CASEIN: Cascading Explicit and Implicit Control for Fine-grained Emotion Intensity Regulation

Yuhao Cui[1], Xiongwei Wang[1], Zhongzhou Zhao[1], Wei Zhou[1], Haiqing Chen[1]

[1]Alibaba Group, China

cyhbrilliant@gmail.com, xiongwei.wangxw@gmail.com, zhongzhou.zhaozz@alibaba-inc.com,
zhouwei546138922@126.com, haiqing.chenhq@alibaba-inc.com

Abstract

Existing fine-grained intensity regulation methods rely on explicit control through predicted emotion probabilities. However, these high-level semantic probabilities are often inaccurate and unsmooth at the phoneme level, leading to bias in learning. Especially when we attempt to mix multiple emotion intensities for specific phonemes, resulting in markedly reduced controllability and naturalness of the synthesis. To address this issue, we propose the CAScaded Explicit and Implicit coNtrol framework (CASEIN), which leverages accurate disentanglement of emotion manifolds from the reference speech to learn the implicit representation at a lower semantic level. This representation bridges the semantical gap between explicit probabilities and the synthesis model, reducing bias in learning. In experiments, our CASEIN surpasses existing methods in both controllability and naturalness. Notably, we are the first to achieve fine-grained control over the mixed intensity of multiple emotions.

Index Terms: Emotional TTS, fine-grained control, emotion intensity control, mixed emotions, manifold learning

1. Introduction

In controllable expressive end-to-end speech synthesis, two paradigms for expression control are utilized: *explicit control* and *implicit control*. Explicit control methods [1, 2, 3, 4, 5, 6, 7] rely on a direct representation, such as emotion class and emotion probability, that the user can manipulate. This high-level semantic representation is ideal for emotion control tasks since it accurately reflects the semantics of the control signal. Implicit control methods [8, 9, 10, 11], however, are based on an implicit representation derived from reference speech, such as style-related and prosody-related vectors. This low-level semantic representation retains more raw information, making it suitable for style transfer tasks.

Our study focuses on fine-grained emotion intensity regulation, which involves continuously adjusting the intensity of a specific emotion expression from 0% to 100%, with control over each individual phoneme. We also explore a fine-grained emotion mixing task, which enhances expressiveness by allowing the simultaneous regulation of multiple emotions on each phoneme, beyond just the primary emotion.

Zhu *et al.* [2] used Relative Attributes [12] to learn a rank function, which extracts primary emotion probability as the intensity. Other techniques [4, 5] use deep neural networks to predict emotion probability for accurate emotion synthesis. The concept of mixed emotions is introduced in psychology [13, 14], represented as a combination of other emotions, e.g., Proud=90%Happiness+45%Surprise. Zhou *et al.* [3] built upon this concept by formulating a mixed emotion intensity regulation task, predicting intensity distribution over multiple emo-
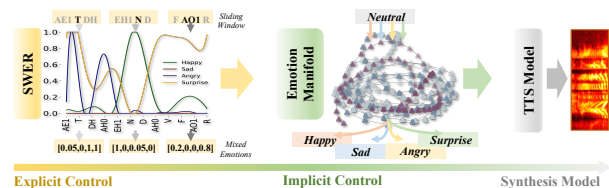

Figure 1: *The overview of cascaded control paradigms.*

tions simultaneously instead of single emotion intensity.

However, such coarse-grained emotion control techniques suffer from the limitation of inconsistent and uncontrollable preceding and following segments, especially in mixed emotion synthesis, where coarse-grained control can result in a patchwork of emotions rather than a genuine blend of them. Recent works [6, 7] refine the control granularity by using Relative Attributes to extract emotion intensity on each phoneme separately, achieving fine-grained control over the phoneme sequence and enabling more precise emotion synthesis. But our experiments showed that fine-grained control resulted in less natural speech synthesis compared to coarse-grained control. Additionally, regulating mixed emotions using fine-grained control was especially challenging due to the short duration of phonemes, resulting in inaccurate emotion intensity prediction. This difficulty in accurately capturing emotions made it challenging for the synthesis model to bridge the semantic gap and align with high-level semantic representations using explicit fine-grained control.

Fine-grained implicit control involves low-level semantic information, which reduces the semantic gap between the representation and synthesis model. However, prior researches [10, 11, 15] have primarily focused on prosody or style representations that lack direct emotion control by users. While Um *et al.* [16] demonstrated the conversion of emotion intensity to a prosody-related vector through clustering interpolation calculations, the non-linear relationship between emotion and prosody often results in imprecise control in implicit schemes.

We propose the **CAS**caded **E**xplicit and **I**mplicit co**N**trol framework, abbreviated as **CASEIN**, to enhance fine-grained emotion controllability as shown in Figure 1. Our approach introduces an implicit **Emotion Manifold** representation that implies the emotion distribution of the phoneme sequence, allowing for accurate and natural speech synthesis. Inspired by Deep Factorization [17], we disentangle emotion manifolds from reference emotional speech through Vector-Quantized VAE [18] reconstruction, using paired emotion-free speech as the "linguistic part" and speaker identity as the "speaker part". Unlike prosody- or style-related representations, our visualization experiments show that the emotion manifold has a linear relationship with emotional intensity. Furthermore, to obtain a more accurate explicit emotion distribution on the phoneme se-
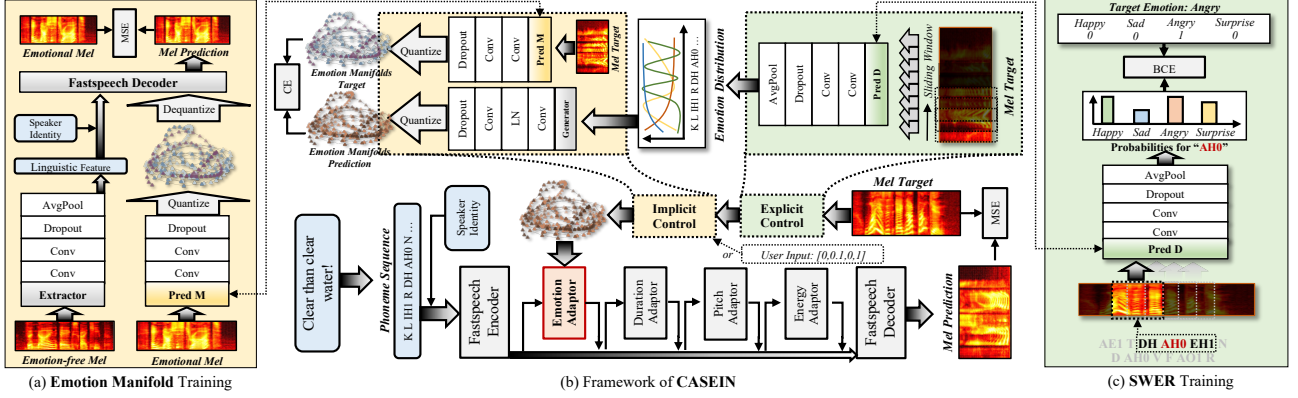
Figure 2: *The pipeline for CASEIN training. The Emotion Manifold Predictor (Pred M) and Emotion Distribution Predictor (Pred D) are trained separately and then cascaded during synthesis. The BCE, CE and MSE are loss functions.*

quence, We propose the SWER (i.e. **S**liding **W**indow **E**motion **R**ecognizer). The SWER is a convolutional neural network equipped with the sliding window mechanism, which enhances the temporal context of the input when recognizing emotions at the phoneme level, thus addressing the issue of emotions not being present within individual phonemes. By cascading the proposed SWER with the Emotion Manifold learning module, we bridge the semantic gap between the synthesis model and explicit emotion labels. Our CASEIN framework demonstrates significant advantages over reproducible SoTA methods [3, 6] in extensive experiments on fine-grained emotion intensity regulation and mixed emotions.

2. Cascaded control learning

This section is dedicated to discussing the learning process of the explicit control module Emotion Manifold (Figure 2(a)) and the implicit control module SWER (Figure 2(c)). The following section will detail the integration of these two modules in training the speech synthesis model CASEIN.

2.1. Implicit Emotion Manifold learning

Emotion manifold learning aims to extract emotion-related information from emotional speech mel-spectrograms M_e at a fine-grained level. Disentangling the linguistic and speaker parts is necessary to obtain the emotional part, following the Deep Factorization [17]. We use speaker identity embedding for the speaker part and paired[1] mel-spectrogram of emotion-free speech M_n instead of text for the linguistic part. This improves linguistic part disentanglement accuracy in the mel-spectrogram space. VQVAE [18] is employed to disentangle each part, ensuring that the learned emotion part accurately represents the original mel-spectrogram.

First, we briefly introduce the Variational Autoencoder with Vector Quantization (VQVAE) framework. Given a feature to be reconstructed X, the continuous latent variables z_c are obtained by encoding X using an encoder f_{enc}. A codebook e is constructed by setting b clusters denoted by $e^1, e^2, ..., e^b$, and clustering z_c on e yields the discrete latent variables z_d. The objective is to reconstruct z_d back to X using a decoder f_{dec}. To perform disentanglement using VQVAE, we additionally include a condition C to be reconstructed with z_d, enabling z_d to learn a distinct part from C.

In our framework, X corresponds to the mel-spectrogram M_e. We construct the manifold predictor, denoted as **Pred M**,

which consists of two convolutional layers serving as the f_{enc}[2]. By inputting M_e to Pred M and clustering, we obtain the latent variables z_d. We adopt the decoder from Fastspeech2 as f_{dec} to accelerate the convergence. The condition C includes the linguistic and speaker parts, as our objective is to disentangle emotion by removing these parts. The speaker part is obtained by embedding the speaker identity of M_e, while the linguistic part is extracted from the pronunciation of the emotion-free speech mel-spectrogram M_n. The extractor comprises two convolutional layers, followed by average pooling to retain only crucial information, such as pronunciation. Our customized VAE allows us to accurately learn the fine-grained implicit representation z_d, which is highly associated with emotion. As the learned z_d exhibits continuously varying emotional periods, as shown in the visualization section, we refer to it as **Emotion Manifold**.

2.2. Explicit SWER learning

The Sliding Window Emotion Recognizer aims to extract the emotion probability of each phoneme explicitly, which requires aligning the corresponding portion of each phoneme in M_e. Following the approach in Fastspeech2 [19], we use the Montreal forced alignment (MFA) [20] tool to obtain the boundaries for each phoneme. We then use these boundaries to slice M_e into phoneme-level M_p representations denoted by $M_p^1, M_p^2, ..., M_p^t$. However, as shown in our previous analysis, the length of the spectrum contained in M_p^i is too short, resulting in low emotion classification accuracy. To address this issue, we propose using the Sliding Window technique to increase the spectrum length of each phoneme while simultaneously smoothing the distribution of the predictions. Specifically, we obtain M_s by mapping sliding windows of radius w to each phoneme in M_p^i. The M_s^i, which represents the i-th element of M_s, is the spectrum that contains the range from M_p^{i-w} to M_p^{i+w} centered at M_p^i. During training, we treat each element in M_s as an individual sample to train the emotion classification model. The class label $p \in \mathbb{R}^n$ represents the emotion category of M_e, where n is the number of emotion categories. We introduce an emotion distribution predictor, referred to as **Pred D**, to predict the emotion probability $\tilde{p}$ of M_s^i. The predictor consists of two convolutional layers and one average pooling layer. To calculate the loss between p and $\tilde{p}$, we use element-wise binary cross-entropy. With the learned Pred D, we are able to classify each sliding window of the phoneme sequence and generate a fine-grained distribution of emotions denoted by $D \in \mathbb{R}^{t \times n}$.

---

[1]The paired refers to two speeches with the same text and speaker.

[2]To align the length of the phoneme sequence with that of the mel-spectrogram, duration pooling is applied after the last layer.
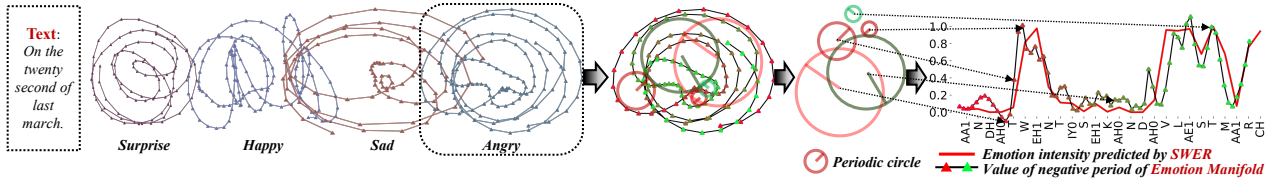
Figure 3: *Visualization of the Emotion Manifold. The left side shows the manifold compressed by PCA, and the right side visualizes the period of each point on the manifold. The circumference of a periodic circle represents the period.*

Table 1: *MCD and MOS score of the emotion restoration task. MCD is the lower the better and MOS is the higher the better.*

model	MOS ↑	MCD ↓
GT.	4.82 ± 0.05	/
MixedEmotion [3]	3.50 ± 0.21	5.66
FET [6]	3.71 ± 0.13	5.15
CASEIN (proposed)	$\mathbf{4.02 \pm 0.15}$	**4.52**

3. Framework of CASEIN

This section describes the process of cascading the pre-trained explicit and implicit control modules to control speech synthesis. Our speech synthesis pipeline, as shown in Figure 2(b), is nearly identical to Fastspeech2, with the addition of an emotion adapter that takes as input the Emotion Manifold containing rich and accurate emotion information. Assuming that the input phoneme sequence of the synthesis model is T and the synthesis target is mel-spectrogram M_e.

As the emotion manifold z_d cannot be directly extracted during inference, we use $\tilde{z}_d$ prediction as a proxy task to train the implicit control module. To obtain z_d, we input M_e to the pre-trained *Pred M* f_{predM} (described in §2.1), and cluster the output to obtain z_d. On the other hand, $\tilde{z}_d$ is generated by the emotion distribution D, which can be predicted by the explicit control module during training or manually input by the user during inference. The emotion distribution D is transformed to the emotion manifold $\tilde{z}_d$ using a two-layer convolutional generator f_{gen}. The training objective of the implicit control can be formulated as follows.

$$z_d^i = e^k, \ k = \text{argmin}_j \|f_{predM}(M_e)^i - e^j\|_2$$
$$\tilde{z}_d^i = e^k, \ k = \text{argmin}_j \|f_{gen}(D)^i - e^j\|_2$$
$$\mathcal{L}_{imp}(\tilde{z}_d, z_d) = \frac{1}{t}\sum_{i=1}^{t}\|\tilde{z}_d^i - z_d^i\|_2 \quad (1)$$

where t denotes the length of the T and e denotes the codebook.

Predicting the emotion distribution D is a crucial step in training, as it determines the accuracy of the predicted emotion manifold and the ability to control synthesis with user input during inference. Following the SWER learning, we align M_e to the phoneme boundaries of T, extract sliding windows corresponding to each phoneme, and obtain M_s. Applying the pre-trained *Pred D* f_{predD} (described in §2.2) to each window in M_s yields D (denoted by D^1, D^2, ..., D^t) as $D^i = f_{predD}(M_s^i)$.

The predicted emotion manifold $\tilde{z}_d$ is fed into the two-layer convolutional emotion adapter f_{ada} and then converged with the synthesis model f_{syn} (i.e., Fastspeech2) to generate the mel-spectrogram $\tilde{M}_e$. The synthetic loss is the same as the original Fastspeech2 and can be represented as follows.

$$\tilde{M}_e = f_{syn}(T, f_{ada}(\tilde{z}_d))$$
$$\mathcal{L}_{syn}(\tilde{M}_e, M_e) = \frac{1}{v}\sum_{i=1}^{v}\|\tilde{M}_e^i - M_e^i\|_2 \quad (2)$$

Figure 4: *ABX preference test of the emotion restoration task.*

where v denotes the length of the mel-spectrogram M_e. Eventually the loss of CASEIN training can be expressed as $\mathcal{L}_{casein} = \mathcal{L}_{syn} + \lambda\mathcal{L}_{imp}$, where λ is empirically set to 0.1.

4. Experiments

4.1. Evaluation setup

Consistent with prior research [3, 5], we trained our CASEIN framework using the English segment of the Emotional Speech Dataset (ESD) [21]. This dataset consists of 300 training texts, 20 validation texts, and 30 testing texts, each spoken by 10 English speakers expressing 5 distinct emotions: Neutral, Happy, Sad, Angry, and Surprised. For our experiment, we focused on a subset of the ESD dataset called '0015', which consists of a total of 1750 utterances from a female speaker. This subset is one of the most frequently used subsets in the ESD dataset. We extracted mel-spectrograms using parameters of 80 channels, 256 hop length and 16000Hz sampling rate.

Our training pipeline consists of three phases: emotion manifold learning, SWER learning, and CASEIN training. In the emotion manifold learning phase, we treat the Neutral as emotion-free speech and the other four emotions as emotional speech. We set the number of clusters to 256, with each cluster having a dimension of 512. The second convolutional layer of the *Pred M* module maintains the same hidden size as the dimension of the cluster, while the other convolutional layers have hidden sizes, kernel sizes, and dropout ratios of 256, 9, and 0.2, respectively. In the SWER learning phase, we specify the sliding window size of 5. In the CASEIN training phase, we employ Fastspeech2 as our base model, with weights pre-trained on LibriTTS [22]. HiFiGAN [23] is used as the vocoder for audio synthesis. We train the model using the Adam optimizer with a learning rate of 5e-4 and betas of (0.9,0.98). A linear learning rate decay scheduler is used, dropping from the maximum to 0 over 100 epochs. The entire training process takes 5 hours on a single Tesla V100 GPU.

We used objective and subjective metrics in our experiments, and 21 native speakers participated in the subjective evaluation. For objective metrics, we used Mean Cepstral Distortion (**MCD**) to estimate speech quality by calculating the MFCC of the synthesized speech and the ground truth speech. For subjective metrics, we used Mean Opinion Score (**MOS**), a 5-point scale widely used in TTS to assess speech quality. In addition, we conducted an **ABX** preference test where judges chose which speech sample was more similar to the reference in emotion expression, with three choices: (1) 1st is better; (2) 2nd is better; (3) no preference. To measure regulation ability in more detail, we proposed a Reference-Free ABX (**RF-ABX**) preference test, since the ABX metrics require reference audio, which may not always be available. In this test, judges rated

Table 2: *The intensity curve from 0% to 100% means that the intensity of the first phoneme is 0%, the last is 100%, and each intermediate phoneme is obtained by linear interpolation.*

Emotion	Intensity Curve	CASEIN	no prefer	FET
Happy	0% - 100%	**35%**	41%	24%
	100% - 0%	23%	48%	**29%**
Sad	0% - 100%	**51%**	31%	18%
	100% - 0%	**48%**	28%	24%
Angry	0% - 100%	**69%**	21%	10%
	100% - 0%	**61%**	23%	16%
Surprise	0% - 100%	**70%**	21%	9%
	100% - 0%	**59%**	29%	12%

CASEIN 52.8%	no prefer 30.2%	MixedEmo. 17.0%	*Utterance*
CASEIN 77.3%	no prefer 17.6%	MixedEmo 5.1%	*The first half*
CASEIN 69.3%	no prefer 22.2%	MixedEmo 8.5%	*The second half*

Figure 5: *RF-ABX preference test of the emotion mixing task.*

which speech sample was more consistent with the given description of emotional expression.

We compared our method with the reproducible (SoTA) baseline methods. For the fine-grained emotional intensity regulation task, we used the **FET** method proposed by Lei et al. [6]. For the emotions mixing task, we used the **MixedEmotion** method proposed by Zhou et al. [3]. While some studies [7, 5] were proposed after the baseline methods, we did not use them as baselines since they are not open source and their contributions do not overlap with ours.

4.2. Evaluation

As shown in Figure 3, we investigated the nature of the learned emotion manifold and found that it captures unique characteristics for each emotion. By extracting emotion manifolds from a pair of emotional speeches and projecting them onto a 2-dimensional space using PCA, we observed that the manifold for sadness is smoother than the manifold for happiness, demonstrating the manifold's ability to model the essence of emotions. We also observed a strong correlation between the fine-grained emotional intensity and the periodicity of the manifold. Using the shift of the tangent angle to approximate the period, we found that the negation of the computed period closely overlapped with the emotional intensity curve, with the former preserving more nuanced information. This discovery provides insight into the underlying mechanism of the emotion manifold's efficacy.

Our CASEIN model can regulate fine-grained emotion intensity while synthesizing more natural speech, as stated in section §1. To validate this, we designed a speech restoration task where only the text and explicit emotion probabilities extracted from the target speech are input to synthesize the speech. We compared the performance of our model with SoTA models using objective and subjective metrics, as shown in Table 1 and Figure 4. Our model significantly outperformed the SoTA models in all three metrics, indicating that CASEIN can regulate emotions more accurately and synthesize higher-quality speech.

Fine-grained intensity regulation: We tested CASEIN's ability to regulate emotions with single primary emotions in a fine-grained intensity task, using RF-ABX to measure emotional strengthening and attenuation curves. Four emotions were assessed, and judges chose the voice that best represented the gradual strengthening or attenuation of the emotion. Results (see Table 2) showed that CASEIN performed better for intense emotions like Angry and Surprise, with a smaller gap for mod-

Table 3: *To convert A to B, the attenuation curve of A is combined with the strengthening curve of B.*

Convert A	To B	CASEIN	no prefer	FET
Happy	Sad	**57%**	25%	18%
	Angry	**78%**	6%	16%
	Surprise	**83%**	7%	10%
Sad	Happy	**64%**	17%	19%
	Angry	**75%**	13%	12%
	Surprise	**77%**	9%	14%
Angry	Happy	**85%**	6%	9%
	Sad	**84%**	8%	8%
	Surprise	**88%**	4%	8%
Surprise	Happy	**60%**	22%	18%
	Sad	**71%**	15%	14%
	Angry	**78%**	16%	6%

erate emotions. This is because intense emotions provide more information, but accurately modeling the explicit emotion distribution at the phoneme level is difficult. CASEIN was found to be more effective in emotion strengthening than attenuation, likely due to more redundant data for emotion strengthening resulting in better emotion manifold modeling. This can be a limitation of CASEIN.

Emotion mixing: Here we demonstrate the superiority of CASEIN in fine-grained emotion mixing. Our baseline models are MixedEmotion for coarse-grained mixing and FET for fine-grained mixing. We use equal emotion probabilities for each phoneme when comparing with the coarse-grained approach to ensure fairness. The experiments are conducted on the test set of the ESD dataset, and we regulate the intensity of each emotion to synthesize new emotions[3] (i.e. Proud, Disappointed, Devastated) based on psychological theories [14]. To verify emotional consistency, we also conduct experiments on the first and second halves of the sentence. Using RF-ABX metrics, we find that our CASEIN can better express the new emotions, especially in terms of back-and-forth consistency, as shown in Figure 5. This suggests that coarse-grained methods tend to combine multiple emotions, leading to a lack of truly mixed emotions as the training objective is to express average sentiment.

Fine-grained emotion conversion: We conducted emotion conversion as the task for comparison, where we combined the intensity curves of two emotions, such as converting emotion A to emotion B by regulating the intensity curve of A to 100%-0% and that of B to 0%-100%. The results presented in Table 3 demonstrate that our advantage in mixed emotion control is more pronounced compared to primary emotion control. This suggests that explicitly modeling mixed emotions at a fine-grained level is challenging and underscores the significant contribution of our cascaded control approach.

5. Conclusions

Our proposed framework, CASEIN, integrates explicit and implicit emotion control in a cascaded manner using two modules: the Emotion Manifold for implicit emotion modeling and SWER for explicit fine-grained emotion distribution extraction. By cascading these modules, we address the inaccuracy issue of existing explicit control methods in fine-grained emotion modeling, particularly for mixed emotions. We demonstrate the effectiveness of CASEIN in fine-grained emotion regulation and mixing tasks, where it can synthesize high-quality speech and enhance fine-grained emotion control capability.

---

[3]We synthesized three new emotions, including Proud = 90%Happy + 45%Surprise; Disappointed = 70%Sad + 64%Angry; Devastated = 10%Surprise + 93%Sad.

6. References

[1] Y. Lee, A. Rabiee, and S. Lee, "Emotional end-to-end neural speech synthesizer," *CoRR*, vol. abs/1711.05447, 2017. [Online]. Available: http://arxiv.org/abs/1711.05447

[2] X. Zhu, S. Yang, G. Yang, and L. Xie, "Controlling emotion strength with relative attribute for end-to-end speech synthesis," in *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019.* IEEE, 2019, pp. 192–199. [Online]. Available: https://doi.org/10.1109/ASRU46091.2019.9003829

[3] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Speech synthesis with mixed emotions," *CoRR*, vol. abs/2208.05890, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2208.05890

[4] R. Liu, B. Sisman, and H. Li, "Strengthnet: Deep learning-based emotion strength assessment for emotional speech synthesis," *CoRR*, vol. abs/2110.03156, 2021. [Online]. Available: https://arxiv.org/abs/2110.03156

[5] Y. Guo, C. Du, X. Chen, and K. Yu, "Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance," *CoRR*, vol. abs/2211.09496, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2211.09496

[6] Y. Lei, S. Yang, and L. Xie, "Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis," in *IEEE Spoken Language Technology Workshop, SLT 2021, Shenzhen, China, January 19-22, 2021.* IEEE, 2021, pp. 423–430. [Online]. Available: https://doi.org/10.1109/SLT48900.2021.9383524

[7] Y. Lei, S. Yang, X. Wang, and L. Xie, "Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 853–864, 2022. [Online]. Available: https://doi.org/10.1109/TASLP.2022.3145293

[8] Y. Wang, R. J. Skerry-Ryan, Y. Xiao, D. Stanton, J. Shor, E. Battenberg, R. Clark, and R. A. Saurous, "Uncovering latent style factors for expressive speech synthesis," *CoRR*, vol. abs/1711.00520, 2017. [Online]. Available: http://arxiv.org/abs/1711.00520

[9] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018,* ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 5167–5176. [Online]. Available: http://proceedings.mlr.press/v80/wang18h.html

[10] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019.* IEEE, 2019, pp. 5911–5915. [Online]. Available: https://doi.org/10.1109/ICASSP.2019.8683501

[11] V. Klimkov, S. Ronanki, J. Rohnke, and T. Drugman, "Fine-grained robust prosody transfer for single-speaker neural text-to-speech," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019,* G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 4440–4444. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-2571

[12] D. Parikh and K. Grauman, "Relative attributes," in *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011,* D. N. Metaxas, L. Quan, A. Sanfeliu, and L. V. Gool, Eds. IEEE Computer Society, 2011, pp. 503–510. [Online]. Available: https://doi.org/10.1109/ICCV.2011.6126281

[13] Plutchik and Robert, *Theories of emotion.* Theories of emotion, 1980.

[14] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *Seventh International Conference on Affective Computing and Intelligent Interaction, ACII 2017, San Antonio, TX, USA, October 23-26, 2017.* IEEE Computer Society, 2017, pp. 248–255. [Online]. Available: https://doi.org/10.1109/ACII.2017.8273608

[15] X. Li, C. Song, J. Li, Z. Wu, J. Jia, and H. Meng, "Towards multi-scale style control for expressive speech synthesis," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021,* H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 4673–4677. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-947

[16] S. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H. Kang, "Emotional speech synthesis with rich and granularized control," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020.* IEEE, 2020, pp. 7254–7258. [Online]. Available: https://doi.org/10.1109/ICASSP40776.2020.9053732

[17] L. Li, D. Wang, Y. Chen, Y. Shi, Z. Tang, and T. F. Zheng, "Deep factorization for speech signal," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018.* IEEE, 2018, pp. 5094–5098. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8462169

[18] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA,* I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 6306–6315. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html

[19] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021.* OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA

[20] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017,* F. Lacerda, Ed. ISCA, 2017, pp. 498–502. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/1386.html

[21] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and ESD," *Speech Commun.*, vol. 137, pp. 1–18, 2022. [Online]. Available: https://doi.org/10.1016/j.specom.2021.11.006

[22] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019,* G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 1526–1530. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-2441

[23] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual,* H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/c5d736809766d46260d816d8dbc9eb44-Abstract.html