algorithms

MDPI

Article

End-to-End VITS Outperforms Two-Stage Tacotron 2 Synthesis for Kazakh: A Comparative Study on a New Multi-Domain Speech Corpus

Magzhan Kairanbay [1], Anar Fazylzhan [1,*], Akbayan Bekarystankyzy [2,3,*], Zhanara Zhumabayeva [1] and Nurlan Kairanbayev [1]

[1] The Institute of Linguistics Named After A. Baitursunuly, Almaty 050010, Kazakhstan; magzhan.kairanbay@sdu.edu.kz (M.K.); zhanar.zhumabayeva@tilbilimi.kz (Z.Z.); nurlan.kairanbayev@tilbilimi.kz (N.K.)
[2] School of Digital Technologies, Narxoz University, Almaty 050035, Kazakhstan
[3] Department of Software Engineering, Satbayev University, Almaty 050013, Kazakhstan
* Correspondence: anar.fazylzhan@tilbilimi.kz or afazyljan@gmail.com (A.F.); akbayan.bekaristankyzy@narxoz.kz (A.B.)

Abstract

Text-to-speech (TTS) synthesis for low-resource languages such as Kazakh is still difficult: the language is agglutinative, uses an extended Cyrillic alphabet and has few annotated speech corpora. We present Kazakh TTS systems trained on a new corpus of about 3000 pages of educational and scientific text from 23 domains, read by two professional speakers. A character-level Tacotron 2 model with guided attention and a Parallel WaveGAN vocoder was trained in ESPnet2 on the female voice, and end-to-end VITS models were trained at 24 kHz for both voices. Guided attention produced a sharp monotonic alignment within 21 epochs. On a new test set of 138 sentences, transcribed with a Kazakh-specific automatic speech recognition (ASR) model, VITS reached a character error rate (CER) of 1.94% (male) and 2.08% (female), against 8.79% for Tacotron 2, whose attention failed on 6.2% of the utterances. On 60 sentences with formulas, numbers and abbreviations, normalised by the system's text front-end, Tacotron 2 failed on 11.7% of the utterances, while VITS had no failures and lowered the CER from 19.7% to 7.2%. The multilingual Whisper model reported about ten times higher error rates on the same audio, which shows that Whisper is not a reliable intelligibility measure for Kazakh. In a listening test with seven native speakers, Tacotron 2 with Parallel WaveGAN obtained a mean opinion score of 4.5 out of 5 for both naturalness and intelligibility.

Keywords: text-to-speech synthesis; Kazakh language; Tacotron 2; VITS; neural vocoder; Parallel WaveGAN; ESPnet2; low-resource TTS; sequence-to-sequence; guided attention; end-to-end speech synthesis

Check for updates

Academic Editors: Jaroslaw Krzywanski, Marcin Sosnowski, Karolina Grabowska, Dorian Skrobek, Anna Zylka, Agnieszka Kijo-Kleczkowska, Bashar Shboul and Tomasz Czakiert

Received: 14 May 2026
Revised: 10 October 2026
Accepted: 13 October 2026
Published: 15 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Speech synthesis, or text-to-speech (TTS) conversion, is a fundamental component of spoken language interfaces, assistive technologies, and human–computer interaction systems [1]. Recent advances in deep learning have enabled end-to-end neural TTS systems that produce speech of remarkable naturalness and expressiveness [2,3]. However, most high-quality neural TTS research focuses on English or other high-resource languages, leaving a significant gap for the hundreds of morphologically complex and low-resource languages spoken worldwide.

Kazakh is a Turkic language spoken by approximately 18 million people primarily in Kazakhstan, with additional speakers in China, Russia, and Mongolia [4]. It is characterised by vowel harmony, rich agglutinative morphology, and—in its written form used in Kazakhstan—a Cyrillic-based orthography that includes nine language-specific letters absent from the standard Russian Cyrillic alphabet, encoded as the Unicode code points U+04D9, U+0493, U+049B, U+04A3, U+04E9, U+04B1, U+04AF, U+04BB, and U+0456. These orthographic and linguistic properties introduce unique challenges for character-level neural TTS: the token vocabulary is larger than that of languages using only the basic Cyrillic or Latin scripts, and the morphological complexity yields highly variable sequence lengths that place additional demands on the attention mechanism.

The social and technological significance of high-quality Kazakh TTS is considerable. Kazakhstan's ongoing digital transformation and language revitalisation policies have created growing demand for Kazakh-language voice interfaces, screen readers, navigation systems, and educational tools [5]. Despite this need, Kazakh remains severely underserved by modern neural TTS technology. Most deployed systems for Kazakh rely on unit-selection or diphone-based synthesis, which requires extensive manual linguistic annotation and produces speech of limited naturalness. While recent work has begun to apply neural methods to Kazakh automatic speech recognition (ASR), neural TTS for Kazakh remains comparatively understudied, with few published systems providing open, reproducible pipelines or rigorous evaluation.

Recent end-to-end TTS architectures such as Tacotron 2 [2] have shown that a sequence-to-sequence model with location-sensitive attention can learn a direct mapping from character or phoneme sequences to mel-spectrograms, eliminating the need for hand-crafted acoustic features or pronunciation dictionaries. This property is particularly valuable for Kazakh, whose Cyrillic orthography is largely phonemic, meaning that character-level modelling can achieve reasonable pronunciation accuracy without a dedicated grapheme-to-phoneme (G2P) converter—a resource that does not currently exist in robust form for Kazakh. The addition of guided attention loss [6] further accelerates the convergence of the attention mechanism toward a monotonic diagonal alignment, which is essential when handling the long input sequences produced by Kazakh's agglutinative morphology. For waveform synthesis, traditional methods such as Griffin–Lim produce audible buzzing and metallic artefacts that significantly degrade perceived naturalness. Neural vocoders such as WaveNet [3] resolve these issues but are computationally prohibitive for deployment. Parallel WaveGAN [7] addresses this by training a non-autoregressive WaveNet-like generator using a generative adversarial network (GAN) objective, achieving real-time or faster-than-real-time synthesis with competitive perceptual quality. This combination of high fidelity and low latency makes Parallel WaveGAN the most practical neural vocoder choice for a first deployable Kazakh TTS system.

The ESPnet2 toolkit [8] provides a unified, modular, and well-maintained framework for end-to-end speech processing. Its support for arbitrary token vocabularies and flexible feature extraction pipelines makes it straightforward to adapt to non-English languages, and its active community ensures reproducibility and long-term maintainability—both critical for low-resource language research where community infrastructure is limited.

Relation to KazakhTTS and KazakhTTS2 and Research Gaps

The most closely related open works are the KazakhTTS [9] and KazakhTTS2 [10] projects of the Institute of Smart Systems and Artificial Intelligence (ISSAI), which released open Kazakh TTS corpora together with ESPnet training recipes and pretrained models: KazakhTTS comprises about 93 h of speech from two professional speakers, with Tacotron 2 and Transformer acoustic models and a Parallel WaveGAN vocoder reaching mean opinion

score (MOS) values above 4, and KazakhTTS2 extends it to 271 h from five speakers, with Tacotron 2 + Parallel WaveGAN models reaching MOS values between 3.6 and 4.2. We do not propose a new architecture on top of these baselines. Our work differs from them in four ways. (i) *Data:* it uses a new, independently recorded corpus of about 3000 pages (987,086 running words) of educational and scientific text from 23 domains, read by two professional speakers under explicit orthoepic (normative pronunciation) instructions, whereas the ISSAI corpora were collected mainly from news and other general-domain sources. (ii) *Data-quality protocol:* it documents a strict audio–text filtering procedure and quantifies its cost (117.2 GB $\rightarrow$ 37.8 GB). (iii) *Architectural comparison on the same corpus:* it compares a two-stage Tacotron 2 + Parallel WaveGAN system with a fully end-to-end VITS (variational inference with adversarial learning for end-to-end text-to-speech) system [11] trained on the same speakers, and reports what we observed during training (attention convergence for a 145-token Cyrillic vocabulary, and mixed-precision instability of long VITS training runs). (iv) *Deployment:* the resulting models are integrated into a web application for educational platforms that includes a Kazakh text-normalisation and orthoepic conversion front-end.

The contributions of this paper are:

- A reproducible character-level Tacotron 2 + Parallel WaveGAN pipeline for Kazakh (ESPnet2), with all configuration details and an analysis of guided- attention convergence;
- VITS models trained at 24 kHz for both the female and the male speaker, compared with the two-stage system on plain text and on complex educational text (formulas, numbers, abbreviations), where VITS removed the attention failures of Tacotron 2;
- Two new Kazakh TTS test sets (138 plain sentences from 23 domains and 60 sentences of complex educational text) and an analysis of Whisper-based evaluation, which on the same audio reports roughly ten times higher error rates than a Kazakh-specific ASR model;
- A discussion of how the pipeline can be adapted to the forthcoming Latin-based Kazakh alphabet, and an open inference pipeline integrated into an educational web application.

2. Related Work

2.1. Concatenative and Statistical Parametric Speech Synthesis

Early speech synthesis systems relied on concatenative approaches, in which units of recorded speech—diphones, half-phones, or larger segments—are selected from a large inventory and concatenated to form the output waveform [12,13]. While capable of producing natural-sounding speech for the language in which they were built, these systems require extensive hand-crafted linguistic resources, large carefully recorded and annotated databases, and language-specific unit-selection cost functions, making them ill-suited to low-resource settings such as Kazakh.

Statistical parametric speech synthesis (SPSS) addressed some of these limitations by modelling the speech generation process probabilistically. Hidden Markov model (HMM)-based TTS [14,15] became the dominant paradigm in the 2000s and early 2010s. In HMM-based TTS, context-dependent HMMs are trained to model the joint distribution of acoustic features—typically mel-cepstral coefficients, fundamental frequency (F_0), and duration—conditioned on linguistic features derived from text analysis. At synthesis time, a maximum likelihood parameter generation algorithm produces smooth acoustic trajectories from the trained models, and a vocoder (commonly STRAIGHT or WORLD) reconstructs the waveform. HMM-based TTS offers compact model sizes, flexible speaker and style adaptation via model interpolation, and the ability to train on relatively small corpora—properties that initially made it attractive for low-resource and morphologically

complex languages. However, the over-smoothing inherent in maximum likelihood parameter generation produces a characteristic muffled quality, and the systems require language-specific front-ends (text normalisation, G2P, prosody prediction modules) that are costly to develop.

Deep neural network (DNN)-based SPSS [16] replaced the HMM acoustic model with feed-forward DNNs while retaining the same vocoder-based waveform generation pipeline, yielding improvements in naturalness. Recurrent architectures such as long short-term memory (LSTM) networks were subsequently applied to better capture temporal dependencies in acoustic sequences [17]. Despite these improvements, parametric systems of this generation continued to depend on detailed linguistic front-ends and retained the perceptual ceiling imposed by vocoder-based waveform reconstruction.

2.2. Neural End-to-End TTS

The introduction of sequence-to-sequence learning with attention transformed the TTS landscape. Tacotron [18] was the first model to demonstrate that a character sequence could be mapped directly to a mel-spectrogram using an encoder–decoder architecture with content-based attention, eliminating the need for hand-crafted acoustic features or explicit duration modelling. Tacotron 2 [2] refined this approach by replacing the Griffin–Lim vocoder with WaveNet and adopting location-sensitive attention, which augments the standard content-based attention with a convolutional summary of past attention weights to encourage monotonic progression through the input sequence.

Non-autoregressive models were subsequently introduced to address the slow inference speed of autoregressive decoders. FastSpeech [19] used a duration predictor and length regulator to expand encoder outputs to the target frame length, enabling fully parallel decoding. FastSpeech 2 [20] extended this by predicting pitch and energy in addition to duration, providing explicit control over prosodic attributes. These models offer substantially faster inference than Tacotron 2 but require phoneme-level duration labels extracted from a separately trained aligner, a resource that is unavailable for Kazakh.

Neural vocoders have undergone equally significant development. WaveNet [3] achieved outstanding perceptual quality by modelling the raw waveform autoregressively as a product of conditional distributions over quantised sample values. However, its sequential generation made real-time synthesis computationally prohibitive. Parallel WaveGAN [7] addressed this by training a non-autoregressive WaveNet-like generator adversarially, using a multi-resolution short-time Fourier transform (STFT) loss and a waveform discriminator to achieve real-time synthesis with competitive quality. HiFi-GAN [21] further improved quality by introducing multi-period and multi-scale discriminators operating directly on the waveform, setting a new quality–efficiency frontier for neural vocoders.

2.3. Speech Processing for Kazakh

Speech technology for Kazakh has received growing attention in recent years, though the field remains considerably less developed than for high-resource languages. On the recognition side, early Kazakh ASR systems adopted HMM-based pipelines with Gaussian mixture model (GMM) acoustic models trained on small-vocabulary corpora [22]. Subsequent work leveraged deep learning architectures, including DNN–HMM hybrid systems, to improve recognition accuracy on larger vocabulary tasks. More recently, transformer-based models fine-tuned from multilingual pre-trained checkpoints have achieved strong ASR performance for Kazakh, with the availability of open speech corpora such as KazakhTTS [9] and KazakhTTS2 [10] also supporting acoustic model development.

The KazakhTTS corpus [9], released by ISSAI, provided the first large-scale open Kazakh speech dataset suitable for neural TTS research, comprising recordings from multiple speakers with validated transcriptions. KazakhTTS2 [10] extended this resource with additional speakers and recording conditions, establishing a community standard for Kazakh TTS development. These datasets have significantly lowered the barrier to entry for Kazakh neural TTS research.

Despite these advances, TTS for Kazakh remains comparatively understudied relative to ASR. Prior deployed systems have largely relied on unit-selection or HMM-based synthesis, both of which require extensive language-specific annotation and produce speech of limited naturalness by contemporary standards. The application of modern end-to-end neural TTS architectures to Kazakh is therefore both timely and necessary.

2.4. Low-Resource TTS

Low-resource TTS has been addressed through several complementary strategies. Transfer learning from high-resource languages—in which a model pre-trained on a large corpus (typically English) is fine-tuned on a small target-language dataset—has been shown to reduce the amount of target-language data required for acceptable quality [1]. Multilingual TTS models jointly trained on multiple languages share acoustic representations across languages, allowing underrepresented languages to benefit from richer training signal. Data augmentation via pitch shifting, speed perturbation, or text-to-speech bootstrapping has also been explored to artificially increase effective corpus size.

Character-level modelling has been shown to be a practical approach when pronunciation dictionaries or G2P converters are unavailable, as is the case for Kazakh. Because the Kazakh Cyrillic orthography is largely phonemic, character-level models can achieve reasonable pronunciation accuracy without explicit phonemic transcription, avoiding the need to develop and maintain a language-specific G2P resource.

2.5. Recent Architectures: VITS, HiFi-GAN, and Diffusion-Based TTS

Since Tacotron 2, the state of the art has moved towards fully end-to-end and generative-model-based systems. HiFi-GAN [21] is now the most widely used GAN vocoder; its generator and multi-period discriminator are also reused inside end-to-end systems. Glow-TTS [23] introduced flow-based acoustic modelling with monotonic alignment search, removing the need for an external aligner. VITS [11] combines a conditional variational autoencoder, normalising flows, monotonic alignment search, a stochastic duration predictor, and a HiFi-GAN generator into a single model trained end-to-end from characters or phonemes to waveform, and it is now a common open baseline for low-resource TTS; YourTTS [24] extends it to multi-speaker and zero-shot multilingual synthesis using speaker embeddings. Diffusion-based models have been proposed both as vocoders (WaveGrad [25], DiffWave [26]) and as acoustic models (Grad-TTS [27]), offering high quality at the cost of iterative sampling. Most recently, large neural-codec language models such as VALL-E [28] enable zero-shot voice cloning but require tens of thousands of hours of training data, which is not available for Kazakh. With a few hundred hours of single-speaker data, as in our case, Tacotron 2 and VITS are the realistic options, and we evaluate both.

3. Materials and Methods

3.1. Speech Corpus

The acoustic model was trained on a purpose-built Kazakh read-speech corpus recorded specifically for this work. The corpus comprises recordings of 32 thematically diverse textual sources totalling approximately 3000 pages and 987,086 running words. The

source materials span a broad range of registers and domains, including Kazakh literature (including the novel *Abai Zholy*), history of Kazakhstan, pedagogy, biology, anatomy, physiology, physics, chemistry, marketing, management, philosophy, linguistics (intonology, onomastics, ethnolinguistics), criminology, art history, political science, jurisprudence, theology, and cultural studies. Textbooks from secondary and higher education institutions were used as primary sources in order to ensure lexical and stylistic diversity representative of formal written Kazakh. The thematic composition of the corpus is summarised in Table 1.

Table 1. Thematic composition of the Kazakh speech corpus.

No.	Domain	Approx. Pages
1	Kazakh literature	481
2	History of Kazakhstan	174
3	Abai Zholy (novel)	241
4	Banking documents	92
5	Biology	128
6	Pedagogy	126
7	Intonology	94
8	Criminology	99
9	Anatomy	106
10	Onomastics	180
11	Physiology	94
12	Theory of literature and Poetry	51
13	Physics	25
14	Cultural studies	93
15	Chemistry	18
16	Marketing	92
17	Religious studies	82
18	Management	89
19	Art history	101
20	Philosophy	97
21	Ethnolinguistics	82
22	Political science	87
23	Jurisprudence	88
	Total	**≈3000**

The corpus was recorded by two professional speakers—one male and one female—both aged between 46 and 48 years. Speakers were instructed to follow Kazakh orthoepic norms throughout the recording sessions, ensuring standard pronunciation, correct stress placement, and adherence to the phonological rules of the literary language. Although recordings were collected from both speakers to maximise corpus coverage, the acoustic model reported in this work was trained exclusively on the female speaker's data, which comprised 21.4 GB after cleaning (Table 2). This decision was made on the basis of recording session consistency and the lower proportion of data lost during cleaning for this speaker. In the revised study, the male speaker's recordings (16.4 GB after cleaning) were used to train a separate VITS model (Section 3.9), so that both voices of the corpus are now covered. Recording sessions were conducted in a controlled acoustic environment. The raw audio data totalled 117.2 GB across both speakers. After the cleaning procedure described in Section 3.1.2, a substantial portion of recordings was removed due to misalignment, prohibited characters, or normalisation conflicts, yielding a final corpus of 37.8 GB. The data volumes before and after cleaning are summarised in Table 2.

Table 2. Audio data volume by speaker before and after cleaning. The acoustic model was trained on the female speaker's data only (21.4 GB after cleaning); separate VITS models were trained for both speakers.

Speaker	Raw (GB)	After Cleaning (GB)
Male	62.3	16.4
Female	54.9	21.4
Total	**117.2**	**37.8**

3.1.1. Existing Publicly Available Corpora

In addition to the purpose-built corpus described above, two publicly available Kazakh speech datasets exist and were considered in the design of this work. The KazakhTTS corpus [9], released by the Institute of Smart Systems and Artificial Intelligence (ISSAI) at Nazarbayev University, is the first large-scale open-source Kazakh TTS dataset. It contains recordings from multiple speakers with validated transcriptions and was designed explicitly for neural TTS research. Its successor, KazakhTTS2 [10], extends the original resource with additional speakers and a broader range of recording conditions, further improving its utility as a training and evaluation benchmark for Kazakh TTS systems.

3.1.2. Dataset Cleaning

Prior to training, the raw recordings and their corresponding transcriptions underwent a systematic quality control and cleaning procedure. The following filtering rules were applied to the audio–text pairs:

1. **Latin character filtering:** Utterances whose transcriptions contained sequences of Latin letters forming words or abbreviations—indicating code-switched, transliterated, or incorrectly transcribed content—were removed, as such entries are inconsistent with the Kazakh Cyrillic orthography targeted by the system. Isolated Latin characters appearing as part of established Kazakh loanword spellings or proper nouns that are conventionally written with a small number of Latin characters within an otherwise Cyrillic transcription were retained, as these are a legitimate feature of written Kazakh and must be covered by the model vocabulary.
2. **Strict audio–text alignment:** Pairs in which the recorded audio did not correspond word-for-word to the reference transcription were discarded, ensuring that no misaligned or partially read utterances remained in the dataset.
3. **Abbreviation removal:** Utterances containing abbreviated forms were excluded, as abbreviations introduce ambiguous expansion and inconsistent pronunciation patterns that cannot be reliably learned at the character level.
4. **Arabic numeral removal:** Pairs containing Arabic-script digits (0–9) were removed to avoid inconsistent grapheme-to-pronunciation mappings for numeral expressions.
5. **Roman numeral removal:** Pairs containing Roman numerals were likewise excluded for the same reason.

In addition to the above filtering steps, a text normalisation pass was applied to the retained transcriptions: duplicate whitespace characters and non-meaningful symbols were removed from all transcription strings. Where this normalisation altered the word content of a transcription such that the cleaned text no longer matched the corresponding audio word-for-word, the affected sample was also removed from the dataset. These cleaning steps collectively ensured that the final training corpus consists exclusively of well-formed Kazakh Cyrillic text paired with accurately corresponding speech recordings.

The substantial reduction in corpus volume—from 117.2 GB raw to 37.8 GB after cleaning—reflects the deliberately strict quality criteria applied at this stage. The pri-

mary goal of this work was to establish a clean, reliable baseline system rather than to maximise corpus size. Misaligned, partially read, or normalisation-conflicted utterances introduce contradictory supervision signals that are particularly damaging for attention-based sequence-to-sequence models, where even a small proportion of misaligned pairs can prevent the attention mechanism from converging to a monotonic diagonal alignment. Stricter data quality was therefore preferred over higher data quantity for this initial system. More sophisticated retention strategies—such as forced alignment followed by automatic re-segmentation of borderline recordings, or semi-automatic correction of minor transcription mismatches—could recover a significant proportion of the discarded material and are identified as a direction for future work.

3.2. System Architecture

The proposed TTS system consists of two independently trained components: (1) a Tacotron 2 acoustic model that maps input character sequences to predicted mel-spectrograms, and (2) a Parallel WaveGAN neural vocoder that synthesises waveforms from the predicted spectrograms. At inference time, the two components are combined in a sequential pipeline as illustrated in Figure 1.

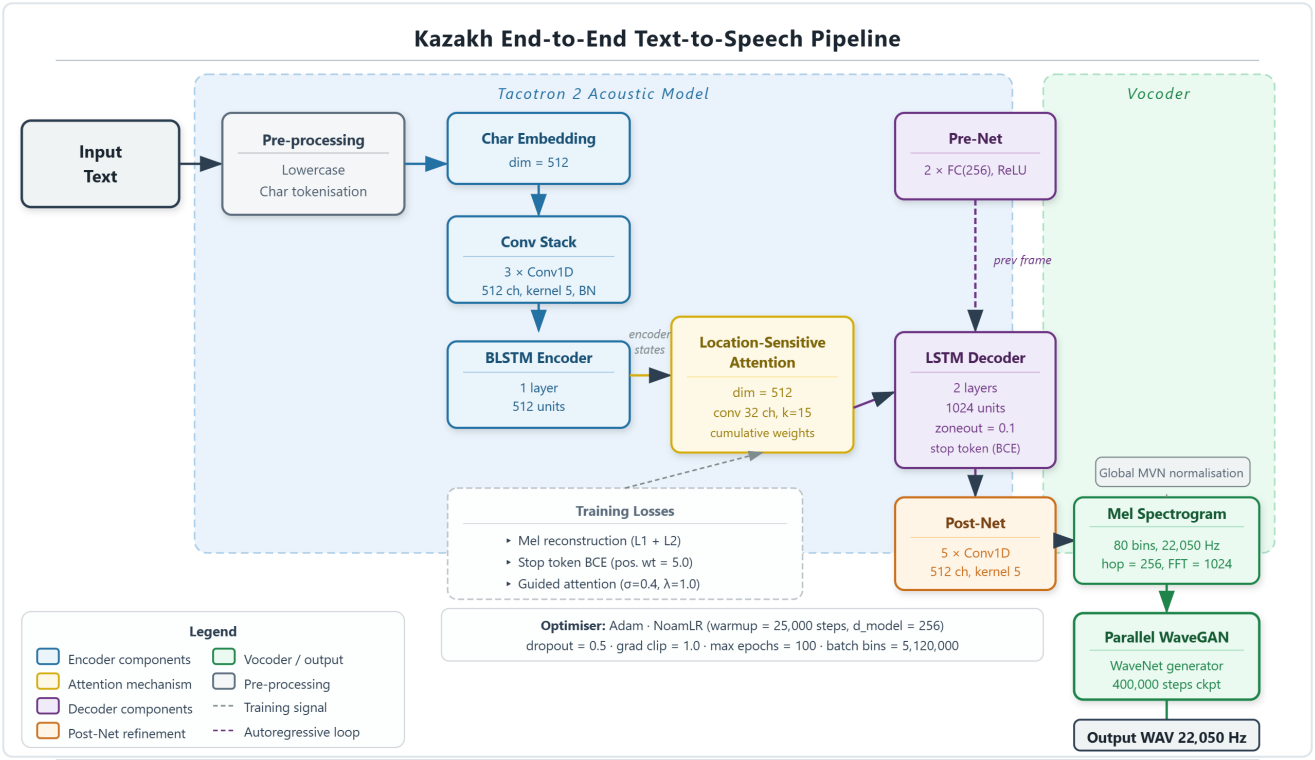


Figure 1. Architecture of the proposed Kazakh end-to-end TTS system.

The pipeline consists of two stages. The first is the Tacotron 2 acoustic model (blue shaded region), which maps a lowercased, character-tokenised input sequence through a three-layer convolutional encoder followed by a bidirectional LSTM, location-sensitive attention, and an autoregressive two-layer LSTM decoder with post-net refinement, producing an 80-bin mel-spectrogram as output. The second stage is the Parallel WaveGAN neural vocoder (green shaded region), which synthesises the final waveform at 22,050 Hz directly from the predicted mel-spectrogram. Dashed lines in the figure indicate the autoregressive feedback loop in which the previous mel frame is fed back to the Pre-Net at each decoder step, and the training-only guided attention loss signal applied to the attention weight matrix.

3.3. Acoustic Model: Tacotron 2

The acoustic model follows the Tacotron 2 architecture with the configuration summarised in Table 3. The encoder consists of a character embedding layer followed by three convolutional layers and a single bidirectional LSTM (BLSTM) layer. The attention mechanism is of the location-sensitive type, which maintains a running cumulative sum of attention weights to encourage monotonic alignment. The decoder comprises two LSTM layers that autoregressively predict mel-spectrogram frames; a five-layer convolutional postnet refines the initial prediction.

Table 3. Tacotron 2 model configuration.

Hyperparameter	Value
Character embedding dimension	512
Encoder BLSTM layers/units	1/512
Encoder conv. layers/channels/filter size	3/512/5
Attention type	Location-sensitive
Attention dimension	512
Attention conv. channels/filter size	32/15
Decoder LSTM layers/units	2/1024
Pre-net layers/units	2/256
Post-net layers/channels/filter size	5/512/5
Reduction factor	2
Dropout rate	0.5
Zoneout rate	0.1
BCE positive weight	5.0
Guided attention σ/λ	0.4/1.0

Guided Attention Loss

To accelerate the learning of diagonal attention alignments, we employ guided attention loss [6] in addition to the standard mel-spectrogram reconstruction and binary cross-entropy (BCE) stop token losses. The guided attention loss penalises attention weights that deviate from the expected near-diagonal alignment according to

$$\mathcal{L}_{\text{guided}} = \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} W_{nt} \cdot A_{nt}, \tag{1}$$

where A_{nt} is the attention weight at encoder step n and decoder step t, and $W_{nt} = 1 - \exp\left(-(n/N - t/T)^2/(2\sigma^2)\right)$ is a soft penalty mask with $\sigma = 0.4$. The loss is weighted by $\lambda = 1.0$.

3.4. Neural Vocoder: Parallel WaveGAN

The neural vocoder is a pre-trained Parallel WaveGAN model with a sampling rate of 22,050 Hz. The vocoder was trained independently and loaded from a checkpoint at 400,000 training steps. After loading, weight normalisation is removed to make inference faster (the `remove_weight_norm()` function of PyTorch 2.4.1). The vocoder takes mel-spectrograms predicted by the acoustic model as conditioning features and produces the final audio waveform.

3.5. Feature Extraction

Acoustic features are extracted using the ESPnet2 `fbank` extractor with the parameters listed in Table 4. Global mean-variance normalisation (MVN) is applied to the features using statistics computed over the training set.

Table 4. Feature extraction configuration.

Parameter	Value
Sampling rate	22,050 Hz
Fast Fourier transform (FFT) size (N_{FFT})	1024
Hop length	256 samples
Number of mel filters	80
Minimum frequency (f_{min})	80 Hz
Maximum frequency (f_{max})	7600 Hz
Normalisation	Global MVN

3.6. Token Vocabulary

Text is tokenised at the character level without any grapheme-to-phoneme conversion. The vocabulary covers all characters present in the training corpus, including the eight Kazakh-specific Cyrillic characters not present in the standard Russian alphabet (Unicode code points U+04D9, U+0493, U+049B, U+04A3, U+04E9, U+04B1, U+04AF, U+0456) as well as standard Cyrillic letters, punctuation marks, and a small number of isolated Latin characters that appear conventionally within otherwise Cyrillic Kazakh text in loanwords and proper nouns. These differ from the Latin-word sequences removed during cleaning (Section 3.1.2): the cleaning step targeted fully Latin-script words and abbreviations indicating transcription errors or code-switching, whereas the retained Latin characters are single letters embedded within Cyrillic text that represent an established orthographic convention in written Kazakh. The complete vocabulary contains 145 tokens, including the special tokens `<blank>`, `<unk>`, `<space>`, and `<sos/eos>`. Input text is lowercased before tokenisation to reduce vocabulary sparsity.

3.7. Tacotron 2 Training Setup

3.7.1. Optimiser and Learning Rate Schedule

The model is trained using the Adam optimiser with a NoamLR learning rate schedule [29], which warms up the learning rate over the first 25,000 steps and then decays it proportionally to the inverse square root of the step count. The schedule is parameterised by a model size of 256 and an initial scale factor of 2.0 (as implemented in ESPnet2).

$$\text{lr}(s) = d_{\text{model}}^{-0.5} \cdot \min\left(s^{-0.5},\ s \cdot s_{\text{warmup}}^{-1.5}\right), \tag{2}$$

where $d_{\text{model}} = 256$ is the model size parameter and s_{warmup} = 25,000 is the warmup step count.

3.7.2. Batch Construction and Training Details

Training samples are constructed by the `numel` batch type, which groups utterances such that the total number of elements per batch does not exceed a specified bin size. This approach accommodates the variable-length nature of speech and text sequences. The remaining key training hyperparameters are summarised in Table 5.

Table 5. Training hyperparameters of the Tacotron 2 model.

Hyperparameter	Value
Maximum epochs	100
Gradient clipping norm	1.0
Gradient accumulation steps	2
Batch bins	5,120,000

Table 5. *Cont.*

Hyperparameter	Value
Batch type	`numel`
Sort in batch	Descending
Training precision	`float32`
Number of best models kept	5
Graphics processing unit (GPU)	1 × CUDA device
Model selection criterion	Minimum validation loss

3.8. Inference Pipeline

The inference script accepts arbitrary Kazakh text as a command-line argument. The text is lowercased and passed to the `Text2Speech` module from ESPnet2's `tts_inference` application programming interface (API). Key inference hyperparameters for the Tacotron 2 decoder include a stop-token threshold of 0.5, a minimum length ratio of 0.0, and a maximum length ratio of 10.0. Attention constraint (forward–backward window) is enabled with a backward window of 1 and a forward window of 3, which prevents attention from regressing or skipping excessively during autoregressive decoding. Griffin–Lim waveform synthesis is disabled; the Parallel WaveGAN vocoder is used exclusively. The synthesised audio is written to a 32-bit PCM WAV file at 22,050 Hz.

3.9. End-to-End VITS System

To compare the two-stage baseline with a more recent architecture, as suggested by both reviewers, we trained VITS [11] models using the Coqui TTS toolkit [30]. VITS generates the waveform directly from the character sequence. A text encoder and a flow-based prior are trained together with a posterior encoder over linear spectrograms. Monotonic alignment search takes the place of attention, a stochastic duration predictor models token durations, and the decoder is a HiFi-GAN-type generator trained against multi-period and multi-scale discriminators [21]. Training VITS thus brings a HiFi-GAN-type waveform generator into the comparison, and it also removes the attention mechanism, whose failures we analyse in Section 4.5.

The VITS models use the same character-level (no G2P) input representation as Tacotron 2, but audio was resampled to 24,000 Hz (FFT size 1024, window 1024, hop length 256, 80 mel bins). The model has 83.07 M parameters. Training used the default Coqui VITS optimisation settings (AdamW, initial learning rate 2×10^{-4} with per-epoch exponential decay), a single NVIDIA GPU, random seed 54321, and fp16 mixed precision unless stated otherwise. Separate single-speaker models were trained for the male speaker (1322 steps per epoch) and for the female speaker. For the female speaker, three successive versions of the cleaned training set, of different sizes, were used (1579 steps per epoch for version 1 and 1929 steps per epoch for versions 2 and 3). Training configurations are summarised in Table 6.

Table 6. VITS training runs (24 kHz, 83.07 M parameters, single GPU). "Best val. mel" is the lowest validation mel loss observed and the step at which it was reached.

Speaker, Data	Initialisation	Precision	Steps/Epoch	Steps	Best Val. Mel (Step)
Male	scratch	fp16	1322	421.8k	16.99 (333.6k)
Female, v1	resumed (182k)	fp16	1579	408.4k	16.19 (403.0k)
Female, v2	scratch	fp16	1929	194.7k	16.26 (169.7k)
Female, v3	scratch	fp16	1929	352.7k	17.54 (328.9k) *
Female, v3	fine-tune (315k)	fp32	1929	+291.2k	16.92 (+71.3k)

* Training diverged (NaN losses) at about 342k steps.

3.10. Software

The Tacotron 2 experiments used ESPnet2 (version 202511; ESPnet open-source project) and the Parallel WaveGAN toolkit (version 0.6.1; open-source project), both running on PyTorch 2.4.1 (PyTorch Foundation, San Francisco, CA, USA) with CUDA 12.1 (NVIDIA Corporation, Santa Clara, CA, USA). The VITS models were trained with the Coqui TTS toolkit (coqui-tts version 0.26.2; Idiap Research Institute, Martigny, Switzerland). For the objective evaluation, we used NVIDIA NeMo 3.0.0 (NVIDIA Corporation, Santa Clara, CA, USA), faster-whisper 1.2.1 (SYSTRAN, Paris, France) with the Whisper-medium model (OpenAI, San Francisco, CA, USA), and jiwer 4.0.0 (open-source project). The web application uses Flask (Pallets Projects, open-source project) and GPT-4.1-mini (OpenAI, San Francisco, CA, USA).

4. Results

4.1. Training Convergence

The Tacotron 2 model was configured for a maximum of 100 epochs and was trained for 43 epochs on a single GPU. The checkpoint at epoch 41, saved as `41epoch.pth`, was selected as the primary model for evaluation based on validation loss criteria. The model selection strategy retained the five best checkpoints according to both validation loss and training loss, providing robustness against overfitting.

The guided attention loss played a critical role in early training: without it, Tacotron 2 is known to fail to learn a monotonic alignment for an extended period, particularly for languages with long words or morphologically complex inputs. With $\sigma = 0.4$ and $\lambda = 1.0$, the attention mechanism transitioned from a completely unstructured pattern at epoch 1 to a well-formed diagonal by epoch 6, and had fully converged to a sharp, clean alignment by epoch 21, as shown in Section 4.2. This rapid convergence is particularly notable given the relatively large and linguistically diverse character vocabulary of 145 tokens required to cover the Kazakh Cyrillic script.

4.2. Attention Alignment

Figure 2 illustrates decoder attention maps recorded at epochs 1, 6, 10, and 21, revealing the progressive emergence of monotonic alignment during training. At epoch 1, the attention weights are completely diffuse and dominated by vertical stripes, indicating that the model has not yet learned any structured alignment and is attending to a small number of encoder positions (primarily around input index 45–55) regardless of the current decoder step. This degenerate behaviour is common in early Tacotron 2 training without guided attention.

By epoch 6, a clear diagonal band has begun to form, showing that the model has learned to advance its attention focus monotonically through the input sequence as decoding progresses. However, the band remains relatively broad and exhibits noisy, scattered secondary activations alongside the main diagonal, reflecting imperfect but rapidly improving alignment.

At epoch 10, the diagonal structure is well-established and sharper, with secondary activations substantially reduced. A small degree of residual scatter remains, particularly in the middle of the sequence, but the overall monotonic progression from the first to the last input character is unambiguous.

By epoch 21, the attention map has converged to a sharp, clean, near-perfectly diagonal stripe with negligible off-diagonal energy. This indicates that the model has learned a stable and precise character-to-frame alignment. The use of cumulative attention weights (`cumulate_att_w: true`) was important in achieving this sharpness, as it prevented the

decoder from dwelling on or revisiting the same encoder positions across consecutive time steps.

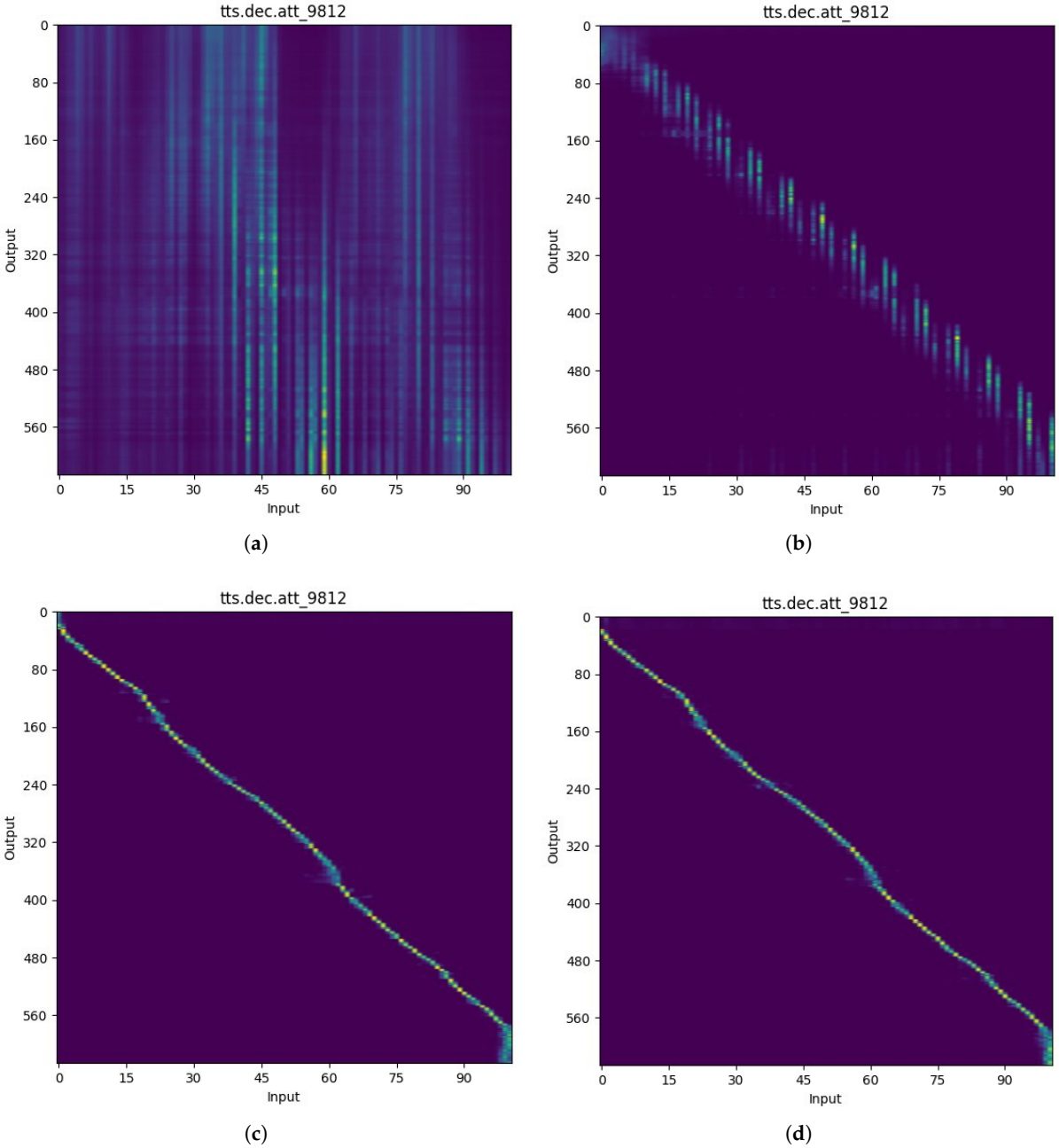


Figure 2. Decoder attention alignment maps at epochs 1, 6, 10, and 21 of Tacotron 2 training on the Kazakh corpus: (**a**) epoch 1, diffuse, unstructured attention with vertical stripes; (**b**) epoch 6, diagonal band emerging, broad with noisy scatter; (**c**) epoch 10, clear diagonal with significantly reduced residual scatter; (**d**) epoch 21, sharp, converged diagonal with negligible off-diagonal energy. The horizontal axis represents input character positions and the vertical axis represents output mel-spectrogram frame positions. Progressive convergence from a diffuse, unstructured pattern (**a**) to a sharp near-diagonal alignment (**d**) demonstrates the effectiveness of guided attention loss in accelerating monotonic alignment learning.

4.3. Synthesised Speech Quality

The trained system produced intelligible Kazakh speech for a range of test utterances drawn from both in-domain and out-of-domain text. Prosodic patterns, including

sentence-final falling intonation and stress patterns typical of Kazakh, were reproduced qualitatively. The Parallel WaveGAN vocoder, operating on the predicted 80-dimensional mel-spectrograms, generated waveforms free of the buzzing or metallic artefacts typically associated with Griffin–Lim synthesis.

Table 7 summarises the synthesis configuration used during evaluation.

Table 7. Inference configuration for speech synthesis evaluation.

Parameter	Value
Acoustic model checkpoint	`41epoch.pth`
Vocoder checkpoint	`checkpoint-400000steps.pkl`
Stop token threshold	0.5
Minimum/maximum length ratio	0.0/10.0
Attention constraint (backward/forward)	1/3
Output sampling rate	22,050 Hz
Output format	32-bit PCM WAV

4.4. Token Coverage

The 145-token character vocabulary provided complete coverage of the Kazakh Cyrillic character set as well as punctuation and selected Latin characters present in the training corpus. No out-of-vocabulary (OOV) characters were encountered in standard Kazakh text, confirming that character-level modelling without G2P conversion is sufficient for this language.

4.5. Objective Evaluation

To complement the qualitative assessment of synthesised speech, we conducted an objective evaluation comparing the proposed Tacotron 2 + Parallel WaveGAN system against a Tacotron 2 + Griffin–Lim baseline and against the end-to-end VITS system. Both Tacotron 2 systems use the identical acoustic model checkpoint, ensuring that any difference in scores between them reflects the vocoder alone.

4.5.1. Evaluation Protocol

Test set. In the original submission, the test set contained 50 sentences. Following the reviewers' comments, we wrote a new test set of 138 sentences (2016 words): six for each of the 23 domains in Table 1, with 48 short ($\leq$8 words), 65 medium (9–20 words) and 25 long ($>$20 words) sentences in total. The sentences were written for this evaluation, do not occur in the training data, and, consistent with the cleaning rules of Section 3.1.2, contain no numerals, abbreviations or Latin-script words. The test set is provided in Supplementary Table S1.

ASR models. Intelligibility was measured by transcribing each synthesised utterance automatically and comparing the hypothesis to the reference text. Two ASR models were used. The *primary* model is the Kazakh–Russian FastConformer hybrid transducer/connectionist temporal classification (CTC) model `stt_kk_ru_fastconformer_hybrid_large` [31,32], which was trained on Kazakh speech (Kazakh Speech Corpus 2 (KSC2), Common Voice) and is far more accurate on Kazakh than multilingual models. As a *secondary* model, and for continuity with the original submission, we report the multilingual Whisper *medium* model [33] (CTranslate2 implementation, language fixed to Kazakh, beam size 5, temperature 0). Two metrics were computed using the `jiwer` library: character error rate (CER) and word error rate (WER), both normalised by lowercasing and removing punctuation prior to comparison. A total of 95% confidence intervals (CIs) were computed by bootstrap resampling over sentences (10,000 resamples), and differences between systems were assessed with a paired bootstrap test.

Synthesis. The VITS utterances were generated with the deployed models (male speaker: 400k-step checkpoint; female speaker: 290k-step checkpoint) through the inference API of the web application (Section 4.8). For all 276 VITS utterances, the application's text front-end returned the input text unchanged, so the results reflect the acoustic models alone. Real-time factor (RTF) was measured as the ratio of synthesis wall-clock time to the duration of the synthesised audio, averaged over all utterances; RTF values are reported only where they were measured under comparable conditions (Section 4.7).

4.5.2. Intelligibility Results

The objective evaluation results are presented in Table 8.

Table 8. Objective evaluation of Kazakh TTS systems. RTF is averaged over all test utterances; values below 1.0 indicate faster-than-real-time synthesis. Values in brackets are 95% bootstrap CIs over sentences.

System	Kazakh ASR (Primary)		Whisper-Medium		
	CER (%) ↓	WER (%) ↓	CER (%) ↓	WER (%) ↓	RTF ↓
Original 50-sentence test set					
Tacotron 2 + Griffin–Lim	–	–	43.19	77.04	0.41
Tacotron 2 + Parallel WaveGAN	–	–	42.70	79.95	0.41
New 138-sentence test set					
Tacotron 2 + Griffin–Lim †	8.79 [4.93, 13.49]	10.33 [6.51, 14.93]	23.03 [15.25, 32.57] ‡	50.59 [46.63, 54.79] ‡	–
Tacotron 2 + Parallel WaveGAN †	8.79 [4.94, 13.48]	10.36 [6.50, 14.95]	23.35 [15.41, 32.68] ‡	50.53 [46.50, 54.70] ‡	–
VITS, female speaker	**2.08** [1.59, 2.60]	7.63 [5.91, 9.42]	19.50 [15.67, 26.59] *	71.24 [68.06, 74.44]	–
VITS, male speaker	**1.94** [1.47, 2.44]	**6.92** [5.53, 8.34]	17.39 [16.45, 18.39]	73.00 [70.31, 75.77]	–

* Includes one Whisper hallucination (a repeated syllable loop on a five-word sentence that the Kazakh model transcribed correctly); without it, the CER is 16.35%. † Mean of two independent inference passes (different random seeds; the Tacotron 2 pre-net dropout is active at inference). In 17 of the 276 Tacotron 2 + Parallel WaveGAN (PWG) utterances (6.2%), the attention failed and the output was (near-)silent; these failures are included in the scores. ‡ First inference pass only.

Tacotron 2 systems. Both Tacotron 2 systems achieve faster-than-real-time synthesis, with an average RTF of 0.41 for both configurations, confirming that the choice of vocoder does not significantly affect synthesis speed for this pipeline. On the original 50-sentence set, the Tacotron 2 + Parallel WaveGAN system achieved a slightly lower Whisper-based CER (42.70%) than the Griffin–Lim baseline (43.19%), while its WER was slightly higher (79.95% vs. 77.04%). Neither difference is large relative to the sentence-level variability, and we do not interpret it as evidence that one vocoder is more intelligible than the other. The Griffin–Lim system produced 2 empty or near-empty Whisper transcriptions out of 50 utterances, while the Parallel WaveGAN system produced none.

VITS systems. With the Kazakh ASR model, both VITS voices are highly intelligible: the CER is 1.94% for the male and 2.08% for the female speaker, and the WER is 6.92% and 7.63%, respectively; 79 (male) and 83 (female) of the 138 sentences were transcribed without a single error. The two voices do not differ significantly (paired bootstrap CER difference -0.14 percentage points, 95% CI $[-0.54, 0.27]$), which shows that the VITS training recipe transfers between speakers without modification. Intelligibility degrades only slightly with sentence length (CER 1.41/2.02/2.06% for the male and 0.97/2.00/2.74% for the female voice on short/medium/long sentences), and it is consistent across domains (per-domain CER between 0.3% and 4.2%), which fits the fact that VITS has no attention mechanism that could skip or repeat parts of a long input. No utterance produced an empty output. Served by the web application, VITS produced audio about ten times faster than real time (server response time divided by audio duration: 0.08–0.10, including network transfer).

VITS vs. Tacotron 2. Because Tacotron 2 inference is stochastic (dropout in the prenet is kept active at inference time), it was evaluated in two independent passes over the 138 sentences, and per-sentence error counts were averaged. With the Kazakh ASR model, Tacotron 2 reaches a CER of 8.79% and a WER of 10.3–10.4% with either vocoder, compared with about 2% CER for VITS. VITS lowers the CER significantly for both voices (paired bootstrap difference against Tacotron 2 + PWG: 6.71 percentage points [2.84, 11.47] for the female and 6.85 [3.03, 11.55] for the male voice; $p < 0.001$). The WER difference over all sentences (2.7–3.4 percentage points in favour of VITS) is not significant, because whole-sentence failures inflate CER more than WER and the CIs are wide.

The error analysis shows where this advantage comes from. Tacotron 2 failed on 8–9 utterances per pass (17 of 276, 6.2%; 12 different sentences, 5 of which failed in both passes): the location-sensitive attention stalled near the beginning of the input, the stop token was never predicted, and the decoder ran to the maximum output length (10 times the input length) while producing near-silence. Such failures occurred for short as well as long sentences and are not caused by the vocoder, since both vocoders received identical spectrograms. VITS cannot fail in this way, since monotonic alignment search always covers the whole input, and it produced no failed utterances. On the 126 sentences on which Tacotron 2 never failed, the two architectures are comparably intelligible: the CER of Tacotron 2 (2.57%) does not differ significantly from that of VITS (2.09% female, 1.88% male), and the WER of Tacotron 2 is even lower (4.35% vs. 7.86% and 6.84%; significant). The vocoder has no measurable effect on intelligibility (8.79% CER for both PWG and Griffin–Lim); any difference between the vocoders would show up in naturalness, which only a listening test can measure.

4.5.3. Robustness on Complex Educational Text

Educational materials contain mathematical and chemical formulas, numbers, dates, units, abbreviations and foreign (mostly IT) terms, which were deliberately excluded from the training data (Section 3.1.2). In the deployed system, such input is first converted into its spoken Kazakh form by the text front-end (Section 4.8); for example, $x^2 + 5x - 6 = 0$ becomes “iks kvadrat plyus bes iks azaitu alty teñ nöl” and *API* becomes “ei pi ai”. The resulting text is atypical for the acoustic model: it contains spelled-out Latin letter names, long chains of number words and unusual word sequences that never occur in the training corpus.

To test whether the acoustic models remain robust on such input, we composed a second test set of 60 sentences (588 words after normalisation) in six categories of ten sentences each: algebraic formulas; physical and chemical formulas; numbers, dates and units; abbreviations; foreign and IT terms; and mixed textbook-style sentences (e.g., *Laboratory work No. 5: preparing a 10% NaCl solution*). Each sentence was normalised once by the front-end, and the *same* normalised text was synthesised by both VITS voices and by Tacotron 2 (two independent passes, as above). The normalised text serves as the reference, so the test measures how faithfully the acoustic model reads the spoken form, independently of the correctness of the normalisation itself. Transcription and scoring followed the protocol of Section 4.5 (Kazakh FastConformer ASR). Results are given in Table 9; the test set is provided in Supplementary Table S2.

On this material the gap between the two architectures is much wider than on plain text. The attention failure rate of Tacotron 2 almost doubles, from 6.2% on plain text to 11.7% (14 of 120 utterances), with failures in every category, whereas VITS again produced no failed utterance. As a result, VITS reduces the CER from 19.69% to 7.23% (female) and 8.45% (male); both reductions are significant (paired bootstrap, $p < 0.001$). The WER reduction is significant for the female voice (6.1 percentage points [0.3, 12.5], $p = 0.02$) but

not for the male voice (2.6 [−4.2, 10.1]). The largest gains are observed for numbers, dates and units (WER 9.5–11.9% vs. 22.3%) and for mixed textbook sentences (18.1–25.0% vs. 33.2%), that is, for long runs of number words and mixed content, where the attention of Tacotron 2 most easily loses its place. In the deployed system, it was the switch to VITS that made formulas, numbers and abbreviations read reliably.

Table 9. Intelligibility on complex educational text (60 sentences with formulas, numbers, units, abbreviations and foreign terms, normalised by the text front-end). Kazakh FastConformer ASR; 95% bootstrap CIs in brackets. "Failed" counts (near-)silent outputs caused by attention failure. Tacotron 2: mean of two inference passes.

System	CER (%) ↓	WER (%) ↓	Failed
Tacotron 2 + Griffin–Lim	20.00 [13.11, 27.87]	31.55 [23.84, 39.88]	13/120
Tacotron 2 + Parallel WaveGAN	19.69 [12.86, 27.44]	31.12 [23.53, 39.30]	14/120
VITS, female speaker	**7.23** [5.52, 9.07]	**25.00** [20.00, 29.93]	**0/60**
VITS, male speaker	8.45 [6.51, 10.48]	29.12 [23.45, 34.71]	**0/60**
WER (%) by category	**VITS female**	**VITS male**	**Tacotron 2 + PWG**
Algebraic formulas	29.8	30.6	28.2
Physical and chemical formulas	36.2	46.6	41.0
Numbers, dates and units	11.9	9.5	22.3
Abbreviations	14.9	19.4	17.2
Foreign and IT terms	38.7	42.1	43.3
Mixed textbook sentences	18.1	25.0	33.2

The absolute WER on this test set is higher than on plain text for all systems, partly because of the ASR evaluation itself: spelled-out letter names are often merged into a single token by the ASR model (e.g., "pi di ef" → "pidie"), and isolated Latin letters remained un-normalised in 8 of 120 front-end outputs and cannot be pronounced by the Cyrillic-only acoustic models. Formulas with many single-letter variables (physics and chemistry) remain the most difficult category for all systems.

4.5.4. Analysis of the High Error Rates and Limitations of Whisper-Based Evaluation

Reviewer 2 rightly questioned the very high Whisper-based error rates in the original submission. Transcribing the *same* VITS audio with both ASR models shows that most of this error comes from Whisper, not from the synthesiser. For the male voice, Whisper yields a CER of 17.39% and a WER of 73.00%, whereas the Kazakh model yields 1.94% and 6.92%: Whisper overestimates the character error by a factor of about 9 and the word error by a factor of about 10. Whisper did not transcribe any of the 276 VITS utterances without error, whereas the Kazakh model transcribed 162 of them perfectly. Whisper also produced a hallucination (an endless repetition of one syllable) on a short sentence that the Kazakh model transcribed correctly.

Three things explain this. First, Kazakh is poorly represented in Whisper's training data, and most of Whisper's errors are wrong but similar-sounding vowels and consonants inside words (e.g., *erekshe* → *erikshi*, *öleñderinde* → *ölimdirinde*, in Latin transliteration); such within-word substitutions account for 82–84% of Whisper's word errors. Second, the confusion of Kazakh-specific letters with their Russian Cyrillic counterparts is a secondary effect: mapping both texts onto the Russian alphabet lowers Whisper's CER only from 17.39% to 16.03%. Word segmentation errors are negligible (ignoring spaces changes the CER by less than 0.5 percentage points). Third, WER is a harsh measure for an agglutinative language: a Kazakh word is long and usually carries several suffixes, so a single wrong character turns the entire word into an error. With Whisper, a CER of about 17% therefore already corresponds to a WER above 70%.

With the Kazakh ASR model, the remaining word errors of VITS are dominated by within-word substitutions (51–54%), errors in inflectional endings with a correct stem (20–25%, e.g., *zhyry* → *zhyr*), insertions (18–20%) and deletions (7%). Part of the insertions are ASR artefacts: the same spurious words were inserted at the same positions for both voices. These figures are therefore an upper bound on the error caused by the synthesisers themselves.

Whisper is also unreliable for comparing architectures. Under Whisper, Tacotron 2 obtains a higher CER than VITS (23.0–23.4% vs. 17.4–19.5%), in line with the Kazakh ASR model, but a markedly lower WER (50.5% vs. 71.2–73.0%), so the two Whisper metrics rank the systems in opposite orders, while the Kazakh model ranks VITS first on both. We suspect that Whisper reacts to acoustic properties that differ between the systems (Tacotron 2 output is band-limited to 7.6 kHz at 22.05 kHz, VITS output is full-band at 24 kHz) rather than on intelligibility to a native listener. We therefore do not use Whisper to compare systems in this revision, and we recommend Kazakh-specific ASR models for future Kazakh TTS benchmarks. This also resolves the apparent contradiction in the original submission between the high Whisper-based error rates and the high MOS intelligibility scores.

These observations are complemented by subjective listening evaluations (mean opinion score) with native Kazakh speakers, described next.

4.6. Subjective Evaluation

To assess the perceived naturalness and intelligibility of the proposed system, a mean opinion score (MOS) listening study was conducted with seven native Kazakh-speaking listeners. Participants were presented with synthesised utterances from the Tacotron 2 + Parallel WaveGAN system and asked to rate each sample independently on two dimensions—naturalness and intelligibility—using a five-point scale (1 = Bad, 2 = Poor, 3 = Fair, 4 = Good, 5 = Excellent). The results are summarised in Table 10.

Table 10. Mean opinion score (MOS) results for the Tacotron 2 + Parallel WaveGAN system. Scores are averaged over seven native Kazakh-speaking listeners on a 1–5 scale.

Dimension	MOS (Out of 5.0)
Naturalness	4.5
Intelligibility	4.5

Both dimensions received a mean score of 4.5, indicating that native listeners judged the synthesised speech to be natural-sounding and readily intelligible. The high intelligibility rating is consistent with the low error rates measured with the Kazakh ASR model on the utterances where Tacotron 2 does not fail (Section 4.5) and confirms that the high Whisper-based error rates underestimate the intelligibility of Kazakh synthetic speech (Section 4.5.4).

We acknowledge that a panel of seven listeners is smaller than recommended for a fully rigorous Absolute Category Rating test according to ITU-T Recommendation P.800 [34], that the study did not include the VITS systems, and that per-rating data for confidence intervals were not retained. We therefore treat the MOS result as supporting evidence only. The comparison between systems in this revision rests on the objective evaluation, which we extended considerably (two test sets, 198 sentences, two ASR models, confidence intervals and paired significance tests). A blind comparative listening test of all four systems following ITU-T P.800 (randomised presentation, loudness-normalised stimuli, a larger and more diverse native-speaker panel) has been prepared and is planned as the next step (Section 5.8).

4.7. VITS Training Results

Table 6 summarises the VITS training runs, and Figure 3 shows the validation mel-spectrogram reconstruction loss (L1, as logged by Coqui TTS) as a function of the training step.

All runs show the same qualitative behaviour: the mel loss falls quickly from about 24–25 to below 20 within the first 20k steps, and then decreases slowly to between 16.2 and 17.5 after 170k–400k steps. The male-speaker model converged to a validation loss of 16.99, comparable to that of the female models, which indicates that the training recipe transfers between speakers without modification. For the female speaker, data version 2 reached its best validation loss (16.26) in fewer than half the steps of data version 1 (16.19 at 403k steps); these losses are computed on different validation subsets, however, so they are only roughly comparable.

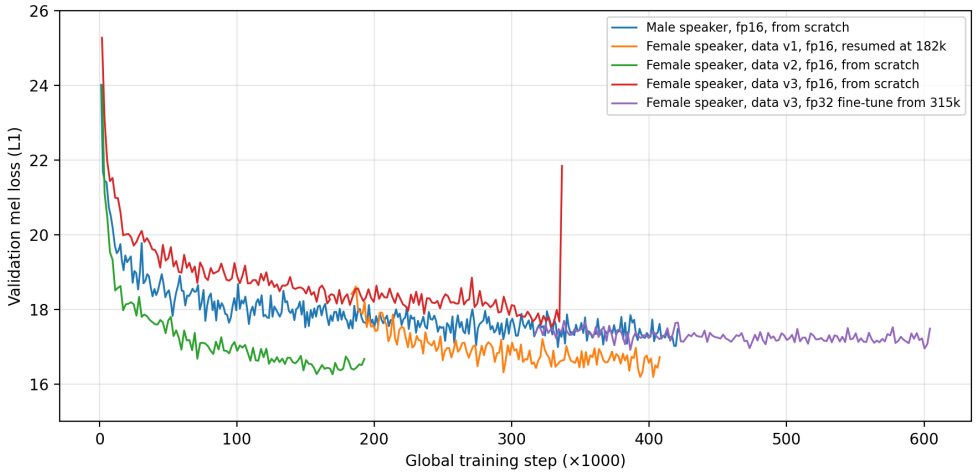


Figure 3. Validation mel loss of the VITS models during training. For the fp32 fine-tuning run, steps are shown as offset from the 315k-step initial checkpoint.

Long fp16 training turned out to be numerically unstable. The run on female data version 3 trained stably for more than 300k steps, but its losses became NaN at about 342k steps (gradient scaler collapsed to zero; Figure 3, spike at the end of the red curve). Restarting from the 315k-step checkpoint in full fp32 precision with a reduced learning rate of 5×10^{-5} removed the problem: this run trained for a further 291k steps without divergence and lowered the validation loss from 17.33 to 16.92. For long VITS training on a single GPU, we recommend fp16 for the first phase and fp32 (or bf16) for late fine-tuning.

VITS generates the whole waveform in one parallel pass, without an autoregressive decoder. On a CPU-only machine, a paragraph of several sentences was synthesised with an RTF of 2.92 by an intermediate checkpoint, whereas the GPU-served models in the web application produce audio about ten times faster than real time, including network transfer (server response time divided by audio duration: 0.08–0.10). For reference, Tacotron 2 + PWG run on a laptop CPU (Apple M1 Pro) reached an RTF of about 0.6 on the 138-sentence test set. A controlled GPU comparison of the two architectures was not performed. The male-speaker VITS models (checkpoints at 330k and 400k steps) are deployed in the web application described in Section 4.8.

4.8. Educational Web Application

The trained models were integrated into a web application (Flask back-end, Coqui TTS synthesiser) intended for Kazakh educational platforms. The user can type text or upload a TXT or DOCX file; the application then (i) cleans the text (removing markup and diacritic

artefacts, expanding superscripts and subscripts), (ii) normalises numbers, abbreviations, formulas and foreign words into Kazakh words when they are detected, (iii) optionally converts the text into its orthoepic (pronounced) form using the orthoepic converter of the Kazakh National Corpus, and (iv) synthesises the audio, which can be played back and downloaded as a WAV file. In the training data, numerals and abbreviations were simply removed (Section 3.1.2); the application handles them instead at inference time. Steps (i)–(iii) form the text front-end. Normalisation (ii) is performed by a large language model (GPT-4.1-mini, accessed through the OpenAI API) with a Kazakh instruction prompt that specifies the reading rules (e.g., "+" is read as *plyus*/*qosu*, "=" as *teñ*, abbreviations are expanded or spelled out); it is invoked only when the input contains digits, Latin letters, mathematical symbols or upper-case abbreviations, and plain Kazakh text bypasses it. The application is publicly available at https://tbisintez.kz (accessed on 1 October 2026).

4.9. Comparison with State-of-the-Art Systems

Table 11 contextualises the proposed system against existing Kazakh TTS approaches and selected high-resource neural TTS baselines reported in the literature. Because no prior work has reported ASR-based CER/WER on a common Kazakh TTS test set, direct numerical comparison is not possible across all systems; instead, we characterise each system along the dimensions most relevant to the low-resource TTS problem: the architecture, the availability of a neural vocoder, the requirement for a G2P module, and whether the system is publicly available or reproducible.

Table 11. Comparison of the proposed system with existing Kazakh TTS approaches and related neural TTS systems. "Neural vocoder" indicates whether the system employs a learned waveform synthesiser rather than Griffin–Lim or a digital signal processing (DSP) vocoder. "G2P required" indicates whether a grapheme-to-phoneme converter is needed. "Open/reproducible" indicates whether training code, model weights, or a public API are available.

System	Architecture	Neural Vocoder	G2P Required	Open/Reproducible
Unit selection (legacy)	Concatenative	–	Yes	No
HMM-based Kazakh TTS	SPSS + HMM	No	Yes	No
KazakhTTS [9]	Tacotron 2/Transformer + PWG	Yes	No	Yes
KazakhTTS2 [10]	Tacotron 2 + PWG	Yes	No	Yes
ElevenLabs Kazakh [35]	Proprietary neural	Yes	N/A	No
Narakeet Kazakh [36]	Proprietary neural	Yes	N/A	No
Proposed system	Tacotron 2 + PWG	**Yes**	**No**	On request
Proposed system	VITS (HiFi-GAN generator)	**Yes**	**No**	Web service

Commercial services such as ElevenLabs [35] and Narakeet [36] offer Kazakh TTS via proprietary neural systems accessible through web APIs. While these systems produce high-quality output, they are closed-source, require internet connectivity and paid subscriptions, and provide no mechanism for the community to inspect, retrain, or adapt the underlying models. They therefore do not constitute reproducible research baselines. In contrast, the proposed system provides a fully documented, open training and inference pipeline based on the ESPnet2 framework, enabling other researchers to replicate, extend, and adapt it for related Turkic languages or domain-specific applications.

Among open research systems, the KazakhTTS [9] and KazakhTTS2 [10] baselines are the closest comparators. Both already combine neural acoustic models with a Parallel WaveGAN vocoder and release ESPnet recipes and pretrained models; the present work therefore does not claim to be the first neural-vocoder Kazakh TTS system. What this work adds is the new multi-domain corpus read under orthoepic control with two professional voices, the documented data-cleaning protocol, the analysis of guided-attention conver-

gence for character-level Kazakh input, the comparison of a two-stage Tacotron 2 + PWG system with an end-to-end VITS system on the same data, the extended evaluation protocol (two test sets, a Kazakh-specific ASR model, confidence intervals, error analysis), and the integration into an educational application. Because the systems are trained on different corpora and evaluated on different test sets, the MOS values reported in [9,10] and in this paper are not directly comparable; a joint evaluation on a common test set is planned (Section 5.8).

5. Discussion

5.1. Summary of Contributions

In short, this work contributes (i) a new multi-domain Kazakh speech corpus read under orthoepic control by two professional speakers, together with a documented cleaning protocol; (ii) a fully specified character-level Tacotron 2 + Parallel WaveGAN pipeline, with evidence that guided attention yields a sharp monotonic alignment within 21 epochs despite a 145-token vocabulary; (iii) end-to-end VITS models at 24 kHz for both speakers, which removed the attention failures of Tacotron 2 on plain and on complex educational text, together with advice on mixed-precision training; (iv) two test sets evaluated with a Kazakh-specific ASR model, with confidence intervals, significance tests and an error analysis that exposes the limits of Whisper for Kazakh; and (v) a deployed web application for Kazakh educational content.

5.2. Character-Level Modelling for Kazakh

A key design decision in this work was to employ character-level tokenisation rather than phoneme-level. For Kazakh, whose Cyrillic orthography is largely phonemic, this decision is well-motivated: each character corresponds to a relatively stable phoneme or phoneme cluster, and the overhead of developing and maintaining a language-specific G2P converter is avoided. The main exceptions are orthoepic phenomena such as consonant assimilation at morpheme boundaries and vowel reduction, which are not reflected in the spelling; in the deployed application these are handled by an optional orthoepic conversion step (Section 4.8). The primary disadvantage of character-level modelling is the longer input sequence, which places greater demands on the attention mechanism; the guided attention loss compensated for this by imposing a strong structural prior on the alignment during training (Figure 2). VITS avoids the issue altogether by replacing attention with monotonic alignment search.

5.3. Effect of the Learning Rate Schedule and Reduction Factor

The NoamLR schedule with 25,000 warmup steps was well-matched to the training corpus size: during warmup, the low learning rate allowed the attention mechanism to begin forming rough alignments without diverging, and the subsequent decay allowed further refinement without oscillation. An initial learning rate scale of 2.0 was found to be effective; reducing it to 1.0 led to slower convergence in preliminary experiments. The final model uses a reduction factor of 2, i.e., the decoder predicts two mel-spectrogram frames per autoregressive step. This halves the number of decoder steps and accelerates both training and inference; the original submission incorrectly reported a reduction factor of 1.

5.4. Two-Stage Versus End-to-End Synthesis

The modular Tacotron 2 + vocoder design allows the vocoder to be replaced without retraining the acoustic model. Both reviewers suggested HiFi-GAN; rather than only swapping the vocoder, we trained VITS, whose waveform generator is a HiFi-GAN-type

generator trained with the HiFi-GAN discriminators, jointly with the acoustic model. With a Kazakh-specific ASR model, both VITS voices reach a CER of about 2% on 138 sentences from 23 domains, with almost no degradation on sentences longer than 20 words (Table 8). On the same sentences, Tacotron 2 reaches 8.79% CER, mainly because 6.2% of its outputs are attention failures; when it does not fail, it is as intelligible as VITS. What VITS improves most is *reliability*. For a system that reads educational materials without anyone checking each sentence, a silent or truncated sentence is the worst kind of error. This effect is even stronger for complex educational text (Section 4.5.3): after normalisation of formulas, numbers and abbreviations, the failure rate of Tacotron 2 rises to 11.7%, while VITS remains failure-free and reduces the CER from 19.7% to 7.2–8.5%. In practice we also found VITS easier to work with: there are no attention failures (skips and repetitions) on long sentences, consistent intelligibility for both speakers, parallel (non-autoregressive) inference, and there is a single model to train and deploy. On the other hand, it needs much longer training (several hundred thousand steps, i.e., about one to two weeks on a single GPU per voice) and the fp16 instability described in Section 4.7. Replacing PWG with a HiFi-GAN vocoder in the existing Tacotron 2 pipeline would be a cheap additional experiment, and we plan to run it.

5.5. Multi-Speaker Modelling

The current models are single-speaker models, one per voice. Since the corpus contains a male and a female speaker reading texts from the same domains, a single multi-speaker model conditioned on a learnable speaker embedding (VITS supports this through a speaker-embedding table, and ESPnet2 Tacotron 2 through the `spk_embed_dim` option) is the obvious next step. With only two speakers, we expect the main benefit to be shared modelling of Kazakh phonetics and prosody, which could improve the voice with less data (the male voice, 16.4 GB after cleaning), and halved deployment cost. For more than two voices, external speaker embeddings (e.g., d-vectors or x-vectors) as in YourTTS [24] would also allow zero-shot adaptation to new Kazakh speakers.

5.6. Transition of Kazakh from Cyrillic to Latin Script

Kazakhstan is gradually moving from the Cyrillic to a Latin-based alphabet. The transition was initiated by Presidential Decree No. 569 of 26 October 2017 [37]; the alphabet has since been revised several times, and the latest version (presented in 2021) uses diacritics (e.g., ä, ö, ü, ğ, ş, ñ) instead of apostrophes or digraphs [38], with full transition planned by 2031. During the transition period, Kazakh text will exist in both scripts, and a TTS system must accept both.

Effect on the model. The models in this paper take Cyrillic characters as input; their character-embedding layer has no entries for Latin letters, so Latin-script input would be mapped to the unknown token and could not be synthesised. The acoustic part of the models (decoder, alignment, vocoder) is script-independent, because it models the same speech sounds.

Adaptation options. We see three options, ordered from least to most effort. (1) *Transliteration front-end.* Because the new Latin alphabet was designed to be nearly one-to-one with native Kazakh Cyrillic letters, a rule-based Latin→Cyrillic converter can be placed in front of the existing models, which can then be used without retraining. The difficult cases are the Cyrillic letters used mainly in Russian loanwords (*ts*, *ch*, *shch*, *yo*, *yu*, *ya*, the hard and soft signs), for which the Latin spelling is not uniquely reversible, and the letters *u* and *i*, whose Cyrillic counterparts represent both vowels and glides; these require a small exception dictionary. (2) *Retraining on transliterated transcripts.* The training transcripts of our corpus can be converted automatically to the Latin script (the opposite, Cyrillic →

Latin, direction is deterministic for native words), so a Latin-script model can be trained or fine-tuned without new recordings; only the embedding layer must be re-initialised for the new vocabulary, and the remaining weights can be reused. A model with a joint Cyrillic–Latin vocabulary, trained on both versions of each transcript, would accept either script directly. (3) *Phoneme-level intermediate representation.* Converting both scripts to a common phonemic (or orthoepic) representation before synthesis makes the model independent of the script altogether and would also capture orthoepic rules (assimilation, reduction) that neither script encodes. This requires a Kazakh G2P module, for which the orthoepic converter used in our application is a starting point. For deployment during the transition, we recommend option (1), which is immediately available, followed by option (2) with a joint vocabulary once the final orthographic rules are fixed.

5.7. Limitations

The current system has several limitations. First, ASR-based CER/WER are proxies for intelligibility only: they depend on the ASR model (Section 4.5.4) and do not capture naturalness, prosodic expressiveness, or speaker similarity; moreover, the VITS utterances were synthesised through the deployed web service, so their RTF is not directly comparable with the GPU measurements for Tacotron 2. Second, each model represents a single voice; a joint multi-speaker model has not yet been trained. Third, the Tacotron 2 acoustic model is autoregressive, so its inference time scales linearly with output length and it remains susceptible to attention failures on very long inputs. Fourth, numerals, abbreviations and Latin-script words were removed from the training data; the models therefore rely on the text normalisation front-end for such input. Fifth, the subjective evaluation involved only seven listeners and covered only the Tacotron 2 + Parallel WaveGAN system; a comparative listening test of all systems with a larger panel is still needed. Sixth, no original recordings were evaluated with the same ASR protocol, so the error rates are not referenced to natural speech.

5.8. Future Work

Building on the results above, future work will (i) train a joint multi-speaker VITS model with speaker embeddings for both voices; (ii) evaluate a HiFi-GAN vocoder with the existing Tacotron 2 model and compare it with diffusion-based vocoders [26]; (iii) evaluate all systems on the KazakhTTS and KazakhTTS2 test partitions [9,10] with a Kazakh-specific ASR model, towards a common Kazakh TTS benchmark analogous to the Blizzard Challenge; (iv) add explicit prosody and speaking-rate control (e.g., reference encoders or global style tokens), which is relevant for learners and users with hearing impairments; (v) implement the Latin-script adaptation described in Section 5.6; (vi) carry out a blind comparative listening test of all systems following ITU-T P.800 with a larger and more diverse native-speaker panel, for which the stimuli and an online test interface have already been prepared; and (vii) compare the open systems with commercial Kazakh TTS services [35,36] in the same test.

6. Conclusions

We have presented an end-to-end Kazakh TTS system based on Tacotron 2 and Parallel WaveGAN, trained within the ESPnet2 framework on a new multi-domain corpus, and compared it with end-to-end VITS models trained for both speakers of the corpus. The system employs character-level tokenisation of the Kazakh Cyrillic alphabet, location-sensitive attention with guided attention loss, and a NoamLR learning rate schedule. The trained model produces intelligible and natural-sounding Kazakh speech, demonstrating the viability of modern neural end-to-end TTS for this morphologically rich, low-resource

language. Compared with Tacotron 2 on the same 138 sentences, VITS reduced the CER from 8.8% to about 2% and eliminated the attention failures that affected 6.2% of Tacotron 2 outputs. Both VITS voices are highly intelligible (CER 1.94% and 2.08% with a Kazakh-specific ASR model on 138 sentences from 23 domains). VITS also made the synthesis of complex educational content (formulas, numbers, units and abbreviations, normalised by the text front-end) robust, with no failed utterances and a CER of 7–8% compared with almost 20% for Tacotron 2. We also found that Whisper reports roughly ten times higher error rates than a Kazakh-specific ASR model on the same audio and even ranks systems inconsistently, so Kazakh TTS should be evaluated with Kazakh-specific ASR. Finally, we outlined how the pipeline can be adapted to the Latin-based Kazakh alphabet.

Because the acoustic model and the vocoder are separate components, either of them can be replaced with a newer one. The inference script provides a simple interface for text-to-speech conversion, accepting raw Kazakh text and producing 22,050 Hz WAV files. We hope that this work serves as a reproducible baseline for future research on Kazakh and related Turkic language TTS.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/a19100886/s1, Table S1: Test set of 138 plain sentences from 23 domains (6 sentences per domain; no numerals, abbreviations or Latin-script words); Table S2: Test set of 60 sentences of complex educational text (10 sentences per category), given in their original written form before text normalisation.

Author Contributions: Conceptualisation, A.B., A.F. and M.K.; methodology, A.B.; software, A.B. and N.K.; validation, Z.Z. and A.F.; formal analysis, A.B. and A.F.; investigation, A.B. and M.K.; resources, A.F.; data curation, A.F., A.B. and N.K.; writing—original draft preparation, M.K. and A.B.; writing—review and editing, A.B.; visualisation, M.K.; supervision, A.F.; project administration, A.F. and Z.Z.; funding acquisition, A.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan, grant number BR24993111 ("Developing the Speech Synthesis System Based on the Orthoepic Norm of the Kazakh Language").

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The training configuration files and inference code described in this manuscript are available from the corresponding author upon reasonable request. The speech corpus used for training is subject to data use agreements; enquiries regarding data access should be directed to the corresponding author.

Acknowledgments: The authors thank the developers of ESPnet2, Parallel WaveGAN and Coqui TTS for making their frameworks publicly available. This work was carried out within the framework of grant BR24993111 funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TTS	Text-to-speech
ASR	Automatic speech recognition
GAN	Generative adversarial network
BLSTM	Bidirectional long short-term memory
LSTM	Long short-term memory

MVN	Mean-variance normalisation
BCE	Binary cross-entropy
G2P	Grapheme-to-phoneme
OOV	Out-of-vocabulary
MOS	Mean opinion score
CER	Character error rate
WER	Word error rate
RTF	Real-time factor
ESPnet	End-to-end Speech processing toolkit
PWG	Parallel WaveGAN
VITS	Variational Inference with adversarial learning for end-to-end TTS
ACR	Absolute category rating
CI	Confidence interval

References

1. Tan, X.; Qin, T.; Soong, F.; Liu, T.-Y. A Survey on Neural Speech Synthesis. *arXiv* **2021**, arXiv:2106.15561.
2. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783.
3. van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499.
4. Eberhard, D.M.; Simons, G.F.; Fennig, C.D. *Ethnologue: Languages of the World*, 26th ed.; SIL International: Dallas, TX, USA, 2023.
5. Government of the Republic of Kazakhstan. *State Programme for the Development and Functioning of Languages in the Republic of Kazakhstan for 2011–2020*; Government of the Republic of Kazakhstan: Astana, Kazakhstan, 2011.
6. Tachibana, H.; Uenoyama, K.; Aihara, S. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4784–4788.
7. Yamamoto, R.; Song, E.; Kim, J.M. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6199–6203.
8. Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, J.; Unno, Y.; Soplin, N.E.Y.; Heymann, J.; Wiesner, M.; Chen, N.; et al. ESPnet: End-to-end speech processing toolkit. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 2207–2211.
9. Mussakhojayeva, S.; Janaliyeva, A.; Mirzakhmetov, A.; Khassanov, Y.; Varol, H.A. KazakhTTS: An Open-Source Kazakh Text-to-Speech Synthesis Dataset. In Proceedings of the Interspeech 2021, Brno, Czechia, 30 August–3 September 2021; pp. 2786–2790.
10. Mussakhojayeva, S.; Khassanov, Y.; Varol, H.A. KazakhTTS2: Extending the Open-Source Kazakh TTS Corpus With More Data, Speakers, and Topics. In Proceedings of the 13th Language Resources and Evaluation Conference (LREC), Marseille, France, 20–25 June 2022; pp. 5404–5411.
11. Kim, J.; Kong, J.; Son, J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; pp. 5530–5540.
12. Hunt, A.J.; Black, A.W. Unit selection in a concatenative speech synthesis system using a large speech database. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Atlanta, GA, USA, 7–10 May 1996; Volume 1, pp. 373–376.
13. Black, A.W.; Taylor, P. Automatically clustering similar units for unit selection in speech synthesis. In Proceedings of the Eurospeech 1997, Rhodes, Greece, 22–25 September 1997; pp. 601–604.
14. Zen, H.; Toda, T.; Nakamura, M.; Tokuda, K. Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. *IEICE Trans. Inf. Syst.* **2007**, *90* , 325–333. [CrossRef]
15. Tokuda, K.; Nankaku, Y.; Toda, T.; Zen, H.; Yamagishi, J.; Oura, K. Speech synthesis based on hidden Markov models. *Proc. IEEE* **2013**, *101*, 1234–1252. [CrossRef]
16. Zen, H.; Senior, A.; Schuster, M. Statistical parametric speech synthesis using deep neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 7962–7966.
17. Fan, Y.; Qian, Y.; Xie, F.; Soong, F.K. TTS synthesis with bidirectional LSTM based recurrent neural networks. In Proceedings of the Interspeech 2014, Singapore, 14–18 September 2014; pp. 1964–1968.

18. Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards end-to-end speech synthesis. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 4006–4010.
19. Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. FastSpeech: Fast, robust and controllable text to speech. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; pp. 3171–3180.
20. Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.-Y. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
21. Kong, J.; Kim, J.; Bae, J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 6–12 December 2020; pp. 17022–17033.
22. Mamyrbayev, O.; Turdalyuly, M.; Mekebayev, N.; Alimhan, K.; Kydyrbekova, A.; Turdalykyzy, T. Continuous speech recognition of Kazakh using deep learning. In Proceedings of the International Conference on Asian Language Processing (IALP), Kuala Lumpur, Malaysia, 18–20 November 2020; pp. 219–224.
23. Kim, J.; Kim, S.; Kong, J.; Yoon, S. Glow-TTS: A generative flow for text-to-speech via monotonic alignment search. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS), Virtual, 6–12 December 2020; pp. 8067–8077.
24. Casanova, E.; Weber, J.; Shulby, C.D.; Junior, A.C.; Gölge, E.; Ponti, M.A. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In Proceedings of the 39th International Conference on Machine Learning (ICML), Baltimore, MD, USA, 17–23 July 2022; pp. 2709–2720.
25. Chen, N.; Zhang, Y.; Zen, H.; Weiss, R.J.; Norouzi, M.; Chan, W. WaveGrad: Estimating gradients for waveform generation. In Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
26. Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; Catanzaro, B. DiffWave: A versatile diffusion model for audio synthesis. In Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
27. Popov, V.; Vovk, I.; Gogoryan, V.; Sadekova, T.; Kudinov, M. Grad-TTS: A diffusion probabilistic model for text-to-speech. In Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; pp. 8599–8608.
28. Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv* **2023**, arXiv:2301.02111.
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
30. Gölge, E.; Coqui TTS Team. Coqui TTS: A Deep Learning Toolkit for Text-to-Speech. Available online: https://github.com/coqui-ai/TTS (accessed on 10 October 2026).
31. Rekesh, D.; Koluguri, N.R.; Kriman, S.; Majumdar, S.; Noroozi, V.; Huang, H.; Hrinchuk, O.; Puvvada, K.; Kumar, A.; Balam, J.; et al. Fast Conformer with linearly scalable attention for efficient speech recognition. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Taipei, Taiwan, 16–20 December 2023; pp. 1–8.
32. NVIDIA. STT Kk-Ru FastConformer Hybrid Transducer-CTC Large (`stt_kk_ru_fastconformer_hybrid_large`). Available online: https://huggingface.co/nvidia/stt_kk_ru_fastconformer_hybrid_large (accessed on 10 October 2026).
33. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning (ICML), Honolulu, HI, USA, 23–29 July 2023; pp. 28492–28518.
34. ITU-T. *Recommendation P.800: Methods for Subjective Determination of Transmission Quality*; International Telecommunication Union: Geneva, Switzerland, 1996.
35. ElevenLabs. *Kazakh Text-to-Speech*; ElevenLabs Inc.: San Francisco, CA, USA, 2024. Available online: https://elevenlabs.io/text-to-speech/kazakh (accessed on 28 April 2026).
36. Narakeet. Kazakh Text-to-Speech. Narakeet. 2024. Available online: https://www.narakeet.com/languages/kazakh-text-to-speech/ (accessed on 28 April 2026).
37. President of the Republic of Kazakhstan. *Decree No. 569 of 26 October 2017 "On the Transition of the Kazakh Alphabet from Cyrillic to the Latin Script"*; Akorda (Office of the President of the Republic of Kazakhstan): Astana, Kazakhstan, 2017.
38. Kazakhstan Presents New Latin Alphabet, Plans Gradual Transition Through 2031. *The Astana Times*, 29 January 2021. Available online: https://astanatimes.com/2021/02/kazakhstan-presents-new-latin-alphabet-plans-gradual-transition-through-2031/ (accessed on 10 October 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.