KazEmoTTS: A Dataset for Kazakh Emotional Text-to-Speech Synthesis

Adal Abilbekov, Saida Mussakhojayeva, Rustem Yeshpanov, Huseyin Atakan Varol
Institute of Smart Systems and Artificial Intelligence
Nazarbayev University, Astana, Kazakhstan
{adal.abilbekov, saida.mussakhojayeva, rustem.yeshpanov, ahvarol}@nu.edu.kz

Abstract

This study focuses on the creation of the KazEmoTTS dataset, designed for emotional Kazakh text-to-speech (TTS) applications. KazEmoTTS is a collection of 54,760 audio-text pairs, with a total duration of 74.85 hours, featuring 34.23 hours delivered by a female narrator and 40.62 hours by two male narrators. The list of the emotions considered include "neutral", "angry", "happy", "sad", "scared", and "surprised". We also developed a TTS model trained on the KazEmoTTS dataset. Objective and subjective evaluations were employed to assess the quality of synthesized speech, yielding an MCD score within the range of 6.02 to 7.67, alongside a MOS that spanned from 3.51 to 3.57. To facilitate reproducibility and inspire further research, we have made our code, pre-trained model, and dataset accessible in our GitHub repository.

Keywords: dataset, emotional TTS, emotion, Kazakh, TTS

1. Introduction

The demanding challenges of generating high-quality synthesized speech for one or more speakers have been met by rapidly developed text-to-speech (TTS) systems (Shen et al., 2017; Ren et al., 2020; Arik et al., 2017). Yet, synthesized speech still faces significant difficulties in expressing paralinguistic features such as emotions.

In the area of emotional TTS, where the voice synthesized by a TTS system is to convey emotions (e.g., anger, happiness, sadness, etc.), the availability of high-quality labeled datasets remains quite limited.

As far as our knowledge extends, most publicly available emotional speech datasets primarily cover high-resource languages, such as Chinese, English, or French (Adigwe et al., 2018; Costantini et al., 2014; Busso et al., 2008). These datasets typically focus on either distinct emotional states (e.g., anger, happiness, etc.; Costantini et al. 2014) or emotion polarity, ranging from absolutely negative to absolutely positive (Cui et al., 2021). They often include several narrators' speech samples and exhibit variations in terms of total audio duration (Zhou et al., 2021; Cui et al., 2021).

In our study, we have undertaken the pioneering task of creating an emotional TTS dataset for Kazakh—a low-resource language. However, its utility extends beyond this specific domain and can be applied effectively in diverse areas, including speech emotion recognition (SER) and emotional voice conversion tasks. The dataset comprises a total of 74.85 hours of recorded high-quality speech data featuring six distinct emotional categories. The Kazakh emotional TTS (KazEmoTTS) dataset is comprised of contributions from three professional narrators, with 34.23 hours of the data provided by a female narrator and 40.62 hours by two male narrators. Additionally, we introduce a TTS model, trained on KazEmoTTS, with the capability to produce Kazakh speech reflecting six emotional expressions. KazEmoTTS and the model are openly accessible for both academic and commercial purposes, operating under the provisions of the Creative Commons Attribution 4.0 International License in our GitHub repository.[1]

The structure of the paper is as follows: Section 2 offers an overview of previous research in emotional TTS. Section 3 describes the construction of the dataset. Section 4 covers the experimental design and evaluation metrics. Section 5 provides a presentation of the experimental results and a brief summary of the main findings. Section 6 concludes the paper.

2. Related Work

Previous studies into the complex relationships and interactions between distinct emotional states suggest that individuals can potentially experience a wide array of diverse emotions (Plutchik, 2001; Braniecka et al., 2014). Plutchik and Kellerman (2013) distill a set of eight fundamental emotions, including anger, anticipation, disgust, fear, joy, sadness, surprise, and trust, with other emotional states believed to arise from various combinations of these.

That said, Paul Ekman's well-known theory of six basic emotions (Ekman, 1992) proposes the existence of anger, disgust, fear, happiness, sadness,

[1] `https://github.com/IS2AI/KazEmoTTS`

LREC-COLING 2024, pages 9626–9632
20-25 May, 2024.

and surprise, and is frequently invoked in emotional TTS research (Schröder, 2009; Zhou et al., 2020, 2022a).

Most datasets employed for emotional TTS include emotion labels, in contrast to prosody modeling approaches (Du and Yu, 2021; Guo et al., 2022b), which do not rely on preset labels. Presently, emotional TTS research primarily revolves around two methods: (1) synthesizing speech with explicit, predefined emotional labels and (2) regulating the intensity of emotions in speech synthesis.

Employing hard-labeled emotions is generally considered the most straightforward approach. Lee et al. (2017) utilized an attention-based decoder that captures an emotion label vector to generate the desired emotional style in the synthesized speech. In Kim et al. (2021), style embeddings were extracted from both a reference speech sample and a corresponding style tag.

With respect to models that allow for the control of emotional intensity, the prevailing method of determining emotional intensity is relative attributes ranking (RAR) (Parikh and Grauman, 2011). RAR involves the creation of a ranking matrix that is derived through a max-margin optimization problem typically addressed using support vector machines. The solution is subsequently employed for model training. However, it is important to note that this process is manually constructed and therefore can potentially introduce biases into the training process (Guo et al., 2022a).

In Um et al. (2019), the researchers introduced an algorithm designed to increase the gap between emotion embeddings. They also employed the interpolation of this embedding space as a means to control the intensity of emotions. In Im et al. (2022), quantization techniques were introduced to measure the distances between emotion embeddings, enabling the control of emotion intensities.

Similar methods have been applied to intensity control in emotion conversion (Choi and Hahn, 2021; Zhou et al., 2022b). However, even though an autoregressive model (Zhou et al., 2022a) that relies on intensity values derived from RAR to weigh emotion embeddings is implemented, the problem of speech quality degradation persists.

EmoDiff (Guo et al., 2022a), built on the design of GradTTS (Popov et al., 2021), introduces a soft-label guidance approach inspired by the classifier guidance technique, employed in diffusion models (Dhariwal and Nichol, 2021; Liu et al., 2021). The classifier guidance technique is a sampling method that leverages the gradient of a classifier to lead the sampling path when provided with a one-hot class label. The adoption of an alternative approach can be observed in EmoMix (Tang et al., 2023), another GradTTS-based model. This approach combines a diffusion probabilistic model with a pre-trained SER model. The emotion embeddings extracted by the SER model act as an additional condition, enabling the reverse process within the diffusion model to generate primary emotions.

3. Dataset Construction

3.1. Text Collection

Narration materials were drawn from multiple sources. Scientific, computer technology, historical, and international articles were retrieved from Kazakh Wikipedia. News content was collected from reputable Kazakh media outlets. In addition, selections from public domain books, fairy tales, and phrasebooks were included. All the collected texts were split sentence-wise.

3.2. Recording Process

We hired three professional narrators—one female and two males—for the project. The narrators were given the option to record in their personally arranged home studios or within the facilities of our institution. They were also given precise instructions to read the texts in quiet indoor settings while conveying a high degree of emotional expression in line with the specified emotions.

Each sentence was paired with one of the six emotions selected for the study, ensuring an even distribution among all sentences. Our selection of these emotions was informed by their prevalence in prior research (Zhou et al., 2021; Adigwe et al., 2018; Busso et al., 2008) and their discernibility by evaluators (Costantini et al., 2014). As a result, the list of emotions examined in our study included all those proposed in Ekman (1992), with the sole exception being "disgust". Additionally, we introduced a "neutral" option, representing the absence of a specific emotion.

All recorded audio files were either sampled at a rate of 44.1 kHz and stored as 16-bit samples or at a rate of 48 kHz and stored as 24-bit samples. The whole data collection process was facilitated by a messenger application (Telegram) bot, as depicted in Figure 1a.

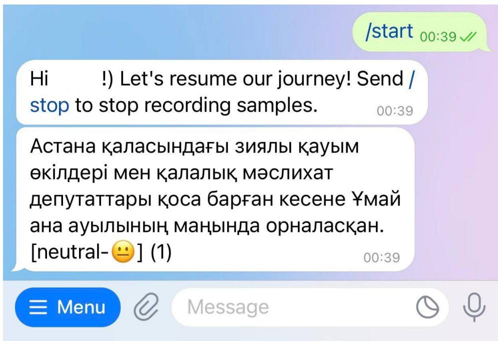
3.3. Audio-to-Text Alignment Verification

We conducted a thorough examination of the recorded audio files using a customized version of the Whisper multilingual automatic speech recognition (ASR) (Radford et al., 2022) system to assess the accuracy of the audio-to-text alignment. The ASR system generated transcriptions based on the audio files, which were subsequently compared with the original texts. Texts exhibiting a character error rate (CER) were identified and subjected to a review process by a team of

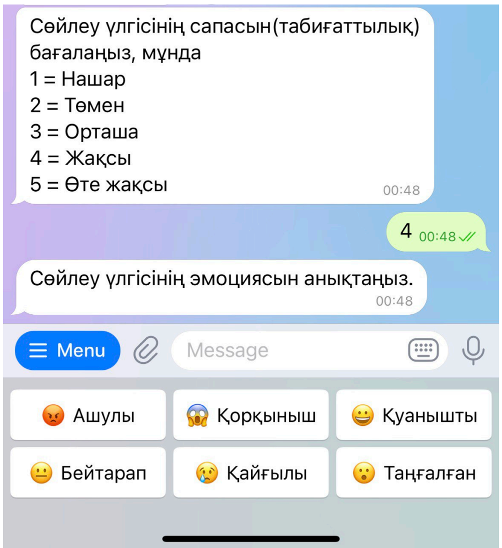
E	Count	F1				M1				M2			
		Total	Mean	Min	Max	Total	Mean	Min	Max	Total	Mean	Min	Max
neu	9,385	5.85	5.03	1.03	15.51	4.54	4.77	0.84	16.18	2.30	4.69	1.02	15.81
ang	9,059	5.44	4.78	1.11	14.09	4.27	4.75	0.93	17.03	2.31	4.81	1.02	15.67
hap	9,059	5.77	5.09	1.07	15.33	4.43	4.85	0.98	15.56	2.23	4.74	1.09	15.25
sad	8,980	5.60	5.04	1.11	15.21	4.62	5.13	0.72	18.00	2.65	5.52	1.16	18.16
sca	9,098	5.66	4.96	1.00	15.67	4.13	4.51	0.65	16.11	2.34	4.96	1.07	14.49
sur	9,179	5.91	5.09	1.09	14.56	4.52	4.92	0.81	17.67	2.28	4.87	1.04	15.81

Note. Each emotion (E) was abbreviated to its first three letters due to space constraints.

Table 1: Recording count and duration statistics: Total (hours), Mean, Max, and Min (seconds)



(a)



(b)

Figure 1: Telegram bot user interfaces: a) Narration functionality, and b) evaluation functionality.

moderators. Recordings were excluded if they contained mispronunciations or significant background noise.

3.4. Dataset Specification

The audio recordings and their corresponding transcriptions are organized into separate folders for each narrator. All audio recordings were downsampled to a rate of 22.05 kHz and saved in the WAV format with a 16-bit per sample configuration. They also underwent a preprocessing step involving the removal of silence and normalization, achieved by dividing the audio by its maximum absolute value. The transcripts were saved as TXT files encoded in UTF-8 variable-length character encoding standard. Both the audio and transcript files share identical filenames, differing only in their file extensions. Each file name comprises the narrator's identifier (ID), emotion, and utterance ID, structured as *narratorID_emotion_utteranceID*.

The dataset contains 8,794 unique sentences and 86,496 unique words with an average sentence length of 10.83 words. Initially, a total of 84,714 audio files were recorded, but following quality checks, the dataset now contains 54,760 audio recordings. These recordings collectively represent an overall duration of 74.85 hours. The duration for the female narrator (F1) is 34.23 hours, with an average segment length of 5.0 seconds. The duration for the first male narrator (M1) is 26.51 hours, with an average audio segment length of 4.8 seconds. For the second male narrator (M2), the duration is 14.11 hours, with an average segment length of 4.9 seconds. More detailed statistics for the dataset are provided in Tables 1 and 2.

Narrator	# recordings	Duration (h)
F1	24,656	34.23
M1	19,802	26.51
M2	10,302	14.11
Total	54,760	74.85

Table 2: Narrator statistics

4. Experimental Setup

4.1. KazEmoTTS Architecture

We built our TTS model based on the design of GradTTS with hard label emotions (Popov et al., 2021), as was done in Guo et al. (2022a) and Tang et al. (2023). The model was trained with the

Adam optimizer and a learning rate of 10^{-4} for 3.7 million steps on one graphics processing unit (GPU) on an NVIDIA DGX A100 machine. To improve the performance of diffusion models (Song et al., 2020), we implemented exponential moving averages for the weights of the model during training. In addition, we removed the dependency on ground truth duration data and adjusted the sampling rate to 22,050 Hz. During the inference phase, we set the guidance level parameter γ to 100. The output of the model was an array of 80-dimensional log mel-filter bank features, representing acoustic features. To transform these acoustic data into time-domain waveform samples, we utilized the HiFiGAN vocoder (Kong et al., 2020). Specifically, we trained it as a multi-speaker vocoder on the KazEmoTTS dataset without providing emotional labels for 1.72 million steps.

4.2. Objective Evaluation

We employed mel-cepstral distortion (MCD) (Kubichek, 1993) as an objective assessment metric to evaluate the quality of the synthesized speech. This approach involves comparing the mel-frequency cepstral coefficient (MFCC) vectors extracted from the generated speech and the ground truth speech, with a lower MCD score suggesting that the generated speech is more similar to the ground truth. To mitigate issues arising from the potentially extreme scaling of MCD due to variations in the two input speech lengths, we adopted the dynamic time warping (DTW) algorithm, as described in Battenberg et al. (2019).

4.3. Subjective Evaluation

To evaluate the quality of the synthesized speech, we conducted a subjective evaluation survey via a messenger application (Telegram) bot. The user interface of the bot was developed in Kazakh, as shown in Figure 1b. To recruit volunteer participants, we distributed the link to the survey on popular social media platforms.

The survey involved a two-fold evaluation process. Participants were first tasked with evaluating the naturalness of a given speech sample, focusing on its degree of human-likeness. The evaluation was conducted using a five-point scale: 1. *bad*, 2. *poor*, 3. *fair*, 4. *good*, and 5. *excellent*. Subsequent to the naturalness evaluation, participants were prompted to identify one of the six distinct emotions with which the speech sample was narrated.

We compiled an evaluation set of 3,600 audio samples that were not included in the training set, from which a random subset of 36 (18 ground truth and 18 synthesized) speech samples was presented to each participant. The samples were selected to ensure an equal representation of each narrator and emotion, amounting to six samples per narrator and one per emotion. Participants were presented with one speech sample at a time. While participants were afforded the opportunity to listen to each sample multiple times, it was emphasized that their selection could not be altered once submitted.

5. Results and Discussion

The evaluation results are provided in Tables 3–5. As can be seen from Table 3, on average, the synthesized speech delivered in a female voice demonstrated a greater likeness to the corresponding ground truth samples compared to the synthesized samples in both male voices. An interesting observation is that synthesized speech samples featuring emotional states typically associated with lower-pitched voices (e.g., neutral, sad, scared) exhibited greater similarity to the corresponding ground truth samples. Conversely, speech samples generated to convey emotional states characterized by higher-pitched voices (e.g., angry, happy, surprised) demonstrated a comparatively lower degree of similarity to the ground truth samples.

N	**MCD**							**MOS**	
	neu	ang	hap	sad	sca	sur	**Avg**	GT	Syn
F1	5.72	6.08	6.2	5.82	5.97	6.34	**6.02**	3.94	3.55
M1	7.63	7.98	7.88	7.19	7.68	7.65	**7.67**	3.95	3.51
M2	6.85	7.56	7.57	6.83	7.13	7.54	**7.24**	4.22	3.57

Note. N: narrators, Avg: average, GT: ground truth, Syn: synthesized

Table 3: MCD and MOS results

As for the evaluation survey, there were a total of 64 participants. The mean opinion score (MOS) for assessing the naturalness of the synthesized speech varied only slightly, ranging from 3.51 to 3.57 (see Table 3). Narrator M2's samples attained the highest MOSs for both the ground truth and synthesized samples, with M2's synthesized samples achieving a slightly higher score than that of F1 by a margin of 0.02. The generated speech of Narrator M1 received the lowest MOS. This underscores the lack of a correlation between MOS and the volume of data accessible for each narrator. Despite the greater volume of data for Narrator F1 in comparison to the other two narrators, it did not translate into a much higher MOS for the female narrator. Similarly, the MOS was not higher for Narrator M1, despite having nearly twice as many data as Narrator M2.

A comparison of MOSs highlights notably higher results in a separate study focused on Kazakh TTS (Mussakhojayeva et al., 2022). For female speakers, ground truth speech evaluations achieved scores within the range of 4.18 to 4.73, while MOSs for the generated speech covered a spectrum from 4.05 to 4.53. In the case of male speakers, MOSs for ground truth ranged from 4.37 to 4.43, while scores for synthesized speech spanned from 3.95 to 4.2.

In English emotional TTS studies, results closely aligned with our findings were reported by Zhou et al. (2022a). They achieved a MOS of 3.45 when the emotion "surprised" was presented with 0% intensity of other emotions. Notably, superior performance was observed with EmoDiff (Guo et al., 2022a), scoring 4.01, and EmoMix (Tang et al., 2023), which attained a MOS of 3.92.

As illustrated in Table 4, sentences delivered in a neutral manner were more accurately recognized as such, achieving an accuracy rate of 65%. In contrast, sentences expressed with anger proved to be the most challenging to identify, with a recognition accuracy of only 22%.

E	F1		M1		M2		F1 & M1 & M2		
	GT	Syn	GT	Syn	GT	Syn	GT	Syn	Overall
neu	0.75	0.60	0.64	0.67	0.53	0.67	0.64	0.65	0.65
ang	0.28	0.07	0.26	0.06	0.58	0.07	0.37	0.07	0.22
hap	0.59	0.35	0.50	0.32	0.64	0.41	0.58	0.36	0.47
sad	0.22	0.32	0.40	0.29	0.37	0.26	0.33	0.29	0.31
sca	0.25	0.35	0.42	0.29	0.28	0.35	0.32	0.33	0.33
sur	0.32	0.19	0.28	0.23	0.33	0.20	0.31	0.21	0.26
Total	0.38	0.31	0.42	0.32	0.46	0.33	0.43	0.32	0.37

Table 4: Results of emotion prediction accuracy

Table 5 displays the percentages of participant responses regarding their choice of emotion and reveals that "neutral" was frequently selected by participants when identifying the emotion of a speech sample. Interestingly, "happy" was the most easily identifiable emotion, chosen by nearly half of all participants, irrespective of the narrator. It is also worth noting that participants faced challenges when distinguishing "angry" speech samples, often mistaking them for "sad" or "scared" expressions. Frequently, samples labelled as "scared" were also erroneously identified as "sad".

In light of the MOS results acquired, it is apparent that despite providing participants with explicit instructions to focus on the emotion conveyed by the delivery of a sentence, rather than its inherent message or content, we acknowledge the cognitive challenge of perceiving a sentence with inherently somber content, such as one related to a grave illness or loss of life, being articulated with a seemingly cheerful emotion. This presents a cognitive challenge that may not have been fully resolved in our current study. In our future emotional TTS studies, we aim to address this challenge more effectively.

Actual emotions	Participant responses (%)						N
	neu	**ang**	**hap**	**sad**	**sca**	**sur**	
neu	66.92	4.62	4.62	13.08	2.31	8.46	F1
	65.32	4.05	6.36	12.14	5.78	6.36	M1
	59.60	15.23	3.31	11.92	4.64	5.30	M2
ang	43.31	16.56	3.18	14.01	14.01	8.92	F1
	39.04	15.75	2.74	18.49	13.70	10.27	M1
	31.33	33.33	2.67	10.00	12.67	10.00	M2
hap	37.75	0.00	45.70	3.97	1.99	10.60	F1
	39.73	8.90	43.84	2.74	10.27	9.15	M1
	28.28	2.07	51.72	1.38	2.76	13.79	M2
sad	35.63	3.13	3.75	25.63	19.38	12.50	F1
	46.43	2.38	0.60	34.52	7.74	8.33	M1
	50.00	0.74	2.94	31.62	8.82	5.88	M2
sca	21.77	17.01	6.12	20.41	29.93	17.01	F1
	23.98	2.92	1.17	26.90	35.09	9.94	M1
	20.95	16.89	0.68	12.84	31.76	16.89	M2
sur	37.04	2.47	12.96	4.94	14.81	27.78	F1
	30.41	5.85	4.09	18.71	15.79	25.15	M1
	37.01	3.15	11.02	12.60	9.45	26.77	M2

Note. N: narrators.

Table 5: Emotion perception results

6. Conclusion

This study aimed to construct the KazEmoTTS dataset for Kazakh emotional TTS applications. The dataset comprises a substantial 54,760 audio-text pairs, covering a total duration of 74.85 hours. This includes 34.23 hours delivered by a female narrator and 40.62 hours by two male narrators. The emotional spectrum within the dataset covers "neutral", "angry", "happy", "sad", "scared", and "surprised" states. In addition, a TTS model was developed through training on the KazEmoTTS dataset. Both objective and subjective evaluations were performed to gauge the synthesized speech quality, resulting in an objective MCD metric ranging from 6.02 to 7.67 and an MOS ranging from 3.51 to 3.57. Our findings are particularly promising, considering that this study represents the first attempt at emotional TTS for Kazakh. To facilitate replicability and further exploration, we have made our code, pre-trained model, and dataset available in our GitHub repository.[1]

7. Acknowledgements

We would like to express our gratitude to the narrators for their contributions and to the anonymous raters for their valuable evaluations.

8. Bibliographical References

Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. 2018. The Emotional Voices Database: Towards Controlling the Emotion Dimension in Voice Generation Systems.

Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. Deep Voice 2: Multi-Speaker Neural Text-to-Speech.

Eric Battenberg, R. J. Skerry-Ryan, Soroosh Mariooryad, Daisy Stanton, David Kao, Matt Shannon, and Tom Bagby. 2019. Location-Relative Attention Mechanisms for Robust Long-Form Speech Synthesis. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6194–6198.

Anna Braniecka, Ewa Trzebińska, Aneta Dowgiert, and Agata Wytykowska. 2014. Mixed Emotions and Coping: The Benefits of Secondary Emotions. *PloS one*, 9:e103940.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Heejin Choi and Minsoo Hahn. 2021. Sequence-to-Sequence Emotional Voice Conversion With Strength Control. *IEEE Access*, 9:42674–42687.

Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. 2014. EMOVO Corpus: an Italian Emotional Speech Database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3501–3504, Reykjavik, Iceland. European Language Resources Association (ELRA).

Chenye Cui, Yi Ren, Jinglin Liu, Feiyang Chen, Rongjie Huang, Ming Lei, and Zhou Zhao. 2021. EMOVIE: A Mandarin Emotion Speech Dataset with a Simple Emotional Text-to-Speech Model.

Prafulla Dhariwal and Alex Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. *ArXiv*, abs/2105.05233.

Chenpeng Du and Kai Yu. 2021. Phone-Level Prosody Modelling with GMM-Based MDN for Diverse and Controllable Speech Synthesis.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Yiwei Guo, Chenpeng Du, Xie Chen, and K. Yu. 2022a. EmoDiff: Intensity Controllable Emotional Text-to-Speech with Soft-Label Guidance. *ArXiv*, abs/2211.09496.

Yiwei Guo, Chenpeng Du, and Kai Yu. 2022b. Unsupervised word-level prosody tagging for controllable speech synthesis.

Chae-Bin Im, Sang-Hoon Lee, Seung bin Kim, and Seong-Whan Lee. 2022. EMOQ-TTS: Emotion Intensity Quantization for Fine-Grained Controllable Emotional Text-to-Speech. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6317–6321.

Minchan Kim, Sung Jun Cheon, Byoung Jin Choi, Jong Jin Kim, and Nam Soo Kim. 2021. Expressive Text-to-Speech Using Style Tag. In *Proc. Interspeech 2021*, pages 4663–4667.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *ArXiv*, abs/2010.05646.

Robert F. Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, 1:125–128 vol.1.

Younggun Lee, Azam Rabiee, and Soo-Young Lee. 2017. Emotional End-to-End Neural Speech Synthesizer.

Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. 2021. More Control for Free! Image Synthesis with Semantic Diffusion Guidance. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 289–299.

Saida Mussakhojayeva, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022. KazakhTTS2: Extending the Open-Source Kazakh TTS Corpus With More Data, Speakers, and Topics.

Devi Parikh and Kristen Grauman. 2011. Relative attributes. In *2011 International Conference on Computer Vision*, pages 503–510.

R. Plutchik and H. Kellerman. 2013. *Theories of Emotion*. Emotion, theory, research, and experience. Elsevier Science.

Robert Plutchik. 2001. 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*, 89(4):344–350.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail A. Kudinov. 2021. Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech. In *International Conference on Machine Learning*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. arxiv. *arXiv preprint arXiv:2212.04356*.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech.

Marc Schröder. 2009. *Expressive Speech Synthesis: Past, Present, and Possible Futures*, pages 111–126. Springer London, London.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2017. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions.

Yang Song, Jascha Narain Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-Based Generative Modeling through Stochastic Differential Equations. *ArXiv*, abs/2011.13456.

Haobin Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao. 2023. EmoMix: Emotion Mixing via Diffusion Models for Emotional Speech Synthesis. *ArXiv*, abs/2306.00648.

Seyun Um, Sangshin Oh, Kyungguen Byun, Inseon Jang, Chung Hyun Ahn, and Hong-Goo Kang. 2019. Emotional Speech Synthesis with Rich and Granularized Control. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7254–7258.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2021. Emotional Voice Conversion: Theory, Databases and ESD.

Kun Zhou, Berrak Sisman, Rajib Kumar Rana, B.W.Schuller, and Haizhou Li. 2022a. Speech Synthesis with Mixed Emotions. *ArXiv*, abs/2208.05890.

Kun Zhou, Berrak Sisman, Rajib Kumar Rana, Björn Schuller, and Haizhou Li. 2022b. Emotion Intensity and its Control for Emotional Voice Conversion. *IEEE Transactions on Affective Computing*, 14:31–48.

Kun Zhou, Berrak Sisman, Mingyang Zhang, and Haizhou Li. 2020. Converting Anyone's Emotion: Towards Speaker-Independent Emotional Voice Conversion.