



NAZARBAYEV
UNIVERSITY

Institute of Smart Systems
and Artificial Intelligence

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

Abilbekov Adal

adal.abilbekov@nu.edu.kz

Saida Mussakhojayeva

saida.mussakhojayeva@nu.edu.kz



issai.nu.edu.kz

Rustem Yeshpanov

rustem.yeshpanov@nu.edu.kz

Huseyin Atakan Varol

ahvarol@nu.edu.kz

Outline

- Introduction
- Related works
- The Kazakh Emotional TTS dataset construction
- TTS Experiments and Results
- Conclusion

Introduction

- In this work we present an open-sourced Emotional TTS dataset in Kazakh.
- Current TTS systems can generate high-quality speech, but remain not human-alike due to lack of emotions. Emotional Text-to-speech is a big part of building human-like dialogue agents.
- 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.

Related Works

Dataset	Language	Emotions	Speakers	Modalities
RAVDESS (Livingstone and Russo, 2018)	EN	Neutral, happy, angry, sad, calm, fear, disgust	24 actors (12 male and 12 female)	Audio/Visual
CREMA-D (Cao et al., 2014)	EN	Neutral, happy, angry, sad, fear, disgust	91 actors (48 male and 43 female)	Audio/Visual
IEMOCAP (Busso et al., 2008)	EN	Neutral, happy, angry, sad, frustrate	10 speakers (5 male and 5 female)	Audio/Visual
EMOVO Corpus (Costantini et al., 2014)	IT	Disgust, joy, fear, anger, surprise, sadness, neutral	6 actors (3 male and 3 female)	Audio
TESS (Pichora-Fuller and Dupuis, 2020)	EN	Neutral, angry, sad, fear, disgust, surprise, pleased	2 speakers (2 female)	Audio
KazakhTTS2 (Mussakhojayeva et al., 2022)	KZ	Neutral	5 speakers (2 male and 3 female)	Audio

Fig. 1. Related works table.

The dataset collection.

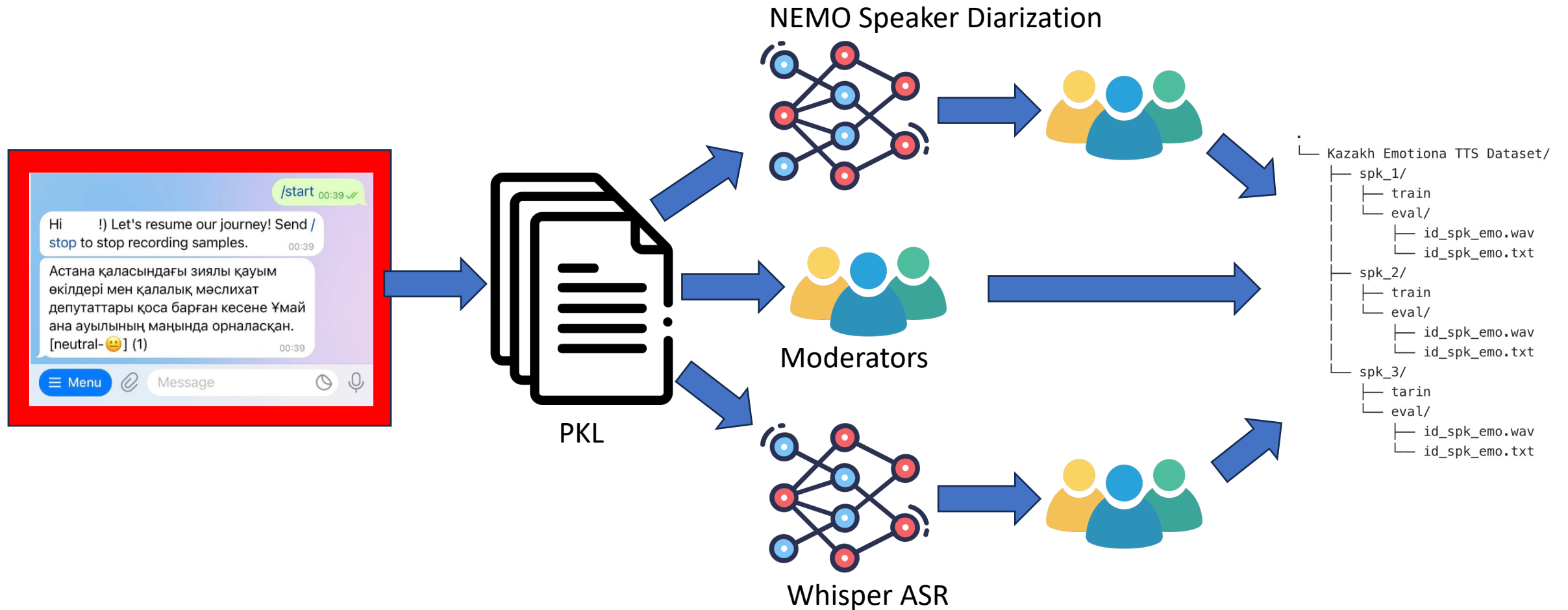


Fig. 2. Diagram of data collection.

The Kazakh Emotional TTS dataset construction

Text:

- 8,794 unique sentences
- 86,496 unique words
- Average sentence length of 10.83 words

Audios:

- 74.85 hours
- 34.23 hours female voices
- 40.62 hours male voices

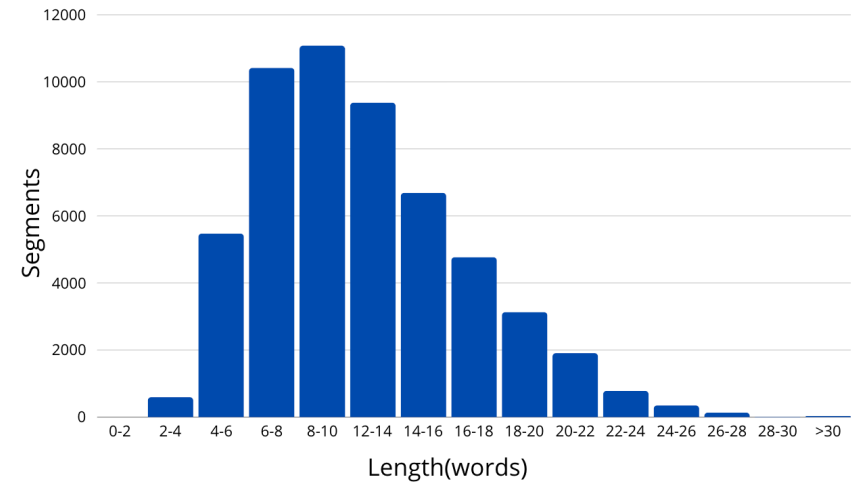


Fig. 4.a. Text segments length.

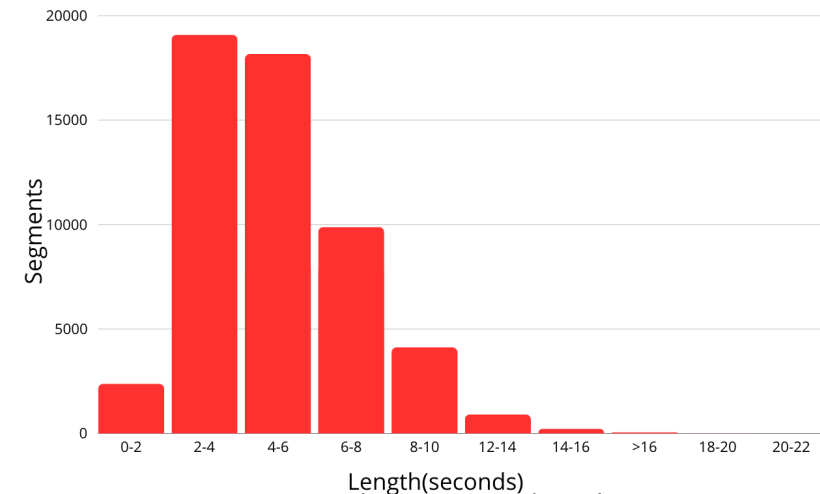


Fig. 4. Audio segments length.

The Kazakh Emotional TTS dataset construction

Emotion	# recordings	Narrator F1				Narrator M1				Narrator M2			
		Total (h)	Mean (s)	Min (s)	Max (s)	Total (h)	Mean (s)	Min (s)	Max (s)	Total (h)	Mean (s)	Min (s)	Max (s)
neutral	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
angry	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
happy	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
scared	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
surprised	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

Fig. 5. Recording count and duration statistics: Total (hours), Mean, Max, and Min (seconds).

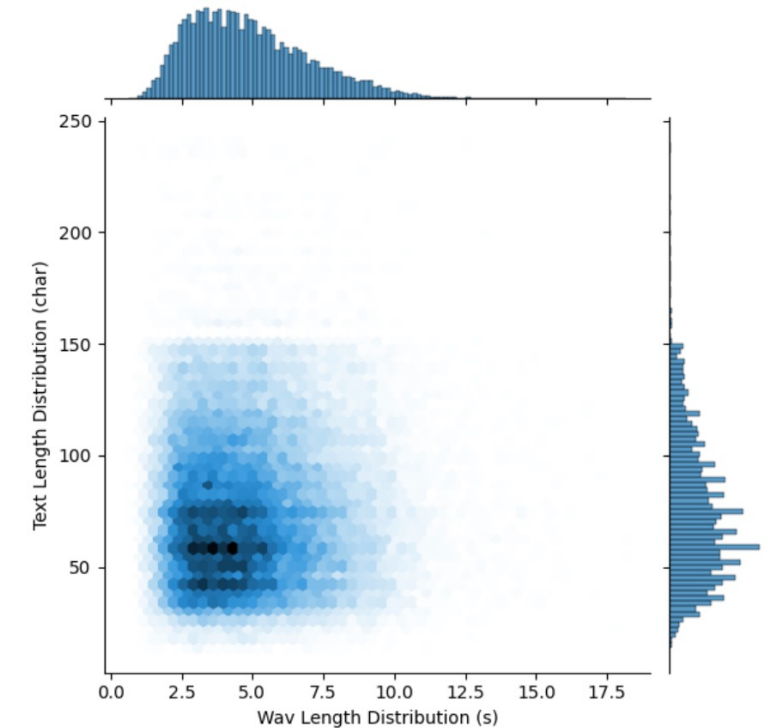


Fig. 6. Joint distribution of audio length (sec.) and text length (car.)

TTS Experiments and Results

Emotional Embedding	X_1	X_2	...	X_63	X_64
Speaker Embedding	X_1	X_2	...	X_63	X_64

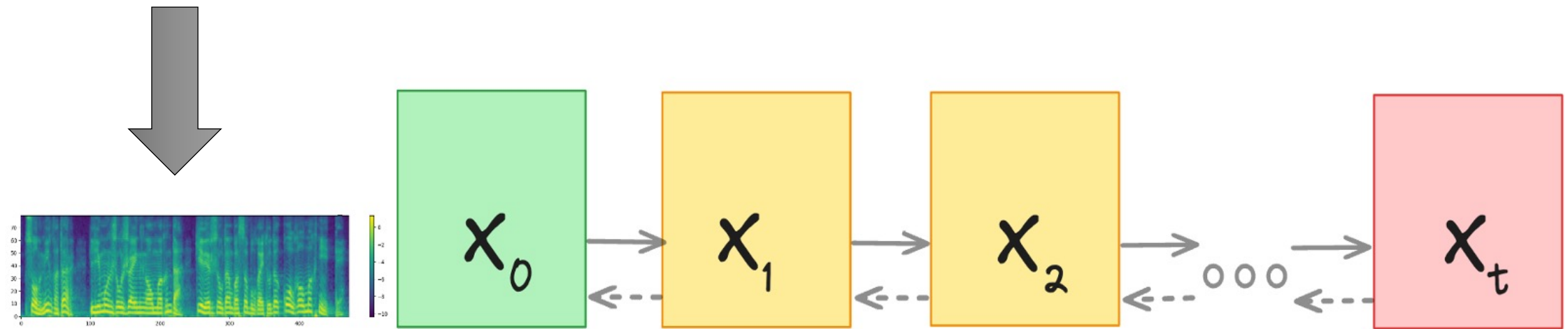


Fig. 7. Diffusion process illustration.

TTS Experiments and Results

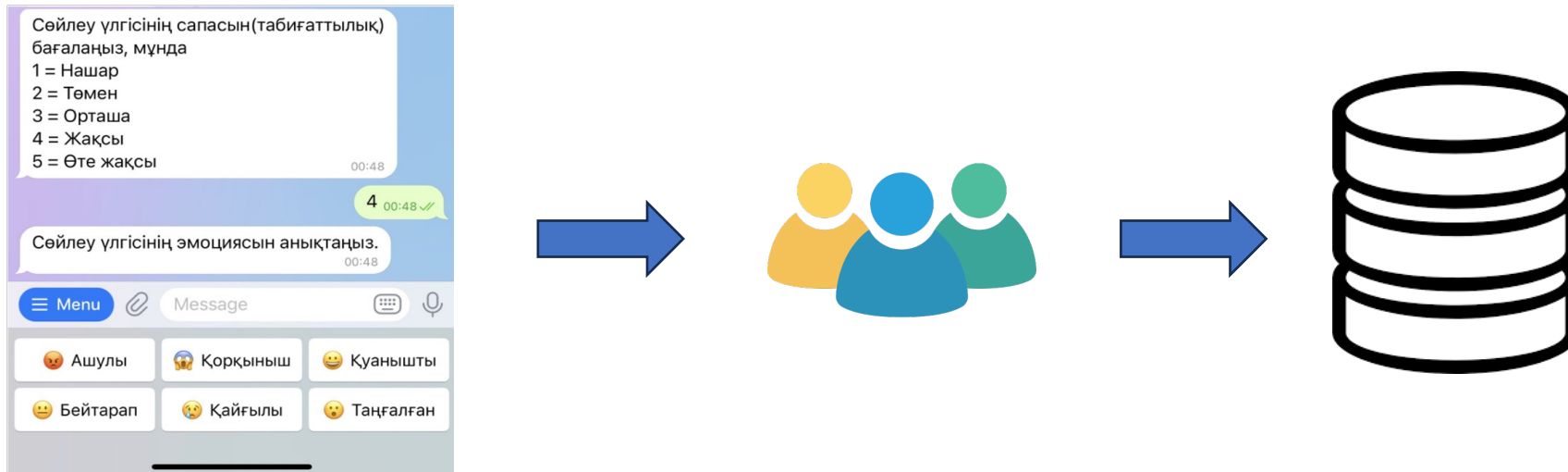


Fig. 8. Evaluation process illustration.

TTS Experiments and Results

		Participant responses (%)						N
		neu	ang	hap	sad	sca	sur	
Actual emotions	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.

Fig. 9. Emotion perception results.

Narrator	Mean Opinion Score	
	Ground Truth	Synthesized
F1	3.94	3.55
M1	3.95	3.51
M2	4.22	3.57

Fig. 10. MOS results.

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

Fig. 11. Results of emotion prediction accuracy

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

References

- Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS One*. 2018 May 16;13(5):e0196391. doi: 10.1371/journal.pone.0196391. PMID: 29768426; PMCID: PMC5955500.
- Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, Verma R. CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. *IEEE Trans Affect Comput*. 2014 Oct-Dec;5(4):377-390. doi: 10.1109/TAFFC.2014.2336244. PMID: 25653738; PMCID: PMC4313618.
- 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).
- 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
- Dupuis, Kate & Pichora-Fuller, M.. (2011). Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set. *Canadian Acoustics - Acoustique Canadienne*. 39. 182-183.
- Saida Mussakhojayeva, Yerbolat Khassanov, and Atakan Varol. 2022. KazakhTTS2: Extending the Open-Source Kazakh TTS Corpus With More Data, Speakers, and Topics. Accepted to LREC 2022.
- 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.
- 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*.

ISSAI

NAZARBAYEV
UNIVERSITY

Institute of Smart Systems
and Artificial Intelligence

Thank You For Your Attention!