ZET-Speech: Zero-shot adaptive Emotion-controllable Text-to-Speech Synthesis with Diffusion and Style-based Models

Minki Kang[1,2], Wooseok Han[1*], Sung Ju Hwang[2], Eunho Yang[1,2]*

AITRICS[1], KAIST[2]

{zzxc1133,hwrg}@aitrics.com,   {sjhwang82,eunhoy}@kaist.ac.kr

Abstract

Emotional Text-To-Speech (TTS) is an important task in the development of systems (e.g., human-like dialogue agents) that require natural and emotional speech. Existing approaches, however, only aim to produce emotional TTS for seen speakers during training, without consideration of the generalization to unseen speakers. In this paper, we propose ZET-Speech, a zero-shot adaptive emotion-controllable TTS model that allows users to synthesize any speaker's emotional speech using only a short, neutral speech segment and the target emotion label. Specifically, to enable a zero-shot adaptive TTS model to synthesize emotional speech, we propose domain adversarial learning and guidance methods on the diffusion model. Experimental results demonstrate that ZET-Speech successfully synthesizes natural and emotional speech with the desired emotion for both seen and unseen speakers. Samples are at https://ZET-Speech.github.io/ZET-Speech-Demo/.

Index Terms: Text-to-Speech Synthesis, Emotional TTS

1. Introduction

Modern Text-To-Speech (TTS) systems can now synthesize high-quality speech of a single or multiple speakers [1, 2, 3]. Beyond synthesizing the speech of seen speakers during training, several works have proposed zero-shot adaptive TTS models based on style-based generators [4, 5, 6] and diffusion models [7, 8], which can generalize to unseen speakers' voice as well. However, the existing methods in adaptive TTS face significant limitations when applied to emotional speech generation, whose objective is to synthesize the speech with the desired emotion, in zero-shot scenarios. Specifically, existing approaches for zero-shot adaptive TTS do not consider the emotion-controllable generation, while the most of existing approaches for emotional TTS [9, 10, 11, 12, 13, 14] do not take zero-shot scenarios into account. These limitations necessitate the development of new methods that can effectively address both requirements in TTS systems, as depicted in Figure 1.

In this work, we tackle the aforementioned limitations in building zero-shot adaptive emotion-controllable TTS systems. We assume that the emotional speeches of multiple speakers with the emotion label are given in the training stage, but only the neutral speech of the target speaker is available in the inference stage. In the context of training, we empirically find that a naive application of the existing emotion control methods [9, 15] is ineffective since the emotion feature and speaker identity are highly entangled in the style vector used by the style-based generator (See Figure 3 for qualitative analysis). In order to overcome this limitation, we use domain adversarial training [16] to dis-
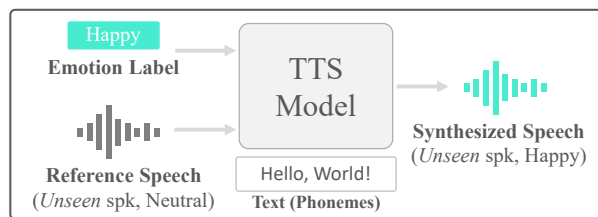
---

*: Equal Contribution

Figure 1: ***Concept.*** *Our objective is to build a zero-shot adaptive TTS model capable of synthesizing emotional speech for both seen and unseen speakers based on emotion labels and neutral reference speech of the target speaker.*

entangle the emotional content from the style vector and make the style-based generator solely pay attention to the specified emotion condition. We further suggest using the guidance methods [17, 18] on the diffusion model in the inference stage, to improve the model's ability to synthesize the emotional speech of the target speaker conditioned on the emotion label.

Combining the above ideas, we propose a **Z**ero-shot adaptive **E**motion-controllable **T**TS model, **ZET-Speech**, that can synthesize the emotional speech of unseen speakers only with their neutral speech. In experiments, we empirically validate that our method enables the zero-shot adaptive TTS model to successfully perform the emotional TTS task.

Our contributions are summarized as follows:

- We propose a zero-shot adaptive emotion-controllable TTS model named *ZET-Speech* which can produce emotional speech given the neutral reference speech of the any speaker and the emotion condition.

- Our proposed approach involves domain adversarial training on the style vector, which enables the TTS model to be controlled by the provided emotional condition. Moreover, we leverage guidance methods on the diffusion model to enhance the level of emotional expression in the synthesized speech.

- Our method demonstrates favorable performance in generating emotional speech, as evidenced by results on the Korean [19, 20] and English [21, 22] speech dataset.

2. Method

As illustrated in Figure 1, the zero-shot adaptive emotion controllable Text-to-Speech (TTS) synthesis task aims to generate the mel-spectrograms of emotional speech $\hat{Y}$ having the target emotion, given the following inputs: the text (phonemes) x, the reference speech Y, and the emotion label e. In this work, we focus on TTS models with the style-based generator [23, 6] and the diffusion model [17]. As shown in Figure 2, our ZET-Speech consists of two ideas upon the TTS models: (1) domain adversar-
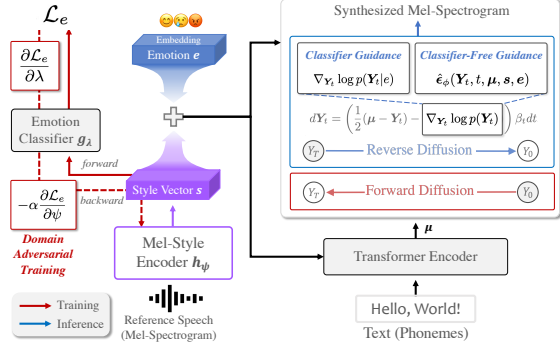
Figure 2: **Method Overview.** *(Left) During training, we use domain adversarial training to separate emotion from style vectors. (Right) In inference, we use guidance methods on the diffusion model to generate speech with more emotional features.*

ial training on the style-based generator in training (§ 2.2) and (2) guidance methods on the diffusion model in inference (§ 2.3). We first briefly recapitulate the basic components in Section 2.1. Then, we introduce our proposed method in Section 2.2, 2.3

2.1. Preliminary: Style-based Generator and Diffusion

A style-based generator [6, 24] consists of two components: the transformer encoder f_θ and the mel-style encoder h_ψ. First of all, the mel-style encoder h_ψ takes a mel-spectrogram of the reference speech Y as input and outputs the style vector $s = h_\psi(Y)$. The style vector s plays an important role in zero-shot adaptive TTS by embedding the reference speech of any speaker into the latent space. Then, the transformer model f_θ takes a phoneme sequence x and the style vector s as input and outputs the reconstructed mel-spectrogram $\mu = f_\theta(x, s)^1$.

Moreover, some previous works [2, 7] propose to generate high-fidelity speech with the diffusion models based on μ from the generator. Specifically, given μ, the reverse diffusion process generates a speech by denoising the noise drawn from the gaussian distribution $Y_T \sim \mathcal{N}(\mu, I)$. For the reverse diffusion, we define the differential equation as follows [2, 17]:

$$dY_t = \left(\frac{1}{2}(\mu - Y_t) - \nabla_{Y_t} \log p(Y_t)\right)\beta_t dt, \quad (1)$$

where the neural network $\epsilon_\phi(Y_t, t, \mu, s)$ estimates the score function of the noisy data distribution $\nabla_{Y_t} \log p_t(Y_t)$ given timestep t, the noisy data Y_t at t, noise prior μ, and the style vector s [7]. For emotion-controllable TTS, we additionally input *emotion vector e* into both the hierarchical transformer encoder $f_\theta(x, s, e)$ and the score estimator $\epsilon_\phi(Y_t, t, \mu, s, e)$.

2.2. Domain Adversarial Training

The conventional approach for incorporating emotional information into the TTS model is to concatenate or add the emotion vector to the style vector, and then forward the combined vector to the style-based generator [15]. However, this approach has proven to be ineffective in style-based generators, primarily due to the fact that the style vector s contains both the speaker identity and emotional features of the reference speech. As a result, the naive method of combining the emotion vector with the style vector fails to isolate the emotional features of the reference speech and also fails to incorporate the desired emotion

---

[1]For brevity, we regard the variance adaptor for the pitch, energy, and duration prediction as a part of the transformer encoder.

condition, leading to suboptimal emotional TTS performance. For instance, as depicted in Figure 1, assume the case where the neutral reference speech of the unseen speaker is given with the desired emotion label 'happy'. If the emotional feature is entangled in the style vector, the TTS model generates neutral speech conditioned on the neutral emotion inside the reference speech, ignoring the given emotion condition 'happy'. Hence, it is of utmost importance to disentangle the speaker identity and emotional features from the style vector, so that the style vector solely encompasses speaker-related information.

To this end, we use domain adversarial training [16] to make the mel-style encoder separate the emotional features of the reference speech from the style vector. Specifically, we introduce an emotion classifier g_λ which is trained to predict the emotion label of the speech given the style vector $p(e|s; \lambda)$, but we insert the gradient reversal layer prior to the emotion classifier which scales the gradients by the hyperparameter $-\alpha$. Therefore, given the emotion classification loss function $\mathcal{L}_e = -\log p(\hat{e}|s; \lambda)$ with the emotion label $\hat{e}$, the gradient for the mel-style encoder parameter ψ with regards to the emotion classification loss $\mathcal{L}_e$ becomes $-\alpha\frac{\partial \mathcal{L}_e}{\partial \psi}$. By employing DAT, the mel-style encoder embeds the speech into the style vector, which contains much less emotional information than it would have otherwise.

2.3. Guidance Methods on the Diffusion Model

Thanks to domain adversarial training on the style vector, we can synthesize emotional speech conditioned on the provided emotion label. The remaining issue is to make the TTS model produce more emotional speech conditioned on the given emotion. Recently, diffusion models have demonstrated state-of-the-art performance in conditional generation [25]. The guidance methods – classifier [26] and classifier-free guidance [18] – which allow the diffusion model to generate better-conditioned data are at the heart of a powerful conditional generation performance.

We incorporate guidance methods so that the diffusion model generates more emotional speech that is controlled by the given condition. For classifier guidance, we train the emotion-unconditional score estimator $\epsilon_\phi(Y_t, t, \mu, s)$ and an emotion classifier on the noisy mel-spectrograms $p(e|Y_t)$ afterwards [27]. Then, we add the classifier gradient into the sampling process:

$$dY_t = \left(\frac{1}{2}(\mu - Y_t) - \nabla_{Y_t} \log p(Y_t|e)\right)\beta_t dt, \quad (2)$$

$$\nabla_{Y_t} \log p(Y_t|e) = \gamma * \nabla_{Y_t} \log p(e|Y_t) + \nabla_{Y_t} \log p(Y_t),$$

where e is the emotion label and γ is a hyperparameter for guidance control.

With regard to classifier-free guidance, we randomly replace the emotion embedding with a null embedding $\varnothing$ in training. The important point to consider is that $\mu = f_\theta(x, s, e)$ also contains emotional information. Therefore, we should also forward $\mu_\varnothing$ to estimate the unconditional noise. We forward null embedding into a transformer encoder to generate $\mu_\varnothing = f_\theta(x, s, \varnothing)$. We then combine the conditional and unconditional score estimation to perform sampling with classifier-free guidance. Formally,

$$\hat{\epsilon}_\phi(Y_t, t, \mu, s, e) = \quad (3)$$
$$\epsilon_\phi(Y_t, t, \mu, s, e) + \gamma(\epsilon_\phi(Y_t, t, \mu, s, e) - \epsilon_\phi(Y_t, t, \mu_\varnothing, s, \varnothing)),$$

where γ is a hyperparameter for guidance control. Then, we modify the differential equation in Equation 1 with score estimation in Equation 3 as follows:

$$dY_t = \left(\frac{1}{2}(\mu - Y_t) - \hat{\epsilon}_\phi(Y_t, t, \mu, s, e)\right)\beta_t dt. \quad (4)$$

Table 1: ***Objective evaluation on Seen speakers (Korean).*** *Experimental results on speakers where the models are trained on. Note that the classifier guidance is not applicable along with the Global Style Token.*

Model	+ Emotion Label			+ Global Style Token (GST)		
	ECA (↑)	CER (↓)	SECS (↑)	ECA (↑)	CER (↓)	SECS (↑)
Grad-StyleSpeech	$16.35_{\pm 8.40}$	$16.84_{\pm 11.28}$	**0.778**	$12.54_{\pm 8.20}$	$17.60_{\pm 11.65}$	**0.778**
+ Domain Adversarial Training (DAT)	$31.27_{\pm 15.78}$	**$11.93_{\pm 9.48}$**	0.757	$24.44_{\pm 16.94}$	**$10.99_{\pm 9.06}$**	0.746
+ Classifier Guidance (CG)	$28.73_{\pm 14.04}$	$21.51_{\pm 20.08}$	0.765	-	-	-
+ Classifier-free Guidance (CFG$_{\gamma=1.75}$)	$19.68_{\pm 10.12}$	$26.38_{\pm 18.15}$	0.776	$13.81_{\pm 9.15}$	$23.21_{\pm 16.33}$	0.776
Ours						
ZET-Speech *(CG)*	**$51.59_{\pm 16.45}$**	$16.22_{\pm 16.97}$	0.759	-	-	-
ZET-Speech *(CFG$_{\gamma=1.25}$)*	$41.75_{\pm 14.64}$	$13.26_{\pm 10.31}$	0.750	$33.65_{\pm 15.40}$	$11.84_{\pm 9.66}$	0.770
ZET-Speech *(CFG$_{\gamma=1.75}$)*	$49.84_{\pm 17.33}$	$16.33_{\pm 11.74}$	0.747	**$38.89_{\pm 13.81}$**	$15.95_{\pm 14.68}$	0.757

Table 2: ***Objective evaluation on Unseen speakers (Korean).*** *We use unseen speakers where the models are **not** trained on.*

Model	+ Emotion Label		+ GST	
	ECA (↑)	CER (↓)	ECA (↑)	CER (↓)
Grad-StyleSpeech	$18.29_{\pm 8.59}$	$10.20_{\pm 9.44}$	$13.72_{\pm 7.26}$	**$9.98_{\pm 9.85}$**
+ DAT	$26.01_{\pm 12.99}$	**$9.32_{\pm 9.79}$**	$23.71_{\pm 13.28}$	$10.15_{\pm 9.35}$
+ CG	$23.29_{\pm 13.65}$	$17.83_{\pm 21.78}$	-	-
+ CFG$_{\gamma=1.75}$	$17.86_{\pm 9.56}$	$15.74_{\pm 13.67}$	$12.43_{\pm 9.40}$	$15.03_{\pm 14.35}$
Ours				
ZET-S. *(CG)*	**$39.86_{\pm 16.57}$**	$12.15_{\pm 15.29}$	-	-
ZET-S. *(CFG$_{\gamma=1.25}$)*	$31.57_{\pm 14.18}$	$11.26_{\pm 10.66}$	$31.29_{\pm 15.61}$	$10.50_{\pm 8.75}$
ZET-S. *(CFG$_{\gamma=1.75}$)*	$34.57_{\pm 16.56}$	$14.01_{\pm 12.21}$	**$32.57_{\pm 16.67}$**	$13.57_{\pm 13.78}$

Table 3: ***Objective evaluations on Unseen speakers (English).*** *We use unseen speakers (LibriTTS Test [22]) for evaluations.*

Model	ECA (↑)	CER (↓)
GSS	$40.50_{\pm 14.82}$	**$13.55_{\pm 9.69}$**
ZET-Speech *(CFG$_{\gamma=1.5}$, ours)*	**$59.75_{\pm 15.65}$**	$21.96_{\pm 13.86}$

ours). Our ZET-Speech model with classifier guidance ($\gamma = 50$). **(6) ZET-Speech (CFG, ours).** Our ZET-Speech model with classifier-free guidance ($\gamma = 1.25, 1.75$).

Evaluation Metric. For objective evaluation, we introduce three metrics. The first one is **Emotion Classifier Accuracy (ECA)** which measures how emotional the synthesized speech is. Specifically, we utilize a pre-built emotion classification model based on the Audio Spectrogram Transformer (AST) [30] and predict the emotion of the synthesized speech. If the AST model successfully predicts the target emotion, we regard this as a successful case of emotional TTS. We measure **Character Error Rate (CER)** to evaluate the quality of the synthesized speech. We leverage Whisper [31] to recognize the text within the synthesized speech. To check whether the synthesized speech reflects the target speaker's identity, we measure **Speaker Embedding Cosine Similarity (SECS)**. For SECS, we measure the similarity between vectors of the synthesized and gold emotional speech. To embed speech into the vector, we use the speaker encoder from the resemblyzer library[3]. If there is no gold emotional speech available, we cannot measure the SECS (unseen speaker case). For subjective evaluation, we use Mean Opinion Score (MOS) to measure the emotional naturalness of the synthesized speech. For MOS, we require the evaluator to assess the degree to which the given emotion is naturally reflected in the synthesized speech. We also measure speaker Similarity MOS (SMOS) for seen speakers, by requesting the evaluator to check the voice similarity between the synthesized speech and the gold emotional speech. For all experiments, we use neutral speech as a reference speech and give emotion conditions as a hard label (Emotion Label) or sampled emotional speech from the training dataset (Global Style Token). We use 9 seen speakers and 10 unseen speakers with 10 scripts and 7 emotions for evaluations. We recruit 20 evaluators for subjective evaluations.

3. Experiment

3.1. Experimental Setup

3.1.1. Implementation Details

We use Grad-StyleSpeech [7] as the base model due to its promising performance based on the style-based generator and diffusion model. We start from the TTS model pre-trained on a neutral-only multi-speaker dataset. Then, we train the pre-trained TTS model on the emotional speech dataset for 200k steps with the batch size of 8 on a single TITAN RTX GPU, using an Adam optimizer with a learning rate of 1e-3 and a linear decay.

3.1.2. Dataset

We mainly target Korean speech synthesis since there are high-quality Korean speech datasets available from the public organization AIHub[2]. We pre-train the model on the large-scale multi-speaker Korean [28] or English [22] dataset. For Korean, we train the model on the multi-emotional dataset [19]. To evaluate the unseen speaker performance, we use a separate Korean dataset [20] not used in training. For English, we use ESD [21] dataset for training and LibriTTS Test [22] for evaluation. We downsample all speech to 16kHz and use HiFi-GAN [29] as a vocoder.

3.1.3. Evaluation Setup

In this subsection, we provide detailed evaluation setups.
Baselines. We use Grad-StyleSpeech [7] as the base model for all baselines for a fair comparison. **(1) Grad-StyleSpeech (GSS).** Vanilla Grad-StyleSpeech model where we input the emotion embedding from the hard emotion label or global style token [9] as input [15]. **(2) GSS + Domain Adversarial Training.** Only applying domain adversarial training on the speaker vector. **(3) GSS + Classifier Guidance.** Only applying classifier guidance on the diffusion model similar to EmoDiff [27]. **(4) GSS + Classifier-free Guidance.** Only applying classifier-free guidance on the diffusion model. **(5) ZET-Speech (CG,**

3.2. Experimental Results

In Table 1 and 2, we present the performance of models in terms of objective evaluation metrics for seen and unseen speakers, respectively. Results show that our ZET-Speech outperforms baselines in terms of ECA, demonstrating that ZET-Speech is able to synthesize the emotional speech conditioned on the given emotion. These findings suggest that our method can be successfully applied to any speaker even if there is only a limited amount of neutral speech available. To elaborate, Domain Adversarial

---

[2]https://aihub.or.kr/

[3]https://github.com/resemble-ai/Resemblyzer

Table 4: ***Subjective evaluations (Korean).*** *Experimental results on seen and unseen speakers with human evaluation.*

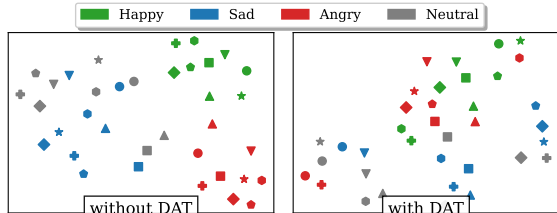|	Seen		Unseen
Model	MOS	SMOS	MOS
GSS	$2.60_{\pm.13}$	$3.11_{\pm.15}$	$2.63_{\pm.11}$
ZET-Speech *(CG, ours)*	$3.02_{\pm.13}$	$3.12_{\pm.14}$	$3.06_{\pm.13}$
ZET-Speech *(CFG, ours)*	$\mathbf{3.44}_{\pm.14}$	$\mathbf{3.22}_{\pm.13}$	$\mathbf{3.31}_{\pm.15}$

Figure 3: ***Style Vector Visualization.*** *To visually analyze the effect of Domain Adversarial Training (DAT), we plot style vectors from various speakers and emotions. The same marker denotes the same speaker, and the same color denotes the same emotion.*

Training (DAT) significantly enhances the model's ability to generate emotional speech. In other words, the model has lower sensitivity to emotion conditions without DAT. Furthermore, the ECA scores obtained by comparing ZET-Speech with the GSS + DAT baseline indicate that guidance methods are beneficial in improving the emotional expressivity of the synthesized speech.

Experimental results show that both guidance methods contribute to improving ECA scores. Although classifier guidance shows better ECA than classifier-free guidance, classifier-free guidance has exclusive merits compared to classifier guidance where it does not require an additional classifier and can be combined with other methods like global style token [9].

Our empirical findings also suggest that incorporating guidance methods may have a negative impact on the quality of the synthesized speech, resulting in a low CER and SECS metric. However, we also observe that there exists a trade-off between emotional expressivity and speech quality, which can be managed by adjusting the relevant hyperparameters γ. We provide an extensive analysis of this phenomenon in Section 3.3.

In Table 3, we briefly present the object evaluations on the English benchmark [21, 22]. We also upload speech samples in English on the demo page (`https://ZET-Speech.github.io/ZET-Speech-Demo/`).

Table 4 shows subjective evaluation results. Most evaluators found ZET-Speech more effective in conveying desired emotoin than Grad-StyleSpeech, consistent with Tables 1 and 2. Interestingly, evaluators found the speech from ours with classifier-free guidance more natural and emotional than the speech from ours with classifier guidance, which differs from Tables 1 and 2.

3.3. Analysis

Effect of Domain Adversarial Training. In Figure 3, we plot the style vectors of the randomly sampled speakers from the training dataset. We sample speech with four major emotions (Happy, Sad, Angry, and Neutral) from each speaker and use t-SNE [32] to reduce the dimension. We plot the style vectors from the model trained with and without Domain Adversarial Training (DAT). The style vectors from the model trained without DAT are clearly clustered based on emotion in Figure 3, demonstrating how the emotional features are entangled with the style vector. On the other hand, the style vectors from the model trained with DAT are more evenly distributed across the emotion and less
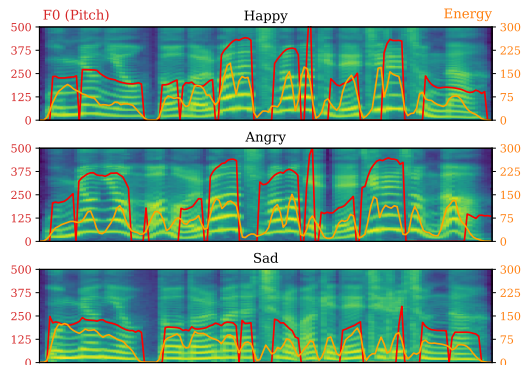

Figure 4: ***Spectrum Visualization.*** *We plot the mel-spectrograms, pitch, and energy contours of the synthesized speech varying the emotion condition, given the same script and unseen speaker.*
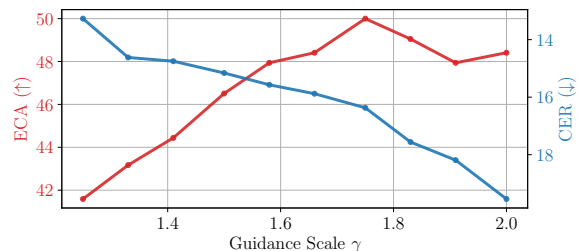

Figure 5: ***Guidance Scale Analysis.*** *In classifier-free guidance, we measure ECA and CER by varying the guidance scale γ.*

tightly clustered. The results demonstrate that emotional features are successfully disentangled from the style vector with DAT.

Spectrum Visualization. In Figure 4, we show the mel-spectrograms of the synthesized speech for different emotions, accompanied by pitch (F0) and energy contours to highlight the variations among them. The visualizations clearly show that the speech generated for Happy and Angry emotions typically has a higher pitch than that of Sad emotion, emphasizing the ability of our model to produce emotionally expressive speech with distinct variations for different emotional conditions.

Guidance Scale. Our analysis indicates that the guidance scale has a significant impact on both the quality and emotional expressiveness of synthesized speech in guidance methods. To evaluate this trade-off, we measure the Emotional Classification Accuracy (ECA) and Character Error Rate (CER) while varying the guidance scale γ from 1.25 to 2.0 in ZET-Speech with classifier-free guidance. As illustrated in Figure 5, we observe that increasing the guidance scale results in higher ECA and CER. This implies that a higher guidance scale improves emotion expressivity, but at the expense of reduced speech quality.

4. Conclusion

We proposed ZET-Speech, a zero-shot adaptive emotion controllable TTS model capable of generating emotional speech in any speaker's voice only with neutral reference speech and emotion labels. Our approach employs domain adversarial training to disentangle emotion features from the style vector during training, allowing the synthesized speech to be controlled solely by the emotion condition. Furthermore, we utilize guidance methods on the diffusion model to enhance the control of emotion in a synthesized speech. Experimental results demonstrated that our method significantly outperforms baselines in objective and subjective measures of emotional expressivity. We hope that our method enables practitioners to synthesize emotional speech for desired speakers without the need for emotional recordings.

5. References

[1] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *ICML*, 2021, pp. 5530–5540.

[2] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. A. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *ICML*, 2021.

[3] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv*, vol. abs/2301.02111, 2023.

[4] S. Ö. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *NeurIPS 2018*, 2018, pp. 10 040–10 050.

[5] Y. Wu, X. Tan, B. Li, L. He, S. Zhao, R. Song, T. Qin, and T. Liu, "Adaspeech 4: Adaptive text to speech in zero-shot scenarios," in *Interspeech 2022*. ISCA, 2022, pp. 2568–2572.

[6] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-stylespeech : Multi-speaker adaptive text-to-speech generation," in *ICML*, 2021.

[7] M. Kang, D. Min, and S. J. Hwang, "Any-speaker adaptive text-to-speech synthesis with diffusion models," *arXiv*, vol. abs/2211.09383, 2022.

[8] S. Kim, H. Kim, and S. Yoon, "Guided-tts 2: A diffusion model for high-quality adaptive text-to-speech with untranscribed data," *arXiv*, vol. abs/2205.15370, 2022.

[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 *ICML 2018*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 5167–5176.

[10] P. Wu, J. Pan, C. Xu, J. Zhang, L. Wu, X. Yin, and Z. Ma, "Cross-speaker emotion transfer based on speaker condition layer normalization and semi-supervised training in text-to-speech," *arXiv*, vol. abs/2110.04153, 2021.

[11] T. Li, X. Wang, Q. Xie, Z. Wang, M. Jiang, and L. Xie, "Cross-speaker emotion transfer based on prosody compensation for end-to-end speech synthesis," in *Interspeech 2022*. ISCA, 2022, pp. 5498–5502.

[12] R. Terashima, R. Yamamoto, E. Song, Y. Shirahata, H. Yoon, J. Kim, and K. Tachibana, "Cross-speaker emotion transfer for low-resource text-to-speech using non-parallel voice conversion with pitch-shift data augmentation," in *Interspeech 2022*. ISCA, 2022, pp. 3018–3022.

[13] S. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H. Kang, "Emotional speech synthesis with rich and granularized control," in *ICASSP 2020*. IEEE, 2020, pp. 7254–7258.

[14] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Speech synthesis with mixed emotions," *arXiv*, vol. abs/2208.05890, 2022.

[15] S. Sivaprasad, S. Kosgi, and V. Gandhi, "Emotional prosody control for speech generation," in *Interspeech 2021*. ISCA, 2021.

[16] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, pp. 59:1–59:35, 2016.

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

[18] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv*, vol. abs/2207.12598, 2022.

[19] "Aihub korean emotional speech dataset 1," https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=466, the page access may be limited in some countries.

[20] "Aihub korean emotional speech dataset 2," https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=485, the page access may be limited in some countries.

[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.

[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, pp. 1526–1530.

[23] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR 2019*,. Computer Vision Foundation / IEEE, pp. 4401–4410.

[24] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao, and T. Liu, "Adaspeech: Adaptive text to speech for custom voice," in *ICLR 2021*. OpenReview.net, 2021.

[25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR 2022*. IEEE, 2022, pp. 10 674–10 685.

[26] P. Dhariwal and A. Q. Nichol, "Diffusion models beat gans on image synthesis," in *NeurIPS 2021*, 2021, pp. 8780–8794.

[27] Y. Guo, C. Du, X. Chen, and K. Yu, "Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance," *arXiv*, vol. abs/2211.09496, 2022.

[28] "Aihub korean multi-speaker speech dataset," https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=542, the page access may be limited in some countries.

[29] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, 2020.

[30] Y. Gong, Y. Chung, and J. R. Glass, "AST: audio spectrogram transformer," in *Interspeech 2021*. ISCA, 2021, pp. 571–575.

[31] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv*, vol. abs/2212.04356, 2022.

[32] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html