotion sample.

3. Emotion Coordinate Shift. To model a spherical coordinate system for each emotion, we shift the emotion coordinates from different emotion sets E_k, obtaining the transformed Cartesian coordinates:

$$\hat{e}_i^k = e_i^k - M_k, \tag{13}$$

where M_k is the centroid coordinate obtained by maximizing the distance ratio between a specific emotion and the neutral emotion:

$$M_k = \underset{M}{\text{argmax}} \frac{E_{e_i^k \in E_k} \left[\left\| M - e_i^k \right\|_2 \right]}{E_{e_i^n \in E_n} \left[\left\| M - e_i^n \right\|_2 \right]}, \tag{14}$$

here e_i^k and e_i^n denote the coordinates of the i-th sample in emotion set E_k and neutral set E_n, respectively.

4. Conversion to Spherical Coordinates. Using the shifted Cartesian coordinates $\hat{e}_i^k = \left(\hat{d}_v, \hat{d}_a, \hat{d}_d\right)$, we convert them into spherical coordinates $s_i^k = (r, \theta, \varphi)$ as follows:

$$r = \sqrt{\hat{d}_v^2 + \hat{d}_a^2 + \hat{d}_d^2}, \quad \theta = arccos\left(\frac{\hat{d}_d}{r}\right), \quad \varphi = arctan\left(\frac{\hat{d}_v}{\hat{d}_a}\right), \tag{15}$$

5. Outlier Processing. To reduce the impact of outliers during model training, we apply the Interquartile Range (IQR) method [57] after coordinate transformation:

$$r_{clamp} = min(max(r, r_{min}), r_{max}), \quad r_{IQR} = \frac{r_{clamp} - r_{min}}{r_{max} - r_{min}}, \tag{16}$$

where $r_{\min}$ and $r_{\max}$ are calculated based on the IQR method. Specifically, $r_{\min}$ is set to the first quartile (Q1) minus 1.5 times the IQR, and $r_{\max}$ is set to the third quartile (Q3) plus 1.5 times the IQR.

Through the above steps, we successfully transform emotion dimension values into Emotion-Adaptive Spherical Vectors, represented as $s_i^k = \left(r_{\text{clamp}}, \theta, \varphi\right)$, which serves as the emotion feature representation for our model.

It is worth noting that, compared to existing methods, the proposed EASV-based modeling approach eliminates the need for additional emotion encoders or style adaptation modules. Specifically, the Emo-VITS model [9] designs dedicated emotion encoders to extract both global and local emotional features from reference audio and fuses them through an attention mechanism. Similarly, GenerSpeech [13] introduces a multi-level style adaptor to simultaneously model global emotion characteristics and local prosodic features. These approaches add new sub-modules to the original TTS framework, thereby significantly increasing model complexity and training cost. In contrast, our method simply leverages a pretrained emotion recognition model during the data processing stage to extract continuous VAD emotion dimensions, and then applies an adaptive coordinate transformation to obtain structurally clear EASV representations, which are directly used as emotion feature inputs. This strategy effectively reduces model parameters and training complexity while maintaining controllability over emotional styles, demonstrating the advantages of structural simplicity, efficiency, and precise emotion control.

2.3.2. Emotion Intensity Modulation via Emotion Arithmetic

From Figure 4, it can be observed that speech samples of different emotions exhibit distinct distribution patterns. Therefore, we infer that the Arousal (A), Valence (V), and Dominance (D) values of an audio sample can effectively reflect changes in emotion intensity. Each emotion occupies a characteristic range within these three dimensions, and as emotion intensity changes, the corresponding V, A, and D values adjust accordingly.

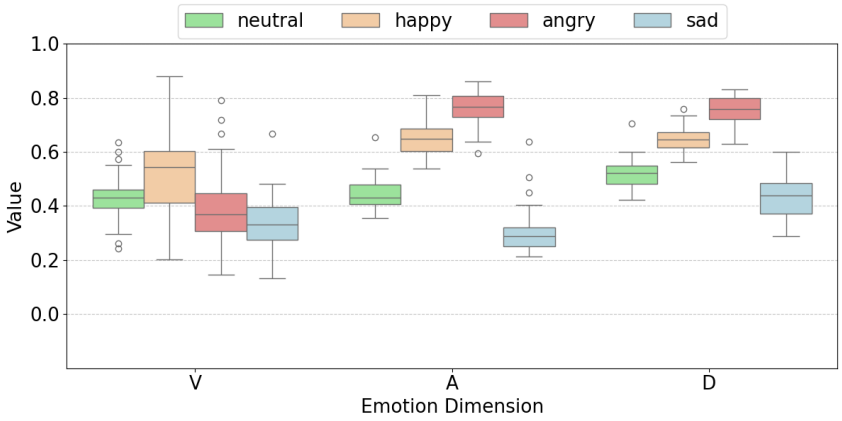

Figure 4. Visualization of 300 speech samples from ground truth recordings across neutral, happy, angry, and sad categories, showing their Valence (V), Arousal (A), and Dominance (D) values. These serve as a reference baseline for analyzing the variation trends.

To achieve precise and flexible control over emotion intensity, this work utilizes an emotion arithmetic-based approach [46]. The core idea of this approach is to build a speech synthesis system that can precisely control emotion intensity by combining pretraining and fine-tuning.

Formally, let $\theta_{pre} \in \mathbb{R}^d$ denote the weights of the pretrained model, and let $\theta_e \in R^d$ represent the weights of the fine-tuned model for a specific emotion e. The emotion vector $\phi_e \in \mathbb{R}^d$ is then computed element-wise as:

$$\phi_e = \theta_e - \theta_{\text{pre}}, \tag{17}$$

To adjust emotion intensity, the emotion vector ϕ_e is incorporated into the base model θ through element-wise addition, controlled by a scaling factor α:

$$\theta_{\text{new}} = \theta + \alpha\phi_e, \tag{18}$$

By modifying α, the model can enhance or attenuate the emotion intensity in synthesized speech. Specifically, when $\alpha = 1$, the new model θ_{new} becomes equivalent to the fine-tuned model for emotion e, the generated speech exhibits the exact emotional characteristics of the fine-tuned model for emotion e. Conversely, when $\alpha < 1$, the intensity of the emotional expression is gradually reduced, and for $\alpha > 1$, the emotional attributes are amplified.

To ensure robust emotion intensity control, the model is initially pretrained on our emotional speech dataset, obtaining a generalized set of pretrained weights θ_{pre}. This pretraining process covers multiple emotion types, such as happiness, sadness, and anger, allowing the model to capture a broad range of emotionally expressive speech features. After pretraining, fine-tuning is conducted for each emotion separately, using emotion-specific datasets. This step yields optimized model weights θ_e, ensuring the model can synthesize speech with accurate emotional expressiveness. The difference between the fine-tuned model weights θ_e and the pretrained model weights θ_{pre} is defined as the emotion vector ϕ_e, which captures the influence of a specific emotion on speech characteristics and provides a controllable parameter for emotion expression.

For flexible and precise control, we generate modified model parameters θ_e^α using:

$$\theta_e^\alpha = \theta_{pre} + \alpha\phi_e. \tag{19}$$

Unlike previous studies that control emotion intensity by training additional emotion intensity predictors, this work adopts an emotion arithmetic-based approach to achieve a more precise and efficient modulation of emotion strength. Taking the WET model [58] as an example, its emotion intensity modeling relies on training a regression network based on the relative attributes method, which learns the relative strength relations from a large number of speech samples and generates intensity labels as conditional inputs. Although effective, this strategy requires additional data annotation, model design, and optimization during training, significantly increasing system complexity and computational cost. In comparison, our method directly modulates emotion intensity at the parameter level without introducing such auxiliary networks, thereby achieving precise control over emotional strength while maintaining lower system complexity.

3. Experiments

3.1. Data

3.1.1. Training Data

The training of high-quality TTS models benefits from a sufficient amount of well-curated data. However, the recording and manual annotation of emotional speech are resource-intensive, making the construction of adequately sized, high-quality emotional datasets challenging. Compared to discrete emotion labels or emotion transfer techniques, our approach leverages an emotion recognition model to extract emotion dimensions and applies an emotion-adaptive coordinate transformation to generate an EASV as the emotion style representation.

In this study, we follow a data curation strategy inspired by [30], filtering and compiling 200 h of high-quality English emotional speech from an initial 600 h corpus that contains abundant expressive content. The construction process consists of the following steps:

1. Emotion Classification and Filtering: We apply the pretrained emo2vec model [59] to assign categorical emotion labels and confidence scores to all audio samples. Only utterances classified as one of the seven target emotions—neutral, happy, angry, sad, disgusted, excited, and fearful—with a confidence score of 1.0 are retained.

2. Audio Quality Control: To ensure perceptual quality, we evaluate each utterance using the DNSMOS model and keep only those with an overall listening rating (OVLR) above 3.0.

3. Emotion Representation Extraction: For each utterance, we extract valence–arousal –dominance (VAD) emotion dimensions and apply the emotion-adaptive spherical transformation to derive the Emotion-Adaptive Spherical Vector (EASV) as a continuous emotional representation.

4. Segment-Level Labeling: To support fine-grained control, we apply a sliding window (0.5 s window, 0.25 s stride) to extract local EASV segments within each utterance.

5. Automatic Transcription: All retained speech samples are transcribed using the Whisper-Large pretrained speech recognition model [60] to generate accurate text–audio pairs.

The final dataset contains 200 h of transcribed emotional speech covering seven categories, with the following approximate distribution: Neutral (23.4%), Happy (14.0%), Angry (10.2%), Sad (14.2%), Disgusted (9.6%), Excited (15.6%), and Fearful (13.0%). The overview of the dataset is shown in Table 1.

Table 1. Overview of the constructed emotional speech dataset.

Voice Quality	Specification
Data Size	200 h
Language	English
Emotion Classes	Neutral, Happy, Angry, Sad, Disgusted, Excited, Fearful
Emotion Annotation Model	Emo2vec
Transcription Model	Whisper-Large

3.1.2. Evaluation Data

To evaluate the model's performance, we conducted experiments using the Emotional Speech Database (ESD), a dataset recorded by 10 native English speakers and 10 native Chinese speakers. The dataset includes 350 parallel sentences, where different speakers recorded the same textual content, producing speech samples with identical text but varying timbre, prosody, or linguistic characteristics. The dataset covers five emotional categories: neutral, happy, angry, sad, and surprised. The recordings were conducted in a controlled acoustic environment, with a total duration of approximately 30 h.

Importantly, the evaluation set consists of speakers who were unseen during model training, enabling a fair assessment of the model's zero-shot generalization ability to novel speaker identities. For evaluation, we randomly sampled 30 speech utterances from the test partition of the English subset in the ESD dataset. To extract emotion feature inputs, we applied a sliding-window approach and processed these speech samples using a pretrained emotion feature extractor to obtain emotion dimension values—valence, arousal, and dominance. These values were then converted into EASV, which were used as emotion feature inputs for the model.

3.2. Experimental Setup

3.2.1. Model Configuration

The proposed model adopts a 22-layer architecture with 16 attention heads and embedding/feed-forward network (FFN) dimensions of 1024/2048, following the DiT

architecture. For ConvNeXt V2, we use a 4-layer structure, with embedding and FFN dimensions set to 512/1024. Additionally, the text feature dimension and emotion feature dimension are set to 512 and 256, respectively, ensuring that the model effectively captures both text semantics and emotional characteristics.

For text processing, the model employs alphabets and symbols, resulting in a character embedding vocabulary size of 212. Regarding audio processing, log mel-filterbank features are extracted with a mel-frequency dimension $F = 100$, using a 24 kHz sampling rate, a hop length of 256, a window length of 1024, and an FFT size of 1024. During infilling task training, 70% to 100% of mel frames are randomly masked. In CFG training, the masked speech input is first dropped with a probability of 0.3, followed by the simultaneous dropping of both masked speech and text input with a probability of 0.2. This two-stage CFG training strategy is designed to enhance the model's ability to learn text–speech alignment.

We utilize the AdamW optimizer [61] with a peak learning rate of 7.5×10^{-5}, employing linear warm-up over 20 k updates, followed by linear decay for the remaining training process. The maximum gradient norm clipping is set to 1. The model was trained for five days on four NVIDIA A100 80 GB GPUs, using a batch size of 307,200 audio frames (approximately 0.91 h per batch) and a total of 66 k updates. Additionally, based on the pretrained model, we further split the entire emotional dataset into individual emotion-specific subsets and conducted fine-tuning. During fine-tuning, the batch size remains 307,200 audio frames, with a peak learning rate of 7.5×10^{-5}, and the model was fine-tuned for 20 k updates.

3.2.2. Hardware and Software Setup

The experiments in this study were conducted on a high-performance computing platform equipped with an AMD EPYC 7F72 24-core processor and NVIDIA A100 80 GB GPUs. The operating system was Ubuntu 20.04 LTS (64-bit). The deep learning framework used was PyTorch 2.5.1 (cu124), with the CUDA driver version set to 12.2, and the programming language was Python 3.10. This environment configuration proved to be stable and reliable, meeting the performance requirements for both the training and inference stages of the speech synthesis model, as summarized in Table 2.

Table 2. Experimental hardware and software configuration.

Component	Configuration
Processor	AMD EPYC 7F72 24-Core Processor
GPU	NVIDIA A100 80 GB
Operating System	Ubuntu 20.04 LTS (64-bit)
CUDA Version	12.2
PyTorch Version	2.5.1 + cu124
Programming Language	Python 3.10

3.3. Evaluation Metrics

3.3.1. Objective Metrics

To evaluate the model's performance, we employ objective metrics that measure both speech intelligibility and the accuracy of emotional expression. These metrics provide a quantitative basis for assessing how well the generated speech aligns with its intended content and emotional attributes:

Word Error Rate (WER): To assess the intelligibility of the generated speech, we use the Whisper-Large model to transcribe the synthesized audio and compute the Word Error Rate (WER). WER quantifies speech intelligibility by comparing the generated transcription with the target text. The WER score is expressed as a percentage, where lower values indicate higher intelligibility.

Aro-Val-Domin SIM: To evaluate the consistency of emotional expression between the generated and target speech, we compute the Arousal–Valence–Dominance (Aro-Val-Domin) sequence similarity. Specifically, we apply a sliding window approach with a window size of 0.5 s and a stride of 0.25 s to extract emotion feature sequences from both the reference and synthesized speech. By computing cosine similarity frame by frame and averaging the results, we obtain the Aro-Val-Domin SIM score. A higher score indicates that the generated speech more accurately reflects the intended emotional expression.

Emo SIM: To assess the similarity of dynamic emotional expression between the generated and reference speech, we compute the Emotion Similarity (Emo SIM) score. Specifically, we apply the emotion2vec model to extract frame-level emotion embeddings for both the synthesized and reference audio. To ensure alignment, we interpolate the embeddings so that both sequences have the same length. Cosine similarity is then computed frame by frame and averaged across all frames. A higher Emo SIM score indicates that the synthesized speech more faithfully preserves the intended fine-grained emotional trajectory.

3.3.2. Subjective Metrics

To assess the quality of the generated speech, we conducted a subjective evaluation experiment using the Mean Opinion Score (MOS), a widely used perceptual evaluation method in speech synthesis research [4,7]. Following these prior works, the MOS evaluation criteria adopted in this study are presented in Table 3.

Table 3. MOS scoring criteria.

Voice Quality	MOS Evaluation Criteria	Score
Very Good	Clear, natural voice with smooth communication.	5
Good	Mostly clear, slight difficulty due to mild noise.	4
Medium	Somewhat unclear, but communication is possible.	3
Bad	Unclear, requires repetition, with noticeable delay.	2
Very Bad	Unintelligible, very difficult to understand.	1

Five university students participated as subjects, all of whom were required to wear professional headphones during testing. Each speech sample was presented 2–3 times to ensure that participants could accurately perceive the acoustic details and provide reliable ratings. The evaluation was conducted by comparing synthesized speech from different models against the original speech as a reference baseline. The final MOS was determined by averaging the ratings assigned by all participants.

For a more comprehensive subjective assessment, we employed the following evaluation metrics:

SMOS (Speaker Similarity Mean Opinion Score): Measures the similarity between the reference audio and the generated speech, ranging from 1 (completely dissimilar) to 5 (highly similar).

NMOS (Naturalness Mean Opinion Score): Assesses the naturalness of the generated speech, ranging from 1 (very poor) to 5 (excellent).

EMOS (Emotion Mean Opinion Score): Evaluates the emotional similarity between the reference audio and the generated speech, ranging from 1 (completely dissimilar) to 5 (highly similar).

3.4. Results and Analyze

To evaluate the performance of different zero-shot emotional speech synthesis models, we compare our proposed ECE-TTS with existing approaches, as outlined below:

GenerSpeech: A high-fidelity zero-shot style transfer model capable of cross-domain speech synthesis without requiring target style data. By integrating a multi-level style

adapter, universal content adapter, and a flow-based post-processing network, GenerSpeech achieves efficient style transfer and exceptional speech quality. The model supports style adaptation across different speakers and emotions while demonstrating strong generalization ability, making it well-suited for unseen data speech synthesis tasks.

EmoSphere++: A zero-shot emotion-controllable TTS model designed for precise emotion style control and intensity adjustment. By incorporating Emotion-Adaptive Spherical Vectors and a multi-level style encoder, it effectively addresses generalization challenges across different speakers and emotions. Additionally, its Conditional Flow Matching decoder enables the synthesis of high-quality speech with enhanced emotional expressiveness.

3.4.1. Results of Emotion Style Control Experiment

Table 4 presents the MOS for speech generated by different models. The first row, "GT" represents the original audio of the selected English text, which serves as a reference for evaluating synthesized speech quality.

Table 4. Subjective evaluation results for speech naturalness, speaker similarity, and emotional expressiveness across different models. Bold values indicate the best scores among all methods for each evaluation metric.

Method	SMOS	NMOS	EMOS
GT	4.02 ± 0.06	4.07 ± 0.06	4.10 ± 0.06
Generspeech [13]	3.84 ± 0.07	3.96 ± 0.07	3.85 ± 0.08
EmoSphere++ [35]	3.91 ± 0.08	$\mathbf{3.98 \pm 0.07}$	$\mathbf{3.86 \pm 0.07}$
ECE-TTS	$\mathbf{3.93 \pm 0.07}$	$\mathbf{3.98 \pm 0.07}$	$\mathbf{3.94 \pm 0.06}$

As shown in Table 4, ECE-TTS achieves a naturalness score consistent with the EmoSphere++ model, both scoring 3.98, while surpassing GenerSpeech. Furthermore, its SMOS of 3.93 is higher than both comparison models, indicating improved speaker similarity and identity consistency in the generated speech. Notably, ECE-TTS achieves the highest EMOS of 3.94, validating that EASV enable precise emotion style control while maintaining natural prosody. These results indicate that the proposed model offers superior emotion style and speech quality control, enabling the generation of high-quality synthetic speech that is more similar to real human speech.

Beyond subjective evaluations, ECE-TTS also excels in objective metrics, as presented in Table 5. ECE-TTS outperforms Generspeech and EmoSphere++ in both aspects. It achieves a WER of 13.91, significantly lower than GenerSpeech (16.34) and EmoSphere++ (15.08), confirming its superior speech intelligibility and semantic accuracy. Moreover, ECE-TTS achieves the highest score on the Aro-Val-Domin SIM metric, with a value of 0.679, compared to 0.627 for GenerSpeech and 0.656 for EmoSphere++. This result indicates its stronger alignment with real speech in terms of valence, arousal, and dominance sequences, further validating its precision in emotion style control. Additionally, ECE-TTS achieves the best performance on the Emo SIM metric (0.594), outperforming GenerSpeech (0.563) and EmoSphere++ (0.578). This demonstrates ECE-TTS's improved consistency with target emotional categories, further validating the effectiveness of our emotion style control strategy. In summary, ECE-TTS excels in both text intelligibility and emotional consistency, establishing its advantage in high-quality emotional speech synthesis.

Table 5. Objective evaluation of different models on WER, Aro-Val-Domin SIM, and Emo SIM. Bold values indicate the best scores among all methods for each evaluation metric.

Method	WER	Aro-Val-Domin SIM	Emo SIM
GT	12.25	1	1
Generspeech [13]	16.34	0.627	0.563
EmoSphere++ [35]	15.08	0.656	0.578
ECE-TTS	**13.91**	**0.679**	**0.594**

3.4.2. Results of Emotion Intensity Control Experiment

Figure 5 illustrates the impact of scaling factor α on emotion intensity control, showing systematic changes in Valence (V), Arousal (A), and Dominance (D) across different emotions (happiness, sadness, and anger). The results confirm that increasing α effectively modulates emotional expression in a predictable manner. For happiness (Figure 5a), V, A, and D increase as α grows from 0.6 to 1.4, reflecting a stronger perception of happiness. In contrast, for sadness (Figure 5b), all three dimensions decrease, aligning with the subdued nature of sad speech. For anger (Figure 5c), V declines, while A and D rise, indicating more intense and forceful speech. These trends align with expected emotional characteristics, demonstrating that ECE-TTS effectively scales emotion intensity through α adjustment, ensuring smooth and controlled modulation of expressive speech synthesis.

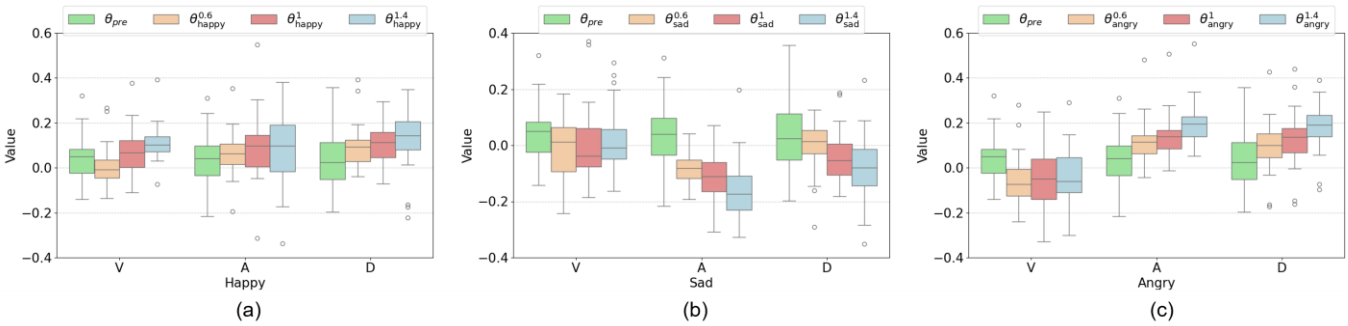

Figure 5. The variations in Valence (V), Arousal (A), and Dominance (D) values compared to ground truth, as the emotional intensity scales for different emotions. (**a**) Happy: V, A, and D increase with intensity; (**b**) Sad: V, A, and D decrease with intensity; (**c**) Angry: V decreases, while A and D increase with intensity.

Following the objective evaluation, a subjective evaluation was conducted to further validate the model's controllability of emotion intensity. Using speech data with the emotion of happiness, participants rated four speech samples on a five-point scale (1 = not happy, 5 = extremely happy). The results, as illustrated in Figure 6, show a clear positive correlation between the emotion intensity scaling factor (α) and perceived happiness ratings. Specifically, the pretrained model θ_{pre} received the lowest happiness rating, with a median score close to 2.0. As the intensity factor α increased, participants consistently perceived higher happiness levels. For example, at $\alpha = 0.6$, the median happiness rating increased to approximately 3.0, indicating a moderate perception of happiness. Similarly, at $\alpha = 1.0$, the perceived happiness continued to rise, and at $\alpha = 1.4$, it reached the highest level, with a median score close to 4.0. This demonstrates that higher α values effectively amplify the perceived emotional intensity in synthesized speech. The results demonstrate a clear correlation between increasing α and higher happiness ratings.

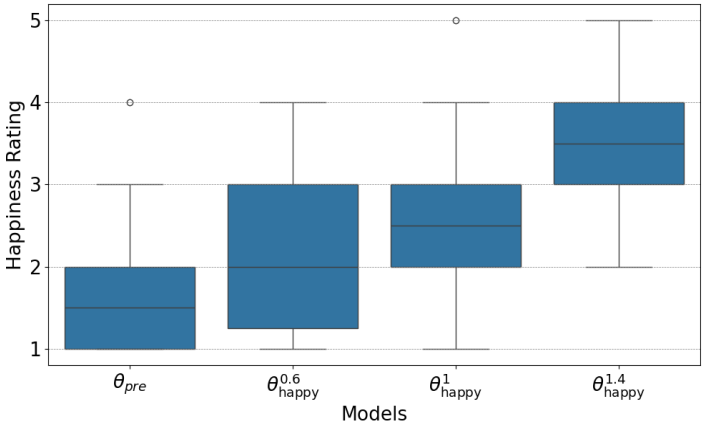

Figure 6. Subjective evaluation of happiness intensity. Higher emotion intensity scaling factor α values correspond to increased perceived happiness ratings.

The results confirm that the emotion arithmetic vector-based intensity control mechanism effectively modulates variations in V, A, and D values, enabling precise adjustment of emotional intensity. By leveraging simple arithmetic operations on model weights, this approach achieves efficient and accurate intensity modulation, offering a more straightforward and effective solution compared to traditional emotion intensity control methods.

4. Conclusions

In this study, we introduced ECE-TTS, a zero-shot emotion-controllable TTS system designed to simplify model architecture while achieving precise and flexible emotional control. ECE-TTS reduces model complexity by eliminating the reliance on phoneme alignment, duration predictors, and text encoders. Additionally, it leverages ConvNeXt V2 modules to refine text representations, improving text–speech alignment and enhancing the naturalness of synthesized speech. By integrating Emotion-Adaptive Spherical Vectors, we achieve a structured representation of continuous emotional dimensions, enabling more fine-grained and flexible emotional style control. For emotion intensity control, we adopted the emotion arithmetic vector method, which allows direct and efficient modulation of intensity without requiring additional ranking-based processes or complex embedding manipulations. Furthermore, to enhance model performance and data availability, we constructed a 200 h emotional speech dataset through an automatic annotation pipeline, eliminating the need for manual labeling while scaling the dataset size. Experimental results demonstrate that ECE-TTS outperforms existing models in both subjective (MOS) and objective (WER, Aro-Val-Domin SIM) evaluations, achieving superior speech naturalness, expressive accuracy, and robust emotion intensity control. These results validate ECE-TTS as a practical and efficient solution for zero-shot emotional speech synthesis.

While ECE-TTS demonstrates strong performance in emotional speech synthesis, its evaluation has been limited to an English dataset, restricting its validated effectiveness to monolingual settings. Future work will focus on cross-lingual adaptation, enabling ECE-TTS to generalize across diverse linguistic characteristics and broadening its applicability to various speech synthesis tasks.

Author Contributions: Conceptualization, S.L., R.Z. and Q.Y.; methodology, S.L.; software, S.L. and R.Z.; formal analysis, S.L.; investigation, S.L.; resources, R.Z.; data curation, R.Z.; writing—original draft preparation, S.L.; writing—review and editing, S.L., R.Z. and Q.Y.; visualization, Q.Y.; supervision, R.Z. and Q.Y.; project administration, R.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shen, K.; Ju, Z.; Tan, X.; Liu, Y.; Leng, Y.; He, L.; Qin, T.; Zhao, S.; Bian, J. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv* **2023**, arXiv:2304.09116.
2. Ju, Z.; Wang, Y.; Shen, K.; Tan, X.; Xin, D.; Yang, D.; Liu, Y.; Leng, Y.; Song, K.; Tang, S.; et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv* **2024**, arXiv:2403.03100.
3. Lee, S.H.; Choi, H.Y.; Kim, S.B.; Lee, S.W. Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis. *arXiv* **2023**, arXiv:2311.12454.
4. Mehta, S.; Tu, R.; Beskow, J.; Székely, É.; Henter, G.E. Matcha-TTS: A fast TTS architecture with conditional flow matching. In Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 11341–11345.
5. Wang, Y.; Zhan, H.; Liu, L.; Zeng, R.; Guo, H.; Zheng, J.; Zhang, Q.; Zhang, X.; Zhang, S.; Wu, Z. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv* **2024**, arXiv:2409.00750.
6. Anastassiou, P.; Chen, J.; Chen, J.; Chen, Y.; Chen, Z.; Chen, Z.; Cong, J.; Deng, L.; Ding, C.; Gao, L.; et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv* **2024**, arXiv:2406.02430.
7. Chen, Y.; Niu, Z.; Ma, Z.; Deng, K.; Wang, C.; Zhao, J.; Yu, K.; Chen, X. F5-tts: A fairytale that fakes fluent and faithful speech with flow matching. *arXiv* **2024**, arXiv:2410.06885.
8. Triantafyllopoulos, A.; Schuller, B.W. Expressivity and speech synthesis. *arXiv* **2024**, arXiv:2404.19363.
9. Zhao, W.; Yang, Z. An emotion speech synthesis method based on vits. *Appl. Sci.* **2023**, *13*, 2225. [CrossRef]
10. Jung, W.; Lee, J. E3-VITS: Emotional End-to-End TTS with Cross-Speaker Style Transfer. Available online: https://openreview.net/pdf?id=qL47xtuEuv (accessed on 27 April 2025).
11. Casanova, E.; Weber, J.; Shulby, C.D.; Junior, A.C.; Gölge, E.; Ponti, M.A. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 2709–2720.
12. Kim, J.; Kong, J.; Son, J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In Proceedings of the International Conference on Machine Learning, PMLR, Online, 18–24 July 2021; pp. 5530–5540.
13. Huang, R.; Ren, Y.; Liu, J.; Cui, C.; Zhao, Z. Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 10970–10983.
14. Cho, D.H.; Oh, H.S.; Kim, S.B.; Lee, S.H. EmoSphere-TTS: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech. *arXiv* **2024**, arXiv:2406.07803.
15. Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv* **2020**, arXiv:2006.04558.
16. Peebles, W.; Xie, S. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 4195–4205.
17. Lee, K.; Kim, D.W.; Kim, J.; Cho, J. Ditto-tts: Efficient and scalable zero-shot text-to-speech with diffusion transformer. *arXiv* **2024**, arXiv:2406.11427.
18. Eskimez, S.E.; Wang, X.; Thakker, M.; Li, C.; Tsai, C.H.; Xiao, Z.; Yang, H.; Zhu, Z.; Tang, M.; Tan, X.; et al. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In Proceedings of the 2024 IEEE Spoken Language Technology Workshop (SLT), Macao, China, 2–5 December 2024; pp. 682–689.
19. Liu, Z.; Wang, S.; Zhu, P.; Bi, M.; Li, H. E1 tts: Simple and fast non-autoregressive tts. *arXiv* **2024**, arXiv:2409.09351.
20. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.S.; Xie, S. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 16133–16142.
21. Lee, Y.; Rabiee, A.; Lee, S.Y. Emotional end-to-end neural speech synthesizer. *arXiv* **2017**, arXiv:1711.05447.
22. Tits, N.; El Haddad, K.; Dutoit, T. Exploring transfer learning for low resource emotional tts. In *Intelligent Systems and Applications, Proceedings of the 2019 Intelligent Systems Conference (IntelliSys), London, UK, 5–6 September 2019*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 1, pp. 52–60.

23. Tits, N.; Wang, F.; Haddad, K.E.; Pagel, V.; Dutoit, T. Visualization and interpretation of latent spaces for controlling expressive speech synthesis through audio analysis. *arXiv* **2019**, arXiv:1903.11570.
24. Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards end-to-end speech synthesis. *arXiv* **2017**, arXiv:1703.10135.
25. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerry-Ryan, R.; et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783.
26. Skerry-Ryan, R.; Battenberg, E.; Xiao, Y.; Wang, Y.; Stanton, D.; Shor, J.; Weiss, R.; Clark, R.; Saurous, R.A. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 4693–4702.
27. Triantafyllopoulos, A.; Schuller, B.W.; İyimen, G.; Sezgin, M.; He, X.; Yang, Z.; Tzirakis, P.; Liu, S.; Mertes, S.; André, E.; et al. An overview of affective speech synthesis and conversion in the deep learning era. *Proc. IEEE* **2023**, *111*, 1355–1381. [CrossRef]
28. Jing, Y.; Yang, Y.; Feng, Z.; Ye, J.; Yu, Y.; Song, M. Neural style transfer: A review. *IEEE Trans. Vis. Comput. Graph.* **2019**, *26*, 3365–3385. [CrossRef]
29. Wang, Y.; Stanton, D.; Zhang, Y.; Ryan, R.S.; Battenberg, E.; Shor, J.; Xiao, Y.; Jia, Y.; Ren, F.; Saurous, R.A. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 5180–5189.
30. Wu, H.; Wang, X.; Eskimez, S.E.; Thakker, M.; Tompkins, D.; Tsai, C.H.; Li, C.; Xiao, Z.; Zhao, S.; Li, J.; et al. Laugh Now Cry Later: Controlling Time-Varying Emotional States of Flow-Matching-Based Zero-Shot Text-To-Speech. In Proceedings of the 2024 IEEE Spoken Language Technology Workshop (SLT), Macao, China, 2–5 December 2024; pp. 690–697.
31. Inoue, S.; Zhou, K.; Wang, S.; Li, H. Hierarchical emotion prediction and control in text-to-speech synthesis. In Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 10601–10605.
32. Inoue, S.; Zhou, K.; Wang, S.; Li, H. Fine-grained quantitative emotion editing for speech generation. In Proceedings of the 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Macao, China, 3–6 December 2024; pp. 1–6.
33. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161. [CrossRef]
34. Wagner, J.; Triantafyllopoulos, A.; Wierstorf, H.; Schmitt, M.; Burkhardt, F.; Eyben, F.; Schuller, B.W. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10745–10759. [CrossRef]
35. Cho, D.H.; Oh, H.S.; Kim, S.B.; Lee, S.W. EmoSphere++: Emotion-Controllable Zero-Shot Text-to-Speech via Emotion-Adaptive Spherical Vector. *arXiv* **2024**, arXiv:2411.02625. [CrossRef]
36. Zhou, K.; Sisman, B.; Rana, R.; Schuller, B.W.; Li, H. Emotion intensity and its control for emotional voice conversion. *IEEE Trans. Affect. Comput.* **2022**, *14*, 31–48. [CrossRef]
37. Parikh, D.; Grauman, K. Relative attributes. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 503–510.
38. Zhu, X.; Yang, S.; Yang, G.; Xie, L. Controlling emotion strength with relative attribute for end-to-end speech synthesis. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 192–199.
39. Lei, Y.; Yang, S.; Xie, L. Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021; pp. 423–430.
40. Schnell, B.; Garner, P.N. Improving emotional TTS with an emotion intensity input from unsupervised extraction. In Proceedings of the 11th ISCA Speech Synthesis Workshop, Budapest, Hungary, 26–28 August 2021; pp. 60–65.
41. Lei, Y.; Yang, S.; Wang, X.; Xie, L. Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 853–864. [CrossRef]
42. Zhou, K.; Sisman, B.; Rana, R.; Schuller, B.W.; Li, H. Speech synthesis with mixed emotions. *IEEE Trans. Affect. Comput.* **2022**, *14*, 3120–3134. [CrossRef]
43. Um, S.Y.; Oh, S.; Byun, K.; Jang, I.; Ahn, C.; Kang, H.G. Emotional speech synthesis with rich and granularized control. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7254–7258.
44. Im, C.B.; Lee, S.H.; Kim, S.B.; Lee, S.W. Emoq-tts: Emotion intensity quantization for fine-grained controllable emotional text-to-speech. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Online, 7–13 May 2022; pp. 6317–6321.
45. Guo, Y.; Du, C.; Chen, X.; Yu, K. Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance. In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.

46. Kalyan, P.; Rao, P.; Jyothi, P.; Bhattacharyya, P. Emotion arithmetic: Emotional speech synthesis via weight space interpolation. In Proceedings of the Interspeech 2024, Kos Island, Greece, 1–5 September 2024; pp. 1805–1809.
47. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124–129. [CrossRef]
48. Zhou, K.; Sisman, B.; Liu, R.; Li, H. Emotional voice conversion: Theory, databases and esd. *Speech Commun.* **2022**, *137*, 1–18. [CrossRef]
49. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
50. Nguyen, T.A.; Hsu, W.N.; d'Avirro, A.; Shi, B.; Gat, I.; Fazel-Zarani, M.; Remez, T.; Copet, J.; Synnaeve, G.; Hassid, M.; et al. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv* **2023**, arXiv:2308.05725.
51. Plutchik, R. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *Am. Sci.* **2001**, *89*, 344–350. [CrossRef]
52. Ho, J.; Salimans, T. Classifier-free diffusion guidance. *arXiv* **2023**, arXiv:2207.12598.
53. Lipman, Y.; Chen, R.T.; Ben-Hamu, H.; Nickel, M.; Le, M. Flow matching for generative modeling. *arXiv* **2022**, arXiv:2210.02747.
54. Le, M.; Vyas, A.; Shi, B.; Bakhturina, E.; Kim, M.; Min, S.; Lee, J.; Mustafa, B.; Pang, R.; Tur, G.; et al. Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 14005–14034.
55. Chen, R.T.; Rubanova, Y.; Bettencourt, J.; Duvenaud, D.K. Neural ordinary differential equations. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1–13. [CrossRef]
56. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
57. Liu, S.; Su, D.; Yu, D. Diffgan-tts: High-fidelity and efficient text-to-speech with denoising diffusion gans. *arXiv* **2022**, arXiv:2201.11972.
58. Li, W.; Minematsu, N.; Saito, D. WET: A Wav2vec 2.0-Based Emotion Transfer Method for Improving the Expressiveness of Text-to-Speech. In Proceedings of the Speech Prosody 2024, Leiden, The Netherlands, 2–5 July 2024; pp. 1235–1239.
59. Ma, Z.; Zheng, Z.; Ye, J.; Li, J.; Gao, Z.; Zhang, S.; Chen, X. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv* **2023**, arXiv:2312.15185.
60. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Honolulu, HI, USA, 23–29 July 2023; pp. 28492–28518.
61. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.