INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Multilingual Text-to-Speech Synthesis for Turkic Languages Using Transliteration

Rustem Yeshpanov, Saida Mussakhojayeva, Yerbolat Khassanov

Institute of Smart Systems and Artificial Intelligence, Nazarbayev University, Kazakhstan

{rustem.yeshpanov, saida.mussakhojayeva, yerbolat.khassanov}@nu.edu.kz

Abstract

This work aims to build a multilingual text-to-speech (TTS) synthesis system for ten lower-resourced Turkic languages: Azerbaijani, Bashkir, Kazakh, Kyrgyz, Sakha, Tatar, Turkish, Turkmen, Uyghur, and Uzbek. We specifically target the zero-shot learning scenario, where a TTS model trained using the data of one language is applied to synthesise speech for other, unseen languages. An end-to-end TTS system based on the Tacotron 2 architecture was trained using only the available data of the Kazakh language. To generate speech for the other Turkic languages, we first mapped the letters of the Turkic alphabets onto the symbols of the International Phonetic Alphabet (IPA), which were then converted to the Kazakh alphabet letters. To demonstrate the feasibility of the proposed approach, we evaluated the multilingual Turkic TTS model subjectively and obtained promising results. To enable replication of the experiments, we make our code and dataset publicly available in our GitHub repository[1].

Index Terms: speech synthesis, TTS, Turkic, IPA, lower-resourced, zero-shot, transliteration

1. Introduction

Text-to-speech (TTS) synthesis has proved successful for higher-resourced languages, for which a large amount of labelled data is available [1, 2, 3]. The generated speech is of high quality and sounds human-like, finding its use in many commercial applications. Importantly, TTS has substantial social impact on assistive technologies for people with disabilities, such as speech and vision impairments, making it an essential speech processing technology for any language. However, for the majority of the world's languages, TTS systems remain inaccessible. This is mainly due to the lack of labelled data of sufficient size, as TTS data collection is considered highly laborious [4].

In this work, we aim to build a multilingual TTS model supporting ten Turkic languages, including Azerbaijani, Bashkir, Kazakh, Kyrgyz, Sakha, Tatar, Turkish, Turkmen, Uyghur, and Uzbek. These languages are considered lower-resourced, and, to the best of our knowledge, this work is the first attempt to develop an end-to-end (E2E) TTS system for most of them. Our study became feasible thanks to the recent work of Mussakhojayeva et al. [5], who have addressed data scarcity in the Kazakh language, by developing a large-scale and open-source speech corpus called KazakhTTS2. This corpus unlocks new opportunities for the hitherto understudied—in regard to spoken language technology—family of Turkic languages. Thus, by taking advantage of KazakhTTS2, we investigate the Kazakh transliteration of other Turkic languages. Specifically, to bring together the target languages under the same input space, we employed the International Phonetic Alphabet (IPA) [6], manually mapping all letters of the Turkic alphabets onto their IPA representations.

[1] https://github.com/IS2AI/TurkicTTS

Table 1: *The characteristics of the ten Turkic languages*

Language	Branch	Native speakers	Main writing system
Azerbaijani	Oghuz	33M	Latin
Bashkir	Kipchak	1.5M	Cyrillic
Kazakh	Kipchak	14M	Cyrillic
Kyrgyz	Kipchak	5M	Cyrillic
Sakha	Siberian	0.4M	Cyrillic
Tatar	Kipchak	5.5M	Cyrillic
Turkish	Oghuz	83M	Latin
Turkmen	Oghuz	7M	Latin
Uyghur	Karluk	11M	Perso-Arabic
Uzbek	Karluk	27M	Latin

The ten Turkic languages under consideration form a comprehensive language family with over 150 million native speakers [7]. Spoken across a wide geographical area stretching from the Balkans through Central Asia to northeastern Siberia, these languages can be divided into four branches, as shown in Table 1. The languages share a wide range of common linguistic features, such as vowel harmony, extensive agglutination, subject-object-verb order, and the absence of grammatical gender and articles, although the intensity of each feature may vary from language to language. We believe that the geographical proximity and similarities in phonology, morphology, and syntax among Turkic languages should further enhance the efficiency of our approach, as was observed in [8, 9].

We evaluated the developed multilingual Turkic TTS system based on the Tacotron 2 architecture [2] using subjective tests and obtained promising results for all the languages. Specifically, we evaluated (1) the overall quality, using the mean opinion score (MOS) measure, (2) the comprehensibility, as well as (3) the intelligibility of the synthesized speech. The obtained results indicate that our TTS model is suitable for most real-world applications.

The main contributions of the work:

- We investigate an IPA-based approach to build a multilingual E2E TTS system for ten Turkic languages under the zero-shot learning scenario;
- We evaluate the utility of the developed Turkic TTS system by subjectively assessing the overall quality, comprehensibility, and intelligibility of the synthesised speech;
- The implemented codes, including the TTS model and the IPA converters, as well as the dataset used, are made publicly available in our GitHub repository[1].

10.21437/Interspeech.2023-249

The rest of the paper is structured as follows: Section 2 briefly reviews lower-resourced TTS approaches and previous work on Turkic TTS. Section 3 provides a thorough overview of the proposed pipeline. Section 4 describes the experimental setup, and Section 5 presents the evaluation results. Section 6 discusses the challenges faced, important findings, and potential future work. Section 7 concludes this paper.

2. Prior work on Turkic TTS

Turkic languages are generally considered to be lower-resourced, as publicly available linguistic data are limited. To address this, several recent works have developed high-quality open-source datasets. This trend can especially be observed for Kazakh, where freely available datasets have been constructed for speech and text processing tasks, such as named entity recognition [10], speech recognition [11, 12, 13], and speech synthesis [5, 14]. Of particular note is the KazakhTTS2 corpus, which made our work possible. KazakhTTS2 consists of five voices (three female and two male), with over 270 hours of high-quality transcribed data. The corpus is in the public domain and can be used both academically and commercially.

Another very well-studied Turkic language is Turkish, on which there are many published works in the literature [15, 16]. One of the earliest works [15], for example, developed a diphone inventory for Turkish to construct diphone-based concatenative TTS systems. In a more recent work, Ergün and Yıldırım [16] investigated whether English data could be used as a source language to train a Turkish TTS system and obtained satisfactory results. Despite the abundance of works focusing on Turkish TTS, we were not able to find a large-scale and open-source speech corpus developed for Turkish speech synthesis[2].

Several papers have addressed Uyghur TTS [17, 18]. These studies, however, were usually conducted using proprietary, in-house data. Moreover, most of the datasets used consist of a few hundred utterances, which might be insufficient for building reliable TTS systems. Similarly, we found only a few papers investigating TTS systems for Azerbaijani [19], Sakha [20], Tatar [21], and Uzbek [22]. These works mostly focused on conventional TTS approaches, such as unit selection and concatenation. Furthermore, the datasets used in these works are either unavailable or unsuitable for developing state-of-the-art TTS systems. We could not find any reasonable work dealing with speech synthesis for Bashkir, Kyrgyz, or Turkmen, as the existing research focuses on specific aspects of TTS rather than on the complete speech synthesis process.

3. Methodology

The overall architecture of the proposed approach is shown in Figure 1, which consists of two main modules: an IPA-based converter and a TTS model. First, we train a TTS model using Kazakh as a source language, with the letters of the Kazakh alphabet as an input sequence. The training process is the same as in [5], where a MOS of above four (out of five) was achieved for all voices present in KazakhTTS2. More precisely, for each voice, we train a separate TTS model (five in total). These models will be used as the backbone of the proposed multilingual TTS system for Turkic languages.

To enable speech synthesis for other Turkic languages, we constructed an IPA-based conversion module. The IPA-based converter takes letters from the alphabets of other Turkic languages and converts them into the letters of the Kazakh alphabet. For this purpose, the letters entered are first converted into the corresponding IPA representations. Next, the IPA symbols are converted into the letters of the Kazakh alphabet, which can be used as input for the TTS models constructed.

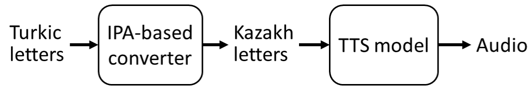

Figure 1: *The multilingual Turkic TTS system overview*

The mappings of the Turkic alphabets onto IPA symbols were manually created based on our expertise, as we could not find a complete mapping that would allow an error-free conversion from Turkic to Kazakh and cover all the languages addressed. Since Kazakh is used as a source language, we selected only 42 IPA symbols corresponding to the 42 letters of the Kazakh alphabet. It is worth mentioning that, of the Turkic languages in question, Kazakh—along with Bashkir—has the most letters and contains a large majority of the phonemes of the target languages. The developed mappings can also be used as a guide for other work aimed at building multilingual systems for Turkic languages, such as speech recognition, speech translation, and so on. The mapping of the Turkic alphabets onto IPA symbols is provided in Table 2.

4. Experimental setup

4.1. Source language data

In our experiments, Kazakh was used as a source language, and we employed the KazakhTTS2 speech corpus. As mentioned earlier, we trained five TTS models corresponding to the five voices present in the corpus. All five models proved capable of synthesising Turkic speech. However, based on our previous experience, we found that raters are usually unwilling to participate in or complete an overly long evaluation session. Therefore, we only used the voice of Speaker M2 to evaluate the proposed approach. Speaker M2's data contain approximately 58 hours of transcribed speech. We did not use any data for the target languages to comply with the zero-shot learning scenario.

4.2. TTS architecture

In our experiments, we trained an E2E TTS model based on the Tacotron 2 [2] architecture using the NVIDIA DGX A100 machines. Specifically, we followed the LJ Speech [23] training recipe implemented within the ESPnet-TTS toolkit [24]. In [5], the speakers were asked to pay attention to punctuation by pausing at commas and using the right intonation when pronouncing sentences ending with question marks and exclamation points. In order to use the intonation and pacing considered in [5], we felt that the input for our TTS model should also include five punctuation symbols ('.', ',', '-','?', '!') in addition to a text sequence of 42 Kazakh letters. The output was a sequence of acoustic features (80 dimensional log Mel-filter bank features). To transform these acoustic features into time-domain waveform samples, we employed WaveGAN [3] vocoder.

In the Tacotron 2-based TTS system, the encoder module was modelled as a single bidirectional LSTM layer with 512 units (256 units in each direction), and the decoder module was modelled as a stack of two unidirectional LSTM layers with 1,024 units. The parameters were optimised using the Adam algorithm with an initial learning rate of 10^{-3} for 200 epochs. To mitigate overfitting, we applied a dropout of 0.5. More details on the model specifications and training procedures can be found in our GitHub repository[1].

[2]It should be stressed that only English language papers were consulted, for which reason a number of publications might have been left out of the study.

Table 2: *The letter-to-symbol mapping used in the study*

IPA	Azerbaijani	Bashkir	Kazakh	Kyrgyz	Sakha	Tatar	Turkish	Turkmen	Uyghur	Uzbek
ɑ	a	а	а	а	а	а	a	a	a	-
æ	ə	ә	ә	-	-	ә	-	ä	e	a
b	b	б	б	б	б	б	b	b	b	b
v	v	в	в	в	в	в	v	w	w	v
ɡ	q	г	г	г	г	г	g	g	g	g
ɣ	ğ	ғ	ғ	-	Ҕ	-	ğ	-	gh	g‘
d	d	д	д	д	д	д	d	d	d	d
j͡e	-	е	е	е	е	е	-	-	-	-
j͡ɔ	-	ё	ё	ё	ё	ё	-	-	-	-
ʒ	j	ж	ж	-	ж	ж	j	ž	zh	-
z	z	з	з	з	з	з	z	-	z	z
i͡j	i	и	и	и	и	и	i	i	i	i
j	y	й	й	й	й	й	y	ý	y	y
k	k	к	к	к	к	к	k	k	k	k
q	-	ҡ	қ	-	-	-	-	-	q	q
l	l	л	л	л	л	л	l	l	l	l
m	m	м	м	м	м	м	m	m	m	m
n	n	н	н	н	н	н	n	n	n	n
ŋ	-	ң	ң	ң	ҥ	ң	-	ň	ng	ng
ɔ	o	о	о	о	о	о	o	o	o	o
ɵ	ö	ө	ө	ө	ө	ө	ö	ö	ö	o‘
p	p	п	п	п	п	п	p	p	p	p
r	r	р	р	р	р	р	r	r	r	r
s	s	с	с	с	с	с	s	-	s	s
t	t	т	т	т	т	т	t	t	t	t
ʊ͡w	u	у	у	у	у	у	u	u	u	u
ʊ	-	-	ұ	-	-	-	-	-	-	-
ʏ	ü	ү	ү	ү	ү	ү	ü	ü	ü	-
f	f	ф	ф	ф	ф	ф	f	f	f	f
x	x	х	х	х	х	х	-	-	x	x
h	h	һ	һ	-	һ	һ	h	h	h	h
t͡s	-	ц	ц	ц	ц	ц	-	-	-	-
t͡ʃ	ç	ч	ч	ч	ч	ч	ç	ç	ch	ch
ʃ	ş	ш	ш	ш	ш	ш	ş	ş	sh	sh
ɕ	-	щ	щ	щ	щ	щ	-	-	-	-
ʔ	-	ъ	ъ	ъ	ъ	ъ	-	-	-	’
ɤ	ı	ы	ы	ы	ы	ы	ı	y	-	-
ɪ	-	-	і	-	-	-	-	-	-	-
ʲ	-	ь	ь	ь	ь	ь	-	-	-	-
e	e	э	э	э	э	э	e	e	ë	e
j͡ʊ͡w	-	ю	ю	ю	ю	ю	-	-	-	-
j͡ɑ	-	я	я	я	я	я	-	-	-	-
d͡ʒ	c	-	-	ж	дь	җ	c	j	j	j
ɡʲ	g	-	-	-	-	-	-	-	-	-
θ	-	ҫ	-	-	-	-	-	s	-	-
ð	-	ҙ	-	-	-	-	-	z	-	-
ɲ	-	-	-	-	нь	-	-	-	-	-
47	**32**	**42**	**42**	**36**	**40**	**39**	**29**	**30**	**32**	**30**

4.3. Evaluation process

To assess the quality of the synthesised recordings, we conducted a subjective evaluation using an online survey on the Qualtrics XM Platform[3]. To recruit volunteer raters, we distributed the link to the survey on popular social media platforms operating in the Turkic languages. We created a separate evaluation questionnaire for each target language.

Informed consent was obtained from raters, certifying they were at least 18 years old and confirming their participation in the project. Individuals younger than 18 years old were not allowed to participate. The survey was first developed in English and later translated into the Turkic languages with the help of Qualtrics and other online translation services. For some languages, we also provided the instructions in English and Russian.

The evaluation survey consisted of three main parts, each assessing different aspects of the synthesised speech. The first part assessed the overall quality of the synthesised speech, with raters presented with ten recordings and their transcripts. Raters were instructed to listen to the recordings and rate them on a five-point Likert scale [25]: 5 for *excellent*, 4 for *good*, 3 for *fair*, 2 for *poor*, and 1 for *bad*. They were allowed to listen to the recordings several times, but could not alter the ratings once submitted. The evaluation recordings were presented to all raters one at a time and in the same order.

The second part assessed the comprehensibility aspect of the synthesised speech. To reduce cognitive effort, raters were presented with five straightforward multiple-choice questions. There were four options per question, with only one being the correct answer. The questions were on average 4.84 words long; the options consisted of one or two short words. A score of one was attached to the right answer.

Finally, the third part assessed the intelligibility aspect of the synthesised speech by using semantically unpredictable sentences (SUS) [26]. The SUS were formed from 23-25 commonly used words. Raters were asked to listen to five sentences and write down what they heard in a field. Raters were informed that the sentences were not meaningful although they contained real words. To evaluate intelligibility, only the sentences that were entirely correct were considered.

5. Evaluation results

The survey results are given in Table 3. Of 572 registered survey participants, 24 were under the age of 18. Raters varied in terms

[3]https://qualtrics.com

Table 3: *The survey statistics for rater number (R), gender (F & M), and age (< 45 & 45+) and the evaluation results of the overall quality (Q), comprehensibility (C), and intelligibility (I) of synthesised speech*

Language	R	F	M	< 45	45+	Q	C	I
Azerbaijani	47	22	25	22	25	2.93	90%	52%
Bashkir	11	8	3	4	7	2.67	92%	47%
Kazakh	151	89	62	120	31	4.18	97%	80%
Kyrgyz	14	12	2	6	8	3.54	86%	43%
Sakha	254	155	99	147	107	2.85	93%	15%
Tatar	15	12	3	3	12	2.82	79%	17%
Turkish	18	6	12	15	3	3.25	91%	61%
Turkmen	6	0	6	6	0	2.37	67%	57%
Uyghur	10	6	4	6	4	3.01	45%	26%
Uzbek	22	2	20	19	3	2.85	80%	45%
Total	548	312	236	348	200	3.25	92%	41%

of both gender and age. The number of raters who answered all survey questions was 435, with at least six participating in a survey per language. The highest number of raters, 254, was observed for Sakha, followed by 151 and 47 raters for Kazakh and Azerbaijani, respectively.

The MOS for overall quality ranges from 2.37 to 4.18, with Kazakh scoring the highest, as expected. The best scores among the target languages were achieved by Kyrgyz, Turkish, and Uyghur. This is remarkable, given the quality of the synthesised recordings was evaluated as above *fair* by speakers of the languages belonging to three Turkic branches. The worst performer was Turkmen, where the quality of the recordings was rated just below average.

In the comprehensibility test, Sakha, Bashkir, Turkish, and Azerbaijani scored the highest among the target languages, with at least 90% of the answers being right. The lowest score was obtained by Uyghur, with raters answering only every second question correctly. Comparing the results of the two tasks, it can be seen that although Sakha speakers—constituting the majority of the survey respondents—evaluated the quality of the synthesised recordings at a MOS of only 2.85, their responses in the comprehensibility test were 93% right. This seems to indicate that recordings synthesised using a Kazakh voice can be relatively easy to understand for speakers of the Siberian Turkic languages.

In the intelligibility test, Turkish (61%) and Turkmen (57%) scored highest among the target languages, while Sakha obtained a result of only 15%. As expected, writing down SUS was the most challenging task, probably due to the greater cognitive load required to correctly recognise a complete sequence of semantically unrelated words. This is consistent with [26], where intelligibility scores were in the range of 10-20%.

Overall, the obtained results appear promising and sufficient for most real-world applications. Moreover, we believe that performance can be further improved for all target languages by fine-tuning the pre-trained models with the seed data of the corresponding language.

6. Challenges and Future Work

In this study, we faced several challenges that should be considered in future work. The first and probably the most important one is the scarcity of speech corpora for TTS in Turkic languages. We found that most of the existing datasets are proprietary, while the majority of available datasets are either of low quality or unsuitable for building state-of-the-art TTS architectures. To facilitate the research and development of TTS systems for Turkic languages, future studies should focus on collecting high-quality and open-source speech corpora. Furthermore, Turkic languages are agglutinative, with rich vocabularies and many characters per word, which requires that the datasets collected be large in size.

Another challenge is loanwords and code-switching. Specifically, the Turkic languages spoken in the former-Soviet countries are often used along with Russian and thus contain many Russian borrowings. Usually, these words retain the orthographic and phonological properties of the original language. Consequently, this might mislead TTS systems, as Russian is different from Turkic languages in many aspects [5].

We also observed the interchangeable use of visually similar/identical characters from different scripts—also known as homoglyphs—in almost all Turkic languages that have transitioned from the Cyrillic to the Latin alphabet. Normally, this does not bother the reader, but can be problematic for computer systems. Therefore, the collected text should be carefully inspected.

Finally, we did not measure the literacy skills of raters; nor did we ensure that raters had the native language keyboard layout on their devices. Information about raters' literacy levels and the provision of a virtual keyboard layout of the required language may have helped us establish a greater degree of accuracy in the intelligibility test.

We believe that addressing these challenges for the Turkic languages will be an interesting direction for future research. We hope that our work will encourage subsequent efforts in this area to solve some of the practical problems encountered in training TTS systems for the Turkic languages. We also hope that our proposed approach will serve as a baseline and as a preliminary solution for real-world applications. We will consider comparative experiments, as we are interested in a more detailed evaluation of the performance of the model, and explore the advantages of using Kazakh speech to build TTS models for Turkic languages. In addition, future work should be extended to other Turkic languages not considered in this work.

7. Conclusion

In this work, we developed a multilingual TTS system for ten Turkic languages. We assumed a zero-shot learning scenario where no target data were used. The proposed approach employed a TTS system trained using Kazakh and an IPA-based converter to translate letters from the target languages into the source language. When evaluating the quality of the synthesised speech over all addressed languges, a MOS of 3.25 was achieved. In the comprehensibility test, 92% of the answers were correct. In the intelligibility test, 41% of the sentences matched the references. Given that this is the first attempt at building a Turkic TTS system, the achieved results are promising. To enable experiment reproducibility, we share our code, pre-trained models, and dataset in our GitHub repository[1].

8. Acknowledgements

We would like to extend our heartfelt thanks to all individuals who contributed to the recruitment of participants for this study. Their efforts were critical to the success of our survey. In particular, we would like to express our deepest appreciation to Viktor Krivogornitsyn for his extraordinary dedication in attracting a substantial number of Sakha speakers. His contribution was invaluable, and we are grateful for his support.

9. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu *et al.*, "Tacotron: Towards End-to-End Speech Synthesis," in *Interspeech*. ISCA, 2017, pp. 4006–4010. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/1452.html

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly *et al.*, "Natural TTS Synthesis by Conditioning Wavenet on MEL spectrogram predictions," in *ICASSP*. IEEE, 2018, pp. 4779–4783. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8461368

[3] R. Yamamoto, E. Song, and J. Kim, "Parallel Wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP*. IEEE, 2020, pp. 6199–6203.

[4] Y.-J. Chen *et al.*, "End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning," in *Interspeech*. ISCA, 2019, pp. 2075–2079. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-2730

[5] S. Mussakhojayeva, Y. Khassanov, and H. A. Varol, "KazakhTTS2: Extending the Open-Source Kazakh TTS Corpus With More Data, Speakers, and Topics," in *LREC*. ELRA, 2022, pp. 5404–5411. [Online]. Available: https://aclanthology.org/2022.lrec-1.578

[6] I. P. Association *et al.*, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[7] Wikipedia contributors, "List of Turkic languages," 2022, [Online; accessed 7-June-2022]. [Online]. Available: https://en.wikipedia.org/wiki/List_of_Turkic_languages

[8] Y.-H. Lin *et al.*, "Choosing Transfer Languages for Cross-Lingual Learning," in *ACL*, 2019, pp. 3125–3135. [Online]. Available: https://doi.org/10.18653/v1/p19-1301

[9] A. Lauscher, V. Ravishankar, I. Vulić, and G. Glavaš, "From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers," in *EMNLP*, 2020, pp. 4483–4499.

[10] R. Yeshpanov, Y. Khassanov, and H. A. Varol, "KazNERD: Kazakh Named Entity Recognition Dataset," in *LREC*. ELRA, 2022, pp. 417–426.

[11] Y. Khassanov, S. Mussakhojayeva, A. Mirzakhmetov, A. Adiyev, M. Nurpeiissov, and H. A. Varol, "A Crowdsourced Open-Source Kazakh Speech Corpus and Initial Speech Recognition Baseline," in *EACL*, 2021, pp. 697–706. [Online]. Available: https://aclanthology.org/2021.eacl-main.58

[12] S. Mussakhojayeva, Y. Khassanov, and H. A. Varol, "A Study of Multilingual End-to-End Speech Recognition for Kazakh, Russian, and English," in *SPECOM*, ser. LNCS, vol. 12997. Springer, 2021, pp. 448–459. [Online]. Available: https://doi.org/10.1007/978-3-030-87802-3_41

[13] S. Mussakhojayeva, Y. Khassanov, and H. A. Varol, "KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus," in *Interspeech*. ISCA, 2022, pp. 1367–1371.

[14] S. Mussakhojayeva, A. Janaliyeva, A. Mirzakhmetov, Y. Khassanov, and H. A. Varol, "KazakhTTS: An Open-Source Kazakh Text-to-Speech Synthesis Dataset," in *Interspeech*, 2021, pp. 2786–2790.

[15] Ö. Salor, B. Pellom, and M. Demirekler, "Implementation and Evaluation of a Text-to-Speech Synthesis System for Turkish," in *EUROSPEECH*, 2003.

[16] E. Ergün and T. Yıldırım, "Is it possible to train a Turkish text-to-speech model with English data?" *Recent Advances in Science and Engineering*, 2022.

[17] W. Silamu, N. Tursun, and P. Saltiniyaz, "Speech processing technology of Uyghur language," in *COCOSDA*. IEEE, 2009, pp. 19–24.

[18] G. Maimaitiaili, Z. Zhang, and A. Rouzi, "TDNN-Based Multilingual Mix-Synthesis with Language Discriminative Training," in *APSIPA*, 2018, pp. 605–610.

[19] S. Rustamov and A. Saadova, "On an Approach to Computer Synthesis of Azerbaijan Speech," *Problems of Cybernetics and Informatics*, vol. 1, pp. 267–270, 2014.

[20] N. Leontiev and A. Sleptzov, "Improving the Quality of Speech Synthesis Using Semi-Syllabic Synthesis," in *FarEastCon*. IEEE, 2020, pp. 1–4.

[21] A. Khusainov, "Design and Creation of Speech Corpora for the Tatar Speech Recognition and Synthesis Tasks," in *TurkLang*, 2015, pp. 475–484.

[22] B. Akmuradov, U. Khamdamov, O. Djuraev, and A. Mukhamedaminov, "Developing a Database of Uzbek Language Concatenative Speech Synthesizer," in *ICISCT*. IEEE, 2021, pp. 1–5.

[23] K. Ito and L. Johnson, "The LJ Speech Dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[24] H. et al., "ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *ICASSP*. IEEE, 2020, pp. 7654–7658.

[25] R. Likert, "A technique for the measurement of attitudes," *Archives of psychology*, 1932.

[26] C. Benoît, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences," *Speech communication*, vol. 18, no. 4, pp. 381–392, 1996.