EmoMix: Emotion Mixing via Diffusion Models for Emotional Speech Synthesis

Haobin Tang[1,2†], *Xulong Zhang*[1†], *Jianzong Wang*[1*], *Ning Cheng*[1], *Jing Xiao*[1]

[1]Ping An Technology (Shenzhen) Co., Ltd., Shenzhen, China
[2]University of Science and Technology of China, Hefei, China

Abstract

There has been significant progress in emotional Text-To-Speech (TTS) synthesis technology in recent years. However, existing methods primarily focus on the synthesis of a limited number of emotion types and have achieved unsatisfactory performance in intensity control. To address these limitations, we propose EmoMix, which can generate emotional speech with specified intensity or a mixture of emotions. Specifically, EmoMix is a controllable emotional TTS model based on a diffusion probabilistic model and a pre-trained speech emotion recognition (SER) model used to extract emotion embedding. Mixed emotion synthesis is achieved by combining the noises predicted by diffusion model conditioned on different emotions during only one sampling process at the run-time. We further apply the *Neutral* and specific primary emotion mixed in varying degrees to control intensity. Experimental results validate the effectiveness of EmoMix for synthesizing mixed emotion and intensity control.

Index Terms: emotional speech synthesis, mixed emotions, denoising diffusion probabilistic model

1. Introduction

As the seq2seq modeling architecture continues to evolve rapidly, reference-based style transfer is an effective method for emotional speech synthesis. Recent researches follow the "say it like this" principle and explored various approaches, such as Global Style Tokens (GST) [1], as well as its updates [2]. Phoneme and Segmental-level prosody embedding are used in Generspeech [3] to capture a wide range of emotional variations across different scales. To model more disentangled prosody embedding some studies [4, 5] use an intermediate embedding extracted from a certain layer of Speech emotion recognition (SER) or automatic speech recognition (ASR) model as deep emotional features. Promising results are achieved in the aspect of emotion expressiveness but the task of achieving intensity controllable emotional TTS remains challenging. Various methods have been proposed to manipulate internal emotion representations, such as scaling [6], interpolation [7], or distance-based quantization [8]. The introduction of relative attributes [9] has been proposed in emotional TTS [10, 11] and emotion voice conversion (EVC) [12] to develop a more properly defined and calculated emotion intensity values.

However, previous methods predominantly concentrate on synthesizing a limited number of emotion types. Few frameworks have explored the correlations and interactions between different emotions. Humans are capable of experiencing approximately 34,000 distinct emotions and even multiple emotional states simultaneously [13, 14]. To address this issue, Plutchik [15] proposed a set of eight primary emotions, namely sadness, disgust, joy, fear, anger, anticipation, surprise, and trust. All other emotions can be considered as derived forms or a mix of these primary emotions. For example, the mixture of *Happy* and *Surprise* can be regarded as *Excitement*. Building a large mixed emotion data set will bring a lot of cost in practical application. Reference-based unseen style transfer may be a solution for mixed emotion synthesis, but it is also difficult to obtain a suitable mixed emotion reference audio set. MixedEmotion [16] study the modeling of mixed emotions in TTS for the first time. MixedEmotion obtains intensity values from relative attributes rank [9] to weight the emotion embedding. The desired mixture of multiple emotions is achieved by manually defining an emotion attribute vector. However, this brings a noticeable degradation in quality. EmoDiff [17] uses a soft-label guidance technique based on the classifier guidance [18] in denoising diffusion probabilistic models (DDPM) [19, 20] to synthesize controllable and mixed emotion. But this classifier must be trained on noisy audio and would be inefficient for high dimensional and unseen primary emotion conditioning.

This study introduces EmoMix, a framework for transferring emotional speech styles that employs a diffusion probabilistic model and a pre-trained speech emotion recognition (SER) model. We use emotion embedding extracted by SER model as an extra condition to enable the reverse process of diffusion model for primary emotion generation. Moreover, to overcome intensity control and mixed emotion issues, we avoid directly modeling the mixed emotion by introducing mix methods [21, 22] from image semantic mixing tasks. Specifically, the noises predicted by DDPM model based on different emotional conditions are combined to synthesize mixed emotions through only one sampling process at the run-time. The intensity is further controlled via *Neutral* and specific emotion combinations. In summary, the primary advantages of EmoMix are:

1. We use high dimensional emotional embedding extracted from reference audio by pre-trained SER model for seen and unseen primary emotion conditioning instead of the discrete emotion label used by classifier guidance.

2. EmoMix achieves mixed emotion generation in terms of a noise combination during only one sampling process at runtime, needless for directly modeling mixed emotion. Furthermore, EmoMix is can generate speech with a specified primary emotional intensity through the process of mixing *Neutral* and target primary emotions in varying proportions.

3. EmoMix can mix and control emotions without posing harm to synthesized voice quality. The generated audio samples exhibit high levels of naturalness and quality.

---

†Equal Contribution
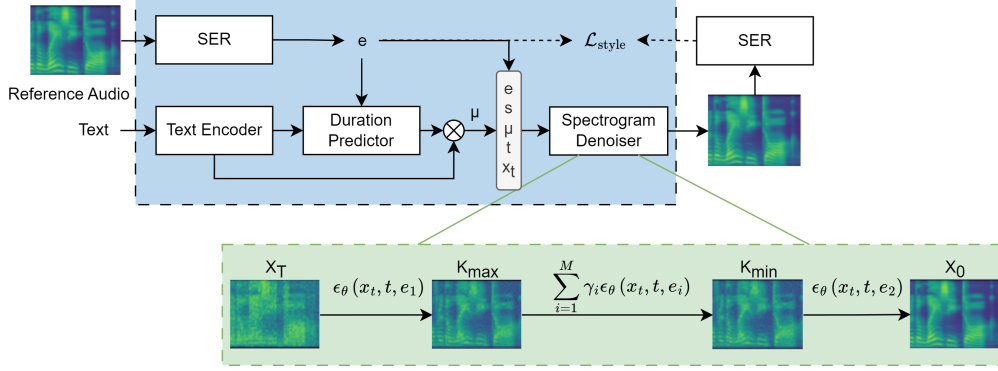*Corresponding author: Jianzong Wang, jzwang@188.com.

Figure 1: *The overview architecture for EmoMix and the green part represents the extended sampling process of mixed emotion synthesis at run-time. The dotted arrows are only used during training. The SER mdoel is pre-trained and their parameters are frozen.*

2. EmoMix

2.1. Preliminary on Score-based Diffusion Model

Denoising Diffusion Probablistic Models (DDPM) [19] has recently achieved great performance in image [21, 22] and audio [23, 24] generation tasks. We propose EmoMix, a novel method for emotion mixing in speech synthesis based on the design of GradTTS [23], which applies stochastic differential equation (SDE) [20] formulation to TTS. GradTTS defines the diffusion process, which can convert any data distribution X_0 to standard normal distribution X_T as follows:

$$dX_t = -\frac{1}{2} X_t \beta_t dt + \sqrt{\beta_t} dW_t, \quad t \in [0, T] \quad (1)$$

where β_t is the pre-defined noise schedule and W_t is the Wiener process. As Anderson proves SDE can formulate the reverse process, which follows the diffusion process's reverse trajectory:

$$dX_t = \left(-\frac{1}{2} X_t - \nabla_{X_t} \log p_t(X_t) \right) \beta_t dt + \sqrt{\beta_t} d\widetilde{W_t} \quad (2)$$

GradTTS uses the following equation, which is a discretized version of the reverse SDE, during sampling to generate data X_0 from standard Gaussian noise X_T

$$X_{t-\frac{1}{N}} = X_t + \frac{\beta_t}{N} \left(\frac{1}{2} X_t + \nabla_{X_t} \log p_t(X_t) \right) + \sqrt{\frac{\beta_t}{N}} z_t, \quad (3)$$

where N is the number of steps in the discretized reverse process, and z_t is the standard Gaussian noise. As GradTTS set T to 1, the size of one step is $\frac{1}{N}$, and $t \in \left\{ \frac{1}{N}, \frac{2}{N}, \ldots, 1 \right\}$.

Given data X_0, X_t is sampled from the distribution derived from Eq. (1) to estimate the score $\nabla_{X_t} \log p_t(X_t)$ as follows:

$$X_t \mid X_0 \sim \mathcal{N}(\rho(X_0, t), \lambda(t)) \quad (4)$$

where $\rho(X_0, t) = e^{-\frac{1}{2} \int_0^t \beta_s ds} X_0$, and $\lambda(t) = I - e^{-\int_0^t \beta_s ds}$. Eq.(3) derives the score, $\nabla_{X_t} \log p_t(X_t \mid X_0) = -\lambda(t)^{-1} \epsilon_t$, where ϵ_t is the standard Gaussian noise used to sample X_t given X_0. To estimate the score $\epsilon_\theta(X_t, t, \mu, s, e)$ is trained for $\forall t \in [0, T]$. Where μ is phoneme-dependent Gaussian mean conditioned on speaker s and emotion e. We always need text and speaker as condition and we focus on emotion throughout the paper. Therefore μ and s in $\epsilon_\theta(X_t, t, \mu, s, e)$ are omitted, and the network is expressed as a simplified notation $\epsilon_\theta(X_t, t, e)$. The following loss is used:

$$\mathcal{L}_{diff} = \mathbb{E}_{\boldsymbol{x}_0, t, \boldsymbol{e}, \epsilon_t} [\lambda_t || \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{e}) + \lambda(t)^{-1} \epsilon_t ||_2^2] \quad (5)$$

2.2. Emotion Conditioning with SER

Previous work [17] only considers single-speaker emotional TTS and uses the gradient of log-probability of classifier to guide the reverse process into a limited number of discrete emotion categories. Inspired by DeepEST [25], we leverage a pretrained SER model that produces a continuous emotion embedding e from a reference utterance that exemplifies the target emotional prosody to condition the model on the desired emotion. The SER architecture is the same as that in [26]. Initially, 3-D CNN layer takes the mel-spectrum and its derivatives as input and extracts a latent representation that encodes the emotional content while filtering out the irrelevant factors. The BLSTM and attention layer then generate an emotion embedding e that represents the utterance-level feature for emotion classification. For speaker conditioning, we use wav2vec 2.0 model [27] in Generspeech [3] to capture the speaker acoustic condition s.

As illustrated in Figure 1, EmoMix is based on GradTTS, except that the predicted duration is conditioned on emotion and speaker. Hidden representations μ reflects the linguistic contents from input text, emotion embedding e and speaker embedding s. Consequently, the spectrogram denoiser iteratively refines μ into mel-spectrograms with target primary emotion and speaker in reference audio.

We use another SER along with the denoiser to further minimize the emotion style gap between the reference speech and the synthesized one. Style loss using gram matrix is a widely used technique in computer vision [28] and has been recently applied to measure mel-spectrogram [2] and emotion embedding [29]. To maintain emotion prosody of reference audio in synthesized speech, we suggest a style reconstruction loss:

$$\mathcal{L}_{\text{style}} = \sum_j || G_j(\hat{m}) - G_j(m) ||_F^2 \quad (6)$$

where $G_j(x)$ is the gram matrix of the j-th layer feature map of the 3-D CNN in SER model for the input x. m and $\hat{m}$ denote reference mel-spectrogram and synthesized mel-spectrogram respectively. This style reconstruction loss enforces synthesized speech to have a similar style to that of the reference audio. The final training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{dur}} + \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{prior}} + \gamma \mathcal{L}_{\text{style}} \quad (7)$$

where $\mathcal{L}_{\text{dur}}$ is the ℓ_2 loss of logarithmic duration, and $\mathcal{L}_{\text{diff}}$ is the diffusion loss as Eq.(5). γ is hyperparameter and empiically set

to 1e-4. We also adopt prior loss $\mathcal{L}_{\text{prior}}$ in GradTTS to encourage converging.

2.3. Run-time Emotion Mixing

Our goal is to synthesize speech with mixed emotions or with a single primary emotion of varying intensities at the run-time. As shown in the green section of Figure 1, we extend the reverse process of the trained DDPM, which was originally conditioned on a single primary emotion in section 2.2, to synthesize mixed emotions. Mix methods [21, 22] was first used in computer vision to solve semantic mixing task which aims to modify the content in a certain part of a given object in image while preserving its layout semantics. It is believed that the sampling procedure of DDPMs first crafts coarse features and details appear last [19].

EmoMix enables emotion mixing of two different emotions by replacing the condition vector after sampling step K_{max} during inference with the purpose of overwriting the base emotion details with the mixed-in emotion. We average the emotion embedding of a set of audio samples that have same primary emotion to avoid the instability of single reference audio [30]. We first synthesize coarse base emotion prosody by denoising from Gaussian noise given the base emotion condition e_1 (e.g., *Happy*) to the intermediate steps K_{max}, followed by denoising on condition of the mixed-in emotion e_2 (e.g., *Surprise*) from K_{min} to obtain a mixture of emotion (e.g., *Excitement*). From timestep K_{max} to K_{min}, we apply the noise combine approach [22] to better preserve elements of the given base emotion and prevent it from being too easily overwritten by the mixed-in emotion. We combine the noises predicted from multiple emotion conditions to synthesize multiple emotion styles through only one sampling process. Specifically, we combine the multiple noises according to the following rule:

$$\epsilon_\theta\left(\boldsymbol{x}_t, t, e_{mix}\right) = \sum_{i=1}^{M} \gamma_i \epsilon_\theta\left(\boldsymbol{x}_t, t, e_i\right) \tag{8}$$

where γ_i is the weight of each condition e_i satisfying $\sum_{i=1}^{M} \gamma_i = 1$, which can be used for controlling the degree of each emotion. M is the number of mixed-in emotion categories and is set to 2 for dual emotion mixture. Therefore, EmoMix allows for flexible mixing of emotions in various combinations, without having to train new models with target mixed emotion condition that is directly modeled. According to Eq.(8), this sampling process can be interpreted as increasing the joint probability of the following conditional distributions:

$$\sum_{i=1}^{M} \gamma_i \epsilon_\theta\left(\boldsymbol{x}_t, t, e_i\right) \propto -\nabla_{\boldsymbol{x}_t} \log \prod_{i=1}^{M} p\left(\boldsymbol{x}_t \mid e_{\text{tar},i}\right)^{\gamma_i} \tag{9}$$

where $e_{\text{tar},i}$ is the specified target emotion condition.

Inspired by the interpolation [7] technique, we intuitively apply the *Neutral* and primary emotion mixture for emotion intensity control. By mixing the noise from the *Neutral* condition and the specified primary emotion with varying γ, we can perform interpolation between the *Neutral* and the target emotion smoothly to control the primary emotion intensity.

3. Experiment

3.1. Experiment Setup

The SER model is trained on IEMOCAP [31] which contains 10k utterances in total to obtain the emotional features e. A subset of IEMOCAP with five emotion types namely *Sad*, *Surprise*, *Happy*, *Neutral* and *Angry*. We use the English part of the ESD dataset [25] which contains 10 speaker with the same five emotions used in IEMOCAP to perform experiments. We adopt the same data partition as the ESD dataset and the *Angry* is left as unseen primary emotion for the experiments.

The noise estimation network ϵ_θ consists of the U-Net and linear attention modules that are identical to those in GradTTS. EmoMix is trained with 32 batch size and Adam optimizer at 10^{-4} learning rate for 1M steps. We provide the alignment of speech and text extracted by Montreal Forced Aligner (MFA) [32] for training duration predictor. We employed Hifi-GAN [33] as the vocoder for subsequent experiments.

3.2. Evaluation of Emotional Speech

We evaluate the quality and similarity of synthesized audio samples of EmoMix against following baseline models:

1. GT and GT (voc.): ground truth audio and wav generated from ground truth mel-spectrogram using HiFi-GAN.
2. MixedEmotion [16]: A relative attribute ranking-based model which pre-compute intensity values for training and can be manually controlled at run time.
3. EmoDiff [17]: EmoDiff is a controllable emotional TTS which can synthesize diverse emotional speech based on DDPM and guided by soft-label.

Note that in this experiment and ablation study, samples from EmoMix are conditioned on single primary emotion during reverse process and mix method is not used.

For subjective evaluation, assessors are asked to rate 20 speech per emotion for naturalness (mean opinion score; MOS) and similarity to the target emotion (similarity mean opinion score; SMOS) measures on a scale from 1 to 5 with 1 point intervals. For objective evaluation, we use mel cepstral distortion (MCD) [34] to evaluate the spectrum similarity between the reference mel-spectrum and generated features. Table 1 presents that the vocoder has negligible effect on speech quality. EmoMix surpasses baselines in SMOS with a significant margin while having comparable MOS and reaching the lowest MCD results. We observe that EmoMix still achieves satisfactory results when synthesizing unseen emotions (*Angry*). The results indicate the superiority of EmoMix over the baselines.

Table 1: *Results for single primary emotion synthesis.*

Model	MOS ↑	SMOS ↑	MCD ↓
GT	4.47 ± 0.08	4.43 ± 0.08	−
GT (voc.)	4.40 ± 0.10	4.38 ± 0.08	2.87
MixedEmotion [16]	3.61 ± 0.08	3.85 ± 0.09	6.17
EmoDiff [17]	4.08 ± 0.12	3.87 ± 0.10	5.76
EmoMix (seen)	4.10 ± 0.10	4.02 ± 0.08	5.29
EmoMix (unseen)	3.92 ± 0.10	3.82 ± 0.12	5.65

3.3. Ablation Study

To assess the contribution of the techniques used in EmoMix, including utilizing SER and the style reconstruction loss, we perform ablation studies and present the results in Table 2. To compare the quality and expressiveness of the generated speeches, comparative mean opinion score (CMOS) and comparative similarity mean opinion score (CSMOS) are used. We use GradTTS (w/ emo label) denotes GradTTS model conditioned on discrete emotion labels instead of the emotion features
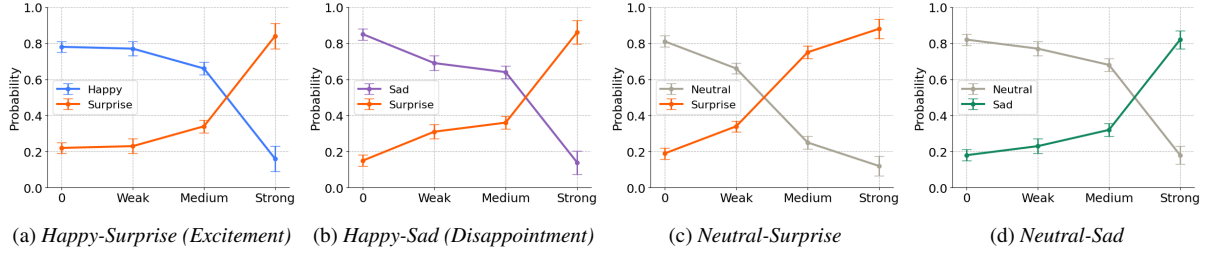
(a) *Happy-Surprise (Excitement)* (b) *Happy-Sad (Disappointment)* (c) *Neutral-Surprise* (d) *Neutral-Sad*

Figure 2: *Classification probabilities obtained from the pre-trained SER model. Each point indicates the mean probability value of 20 utterances with mixed two emotions, while the probability of other emotions is negligible and thus omitted.*

extracted by SER. The drop of quality and similarity scores indicate the importance of modeling emotion style representation in continuous space. We can also observe that the absence of style reconstruction loss lead to obvious decline in CSMOS. This indicates that the style loss can force EmoMix to synthesize target emotion in the reference audio.

Table 2: *CMOS and CSMOS comparison for ablation study.*

Model	CMOS	CSMOS
EmoMix	/	/
GradTTS (w/ emo label)	-0.05	-0.16
EmoMix (w/o $\mathcal{L}_{\text{style}}$)	-0.01	-0.09

3.4. Evaluation of Mixed Emotion

We fix K_{max} and K_{min} to be $0.6T$ and $0.2T$ respectively and vary only γ to control mixed-in emotion intensity where the larger γ indicates a stronger intensity of the mixed-in emotion. T is set to 10 in this part. It is challenging for assessors to perceive the subtle emotion differences under continuous γ settings. Therefore, we set the $\gamma \in [0.1, 0.3], [0.4, 0.6]$ and $[0.7, 0.8]$ to represent the weak, medium and strong emotion intensity level respectively. We set γ no lager than 0.8 to prevent base emotion to be completely overwritten.

Two different mixed emotions namely *Excitement* and *Disappointment* are evaluated. We use *Happy* and *Sad* as base emotion respectively and mix *Surprise* in. These mixed emotions have been studied in emotion theory [15]. We follow MixedEmotion [16] to employ the classification probabilities obtained from the softmax layer of another SER that is pre-trained on ESD dataset to analyze the performance of mixed emotions. We present the evaluation of different mixed-in intensity of each mixed emotion in Figure 2(a), (b). We notice that the probability of *Surprise* consistently rises and that of *Happy* and *Sad* decrease when gradually increase the percentage of noise corresponding to *Surprise*. These observations indicate that by varying the percentage of mixed-in emotion noise, we can easily synthesize and control the desired mixed emotion which can be recognized by a SER. Moreover, we present CMOS result of 10 types of mixed emotion in different mixed-in intensity in Table 3. The mix method of EmoMix brings slightly decreases in speech quality while that of MixedEmotion result in obvious

Table 3: *CMOS comparison for mixed emotion.*

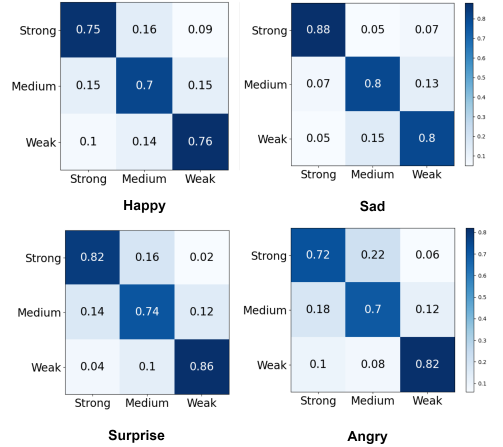| Mix Intensity	EmoMix	MixedEmotion [16]
w/o mix	/	/
Weak	-0.11	-0.23
Medium	-0.07	-0.22
Strong	-0.06	-0.19

Figure 3: *Confusion matrices of synthesized mixed emotion. The vertical and horizontal axes of each subplot indicate the ground truth and perceived emotion intensity, respectively.*

quality degradation. These results indicate that EmoMix can generate and control mixed emotional speech without compromising the quality effectively.

Then *Neutral* is further introduced as base emotion for primary emotion intensity control. An increase of mixed-in emotion similar to that seen in Figure 2(a), (b) appears in Figure 2(c), (d). This demonstrates that *Neutral* and primary emotion mixing will lead to diverse primary emotions. To further evaluate the intensity of primary emotion a human perceptual experiment was conducted. Figure 3 illustrates the confusion matrix for each emotion category obtained from the evaluation. The results clearly demonstrate that the generated samples accurately reflect the desired intensity, which confirms the effectiveness of EmoMix in controlling primary emotion intensity.

4. Conclusion

We propose a controllable emotional TTS framework, EmoMix, to further explore mixed emotion synthesis and intensity control. We avoid explicitly modeling the mixed emotion by introducing mix methods. By manually combining the noise at runtime, EmoMix could produce different emotional mixtures. Evaluations demonstrate the ability to generate speech with various mixed emotions.

5. Acknowledgement

6. References

[1] Y. Wang, D. Stanton, Y. Zhang, R.-S. 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 *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[2] S. Ma, D. Mcduff, and Y. Song, "Neural tts stylization with adversarial and collaborative games," in *International Conference on Learning Representations*, 2019.

[3] R. Huang, Y. Ren, J. Liu, C. Cui, and Z. Zhao, "Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 10 970–10 983.

[4] R. Liu, B. Sisman, G. Gao, and H. Li, "Expressive tts training with frame and style reconstruction loss," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1806–1818, 2021.

[5] T. Li, X. Wang, Q. Xie, Z. Wang, M. Jiang, and L. Xie, "Crossspeaker Emotion Transfer Based On Prosody Compensation for End-to-End Speech Synthesis," in *Proc. Interspeech 2022*, 2022, pp. 5498–5502.

[6] H. Choi and M. Hahn, "Sequence-to-sequence emotional voice conversion with strength control," *IEEE Access*, vol. 9, pp. 42 674–42 687, 2021.

[7] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7254–7258.

[8] C.-B. Im, S.-H. Lee, S.-B. Kim, and S.-W. Lee, "Emoq-tts: Emotion intensity quantization for fine-grained controllable emotional text-to-speech," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6317–6321.

[9] D. Parikh and K. Grauman, "Relative attributes," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 503–510.

[10] Y. Lei, S. Yang, X. Wang, and L. Xie, "Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 853–864, 2022.

[11] X. Zhu, S. Yang, G. Yang, and L. Xie, "Controlling emotion strength with relative attribute for end-to-end speech synthesis," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 192–199.

[12] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Emotion intensity and its control for emotional voice conversion," *IEEE Transactions on Affective Computing*, no. 01, pp. 1–1, 2022.

[13] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[14] A. Braniecka, E. Trzebińska, A. Dowgiert, and A. Wytykowska, "Mixed emotions and coping: The benefits of secondary emotions," *PloS one*, vol. 9, no. 8, p. e103940, 2014.

[15] R. Plutchik and H. Kellerman, *Theories of emotion*. Academic Press, 2013, vol. 1.

[16] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Speech synthesis with mixed emotions," *IEEE Transactions on Affective Computing*, 2022.

[17] Y. Guo, C. Du, X. Chen, and K. Yu, "Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[18] P. Dhariwal and A. Q. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 8780–8794.

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[20] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021.

[21] J. H. Liew, H. Yan, D. Zhou, and J. Feng, "Magicmix: Semantic mixing with diffusion models," *arXiv preprint arXiv:2210.16056*, 2022.

[22] G. Kim, T. Kwon, and J. C. Ye, "Diffusionclip: Text-guided diffusion models for robust image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2426–2435.

[23] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8599–8608.

[24] R. Huang, Z. Zhao, H. Liu, J. Liu, C. Cui, and Y. Ren, "Prodiff: Progressive fast diffusion model for high-quality text-to-speech," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2595–2605.

[25] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 920–924.

[26] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[28] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision (ECCV) Part II 14*, 2016, pp. 694–711.

[29] T. Li, S. Yang, L. Xue, and L. Xie, "Controllable emotion transfer for end-to-end speech synthesis," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.

[30] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5734–5738.

[31] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Lang. Resour. Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[32] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[33] 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*, vol. 33, 2020, pp. 17 022–17 033.

[34] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1. IEEE, 1993, pp. 125–128.