Using Songs to Improve Kazakh Automatic Speech Recognition

Rustem Yeshpanov
Independent Researcher
Astana, Kazakhstan
yeshpanov.rustem@gmail.com

Abstract

Developing automatic speech recognition (ASR) systems for low-resource languages is hindered by the scarcity of transcribed corpora. This proof-of-concept study explores songs as an unconventional yet promising data source for Kazakh ASR. We curate a dataset of 3,013 audio-text pairs (about 4.5 hours) from 195 songs by 36 artists, segmented at the lyric-line level. Using Whisper as the base recogniser, we fine-tune models under seven training scenarios involving Songs, Common Voice Corpus (CVC), and FLEURS, and evaluate them on three benchmarks: CVC, FLEURS, and Kazakh Speech Corpus 2 (KSC2). Results show that song-based fine-tuning improves performance over zero-shot baselines. For instance, Whisper Large-V3 Turbo trained on a mixture of Songs, CVC, and FLEURS achieves 27.6% normalised WER on CVC and 11.8% on FLEURS, while halving the error on KSC2 (39.3% vs. 81.2%) relative to the zero-shot model. Although these gains remain below those of models trained on the 1,100-hour KSC2 corpus, they demonstrate that even modest song-speech mixtures can yield meaningful adaptation improvements in low-resource ASR. The dataset is released on Hugging Face for research purposes under a gated, non-commercial licence.

Keywords: ASR, Kazakh, songs, Common Voice, FLEURS, KSC2, low-resource languages

1. Introduction

Automatic speech recognition (ASR) systems convert speech into text and typically rely on large, diverse, and carefully transcribed corpora to achieve robust performance. Such resources remain scarce for most languages worldwide. Although recent work has improved the situation for Kazakh (a Turkic language spoken by more than 15 million people worldwide), it still qualifies as low-resource by global standards, while several other Turkic languages remain even more under-resourced (Veitsman and Hartmann, 2025).

We hypothesise that songs offer a practical, widely available, and language-agnostic data source that can complement or partially substitute conventional speech corpora for low-resource ASR. Songs are prevalent across many languages and communities; their recordings are often high quality, and accompanying lyrics can serve as approximate transcriptions. At the same time, songs pose challenges for ASR—background music, non-conversational prosody, elongated vowels, and repetitions—which makes their net utility an empirical question. This paper presents a feasibility study that asks: *To what extent can song audio with aligned lyric segments help adapt ASR for Kazakh?*

To investigate this, we curated a small song-based dataset of Kazakh consisting of 3,013 audio-text pairs drawn from 195 songs performed by 36 artists, amounting to approximately 4.5 hours. Using Whisper (Radford et al., 2023) as the base recogniser, we compare zero-shot models against fine-tuned models trained on songs, on each small corpus individually (Common Voice Corpus 17.0 (CVC) (Ardila et al., 2020), Few-shot Learning Evaluation of Universal Representations of Speech (FLEURS) (Conneau et al., 2023)), on each pair, and on all three combined, to systematically assess where songs help. For context, we also compare against an "upper bound" scenario that assumes access to over 1,100 hours of transcribed recordings from Kazakh Speech Corpus 2 (KSC2) (Mussakhojayeva et al., 2022a), a scale of data far beyond what most low-resource languages can provide.

Our experiments show that incorporating songs yields substantial improvements over zero-shot baselines on multiple benchmarks (notably on CVC and FLEURS), and that pairing songs with small speech corpora further strengthens generalisation. However, the gains are uneven across evaluation domains, and song-augmented models do not match the performance of the large-corpus upper bound trained on KSC2. These findings suggest that songs are not a panacea, but they are a promising, broadly accessible resource that can provide measurable benefits when large speech corpora are unavailable.

The remainder of the paper is organised as follows. Section 2 reviews related work on low-resource ASR. Section 3 details the construction of the song dataset. Section 4 describes the experimental setup. Section 5 presents results and analysis, and Section 6 concludes with limitations and directions for future work.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5528–5537
11–16 May 2026.

2. Related Work

Research on Kazakh ASR has accelerated in recent years, supported by several initiatives that created foundational resources. KSC2 (Mussakhojayeva et al., 2022a), for example, provides 1,128 hours of transcribed audio across more than 520,000 utterances. Other efforts, such as KazakhTTS (Mussakhojayeva et al., 2021), its extension KazakhTTS2 (Mussakhojayeva et al., 2022b), and KazEmoTTS (Abilbekov et al., 2024), have produced high-quality text-to-speech datasets that also support ASR research. Despite these contributions, Kazakh resources remain small compared to those available for high-resource languages, highlighting the need for alternative data sources.

Open-source multilingual datasets such as CVC (Ardila et al., 2020) and FLEURS (Conneau et al., 2023) include Kazakh and have proven valuable for training and benchmarking. Yet their scale and diversity are modest relative to corpora in languages like English or Mandarin (Parcollet et al., 2025; Zhang et al., 2021), limiting their effectiveness for robust ASR development.

To address such scarcity, prior work in low-resource ASR has focused on techniques such as data augmentation and transfer learning. Data augmentation has been shown to improve performance by synthesising additional speech data (Baas and Kamper, 2022; Casanova et al., 2023), while transfer learning approaches leverage transliteration (Khare et al., 2021) or cross-lingual speech-to-text translation (Wang et al., 2020) to adapt models from high-resource languages. These methods, however, often still rely on a non-trivial amount of initial speech data, making them less applicable in scenarios where only minimal resources are available.

This motivates exploration of alternative resources that are both accessible and linguistically broad. In this paper, we propose the use of songs as such a resource. Songs exist in nearly every language, are often distributed in high-quality recordings, and are accompanied by lyrics that provide approximate transcriptions. While challenges include background music and non-standard prosody, their universality makes songs a promising but underexplored candidate for low-resource ASR. To our knowledge, no prior work has systematically investigated the use of songs for Kazakh or other low-resource languages, making this study a first step in addressing this gap.

3. Dataset

3.1. Song Selection and Collection

The dataset was collected over four months by the author. To achieve diversity, 195 Kazakh songs performed by 36 artists (14 female and 22 male) were selected. Songs were downloaded from YouTube using the `youtube_dl` library[1]. Songs were selected according to the following criteria:

- **Vocal focus:** Only tracks featuring prominent solo vocals were included; songs dominated by choral arrangements, duets/overlapping lead vocals, or instrumental-only sections were excluded.
- **Genre diversity:** Songs representing a range of mainstream Kazakh music were included to capture stylistic variety (pop, pop-estrada, folk-pop, folk-rock, R&B, and hip-hop).
- **Artist representation:** Popular contemporary artists with publicly available songs were included.
- **Source search strategy:** Songs were identified using artist discographies, popular streaming platforms, and online music catalogues. Keywords included artist names, album titles, and widely recognised track titles.

In the interest of space, we do not list all artists and songs; however, each audio-text pair retains metadata with the corresponding artist name and song title to support reproducibility and future reference. In addition, each segment retains start/end timestamps to make the line-level segmentation reproducible.

3.2. Data Preparation and Validation

Preprocessing involved several steps to prepare high-quality audio-text pairs suitable for ASR. First, vocal tracks were separated from instrumental accompaniment using `Spleeter`[2]. While this process does not perfectly isolate vocals and may leave residual background music, the resulting tracks were deemed sufficiently clear to preserve intelligible speech content. As a quality-control step, we manually audited the separated vocals by listening and discarded segments where the lyrics were not intelligible due to residual accompaniment or separation artefacts.

Lyrics were collected from online repositories and official artist websites. As the accuracy of

[1] https://pypi.org/project/youtube_dl/
[2] https://pypi.org/project/spleeter/

these sources varied, the lyrics were manually reviewed and corrected to match the actual sung content, including repetitions and colloquial pronunciations. Corrected texts retained original casing and punctuation, reflecting the structure of the lyrics rather than applying ASR-style normalisation.

Alignment was performed manually using Audacity[3], an open-source audio editing tool. Each song was segmented at the line level by listening and synchronising with the corrected lyrics, after which the "Export Labels" feature in Audacity was used to generate audio-text pairs.

The resulting dataset comprises 3,013 pairs with a total duration of approximately 4.5 hours. Summary statistics are presented in Table 1. Note that the totals for unique utterances and words are lower than the sum of female and male counts, as some material occurs in songs by both groups; such overlaps were removed in the overall totals. Mean utterance length was nearly identical across male and female artists (5.4 s), an outcome that appears coincidental given the diversity of genres and performers. The dataset, while carefully constructed, also presents several limitations, discussed below.

	Female	Male	Total
Artists	14	22	36
Utterances	1,387	1,626	3,013
Unique utterances	1,369	1,598	2,938
Words	6,504	7,441	13,945
Unique words	3,075	3,388	5,359
Duration (h)	2.1	2.4	4.5
Min (s)	1.0	1.1	1.0
Max (s)	17.0	15.3	17.0
Mean (s)	5.4	5.4	5.4

Table 1: Summary statistics of the Kazakh songs dataset

Limitations

As with any dataset of this scale and nature, it has several limitations. First, its total duration of 4.5 hours is relatively small compared to conventional ASR training corpora, which may restrict model generalisation. Second, although multiple genres were included, the selection is not exhaustive and may underrepresent less common styles or regional variations. Third, despite vocal separation with Spleeter, residual background music remains in some segments, potentially introducing noise. Finally, alignment and lyric correction were performed manually by the author, which, although carefully conducted, may introduce subjective inconsistencies. These limitations highlight that the dataset should be viewed as a proof-of-concept resource rather than a comprehensive corpus.

[3] https://www.audacityteam.org/

4. Methodology

4.1. Training Setup

Given the computational expense of fine-tuning large ASR models, all experiments were conducted on Vast.AI[4] using an NVIDIA RTX 3090 GPU (24 GB VRAM). Across all fine-tuning runs, the total compute cost was approximately $25, underscoring the affordability of this approach relative to large-scale training.

The fine-tuning configuration was kept consistent across experiments. We used an initial learning rate of 5×10^{-6} with 50 warm-up steps, a batch size of 60, and an early stopping criterion with a patience of two epochs. Training was conducted under seven scenarios using the training splits of the respective datasets: (i) Songs only, (ii) CVC only, (iii) FLEURS only, (iv) Songs + CVC, (v) Songs + FLEURS, (vi) CVC + FLEURS, and (vii) Songs + CVC + FLEURS. Validation was performed on the CVC and FLEURS validation sets. This choice reflects the fact that the Songs dataset was intended solely as a training resource, and that the KSC2 corpus differs substantially in format (all lowercase, no punctuation). In contrast, CVC and FLEURS include casing and punctuation, making them more compatible with the song-based training data and more informative for monitoring model generalisation. These format differences underscore the broader challenge of cross-dataset evaluation in Kazakh ASR, where corpora often vary in orthographic conventions and normalisation standards. Final evaluation was conducted on three independent benchmarks: the KSC2 test set (Mussakhojayeva et al., 2022a), the CVC test set (Ardila et al., 2020), and the FLEURS test set (Conneau et al., 2023).

The benchmark datasets are briefly summarised below. KSC2 contains 1,128 hours of transcribed audio, crowdsourced and collected for both ASR and text-to-speech (TTS) development. It spans a wide range of sources, including news broadcasts, radio programs, parliamentary speeches, podcasts, and Kazakh-Russian code-switching utterances. CVC is a large-scale, volunteer-based corpus for multilingual ASR, which includes a Kazakh subset. Inspection of the Kazakh test set revealed that it predominantly consists of sayings and proverbs. FLEURS is a multilingual read-speech benchmark developed by Google Research. It extends the FLoRes machine translation dataset (Goyal et al.,

[4] https://vast.ai/

2022) by adding speech recordings of its sentences, originally derived from English Wikipedia. The Kazakh subset includes some words of foreign origin spelt in Latin script, in contrast to the Cyrillic-only orthography of the other test sets used here.

Before training and evaluation, additional preprocessing was applied for consistency. Two English sentences were identified in the FLEURS training set and removed. Sentences in the KSC2 test set that contained only Russian words were identified and removed. In addition, homoglyphs appearing in both the KSC2 and CVC test sets were replaced with their respective Kazakh letters to maintain script consistency.

Another important aspect of this study is that the datasets differ in their transcription conventions. The Songs dataset and CVC both use Cyrillic script with casing and punctuation, but contain no digits. In contrast, FLEURS includes a mix of Cyrillic and Latin script, digits, casing, and punctuation. The KSC2 corpus is distinct from all of these, being entirely lowercased, Cyrillic-only, and stripped of punctuation. Table 2 summarises transcription format differences across datasets, while Table 3 reports statistics for the specific splits used in this study—namely, the training, validation, and test sets of CVC and FLEURS, and the test set only for KSC2.

Dataset	Script	Casing	Punctuation	Digits
Songs	Cyrillic	Yes	Yes	No
CVC	Cyrillic	Yes	Yes	No
FLEURS	Cyrillic + Latin	Yes	Yes	Yes
KSC2	Cyrillic	No	No	No

Table 2: Transcription conventions across datasets

	CVC			FLEURS			KSC2
	Train	Valid	Test	Train	Valid	Test	Test
Speakers	4	21	106	-	-	-	-
Utterances	548	498	514	3,198	369	856	9,192
Unique utter.	548	498	513	1,493	147	349	9,072
Words	3,324	2,989	3,100	53,520	5,877	15,014	102,035
Unique words	1,990	1,784	1,873	10,358	1,675	3,502	23,678
Duration (h)	0.7	0.6	0.7	11.8	1.5	3.8	15.6
min (s)	2.2	1.9	2.1	3.0	4.7	5.8	1.0
max (s)	10.5	10.5	9.7	36.0	46.0	43.3	26.3
mean (s)	4.9	4.6	5.1	13.3	14.9	16.1	6.1

Table 3: Dataset splits and test set statistics

4.2. Model Selection

We first evaluated whether fine-tuning smaller models on the collected dataset yields measurable performance improvements. Specifically, we experimented with `Whisper-tiny`[5] (39M parameters) and `Whisper-small`[6] (244M parameters), two compact variants from OpenAI's Whisper family (Radford et al., 2023). Due to space constraints, Table 4 reports word error rate (WER) and character error rate (CER) on three test sets (CVC, FLEURS, KSC2) using normalised transcripts, which provide consistency across datasets with differing conventions. Results show that even these small-scale models exhibit clear relative improvements after fine-tuning (FT) on the Songs dataset compared to their pre-trained (PT) baselines, although absolute error rates remain high.

Whisper		CVC		FLEURS		KSC2	
		WER	CER	WER	CER	WER	CER
tiny	PT	190.7	160.7	154.4	110.0	219.2	156.4
	FT	87.0	28.4	99.8	31.1	111.5	46.5
small	PT	179.6	108.9	89.8	48.0	125.0	81.1
	FT	60.7	16.4	63.8	17.4	78.5	27.8

Table 4: Normalised WER and CER performance of Whisper-tiny and Whisper-small

Given the high computational cost of training Whisper Large-V3 (1550M parameters) directly, we instead adopted Whisper Large-V3 Turbo[7], a pruned variant that reduces the number of decoding layers (32 to 4) and achieves substantially faster inference with only minor quality degradation. For additional context, we also included a community fine-tuned version of Whisper Large-V3 Turbo trained on KSC2[8], representing an upper-bound scenario unlikely to be available for most low-resource languages.

4.3. Evaluation Metrics

To assess ASR performance, this study employed WER and CER, both of which are standard evaluation metrics in speech recognition (Morris et al., 2004; MacKenzie and Soukoreff, 2002). WER measures transcription accuracy at the word level, while CER provides finer-grained evaluation at the

[5] https://huggingface.co/openai/whisper-tiny

[6] https://huggingface.co/openai/whisper-small

[7] https://huggingface.co/openai/whisper-large-v3-turbo

[8] https://huggingface.co/abilmansplus/whisper-turbo-ksc2

character level, making it particularly useful for languages with rich morphology such as Kazakh.

Two versions of WER and CER were computed. The first, orthographic WER and CER (WER_{or} and CER_{or}), was calculated using the original transcriptions, preserving casing, punctuation, and extra whitespace. The second, normalised WER and CER (WER_{no} and CER_{no}), was computed after normalising text by lowercasing, removing punctuation, and collapsing extra whitespace. Given the formatting of KSC2 (see Table 2), we report only normalised WER and CER for the respective test set.

5. Results

5.1. Quantitative Evaluation

5.1.1. Baselines

Table 5 reports performance on three evaluation sets (CVC, FLEURS, KSC2). Zero-shot models struggle on Kazakh overall. On CVC (normalised), Whisper Large-V3 achieves 56.5, while the pruned Turbo variant is better at 47.7; however, Turbo is markedly worse on KSC2 (81.2 vs. 58.9). Orthographic scores on CVC show the opposite trend: Turbo degrades from 61.2 to 70.6, reflecting its difficulty with casing and punctuation. In contrast, the community fine-tuned model trained on the 1,128-hour KSC2 corpus establishes a strong upper bound with 12.5 (CVC), 11.3 (FLEURS), and 9.3 (KSC2) normalised WER.

5.1.2. Fine-tuning Whisper Large-V3 Turbo

We fine-tuned the Turbo model on Songs, CVC, and FLEURS—individually and in mixtures.

Single-source fine-tuning. *Songs only* improves CVC and KSC2 over the Turbo baseline (CVC: 37.3 vs. 47.7; KSC2: 45.2 vs. 81.2 normalised WER), but hurts FLEURS (23.7 vs. 21.0). *CVC only* helps in-domain (39.1 on CVC) but remains weak out-of-domain (FLEURS 37.9; KSC2 58.1). *FLEURS only* helps FLEURS strongly (13.6) and also improves KSC2 vs. baseline (46.6), with modest CVC gains (43.6).

Mixtures. Mixtures are consistently stronger than single sources. *CVC* + *FLEURS* yields 28.1 (CVC) and 11.8 (FLEURS) normalised WER, and improves KSC2 to 39.3. The triple mixture (*Songs* + *CVC* + *FLEURS*) is the most balanced overall: 27.6 on CVC (best), 11.8 on FLEURS (ties best), and 39.3 on KSC2 (ties best). *Songs* + *FLEURS* also reaches 11.8 on FLEURS but is weaker on CVC (34.1) and KSC2 (40.4); *Songs* + *CVC* prioritises CVC (29.6) but leaves FLEURS relatively high (23.7).

Orthographic view. Orthographic improvements mirror the normalised trends. The triple mixture reduces orthographic WER on CVC from 70.6 (Turbo baseline) to 32.0 (a $\sim$55% relative reduction), and on FLEURS from 38.0 to 19.5 ($\sim$49% relative reduction).

5.1.3. Fine-tuning the Community (KSC2) Model

Starting from the KSC2-trained upper bound leaves limited headroom and introduces domain-drift risks.

Single-source fine-tuning. *Songs only* degrades performance relative to the KSC2 baseline (CVC 13.9 vs. 12.5; FLEURS 11.7 vs. 11.3; KSC2 10.3 vs. 9.3). *CVC only* largely preserves KSC2 (9.3) and slightly improves FLEURS (10.8), with CVC roughly unchanged (12.6). *FLEURS only* achieves a large gain on FLEURS (7.3) but substantially harms KSC2 (13.9) and worsens CVC (14.4).

Mixtures. *CVC* + *FLEURS* produces the best FLEURS normalised WER 7.1, but with notable forgetting on KSC2 (13.8). *Songs* + *CVC* keeps CVC near the baseline (12.3) while slightly worsening FLEURS (11.5) and KSC2 (10.3). The triple mixture (15.6 on CVC; 7.9 on FLEURS; 16.3 on KSC2) is more balanced than *FLEURS only* but still shows clear degradation on KSC2 relative to the baseline.

To further assess cross-domain generalisation, Table 6 reports normalised WER across six speech domains of the KSC2 test set: crowd-sourced, parliamentary, podcasts, radio, talk-shows, and television news. The largest relative gains are observed for the `Whisper Large-V3 Turbo` model fine-tuned within this study—rather than for the community model already trained on the full 1,100-hour KSC2 corpus. Song-based fine-tuning yields striking improvements in spontaneous and conversational domains such as *podcasts* and *talkshows*, where error rates drop by roughly two-thirds compared to the zero-shot Turbo baseline, and by about half in *parliamentary* speech. More moderate but still consistent reductions are seen for *radio* and *TV news*, indicating that musical data can aid adaptation even across differing acoustic and stylistic conditions.

	CVC				FLEURS				KSC2	
Model / Training	**WER$_{or}$**	**CER$_{or}$**	**WER$_{no}$**	**CER$_{no}$**	**WER$_{or}$**	**CER$_{or}$**	**WER$_{no}$**	**CER$_{no}$**	**WER$_{no}$**	**CER$_{no}$**
Baselines										
Whisper Large-V3 (zero-shot)	61.2	22.0	56.5	20.7	41.1	9.5	33.1	7.8	58.9	23.6
Whisper Large-V3 Turbo (zero-shot)	70.6	28.6	47.7	23.8	38.0	9.0	21.0	6.0	81.2	42.6
Community fine-tuned (KSC2)	56.2	10.9	12.5	3.1	36.0	8.8	11.3	5.1	9.3	3.2
Fine-tuning Whisper Large-V3 Turbo										
+ Songs only	49.8	12.3	37.3	9.3	33.9	7.6	23.7	5.7	45.2	15.2
+ CVC only	42.3	10.0	39.1	9.2	48.1	11.3	37.9	9.2	58.1	19.6
+ FLEURS only	51.6	14.8	43.6	12.8	21.0	4.4	13.6	3.1	46.6	18.7
+ Songs + CVC	33.9	8.7	29.6	7.8	33.8	7.5	23.7	5.6	43.7	14.2
+ Songs + FLEURS	40.9	11.0	34.1	9.5	19.7	3.9	11.8	2.6	40.4	16.0
+ CVC + FLEURS	33.0	7.7	28.1	6.6	19.7	4.0	11.8	2.6	39.3	13.9
+ Songs + CVC + FLEURS	32.0	7.4	27.6	6.5	19.5	3.9	11.8	2.6	39.3	14.4
Fine-tuning Community Model (KSC2)										
+ Songs only	41.1	8.5	13.9	3.5	32.6	7.9	11.7	4.7	10.3	3.5
+ CVC only	56.3	10.9	12.6	3.1	35.7	8.3	10.8	4.5	9.3	3.2
+ FLEURS only	26.4	6.1	14.4	3.7	16.0	3.6	7.3	2.1	13.9	5.8
+ Songs + CVC	19.5	4.5	12.3	3.1	26.5	7.2	11.5	4.8	10.3	3.5
+ Songs + FLEURS	26.8	7.9	17.1	6.0	16.3	4.1	8.4	2.7	15.2	6.2
+ CVC + FLEURS	19.1	4.7	13.5	3.5	15.4	3.4	7.1	2.0	13.8	5.7
+ Songs + CVC + FLEURS	21.0	5.2	15.6	4.1	15.8	3.7	7.9	2.3	16.3	6.6

Table 5: Orthographic and normalised WER/CER (%) on three Kazakh test sets

5.2. Qualitative Error Analysis

Beyond quantitative metrics, we analyse representative outputs in Table 7 to understand how song-based adaptation affects recognition behaviour. Across all test sets, models further fine-tuned with Songs—Whisper Large-V3 Turbo fine-tuned on Songs + CVC + FLEURS (WLT_SCF) and the community KSC2-trained model further fine-tuned on Songs + CVC + FLEURS (CFT_SCF)—exhibit more stable and linguistically coherent transcriptions than the zero-shot baselines.

A key difference is reduced cross-lingual drift. On the KSC2 sample, the zero-shot Whisper Large-V3 output shifts into another language and the turbo variant produces nonsensical tokens, whereas the song-adapted models remain consistently in Kazakh and recover the intended meaning with only minor variation. This suggests that exposure to song data strengthens lexical grounding and decoding stability under acoustically challenging conditions.

Improvements are also evident in lexical and morphological accuracy. On the CVC example, zero-shot models produce unintelligible outputs, while song-adapted models recover the syntactic structure and core vocabulary, with only minor phonetic substitutions (e.g., қ → ұ) that do not affect lexical interpretability. On the FLEURS sample, which represents instructional prose, song-adapted models more reliably preserve key lexical items and suffixes, whereas zero-shot variants exhibit vowel distortions and incorrect substitutions.

Finally, qualitative differences appear in punctuation and sentence segmentation. Song-adapted models more consistently restore clause boundaries and punctuation on the FLEURS example, suggesting improved modelling of prosodic and syntactic cues. This behaviour is consistent with the nature of song data, where lyrical phrasing and rhythmic pauses provide additional boundary information.

Overall, the qualitative evidence indicates that song-based fine-tuning improves not only acoustic robustness but also language stability, lexical recovery, morphological accuracy, and resistance to cross-lingual hallucinations, complementing the quantitative WER improvements.

Model / Training	Crowdsourced	Parliament	Podcasts	Radio	Talkshow	TV News
Baselines						
Whisper Large-V3	47.3	79.9	68.0	72.1	63.6	55.5
Whisper Large-V3 Turbo	30.2	68.5	166.5	88.9	164.3	60.5
Community fine-tuned (KSC2)	5.0	6.0	18.8	15.7	13.7	4.9
Fine-tuning Whisper Large-V3 Turbo						
+ Songs only	32.6	39.7	61.8	62.9	60.4	41.3
+ CVC only	44.7	55.3	73.5	75.7	68.6	56.6
+ FLEURS only	35.6	37.3	64.5	67.3	59.3	40.1
+ Songs + CVC	31.5	42.7	59.4	60.8	52.1	40.3
+ Songs + FLEURS	27.7	33.3	59.0	58.3	56.4	35.2
+ CVC + FLEURS	27.2	37.5	55.5	59.2	51.1	33.9
+ Songs + CVC + FLEURS	28.0	34.6	55.7	57.1	49.9	34.6
Fine-tuning Community Model (KSC2)						
+ Songs only	6.1	6.8	20.6	15.8	15.7	5.4
+ CVC only	5.1	5.9	18.8	15.5	13.8	4.9
+ FLEURS only	12.8	9.1	20.8	16.5	15.4	9.3
+ Songs + CVC	6.0	6.6	20.1	16.5	15.5	5.5
+ Songs + FLEURS	13.4	10.6	23.2	19.6	17.2	10.2
+ CVC + FLEURS	12.5	8.9	20.7	17.1	15.9	9.4
+ Songs + CVC + FLEURS	13.9	11.8	25.3	21.9	18.5	10.9

Table 6: Normalised WER (%) across six speech domains of the KSC2 test set

System output	Test set	Text	WER
Reference	CVC	*Жақсыда жаттық жоқ, жаманда достық жоқ.* Zhaqsyda zhattyq zhoq, zhamanda dostyq zhoq. The good have no strangers; the bad have no friends.	0.0
Whisper Large-V3		Şahsı da jaktozok, samanda toztozok.	100.0
Whisper Large-V3 Turbo		자수다자 도적 삼엔더 도적	100.0
Community fine-tuned		жақсы да жаттық жоқ жаман да достық жоқ	100.0
WLT_SCF		Жақсы да жаттыұ жоқ, жаман да достыұ жоқ.	100.0
CFT_SCF		Жақсыда жаттық жоқ, жаманда достық жоқ.	0.0
Reference	FLEURS	*Матаның тым ыстық болуына жол бермеңіз (бұл қысқаруға немесе күюіне себеп болуы мүмкін).* Matanyng tym ystyq boluyna zhol bermengiz (būl qysqaruğa nemese küyüine sebep boluy mümkin). Do not expose the fabric to excessive heat (this may cause shrinkage or scorching).	0.0
Whisper Large-V3		Матаның тұм ұстық болуына жол бермеңіз. Бұл қысқаруа немесе күйіне себеп болуы мүмкін.	53.85
Whisper Large-V3 Turbo		матаның тым ыстық болуына жол бермеңіз бұл қысқаруға немесе күйеуіне себеп болуы мүмкін	30.77
Community fine-tuned		матаның тым ыстық болуына жол бермеңіз бұл қысқаруға немесе күюіне себеп болуы мүмкін	23.08
WLT_SCF		Матаның тым ыстық болуына жол бермеңіз. Бұл қысқаруа немесе күюіне себеп болуы мүмкін.	30.77
CFT_SCF		Матаның тым ыстық болуына жол бермеңіз. Бұл қысқаруға немесе күюіне себеп болуы мүмкін.	23.08
Reference	KSC2	*осыны мәселені есте ұстауымыз керек* osyny māseleni este ūstauymyz kerek We should keep this issue in mind.	0.0
Whisper Large-V3		основным осиленным из тех стал маскерек	120.0
Whisper Large-V3 Turbo		osnamaðsýlina í stíðstáum skýrek	100.0
Community fine-tuned		осындай мәселені есте ұстауымыз керек	20.0
WLT_SCF		осыны мәселені есте ұстауымыз керек	0.0
CFT_SCF		осындай мәселені есте ұстауымыз керек	20.0

Table 7: Sample ASR outputs across CVC, FLEURS, and KSC2 for zero-shot, community fine-tuned, and song-adapted systems (WLT_SCF and CFT_SCF). WER_{or} is reported for CVC and FLEURS, and WER_{no} for KSC2.

5.3. Discussion

Three conclusions emerge. First, songs alone are not sufficient and may hurt some domains (e.g., FLEURS on Turbo; most domains on the KSC2 model). Second, songs are valuable when combined with modest corpora: the triple mixture achieves the best CVC (27.6) and ties the best FLEURS (11.8) on Turbo, while also halving KSC2 error relative to the Turbo baseline (39.3 vs. 81.2). Third, once a model is trained on approximately 1,100 hours (KSC2 upper bound), additional fine-tuning on songs/small corpora confers marginal gains at best and often induces forgetting.

A further factor likely contributing to the modest gains from song-only fine-tuning is dataset size.

With only 4.5 hours of audio and roughly 3,000 lyric lines, the song corpus represents a very small fraction of the data typically required to meaningfully adapt a billion-parameter model such as Whisper. At this scale, the training signal may have been too limited to shift model parameters substantially, even though it proved complementary when combined with other small corpora. Future work should investigate whether larger and more diverse song collections—potentially including synthetic data—could yield stronger and more consistent adaptation effects.

Orthographic metrics reinforce these findings: while casing and punctuation remain challenging, mixtures (especially those including FLEURS) deliver substantial orthographic error reductions, indicating that the benefits of songs extend beyond pure lexical recognition to improved written-form fidelity.

Ethical and legal considerations. Beyond these technical outcomes, it is important to acknowledge the elephant in the room: the copyrighted nature of the Songs dataset. The recordings used here are copyrighted works, and no explicit permission was obtained from the artists for use in ASR development. This raises a broader question: *Is the absence of prior research on songs as an ASR resource due primarily to lack of exploration, or to the legal and ethical complexities surrounding their use?* This study is exploratory and not intended as a deployment-ready approach; rather, it seeks to assess whether songs have technical merit as a training signal. If the answer is yes, the next step would involve dialogue on how such data might be ethically and legally integrated into ASR development pipelines for low-resource languages—for example, through short excerpts, public-domain materials, collaborations with artists, or structured fair-use frameworks.

Synthetic alternatives. A promising direction for addressing copyright concerns lies in synthetic music generation. Modern tools such as `Suno.com`[9] can generate songs with customisable parameters: lyrics in low-resource languages, stylistic control (e.g., folk, pop, rap), and varied vocal timbres (male/female, solo/chorus). If song-based training proves beneficial, synthetic songs could provide a scalable and legally permissible alternative. They would allow researchers to systematically generate datasets reflecting specific phonetic or prosodic characteristics without relying on copyrighted works. This opens a potential new research avenue: *To what extent can synthetic songs stand in for real-world ones in improving low-resource ASR?*

6. Conclusion

This study investigated the feasibility of using songs as a novel training resource for Kazakh ASR. The results show that while songs alone do not consistently improve recognition performance and can even degrade generalisation, they provide a meaningful signal when combined with modest corpora such as CVC and FLEURS. Mixtures that included songs consistently outperformed single-corpus baselines, with the best results achieved by combining all three datasets (Songs + CVC + FLEURS). In this setting, normalised WER dropped to 27.6 on CVC and 11.8 on FLEURS, marking substantial improvements over the zero-shot Whisper models and narrowing the gap to the community model trained on the 1,100-hour KSC2 corpus. These findings highlight that songs are not a standalone solution but can amplify the value of small existing corpora in low-resource ASR.

The multi-domain evaluation also revealed important limitations. Song-based training does not fully transfer to conversational or broadcast speech, and gains remain modest compared to the large-scale upper bound. Moreover, orthographic errors (casing, punctuation) remain challenging, though the inclusion of songs helped reduce them in some scenarios, suggesting that lyrics-based data may support better modelling of written-form conventions.

From a practical standpoint, the entire set of fine-tuning experiments cost only $25 in compute, underscoring that meaningful exploratory research in low-resource ASR can be conducted with modest resources.

Beyond technical findings, this work raises crucial legal and ethical questions. The Songs dataset consists of copyrighted recordings without explicit permission, underscoring the barriers to directly incorporating music into ASR pipelines. This study is therefore intended as a proof of concept rather than a deployable approach. The Kazakh Songs dataset—containing short vocal excerpts ($\lesssim$ 10 s) aligned with lyric transcriptions—is made available on Hugging Face[10] under a gated, non-commercial research licence. Full recordings are not distributed; access to excerpts is reviewed and granted solely for academic research purposes. Future research must explore ethical pathways, including collaboration with artists, the use of public-domain or fair-use excerpts, and the development of frameworks for responsible data use.

Finally, synthetic alternatives represent a

[9] `https://suno.com/`

[10] `https://huggingface.co/datasets/yeshpanovrustem/kazakh_songs_asr`

promising direction. Modern tools such as Suno.com allow the generation of songs with customisable lyrics, genres, and voices, enabling the creation of corpora in low-resource languages without copyright restrictions. Such synthetic songs could be designed to cover diverse phonetic contexts, prosodic variations, and stylistic registers, offering a scalable complement to natural data. Exploring whether synthetic music can replicate or even surpass the benefits of real songs is a natural next step for advancing low-resource ASR.

7. Bibliographical References

Adal Abilbekov, Saida Mussakhojayeva, Rustem Yeshpanov, and Huseyin Atakan Varol. 2024. KazEmoTTS: A Dataset for Kazakh Emotional Text-to-Speech Synthesis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9626–9632, Torino, Italia. ELRA and ICCL.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Language Resources and Evaluation Conference (LREC)*, pages 4218–4222, Marseille, France. European Language Resources Association.

Matthew Baas and Herman Kamper. 2022. Voice Conversion Can Improve ASR in Very Low-Resource Settings. In *Interspeech 2022*, pages 3513–3517.

Edresson Casanova, Christopher Shulby, Alexander Korolev, Arnaldo Candido Junior, Anderson da Silva Soares, Sandra Aluísio, and Moacir Antonelli Ponti. 2023. ASR data augmentation in low-resource settings using cross-lingual multi-speaker TTS and cross-lingual voice conversion. In *Interspeech 2023*, pages 1244–1248.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. FLEURS: Few-Shot Learning Evaluation of Universal Representations of Speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The FLoRes-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Shreya Khare, Ashish Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. 2021. Low Resource ASR: The Surprising Effectiveness of High Resource Transliteration. In *Interspeech 2021*, pages 1529–1533.

I Scott MacKenzie and R William Soukoreff. 2002. A Character-level Error Analysis Technique for Evaluating Text Entry Methods. In *Proceedings of the second Nordic conference on Human-computer interaction*, pages 243–246.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. In *Interspeech*, pages 2765–2768.

Saida Mussakhojayeva, Aigerim Janaliyeva, Almas Mirzakhmetov, Yerbolat Khassanov, and Huseyin Atakan Varol. 2021. KazakhTTS: An Open-Source Kazakh Text-to-Speech Synthesis Dataset. In *Interspeech 2021*, pages 2786–2790.

Saida Mussakhojayeva, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022a. KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus. In *Proc. Interspeech 2022*, pages 1367–1371.

Saida Mussakhojayeva, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022b. KazakhTTS2: Extending the Open-Source Kazakh TTS Corpus With More Data, Speakers, and Topics. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5404–5411, Marseille, France. European Language Resources Association.

Titouan Parcollet, Yuan Tseng, Shucong Zhang, and Rogier C. van Dalen. 2025. Loquacious Set: 25,000 Hours of Transcribed and Diverse English Speech Recognition Data for Research and Commercial Use. In *Interspeech 2025*, pages 4053–4057.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Yana Veitsman and Mareike Hartmann. 2025. Recent Advancements and Challenges of Turkic Central Asian Language Processing. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages

309–324, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Changhan Wang, Juan Pino, and Jiatao Gu. 2020. Improving Cross-Lingual Transfer Learning for End-to-End Speech Recognition with Speech Translation. In *Interspeech 2020*, pages 4731–4735.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. 2021. WENETSPEECH: A 10000+ Hours Multi-Domain Mandarin Corpus for Speech Recognition. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186.