applied sciences

MDPI

Article

Vowel-Harmony Constraints in Kazakh ASR: A Systematic Investigation of Integration Approaches

Mussa Turdalyuly [1,2,3], Aigerim Yerimbetova [1,2,3,*], Akbayan Bekarystankyzy [1,3], Dina Oralbekova [1], Tolganay Turdalykyzy [1,2,*], Bekturgan Akhmetov [4], Bakzhan Sakenov [1], Nazerke Baisholan [5] and Elmira Daiyrbayeva [1,6]

1 Institute of Information and Computational Technologies, Almaty 050000, Kazakhstan; m.turdalyuly@gmail.com (M.T.); akbayan.bekaristankyzy@narxoz.kz (A.B.); dinaoral@mail.ru (D.O.); sbakzhan22@gmail.com (B.S.); e.daiyrbayeva@satbayev.university (E.D.)
2 School of Engineering and Information Technology, META University, Almaty 050012, Kazakhstan
3 School of Digital Technologies, Narxoz University, Almaty 050035, Kazakhstan
4 Construction and Technical College, Almaty 050062, Kazakhstan; abs83bek83@mail.ru
5 Department of Computer and Information Technology, Purdue University, West Lafayette, IN 47907, USA; nbaishol@purdue.edu
6 Department of Software Engineering, Satbayev University, Almaty 050010, Kazakhstan
* Correspondence: aigerian8888@gmail.com (A.Y.); t.turdalykyzy@meta.edu.kz (T.T.)

Abstract

Kazakh vowel harmony is well established linguistically, but its explicit integration into modern neural automatic speech recognition (ASR) remains insufficiently studied. This work investigates whether a productive front/back vowel-harmony constraint can be operationalized within Whisper large-v3 for Kazakh ASR. Two complementary strategies are evaluated on the ISSAI Kazakh Speech Corpus 2: (1) a differentiable token-level harmony regularizer added during fine-tuning and (2) harmony-based N-best hypothesis re-ranking during inference. We further introduce the Harmony Violation Rate (HVR) as a diagnostic measure and analyze Whisper's byte-level BPE representation of Kazakh vowels. Neither integration strategy produces a statistically significant improvement in test WER. Re-ranking yields a small but statistically significant reduction in HVR, while additional analyses show no recognition disadvantage for byte-fragmented vowels and no systematic reduction in front/back vowel confusions after harmony-constrained fine-tuning. These findings show that simple token-level regularization and sentence-level re-ranking do not adequately model the contextual stem–suffix dependency underlying Kazakh vowel harmony. The study therefore identifies specific limitations of straightforward phonological-constraint integration and provides diagnostic evidence for future representation- and morphology-aware Kazakh ASR.

Keywords: Kazakh ASR; vowel harmony; BPE tokenization; Whisper; N-best re-ranking; low-resource speech recognition; harmony violation rate

Check for updates

Academic Editors: Andrea Prati and Douglas O'Shaughnessy

Received: 10 July 2026
Revised: 14 October 2026
Accepted: 3 November 2026
Published: 7 November 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

Automatic speech recognition (ASR) for low- and medium-resource languages has advanced considerably with the advent of large-scale multilingual models. Whisper [1], trained on 680,000 h of weakly supervised audio across 99 languages, has demonstrated strong out-of-the-box performance for Kazakh—a Turkic language spoken by approximately 13 million people—despite Kazakh representing a small fraction of its training data [2]. Research on Kazakh ASR has progressed from HMM-DNN systems with robust acoustic

features [3–5] to end-to-end transformer architectures [6,7] and non-autoregressive CTC-based decoding [8], and more recently to fine-tuned large-scale multilingual models [9–12]. Comparative evaluations consistently identify vowel harmony and agglutinative morphology as persistent sources of error that standard WER-based evaluation does not adequately characterize [12–14].

Despite these advances, a fundamental limitation remains insufficiently examined: whether modern neural ASR systems are capable of representing and enforcing linguistically grounded phonological constraints, rather than merely approximating acoustic-to-text mappings. For languages with strong phonological regularities, such as Kazakh, this question is not only theoretical but directly observable in system outputs.

Recent studies highlight that the phonological complexity of Kazakh–including vowel harmony and agglutinative morphology–poses persistent challenges for ASR systems, affecting both acoustic and language modeling stages [13,14]. However, these challenges are typically treated as data or modeling issues, without examining whether they arise from deeper representational constraints within the model architecture itself.

A fundamental question underlies this body of work: do state-of-the-art ASR models learn the deep phonological structure of the languages they transcribe, or do they acquire surface-level acoustic-to-text mappings that ignore language-internal constraints? For Kazakh, this question is unusually concrete. Vowel harmony is a productive phonological pattern in Kazakh in which vowels within native stems and many grammatical suffixes typically agree in front/back class. A model that outputs a back-vowel suffix after a front-vowel stem may therefore produce a form that violates the expected harmony pattern. However, loanwords, compounds, and other lexical exceptions mean that mixed-vowel forms should not automatically be interpreted as phonological errors.

1.1. Motivation

Harmony-aware ASR matters for reasons that go beyond phonological correctness. A transcription that outputs ‘балаларға’ as ‘балаларге’ does not merely reflect a substitution error but constitutes a violation of a productive phonological constraint. Such errors are qualitatively distinct from random substitutions, as they generate linguistically ill-formed outputs that are systematically identifiable by native speakers.

This distinction is critical because standard evaluation metrics such as WER assign uniform penalties to all errors, thereby obscuring structured linguistic failures. As a result, systems that systematically violate phonological constraints may appear competitive under conventional evaluation, while producing outputs that are perceptually and linguistically implausible.

Beyond phonological considerations, such errors have downstream consequences. Morphological analysers, machine translation systems, and text-to-speech components that consume ASR output are typically trained on well-formed text, and systematically ill-formed input can propagate errors across the entire processing pipeline. This creates a compounding effect, where phonological inaccuracies degrade performance in subsequent tasks.

Despite the centrality of vowel harmony to Kazakh phonology, existing work has not examined whether such constraints are representable at the level of model internal structure. Prior studies primarily report harmony-related errors or propose improvements at the acoustic or modeling level, without analyzing whether the failure originates from the model’s tokenization and representation mechanisms.

In parallel, recent research demonstrates that subword tokenization can fundamentally limit the encoding of phonological information, leading to systematic degradation of linguistic structure in neural models [15]. However, this line of work has not been connected to ASR systems or to phonological constraints such as vowel harmony.

This gap motivates our central research question: whether phonological constraints can be effectively integrated into neural ASR, or whether their integration is affected by representation choices such as byte-pair encoding.

1.2. Research Questions

The specific problem addressed in this study is not whether Kazakh vowel harmony is linguistically well understood. Rather, it is whether an explicit word-level phonological regularity can be operationalized within a pretrained multilingual ASR system whose decoder operates over subword and byte-level units, and whether such integration produces measurable changes in recognition output. The significance of this problem lies in the mismatch between the linguistic level at which vowel harmony operates and the representation and decoding levels used by modern ASR systems. By testing this mismatch directly, the study aims to identify which straightforward integration strategies are effective, which are not, and why.

We investigate three research questions:

- RQ1. Can vowel-harmony constraints be integrated into the fine-tuning of a large pretrained ASR model, and does this improve transcription quality?
- RQ2. Can harmony-based re-ranking of beam search hypotheses reduce harmony violations in ASR output without retraining?
- RQ3. To what extent does Whisper's byte-pair encoding (BPE) vocabulary affect the representation of Kazakh vowel harmony, and how may this representation influence the integration of linguistic knowledge into ASR systems?

1.3. Contributions

Our contributions are fourfold. First, we formulate a systematic two-stage framework for investigating explicit phonological-constraint integration in Kazakh ASR, separating training-time integration from inference-time hypothesis selection. Second, we operationalize Kazakh vowel harmony through two complementary algorithms: a differentiable token-level harmony regularizer for Whisper fine-tuning and a sentence-level harmony penalty for N-best re-ranking. Third, we analyze how Whisper's BPE representation interacts with explicit harmony integration, including a direct comparison of atomic and byte-fragmented harmony-bearing vowels. Fourth, we provide a multi-level diagnostic evaluation using HVR, cross-class vowel confusion analysis, paired significance testing, and detailed N-best diagnostics. Together, these contributions allow us to identify not only whether straightforward harmony integration improves recognition, but also where its practical limitations arise.

1.4. Paper Organisation

Section 2 provides background on Kazakh vowel harmony, the Whisper architecture, and the BPE tokenization analysis. Section 3 reviews related work. Section 4 describes the ISSAI KSC2 dataset and evaluation metrics including HVR. Section 5 presents HarmonyConstrainedWhisper and its experimental results. Section 6 describes and evaluates harmony re-ranking. Section 7 discusses the implications, and Section 8 concludes.

2. Background

This section provides the linguistic and architectural background necessary to understand the interaction between Kazakh vowel harmony and neural ASR systems, with particular emphasis on how tokenization influences the representation of phonological structure.

2.1. Kazakh Vowel Harmony

Vowel harmony is a phonological process in which vowels within a prosodic domain—typically the word—must share one or more phonological features. In Kazakh, the relevant feature is backness: vowels are divided into two classes, front and back, and all vowels within a native Kazakh word must belong to the same class [16,17]. This constraint applies not only to stems but extends productively to grammatical suffixes, which have front and back allomorphs that agree with the stem vowel class.

Vowel inventory. Kazakh has fourteen vowel phonemes in the Cyrillic orthographic system. For the purposes of vowel harmony, we classify these vowels into three groups following the standard phonological analysis [16], as summarized in Table 1.

Table 1. Kazakh vowel classification for harmony.

Class	Vowels	Description
Front	ү, ә, ө, і	High and mid front vowels
Back	а, о, ұ, ы	Low and mid back vowels
Neutral	е, и, у	Vowels that do not trigger harmony alternations

Neutral vowels do not participate in harmony alternations and are compatible with both front- and back-harmony words. The vowels е and и in particular behave transparently—they can appear in words of either harmony class without triggering violations.

Suffix alternations. The practical consequence of vowel harmony for ASR is most visible in suffixes. The dative suffix, for example, has two allomorphs: '-ға'/'-ге' (back/front) and '-қа'/'-ке' (back/front after voiceless consonants). A model that transcribes 'үйге' ("to the house", front-harmony) as 'үйға' has produced a form that is phonologically impossible in Kazakh—not merely a stylistic variant but a violation of a productive grammatical constraint. Table 2 illustrates selected suffix alternations.

Table 2. Selected Kazakh suffix alternations by vowel-harmony class.

Suffix Function	Back Allomorph	Front Allomorph	Example (Back)	Example (Front)
Dative	-ға/-қа	-ге/-ке	балаға (to child)	үйге (to house)
Plural	-лар	-лер	балалар (children)	үйлер (houses)
Genitive	-ның	-нің	баланың (child's)	үйдің (house's)
Locative	-да/-та	-де/-те	балада (at child)	үйде (at house)
Ablative	-дан/-тан	-ден/-тен	баладан (from child)	үйден (from house)

Exceptions. Vowel harmony in Kazakh is not exceptionless. Loanwords—predominantly from Russian, Arabic, and Persian—frequently violate harmony, as do compound words where the constituents belong to different harmony classes. The word университет ("university"), for example, contains both front and back vowels in a pattern that would be impossible in a native stem. These systematic exceptions account for the 3.21% baseline HVR observed in the ISSAI KSC2 training corpus (Section 4.2).

From the perspective of automatic speech recognition, vowel harmony provides a useful language-internal regularity for examining whether model outputs preserve productive phonological patterns. Because this regularity is systematic but not exceptionless, apparent violations should be interpreted in relation to lexical status and morphological context rather than treated automatically as modeling errors. This motivates harmony-aware diagnostic analysis while also requiring caution when interpreting mixed front/back vowel patterns in ASR output.

2.2. Whisper Architecture and Multilingual ASR

Whisper [1] is a sequence-to-sequence ASR model based on the Transformer architecture [18]. The model consists of a convolutional audio encoder that processes log-mel spectrograms and an autoregressive text decoder that generates transcriptions token by token. Whisper large-v3, used in all experiments reported here, contains 1543M parameters across 32 encoder layers and 32 decoder layers with a model dimension of 1280.

Pretraining. The original Whisper models described by Radford et al. [1] were trained on 680,000 h of weakly supervised audio–transcript pairs across 99 languages. The experiments in this study use the later Whisper large-v3 checkpoint.

Decoding. At inference time, Whisper generates transcriptions using beam search over the token vocabulary. The decoder is conditioned on a language token and a task token (transcription or translation), enabling language-specific decoding without language identification as a preprocessing step. Beam search with width B maintains the B highest-probability partial hypotheses at each decoding step, returning the globally highest-probability complete sequence as the final transcription.

BPE tokenization. Whisper uses byte-pair encoding (BPE) [19] to construct a subword vocabulary of 51,865 tokens from the pretraining corpus. BPE begins with a byte-level vocabulary covering all possible byte values (256 tokens) and iteratively merges the most frequent adjacent byte pairs until the target vocabulary size is reached. For languages with high-resource representation in the pretraining data, common words and subwords receive dedicated tokens. For low-resource languages, and particularly for characters outside the Latin and basic Cyrillic range, the algorithm falls back to byte-level representations.

While BPE tokenization is effective for capturing frequent subword patterns in high-resource languages, it introduces representational challenges for low-resource and morphologically rich languages. In particular, byte-level fallback mechanisms may fragment linguistically meaningful units, such as phonemes or morphemes, into sequences that lack explicit semantic or phonological interpretation.

For explicit token-level harmony constraints, such fragmentation may reduce direct access to vowel-class information, although it does not by itself imply lower vowel recognition accuracy.

The limitations of BPE for morphologically rich agglutinative languages are well-documented for Turkish [20,21], a Turkic language with structural properties closely analogous to Kazakh. Toraman et al. [20] demonstrate that morphological-level tokenization is competitive with BPE for Turkish downstream tasks, while Kaya and Tantuğ [21] show that Turkish words require substantially more subword splits than English under common subword tokenization schemes. These findings extend directly to Kazakh, as we demonstrate in Section 2.3.

2.3. Whisper's BPE Tokenization and Kazakh Vowel Harmony

A critical prerequisite for any token-level linguistic constraint is that the phenomenon of interest must be observable at the token level. Whisper's byte-pair encoding (BPE) vocabulary creates a potential implementation constraint for explicit harmony-aware processing of Kazakh text. Specifically, key harmony-bearing vowels are represented as multi-byte token sequences, which obscures their phonological identity at the level where decoding decisions are made.

This observation is consistent with previous work on Kazakh tokenization. Tukeyev et al. [22] show that standard BPE tokenization may fail to preserve morphologically meaningful boundaries in agglutinative Kazakh, suggesting that subword segmentation can complicate the representation of linguistically relevant structure. Together, these results

suggest that subword tokenization may hinder the preservation of linguistically meaningful structure, including phonological constraints such as vowel harmony.

2.3.1. Byte-Level Encoding of Kazakh Front Vowels

Whisper large-v3 uses a multilingual BPE vocabulary of 51,865 tokens trained on a broad corpus dominated by high-resource languages. Kazakh-specific characters that fall outside the Latin and basic Cyrillic range are encoded at the byte level rather than as dedicated tokens. Table 3 presents the tokenization of Kazakh vowels that are phonologically decisive for vowel harmony.

Table 3. Whisper BPE tokenization of Kazakh vowels.

Character	Unicode	Vowel Class	Token IDs	N Tokens
ү	U+04AF	Front	[142, 107]	2
ә	U+04D9	Front	[143, 247]	2
ө	U+04E9	Front	[143, 102]	2
і	U+0456	Front	[1827]	1
ұ	U+04B1	Back	[142, 109]	2
а	U+0430	Back	[386]	1
о	U+043E	Back	[354]	1
ы	U+044B	Back	[698]	1
е	U+0435	Neutral	[387]	1
и	U+0438	Neutral	[435]	1

Four of the eight harmony-bearing vowels—‘ү’, ‘ә’, ‘ө’, ‘ұ’—are split into two opaque byte tokens. These byte tokens carry no semantic identity: token 142 (<0xD2>) appears as the first byte of both the front vowel ‘ү’ (U+04AF) and the back vowel ‘ұ’ (U+04B1), making it difficult to determine harmony class from a single token in isolation.

2.3.2. Consequences at the Word Level

The byte-level fragmentation propagates to the word level. Table 4 shows the tokenization of representative Kazakh words.

Table 4. Word-level BPE tokenization examples.

Word	Gloss	Tokens	Front Vowels Visible?
мектепке	school-DAT	[‘м’, ‘ект’, ‘еп’, ‘ке’]	–
үйге	home-DAT	[‘<0xD2>’, ‘<0xAF>’, ‘й’, ‘ге’]	No
қазақша	in Kazakh	[‘қ’, ‘аз’, ‘а’, ‘қ’, ‘ша’]	–
сөйлеу	speaking	[‘с’, ‘<0xD3>’, ‘<0xA9>’, ‘й’, ‘ле’, ‘у’]	No
балаларға	children-DAT	[‘б’, ‘ал’, ‘ал’, ‘ар’, ‘<0xD2>’, ‘<0xB1>’, ‘а’]	No

These examples illustrate that phonologically decisive information is distributed across multiple tokens without explicit boundaries, requiring the model to implicitly reconstruct character-level structure before any phonological reasoning can be applied. This additional reconstruction step introduces ambiguity and increases the difficulty of learning consistent phonological patterns during training.

In the word ‘үйге’ (“to home”), the front vowel ‘ү’—the sole harmony trigger—is split across tokens 142 and 107. Neither token alone identifies the vowel as front. Similarly, ‘сөйлеу’ (“speaking”) contains the front vowel ‘ө’ split into tokens 143 and 102. A model operating on these token sequences cannot determine vowel-harmony class without reconstructing the original Unicode codepoint from its constituent bytes.

2.3.3. Implications for Harmony-Constrained Decoding

The tokenization pattern described above has practical implications for explicit harmony-aware modeling. In several cases, a harmony-bearing Kazakh vowel is not available as a single atomic tokenizer entry. For example, token 142 (<0xD2>) is the first byte of both ү (U+04AF) and ұ (U+04B1), while the identity of the complete vowel becomes available only after the second byte token is generated.

This representation does not prevent an autoregressive Transformer from recovering the intended character from the complete byte sequence and its surrounding context. However, it reduces the extent to which the harmony class of some Kazakh vowels is directly represented by individual vocabulary tokens. This is particularly relevant to the training-time regularizer used in this study, because its front- and back-harmony vocabulary masks are defined over decoded tokenizer entries. Byte fragments that do not independently reveal a front/back vowel distinction cannot contribute directly to these masks.

The two approaches examined in this study therefore interact with tokenization differently. During harmony-constrained fine-tuning, the auxiliary regularizer operates directly on decoder probabilities but is limited by the coverage of harmony-relevant vocabulary entries. During N-best re-ranking, harmony is evaluated after complete hypotheses have been decoded into Unicode text, so the full character and word forms are available to the sentence-level harmony-penalty calculation.

Accordingly, byte-level fragmentation is treated here as a potential representational limitation for explicit harmony integration rather than as evidence that Whisper cannot learn vowel harmony. Establishing a causal effect of tokenization would require controlled comparison with alternative character-level or morphology-aware tokenization schemes.

3. Related Work

Research on Kazakh automatic speech recognition has progressed rapidly in recent years, driven by the availability of larger datasets and the adoption of transformer-based architectures. Early work focused on hybrid HMM-DNN systems and acoustic feature engineering [3–5], while more recent studies explore end-to-end neural approaches, including transformer-based ASR and large multilingual pretrained models such as Whisper [6,9–12]. These approaches achieve strong performance in terms of conventional metrics such as WER, but consistently report errors related to agglutinative morphology and vowel harmony [12–14].

Large multilingual speech models have changed the baseline expectations for low-resource ASR [23]. Whisper, trained on large-scale multilingual weakly supervised speech data, has shown strong transfer ability across many languages [1]. Recent studies on Kazakh ASR have also examined Whisper-based and large pretrained speech models, showing that such systems can achieve competitive recognition accuracy while still producing errors related to morphology, suffixes, and language-specific phonological patterns [9–12,14].

Previous work on Turkic ASR further shows that multilingual modeling can be useful for related low-resource languages, but does not remove the need for language-specific evaluation. Kazakh, Uzbek, Kyrgyz, Tatar, and other Turkic languages share agglutinative morphology and, in many cases, vowel harmony, yet they differ in script, orthography, available corpora, and domain coverage [7,12,24]. This makes Turkic ASR a useful setting for studying how multilingual speech models handle language-internal structure beyond standard WER and CER.

Despite these advances, existing work primarily treats such errors as data-driven or modeling limitations. For example, studies on Kazakh ASR typically focus on improving acoustic representations, training strategies, or decoding techniques, without explicitly addressing whether phonological constraints are captured by the model. As a result,

vowel-harmony violations are documented but not systematically analyzed as a structural limitation of the model.

Another relevant line of work concerns the integration of linguistic knowledge into neural recognition systems. In ASR, linguistic constraints are commonly introduced through language model rescoring, lexicon-based decoding, constrained hypothesis selection, or post-processing methods [10,25]. These approaches are useful because they can improve linguistic consistency without fully redesigning the acoustic model. However, their effectiveness depends on whether the relevant linguistic distinction is available in the model output or in the candidate hypotheses produced during decoding.

Tokenizer design is also central for morphologically rich and low-resource languages. Subword tokenizers such as BPE are efficient for frequent patterns, but they may fragment rare characters, suffixes, or morphologically meaningful units [15,19,26]. For agglutinative languages, this can increase the mismatch between tokens and linguistic units such as stems, suffixes, and phonologically relevant vowels. Studies on Turkish and Kazakh tokenization have shown that morphology-aware or language-aware segmentation can better preserve linguistic structure than standard subword segmentation in some settings [20–22].

More broadly, recent work in neural language processing highlights that tokenization is not a neutral preprocessing step but a critical component that shapes the representational capacity of the model. In particular, byte-level tokenization can obscure linguistically meaningful units, especially for underrepresented scripts and phonological systems [15,27]. These findings suggest that certain linguistic phenomena may be inherently difficult to model when they are not explicitly encoded at the token level.

Within ASR, approaches to incorporating linguistic constraints have typically focused on higher-level mechanisms such as language model rescoring, lexicon constraints, or post-processing rules [10,25]. While these methods can improve lexical or syntactic consistency, they assume that the relevant linguistic information is already recoverable from model outputs. This assumption may not hold for phonological constraints if the underlying tokenization fails to preserve the necessary information.

In this context, the present study is diagnostic rather than purely architectural. Harmony-constrained fine-tuning and N-best re-ranking are based on established strategies for incorporating external knowledge into neural systems [10,25]. Our contribution is to use Kazakh vowel harmony as a test case for examining how far these straightforward strategies can go when the linguistic constraint is phonological, word-level, and partially misaligned with the model's byte-level tokenization. This framing positions the study as an analysis of representation, constraint integration, and evaluation for Kazakh ASR.

4. Dataset and Evaluation Metrics

4.1. ISSAI KSC2 Corpus

All experiments are conducted on the ISSAI Kazakh Speech Corpus 2 (KSC2) [2], the largest publicly available transcribed speech corpus for Kazakh. The corpus covers five distinct domains, providing substantial variation in speaking style, vocabulary, and recording conditions. Table 5 summarises the data splits used in this work.

Table 5. ISSAI KSC2 dataset statistics.

Split	Utterances	Domain Coverage
Train	627,824	All 5 domains
Validation	6259	All 5 domains
Test	6733	All 5 domains
Total	640,816	

The five domains differ substantially in their linguistic and acoustic properties. Table 6 provides per-domain statistics for the test split.

Table 6. Per-domain statistics (test split).

Domain	Utterances	Baseline WER (%)	Speaking Style
Crowdsourced	3328	4.75	Read speech, controlled
Parliament	773	7.05	Formal, prepared speech
Talkshow	383	12.69	Spontaneous dialogue
Radio	702	13.53	Broadcast, varied topics
Podcasts	1547	15.45	Conversational, informal

The diversity of domains in KSC2 is particularly important for evaluating phonological constraints, as spontaneous speech introduces greater morphological variability and a higher likelihood of phonological reduction or irregularities. This makes the dataset suitable not only for standard ASR evaluation but also for assessing the robustness of linguistically informed modeling approaches.

The substantial WER variation across domains—from 4.75% on crowdsourced read speech to 15.45% on podcasts—reflects differences in spontaneity, background noise, and vocabulary diversity. Audio is sampled at 16 kHz and transcribed in Kazakh Cyrillic script.

4.2. Harmony Distribution in the Corpus

To characterise the phonological profile of the corpus, we compute vowel-harmony statistics over the training transcriptions, as summarized in Table 7. We classify each content word (length > 1, excluding numeric tokens) according to its dominant vowel class: front, back, or mixed (harmony violation).

Table 7. Vowel-harmony distribution in KSC2 training transcriptions.

Vowel Class	Words	Percentage (of Harmony-Bearing)
Back-harmony words	4,212,287	64.77%
Front-harmony words	2,082,259	32.02%
Mixed (violations)	208,973	3.21%
No harmony vowels	658,141	–
Total content words	7,161,660	

Back-harmony words account for 64.77% of the harmony-bearing words in the training transcriptions, while front-harmony words account for 32.02%. Mixed front/back vowel patterns represent 3.21% of the harmony-bearing words according to the rule-based classification used in this study.

The non-zero rate observed in the reference transcriptions is important for interpreting HVR. Mixed-vowel patterns in reference text may arise from valid lexical items such as loanwords, proper names, compounds, abbreviations, and code-switched expressions, as well as from other corpus-specific or annotation-related factors. Therefore, the 3.21% value should be treated as an empirical reference level for the KSC2 training transcriptions rather than as a lower bound on phonological correctness. The objective of HVR-based analysis is consequently not to drive the metric toward zero, but to compare system outputs with reference patterns while considering WER, CER, and qualitative error analysis.

4.3. Evaluation Metrics

We evaluate all systems using three complementary metrics that capture different aspects of transcription quality. While WER and CER measure lexical and character-level accuracy, they do not account for structured linguistic constraints. Therefore, we introduce an additional metric designed to capture phonological well-formedness.

4.3.1. Word Error Rate (WER)

Word Error Rate is the standard metric for ASR evaluation, defined as:

$$WER = \frac{S + D + I}{N}, \tag{1}$$

where S, D, I are the number of substitutions, deletions, and insertions respectively relative to the reference transcription, and N is the total number of words in the reference. However, WER does not differentiate between types of errors. In particular, it assigns equal cost to phonologically invalid outputs and semantically unrelated substitutions, which limits its ability to capture structured linguistic failures.

4.3.2. Character Error Rate (CER)

Character Error Rate applies the same edit-distance formulation at the character level. For agglutinative languages such as Kazakh, where a single word can encode multiple morphological categories, CER provides a finer-grained measure of transcription accuracy [7]. Nevertheless, CER remains insensitive to higher-level linguistic structure, including phonological constraints such as vowel harmony, as it operates purely at the character level.

4.3.3. Harmony Violation Rate (HVR)

We introduce the Harmony Violation Rate (HVR) as a rule-based diagnostic metric for identifying mixed front/back vowel patterns in Kazakh ASR output. HVR is intended to complement WER and CER rather than to serve as an independent measure of transcription quality or phonological correctness.

Definition. Let W_H denote the set of harmony-bearing content words in a transcription. Content words are defined as words longer than one character with numeric tokens excluded, and a word is considered harmony-bearing if it contains at least one decisive front vowel (ү, ә, ө, і) or back vowel (а, о, ұ, ы). A harmony-bearing word is classified as mixed when it contains at least one vowel from both classes. HVR is defined as the proportion of mixed words among all harmony-bearing words, as shown in Equation (2).

$$HVR = \frac{\left|\{w \in W \ : F(w) \neq \emptyset \bigwedge B(w) \neq \emptyset\}\right|}{\left|\{w \in W \ : F(w) \cup B(w) \neq \emptyset\}\right|} \tag{2}$$

HVR ranges from 0 to 1, with larger values indicating a greater proportion of mixed front/back vowel patterns according to the rule above. Importantly, such patterns are not necessarily phonological errors. Reference transcriptions may contain valid loanwords, proper names, compounds, abbreviations, code-switched expressions, and other forms that satisfy the rule despite being legitimate lexical items.

HVR therefore provides information that is different from WER and CER. It can be used to track changes in a specific class of vowel-pattern outputs across systems, but a lower HVR does not by itself indicate a more accurate or linguistically better transcription. HVR results should consequently be interpreted together with conventional error metrics and, where possible, qualitative inspection of the affected words.

Limitations. The current HVR implementation is rule-based and does not distinguish native Kazakh words from loanwords, proper names, compounds, abbreviations, or code-switched expressions. It also does not model morphological structure and therefore cannot determine whether a mixed-vowel pattern constitutes an actual stem–suffix harmony violation. A future lexicon-aware or morphology-aware version of HVR could restrict the analysis to linguistically verified contexts. In addition, the current formulation uses a simplified front/back classification and does not represent more detailed context-dependent phonological phenomena.

Together, these metrics provide a multi-dimensional evaluation framework, allowing lexical accuracy and harmony-related output patterns to be examined separately.

4.4. Systematic Investigation Framework

The present study is designed as a systematic two-path investigation of explicit vowel-harmony integration in Kazakh ASR. Rather than proposing two unrelated techniques, we evaluate the same linguistic constraint at two different intervention points of the Whisper recognition pipeline.

The first approach, harmony-constrained fine-tuning, addresses the training-time integration problem: whether vowel-class information can directly influence the decoder probability distribution and model parameter updates before discrete transcription is produced. The method therefore tests whether phonological knowledge can be incorporated into the internal optimization process of the ASR model.

The second approach, harmony-based N-best re-ranking, addresses the inference-time selection problem: whether vowel-harmony information can improve the selection of a final transcription when alternative hypotheses have already been generated by the model. Unlike the first approach, it does not modify model parameters and operates only on fully decoded Unicode hypotheses.

The two approaches are therefore complementary rather than competing. The training-time experiment evaluates whether the constraint can influence representation and optimization, whereas the inference-time experiment evaluates whether the same constraint can influence hypothesis selection after decoding. Using the same corpus, model family, and evaluation metrics provides a common basis for evaluating the two integration approaches against their respective baselines, while accounting for differences in decoding configurations.

The systematic investigation consequently consists of three linked analytical levels: (1) representation, through the analysis of Whisper BPE tokenization and atomic versus byte-fragmented vowels; (2) integration, through training-time regularization and inference-time re-ranking; and (3) diagnostic evaluation, through WER, CER, HVR, statistical significance testing, vowel-class confusion analysis, and N-best candidate analysis. The overall research methodology, including the relationship between the two vowel-harmony integration approaches and the diagnostic evaluation procedures, is illustrated in Figure 1.

This design allows the study to determine not only whether explicit vowel-harmony integration improves ASR performance, but also at which stage potential limitations arise: token representation, training-time constraint integration, or inference-time candidate selection.

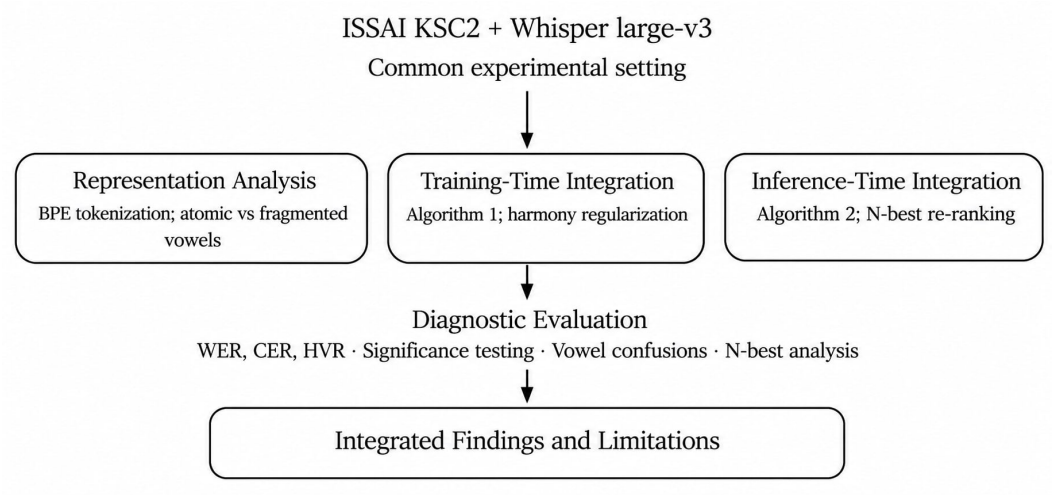


Figure 1. Integrated Research Methodology for Systematic Investigation of Vowel-Harmony Constraints in Kazakh ASR.

5. Approach 1: Harmony-Constrained Fine-Tuning

5.1. Motivation

A natural approach to incorporating vowel harmony into ASR is to integrate phonological constraints directly into the training objective. This approach specifically addresses the training-time integration problem: whether vowel-harmony information can influence decoder probability distributions and parameter updates before discrete decoding. In principle, this allows the model to internalize harmony-consistent patterns during parameter updates. However, this approach assumes that the relevant phonological information is accessible at the representation level used by the model. As discussed in Section 2.3, this assumption may not hold under byte-level tokenization, where phonologically meaningful units are not explicitly represented.

Figure 2 summarizes the two harmony-integration strategies evaluated in this study. The first path corresponds to training-time harmony-constrained fine-tuning, where the standard Whisper cross-entropy objective is supplemented with a differentiable token-level harmony regularization term computed directly from the decoder probability distribution under teacher forcing. Fixed front- and back-harmony vocabulary masks are used to penalize probability mass assigned to tokens belonging to the harmony class opposite to that of the corresponding reference token. The second path corresponds to inference-time N-best re-ranking, where fully decoded hypotheses are scored using both model likelihood and a sentence-level harmony penalty. Thus, the training-time and inference-time approaches operate at different representation levels: the former acts directly on decoder probabilities, whereas the latter evaluates phonological consistency after decoding.

Methodological framework. The methodological contribution of this study is not a new acoustic encoder or ASR backbone. Instead, we operationalize Kazakh vowel harmony at two distinct points of an existing sequence-to-sequence ASR system. The training-time procedure converts vowel-class information into a differentiable probability-mass penalty over the decoder vocabulary, whereas the inference-time procedure converts the same linguistic principle into a sentence-level penalty applied to fully decoded N-best hypotheses. These two procedures constitute complementary algorithms for testing whether phonological knowledge is more effectively integrated before or after discrete decoding.

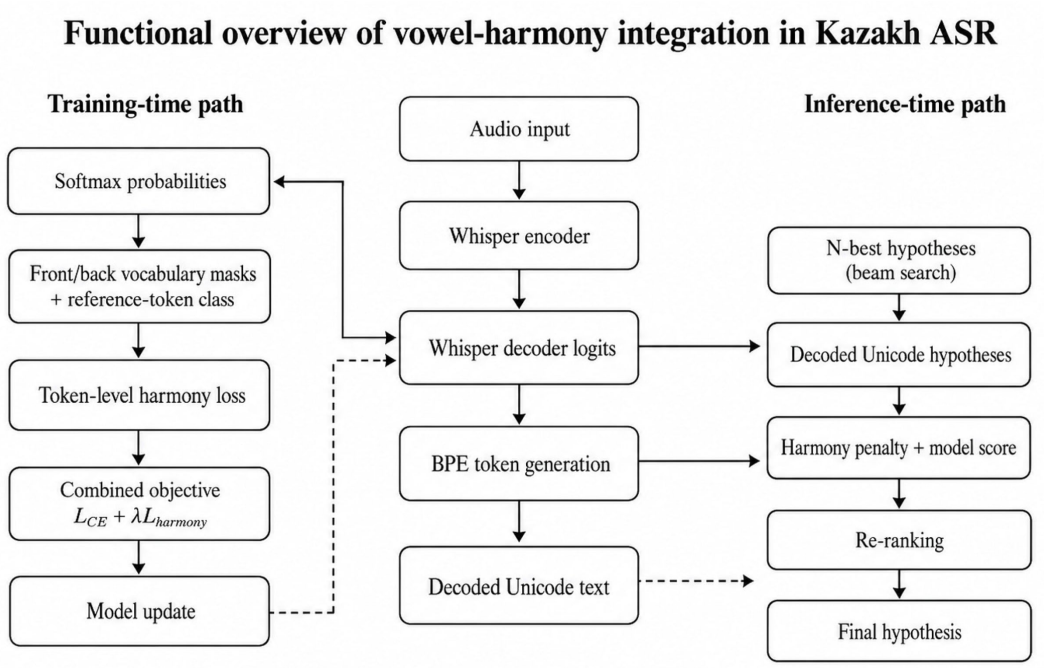


Figure 2. Functional overview of the two vowel-harmony integration strategies evaluated in Kazakh ASR. The central path represents the standard Whisper encoder–decoder pipeline. During training, a differentiable harmony regularization term is computed directly from the decoder probability distribution using fixed front- and back-harmony vocabulary masks and the corresponding reference-token classes. This term is combined with the standard cross-entropy objective and optimized through ordinary backpropagation. During inference, N-best hypotheses are fully decoded into Unicode text and evaluated using the sentence-level harmony penalty together with the model log-probability. The two branches therefore apply vowel-harmony information at different representation levels: token probabilities during training and decoded word strings during re-ranking. Solid arrows indicate the main data and computation flow, while dashed arrows represent the training feedback path and auxiliary connections between processing stages.

5.2. HarmonyConstrainedWhisper

5.2.1. Architecture

HarmonyConstrainedWhisper retains the standard Whisper large-v3 encoder–decoder architecture and augments its training objective with an additional harmony-aware regularization term. No new acoustic encoder, decoder block, or external phonological module is introduced. Instead, the proposed modification operates at the level of the decoder output distribution under teacher forcing.

Let $z_t \in \mathbb{R}^V$ denote the decoder logits at output position t, where V is the tokenizer vocabulary size, and let $p_t(k) = softmax(z_t)_k$ be the probability assigned to token k. To inject vowel-harmony information into training, we construct a fixed vocabulary-level mapping that assigns tokenizer entries to harmony-relevant classes based on their decoded character content. In particular, tokenizer entries containing decisive Kazakh front vowels are assigned to the front class, those containing decisive back vowels are assigned to the back class, and all remaining entries are treated as mixed, neutral, or non-decisive for the purpose of the auxiliary regularizer.

During training, the reference label at each output position determines which opposite-class probability mass should be penalized. If the reference token belongs to the back-harmony class, the model is penalized for assigning probability mass to front-class tokens; if the reference token belongs to the front-harmony class, the model is penalized for assigning

probability mass to back-class tokens. In this way, the regularizer acts directly on the decoder probability distribution rather than on fully decoded word strings.

The total training objective combines the standard cross-entropy loss with the harmony regularization term:

$$L_{total} = L_{CE} + \lambda \times L_{harmony} \quad (3)$$

where λ is the harmony-loss weight. In our experiments, $\lambda = 0.15$. Because $L_{harmony}$ is computed from the softmax probabilities over the decoder logits, the full objective remains differentiable and is optimized using ordinary backpropagation.

This design allows harmony information to influence training without modifying the underlying Whisper architecture. At the same time, the effectiveness of the approach depends on how clearly harmony-relevant distinctions are represented in the tokenizer vocabulary, which motivates the tokenization analysis presented in Section 2.3 and the diagnostic interpretation of the results discussed later in the paper.

5.2.2. Harmony Loss

To implement the auxiliary harmony regularization term, the Whisper vocabulary is first divided into harmony-related classes. Each vocabulary token is decoded and examined for the presence of Kazakh front vowels (ү, ә, ө, і) or back vowels (а, о, ұ, ы). Tokens containing at least one front vowel and no back vowels are assigned to the front-harmony class, while tokens containing at least one back vowel and no front vowels are assigned to the back-harmony class. Tokens that do not provide an unambiguous front/back distinction do not directly contribute to the auxiliary loss.

During training, the model operates under teacher forcing, so the reference token at each decoder position is known. If the reference token belongs to the back-harmony class, the auxiliary loss penalizes probability mass assigned by the decoder to front-harmony tokens at that position. Conversely, if the reference token belongs to the front-harmony class, probability mass assigned to back-harmony tokens is penalized.

The penalties are averaged separately over valid front- and back-harmony reference positions and then combined to obtain the final harmony regularization term. Padding positions are excluded from the calculation. If only one of the two harmony classes is represented in a training batch, only the corresponding component contributes to the auxiliary loss.

This procedure differs from a word-level HVR calculation. The harmony loss is applied directly to the decoder probability distribution before discrete decoding, whereas corpus-level HVR is calculated from fully decoded text and is used as an evaluation measure. N-best re-ranking uses the related sentence-level harmony penalty described in Section 6.2.2. Consequently, the training objective can directly influence model parameters through standard backpropagation.

The training procedure used for HarmonyConstrainedWhisper is summarized in Algorithm 1.

Algorithm 1: Harmony-Constrained Fine-Tuning

1. Decode all Whisper vocabulary entries.
2. Construct fixed front/back harmony masks.
3. Compute decoder logits under teacher forcing.
4. Apply softmax to obtain token probabilities.
5. At each eligible reference position, calculate probability mass assigned to the opposite harmony class.
6. Average the penalties to obtain \(L_{harmony}\).
7. Compute $L_{total} = L_{CE} + \lambda \times L_{harmony}$.
8. Update Whisper parameters by ordinary backpropagation.

5.2.3. Differentiability and Gradient Flow

The harmony regularization term is fully differentiable with respect to the decoder logits. Unlike the corpus-level HVR metric used for evaluation, the training-time harmony loss does not require discrete decoding, detokenization, or word segmentation. It is computed directly from the softmax probability distribution produced by the decoder during teacher-forced training.

The front- and back-harmony vocabulary masks and the reference-token classes are fixed during each forward pass and therefore do not require gradient computation. For reference positions belonging to one harmony class, the loss accumulates the probability mass assigned to tokens of the opposite class. Because these probabilities are obtained directly from the decoder logits through the softmax operation, gradients from the harmony term propagate through the decoder in the same computational graph as the standard cross-entropy loss.

The harmony term is added to the standard cross-entropy objective before backpropagation. Consequently, both components contribute to parameter updates through the ordinary automatic differentiation procedure used during fine-tuning. No straight-through estimator, reinforcement-learning procedure, or other surrogate-gradient approximation is required.

The differentiability of the objective does not, however, guarantee that the auxiliary signal will substantially improve recognition accuracy. Its practical effect depends on the coverage of the harmony-related vocabulary masks, the representation of Kazakh characters in the tokenizer, and the relative strength of the auxiliary term compared with the primary transcription objective. These factors are considered in the analysis of the experimental results.

5.2.4. Training Configuration

Fine-tuning is performed on the ISSAI KSC2 training split (627,824 utterances) using three NVIDIA L40 GPUs (46 GB each). Table 8 summarises the key hyperparameters.

Table 8. HarmonyConstrainedWhisper training configuration.

Hyperparameter	Value
Base model	whisper-large-v3
Total parameters	1543M
Learning rate	1×10^{-5}
Effective batch size	48 (2/GPU × 8 grad. accum. × 3 GPUs)
Harmony-loss weight λ	0.15
Epochs	5
Gradient checkpointing	Yes
Optimizer	AdamW

The harmony-loss weight was fixed at $\lambda = 0.15$ for the reported experiment. This value determines the relative contribution of the auxiliary harmony regularization term to the standard cross-entropy objective. We did not perform a systematic ablation over λ, and therefore do not claim that $\lambda = 0.15$ is an optimal value. The remaining training settings were selected to maintain stable fine-tuning of Whisper large-v3 within the available computational resources.

5.2.5. Implementation and Decoding Details

All experiments were conducted using the Whisper large-v3 multilingual model with its original tokenizer and vocabulary. The tokenizer was not modified during fine-tuning,

which allowed us to directly examine the behavior of the standard byte-level BPE representation for Kazakh. Audio inputs were processed at 16 kHz following the standard Whisper preprocessing pipeline. Text normalization was applied consistently across the baseline, harmony-constrained fine-tuning, and re-ranking experiments.

The training experiments were implemented in Python 3.12.3 using PyTorch 2.10.0 and the Hugging Face Transformers 5.6.2 implementation of Whisper. Fine-tuning was performed on three NVIDIA L40 GPUs with 46 GB of memory each. The model was trained for five epochs using AdamW, a learning rate of 1×10^{-5}, gradient checkpointing, and an effective batch size of 48. The auxiliary harmony-loss weight was fixed at $\lambda = 0.15$. Model selection was based on validation WER.

For decoding, the baseline and harmony-constrained models were evaluated using the same decoding and normalization settings. For the re-ranking experiment, N-best hypotheses were generated using beam search with a beam width of 10, with five candidate hypotheses retained for each utterance. Each candidate was decoded into Unicode text and scored using the sentence-level harmony penalty together with its model score. The final hypothesis was selected using the combined re-ranking score described in Section 6.2. The penalty weight α was selected on the validation set and then fixed for test-set evaluation. The baseline values reported for the fine-tuning experiment (Tables 9 and 10) were obtained using the generation settings employed during model evaluation, whereas the re-ranking experiment was evaluated from separately generated five-best beam-search lists. Consequently, the baseline WER values associated with these two evaluation procedures are not numerically identical.

To support reproducibility, all reported WER, CER, and HVR values were computed on the same validation and test splits of ISSAI KSC2. No additional language model, lexicon, or external post-processing system was used in the main experiments.

5.3. Results

Table 9 presents validation metrics at each training epoch alongside the baseline, while the corresponding test set performance is summarized in Table 10.

Table 9. HarmonyConstrainedWhisper validation metrics per epoch.

System	Epoch	Val WER (%)	Val CER (%)	Eval Loss
Baseline	–	9.62	3.38	–
HarmonyConstrainedWhisper	1	11.05	3.87	0.0985
HarmonyConstrainedWhisper	2	10.53	3.64	0.1018
HarmonyConstrainedWhisper	3	10.01	3.47	0.1087
HarmonyConstrainedWhisper	4	9.73	3.36	0.1174
HarmonyConstrainedWhisper	5	9.25	3.17	0.1256

Table 10. Test set results.

System	Test WER (%)	Test CER (%)	HVR (%)	ΔWER (Percentage Points)
Baseline	9.32	3.25	3.153	–
HarmonyConstrainedWhisper (epoch 5)	9.30	3.24	3.168	−0.02

The observed test-set differences between the baseline and HarmonyConstrainedWhisper are very small: WER changes from 9.32% to 9.30% and CER from 3.25% to 3.24%. These differences are too small to support a claim of improved recognition accuracy under the

present experimental setup. Similarly, HVR changes only from 3.153% to 3.168%, providing no evidence that the auxiliary regularization improves harmony-related output patterns on the test set.

The validation results show a different pattern, with WER decreasing from the baseline value of 9.62% to 9.25% at epoch 5. However, this improvement is not reflected to the same extent on the held-out test set. We therefore interpret the validation–test discrepancy cautiously rather than attributing it directly to improved phonological generalization or to overfitting. Subsequent paired significance testing (Section 5.4.5) confirmed that the 0.021-percentage-point WER difference is not statistically significant ($p = 0.801$).

5.4. Analysis

5.4.1. Validation–Test Discrepancy

The validation and test sets show different patterns. Validation WER decreases from 9.62% to 9.25%, whereas test WER changes only from 9.32% to 9.30%. At the same time, the evaluation loss increases from 0.0985 at epoch 1 to 0.1256 at epoch 5 while validation WER continues to decrease. This divergence indicates that the validation improvement does not translate into a comparable gain on the held-out test set. However, these observations alone are insufficient to establish overfitting or a failure of phonological generalization. We therefore interpret the discrepancy more cautiously as evidence that the effect of the auxiliary harmony regularizer on recognition accuracy is small and may depend on the evaluation split under the present experimental setup.

5.4.2. Why the Harmony Loss Has Limited Effect

The small test-set effect of the auxiliary harmony term should be interpreted in relation to what the regularizer actually optimizes. Although the harmony loss is fully differentiable, it is not a direct word-level harmony constraint and does not explicitly compare a suffix with the vowel class of the preceding stem. Under teacher forcing, the reference token at each eligible position determines the harmony class, and the regularizer penalizes probability mass assigned to vocabulary tokens belonging to the opposite class. The resulting signal therefore encourages local token-class consistency with the reference sequence, while word-level vowel harmony is addressed only indirectly.

Several factors may limit the strength of this signal. First, the effectiveness of the fixed vocabulary masks depends on whether harmony-relevant characters are identifiable in individual tokenizer entries. As shown in Section 2.3, several Kazakh vowels, including ɣ, ə, ө, and ұ, are represented as multi-token byte sequences. In such cases, individual byte-level tokens may not themselves contain a directly identifiable front- or back-harmony vowel, reducing the coverage of the vocabulary-level harmony mapping.

Second, the auxiliary term is applied only at reference positions whose tokens can be assigned unambiguously to the front- or back-harmony classes. Neutral, mixed, and non-decisive tokens do not directly contribute to this term, whereas the standard cross-entropy objective provides supervision at all valid output positions. The harmony regularizer therefore represents a more selective training signal than the primary transcription loss.

Third, the relative strength of the auxiliary term is controlled by λ, which was fixed at 0.15 in the reported experiment. Because no systematic λ ablation was conducted, the present results do not establish whether a different weighting would produce a stronger or weaker effect. Taken together, these factors provide plausible explanations for the limited test-set change, but they should not be interpreted as evidence that vowel harmony cannot be learned by Whisper or that byte-level tokenization is the sole cause of the observed result.

5.4.3. Direct Comparison of Atomic and Byte-Fragmented Vowels

To examine more directly whether byte-level fragmentation is associated with poorer recognition of Kazakh harmony-bearing vowels, we conducted an additional analysis of the baseline test-set predictions. Reference and predicted transcriptions were aligned at the character level using minimum edit distance. For each harmony-bearing vowel in the reference transcription, we recorded whether the aligned output contained the same vowel. The vowels ү, ә, ө, and ұ were treated as byte-fragmented because they are represented by two tokenizer tokens, whereas і, а, о, and ы were treated as atomic according to the tokenizer analysis in Table 3.

Table 11 reports vowel-level recognition accuracy together with descriptive exact-word-match rates for words containing each vowel. The four byte-fragmented vowels achieved an aggregate vowel-level accuracy of 97.455%, compared with 97.313% for the four atomic vowels. The difference was therefore only +0.141 percentage points in favor of the fragmented group. An utterance-level bootstrap analysis produced a 95% confidence interval of −0.124 to +0.403 percentage points for the fragmented-minus-atomic difference, providing no evidence that byte-fragmented vowels are recognized less accurately in the present test set.

Table 11. Direct Comparison of Atomic and Byte-Fragmented Kazakh Vowels in Baseline ASR Output.

Vowel	BPE Representation	Ref. Occurrences	Vowel Accuracy	Words Containing Vowel	Exact Word Match
ү	Fragmented	4311	97.54%	4306	92.48%
ә	Fragmented	3951	96.08%	3925	91.39%
ө	Fragmented	4364	98.46%	4359	93.30%
ұ	Fragmented	4346	97.61%	4311	92.83%
і	Atomic	28,301	95.69%	20,538	91.70%
а	Atomic	59,103	98.66%	36,251	92.53%
о	Atomic	14,725	98.10%	13,736	93.04%
ы	Atomic	37,034	96.09%	27,373	91.24%

The per-vowel results lead to the same general conclusion. For example, the fragmented vowels ү and ө achieved vowel-level accuracies of 97.54% and 98.46%, respectively, whereas the atomic vowel і achieved 95.69%. These comparisons are descriptive and may reflect differences in lexical frequency, phonetic context, and word composition; they should therefore not be interpreted as a controlled tokenizer ablation.

Importantly, this analysis does not support a causal claim that byte-level fragmentation itself increases recognition errors. Its relevance to HarmonyConstrainedWhisper is narrower: fragmented byte tokens may reduce the coverage of the fixed front/back vocabulary masks used by the auxiliary regularizer because individual byte fragments do not necessarily expose the vowel class. We therefore distinguish between recognition accuracy, for which no fragmentation disadvantage is observed here, and accessibility to the explicit token-level harmony regularizer, for which non-atomic representation remains a potential implementation constraint.

5.4.4. Vowel-Class Confusion Analysis

To further examine whether the observed errors are related to the front/back distinction underlying Kazakh vowel harmony, we analyzed cross-class vowel substitutions in the baseline and HarmonyConstrainedWhisper outputs. Reference and predicted transcriptions were aligned at the character level using minimum edit distance. Substitutions in which

a front vowel was replaced by a back vowel, or vice versa, were counted as cross-class vowel confusions.

Table 12 reports the confusion counts and rates for each harmony-bearing vowel. Figure 3 visualizes these rates and shows that the auxiliary harmony regularizer does not reduce cross-class vowel substitutions on the test set. In most cases, the confusion rates remain similar or increase slightly in HarmonyConstrainedWhisper relative to the baseline. The largest confusion rates are observed for ə, i, and ы, whereas a and o show very low cross-class confusion rates in both systems.

Table 12. Cross-Class Vowel Confusions in Baseline and HarmonyConstrainedWhisper Outputs.

Reference Vowel	Baseline Cross-Class Errors	Baseline Rate (%)	HarmonyConstrained-Whisper Errors	HarmonyConstrained-Whisper Rate (%)
ү	13	0.302	15	0.348
ə	70	1.772	67	1.696
ө	14	0.321	17	0.390
i	398	1.406	434	1.534
ұ	8	0.184	11	0.253
a	58	0.098	58	0.098
o	14	0.095	16	0.109
ы	405	1.094	446	1.204

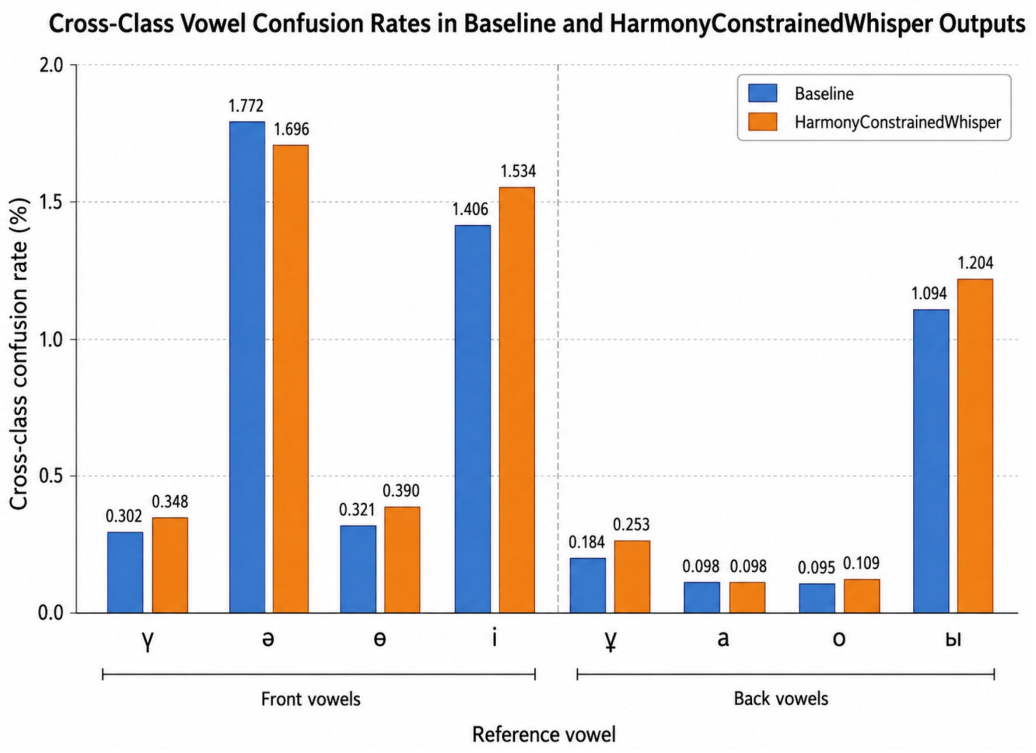


Figure 3. Cross-Class Vowel Confusion Rates in Baseline and HarmonyConstrainedWhisper Outputs. Cross-class substitution rates for the eight Kazakh harmony-bearing vowels in the baseline and HarmonyConstrainedWhisper test outputs. Front vowels are shown on the left and back vowels on the right. HarmonyConstrainedWhisper does not show a systematic reduction in front/back vowel confusion rates relative to the baseline.

These results suggest that the small test-set effect of the auxiliary harmony loss cannot be explained by a general reduction in front/back vowel confusions. Rather, they support

the interpretation that the training-time regularizer acts as a local token-class constraint and does not directly model the stem–suffix dependency that defines vowel harmony in Kazakh.

5.4.5. Statistical Significance of Test-Set Differences

To assess whether the small test-set differences observed between systems are statistically reliable, we performed paired utterance-level significance testing using the saved predictions for all 6733 test utterances. For WER, 95% confidence intervals for paired differences were estimated using 50,000 bootstrap resamples of utterances, and two-sided paired randomization tests were used to assess the null hypothesis of no system difference. The same paired resampling framework was applied to corpus-level HVR.

The results of the paired significance tests are summarized in Table 13.

Table 13. Results of Paired Significance Tests.

Comparison	Metric	Difference (pp)	95% CI (pp)	*p*-Value
HarmonyConstrainedWhisper—Baseline	WER	−0.021	[−0.176, +0.133]	0.801
HarmonyConstrainedWhisper—Baseline	HVR	+0.015	[−0.017, +0.047]	0.360
Re-ranking—N-best Baseline	WER	+0.015	[−0.002, +0.035]	0.154
Re-ranking—N-best Baseline	HVR	−0.039	[−0.055, −0.024]	<0.001

For HarmonyConstrainedWhisper, test WER decreased from 9.320% to 9.299%, corresponding to a difference of −0.021 percentage points. The 95% bootstrap confidence interval for this difference was −0.176 to +0.133 percentage points, and the paired randomization test yielded $p = 0.801$. Thus, the small WER difference is not statistically significant. HVR changed from 3.153% to 3.168%; this difference was also not statistically significant ($p = 0.360$).

For N-best re-ranking, test WER increased from 9.311% to 9.326%, a difference of +0.015 percentage points. The 95% bootstrap confidence interval was −0.002 to +0.035 percentage points, and the paired randomization test yielded $p = 0.154$, again providing no evidence of a statistically significant WER difference.

In contrast, the corpus-level HVR decreased from 3.150% to 3.111% after re-ranking. The paired bootstrap 95% confidence interval for the HVR difference was −0.055 to −0.024 percentage points, and the paired randomization test gave $p < 0.001$. Thus, although the reduction is small in absolute magnitude, it is statistically consistent across the test set. Importantly, this result should not be interpreted as an improvement in overall ASR accuracy, because WER does not improve and HVR remains a rule-based diagnostic measure.

5.5. Summary

HarmonyConstrainedWhisper demonstrates that a differentiable token-level harmony regularizer can be incorporated into the Whisper fine-tuning objective without modifying the underlying encoder–decoder architecture. Under the present experimental setup, however, the additional regularization produces only marginal changes in test WER, CER, and HVR. The results indicate that differentiability alone is not sufficient to ensure effective phonological-constraint integration. The practical contribution of the auxiliary signal also depends on the tokenizer representation, the coverage of harmony-related vocabulary masks, and the relative weighting of the harmony term. These findings motivate the complementary inference-time re-ranking experiment described in Section 6, where harmony is evaluated directly over fully decoded hypotheses.

6. Approach 2: Harmony Re-Ranking

6.1. Motivation

This second approach addresses the complementary inference-time selection problem: whether harmony information can improve the choice among hypotheses already generated by the ASR model without modifying the model parameters. A natural alternative to harmony-constrained fine-tuning is to apply the harmony constraint after decoding, where the full Unicode character sequence is available and the BPE fragmentation problem does not arise. N-best re-ranking offers a training-free mechanism to incorporate linguistic knowledge into ASR without modifying the underlying model, and has been widely used in ASR systems for language model rescoring and constrained hypothesis selection [10,25].

6.2. Method

6.2.1. N-Best Generation

For each utterance, beam search with a beam width of 10 was used to generate candidate sequences from the baseline Whisper large-v3 model, and the five highest-ranked complete hypotheses were retained for re-ranking. Each hypothesis h_i is associated with a sequence-level log-probability score s_i:

$$s_i = \frac{1}{|h_i|}\sum_{t=1}^{|h_i|} \log P(y_t|y_{<t}, x) \tag{4}$$

The hypotheses generally represent closely competing decoder outputs, but their overall diversity alone does not determine whether harmony-based re-ranking can be effective. More important is whether the candidate set contains alternatives that differ in harmony-relevant properties while remaining competitive according to the model score. This aspect is examined quantitatively in Section 6.4.1.

6.2.2. Harmony Penalty and Re-Ranking Score

For each decoded hypothesis h_i, we compute a sentence-level harmony penalty $P_{harm}(h_i)$. The hypothesis is first segmented into eligible content words, defined as words longer than one character with numeric tokens excluded. A word is assigned a violation indicator of 1 if it contains at least one decisive front vowel (ү, ә, ө, і) and at least one decisive back vowel (а, о, ұ, ы), and 0 otherwise. The sentence-level harmony penalty is calculated as the proportion of such mixed-vowel words among all eligible content words in the hypothesis.

This sentence-level penalty is used specifically for N-best hypothesis selection and differs in normalization from the corpus-level HVR defined in Section 4.3.3. The re-ranking score is:

$$s_i = \log P(h_i|x) - \alpha \cdot P_{harm}(h_i) \tag{5}$$

where $logP(h_i|x)$ denotes the sequence-level model score defined in Equation (6), $P_{harm}(h_i)$ is the sentence-level harmony penalty, and α controls the contribution of the harmony penalty to hypothesis selection.

The final transcription is the hypothesis with the highest re-ranking score:

$$\hat{h} = \underset{i}{\text{argmax}} s_i \tag{6}$$

The complete harmony-based N-best re-ranking procedure is summarized in Algorithm 2.

Algorithm 2: Harmony-Based N-best Re-ranking

1. Decode each utterance using beam search.
2. Retain the five highest-ranked complete hypotheses.
3. Convert each hypothesis to Unicode text.
4. Calculate the sentence-level harmony penalty $P_{harm}(h_i)$.
5. Combine it with the model score:
$$s_i = \log P(h_i|x) - \alpha \cdot P_{harm}(h_i)$$
6. Select the hypothesis with maximum s_i.

6.2.3. Alpha Selection

The penalty weight α is selected by grid search on the validation set over

$$\alpha \in \{0.0,\ 0.05,\ 0.1,\ 0.2,\ 0.3,\ 0.5,\ 0.8,\ 1.0,\ 1.5,\ 2.0\}$$

The choice of α reflects a trade-off: increasing its value strengthens the influence of phonological constraints but may override acoustically well-supported hypotheses.

6.3. Results

6.3.1. Alpha Search on Validation Set

The effect of the harmony penalty weight α on validation performance is shown in Table 14. The optimal value is $\alpha = 0.05$, yielding a marginal improvement in WER (-0.010 pp). However, larger values lead to rapid degradation in WER (up to $+0.425$ pp at $\alpha = 2.0$). This asymmetry suggests that harmony-based penalties are not well aligned with the probability distribution learned by the ASR model.

Table 14. Validation WER as a function of harmony penalty weight α.

α	Val WER (%)	Val CER (%)	ΔWER (Percentage Points)
0.00 (baseline)	9.298	3.210	–
0.05	9.288	3.209	−0.010
0.10	9.297	3.212	−0.001
0.20	9.323	3.221	+0.025
0.30	9.337	3.227	+0.039
0.50	9.423	3.281	+0.124
0.80	9.486	3.292	+0.187
1.00	9.533	3.311	+0.235
1.50	9.674	3.382	+0.376
2.00	9.723	3.395	+0.425

6.3.2. Test Set Results

The test set performance of the re-ranking approach is summarized in Table 15.

Table 15. Test set results for harmony re-ranking.

System	Test WER (%)	Test CER (%)	HVR (%)	ΔWER (Percentage Points)	ΔHVR (Percentage Points)
Baseline	9.31	3.25	3.150	–	–
Re-ranking ($\alpha = 0.05$)	9.33	3.25	3.111	+0.02	−0.039

The re-ranking experiment produces a small reduction in HVR but does not improve WER. This confirms that HVR and WER capture different aspects of the output. A hypothesis

may be phonologically more regular according to the rule-based HVR function while still being less accurate with respect to the reference transcription. For this reason, the HVR reduction should be interpreted as a diagnostic effect rather than as an improvement in overall ASR quality.

While re-ranking reduces HVR, the absence of improvement in WER indicates that phonologically consistent hypotheses are not consistently aligned with acoustically optimal predictions. This reveals a tension between linguistic constraints and model likelihood.

Re-ranking reduces HVR by 0.039 percentage points, indicating a small decrease in mixed front/back vowel patterns according to the evaluation metric. However, test WER increases marginally (+0.02 pp).

6.3.3. Per-Domain Results

Per-domain results are presented in Table 16.

Table 16. Per-domain test WER: baseline vs. harmony re-ranking.

Domain	*N*	Baseline WER (%)	Re-Ranked WER (%)	ΔWER (%)
Crowdsourced	3328	4.75	4.76	+0.01
Parliament	773	7.05	7.09	+0.04
Talkshow	383	12.69	12.71	+0.02
Radio	702	13.53	13.57	+0.04
Podcasts	1547	15.45	15.46	+0.01

The per-domain results show the same pattern across all domains: re-ranking changes WER only marginally, with small increases ranging from +0.01 to +0.04 percentage points. This consistency suggests that the harmony penalty does not provide a domain-specific WER advantage under the present decoding setup.

6.4. Analysis

6.4.1. Why Re-Ranking Does Not Improve WER

To examine why harmony-based re-ranking produces only a marginal change in WER, we performed an additional diagnostic analysis of the five-best hypotheses retained for all 6733 test utterances. The top-ranked hypotheses yield a WER of 9.311%, whereas an oracle that selects, for each utterance, the candidate with the lowest reference-conditioned word error achieves a WER of 5.833%. This substantial oracle gap shows that the N-best lists contain meaningful lexical alternatives and therefore cannot be characterized as globally lacking diversity. The main diagnostic statistics of the five-best hypothesis sets are summarized in Table 17.

The five-best lists contain an average of 4.90 unique hypotheses per utterance. Moreover, the exact reference transcription appears among the five candidates for 65.56% of the test utterances. However, diversity in lexical form does not necessarily imply diversity in the harmony property targeted by the re-ranker. Only 23.18% of the N-best lists contain candidates with different sentence-level harmony penalty values.

A more targeted analysis shows that 1755 of the 6733 top-ranked hypotheses (26.07%) contain at least one word identified as a harmony violation by the current rule-based function. Among these cases, only 337 (19.20%) contain another candidate with fewer violation words. Thus, for most top-ranked outputs containing a detected violation, the five-best list does not provide a candidate that directly improves the violation count.

Table 17. Diagnostic analysis of the five-best hypotheses on the test set.

Diagnostic	Result
Test utterances	6733
Top-1 WER	9.311%
Oracle WER@5	5.833%
Mean unique hypotheses	4.90/5
Exact reference present in 5-best	65.56%
Lists with different harmony penalty values	23.18%
Top-1 hypotheses with ≥1 detected violation	1755 (26.07%)
Of these, with a candidate containing fewer violations	337 (19.20%)
Re-ranking changed the top-1 hypothesis	29 (0.43%)
Changed cases: improved/unchanged/worsened WER	6/13/10

With the validation-selected value $\alpha = 0.05$, re-ranking changes the original top-ranked hypothesis in only 29 of 6733 test utterances (0.43%). Among these 29 changes, WER improves in 6 cases, remains unchanged in 13 cases, and worsens in 10 cases. This explains why the overall HVR can decrease slightly while corpus-level WER does not improve: the harmony penalty affects only a small fraction of outputs, and a lower harmony penalty is not consistently associated with a lower reference-conditioned word error.

6.4.2. Qualitative Analysis

Table 18 presents representative cases in which harmony-based re-ranking changes the top-ranked hypothesis. The examples illustrate that a lower harmony penalty does not necessarily correspond to a lower transcription error.

Table 18. Examples of harmony re-ranking corrections.

Reference	Baseline	Re-Ranked	WER Change
қазақша сөйлей алатын… қызықтап қарасын деп	қазір сөйлей алатын… қызықтап қарасың	қазақша сөйлей алатын… қызықтап қарасың	Improved
…бір ақ байқалды	…бірақ байқалдым	…бір ақ байқалдым	Improved
жаздыгүні ақша жинап…	жаздыгүні ақша жинап…	жаздың күні ақша жинап…	Degraded

These examples illustrate the main limitation of a rule-based harmony penalty: a reduction in the detected violation count does not necessarily correspond to a more accurate transcription. Valid lexical and morphological structures may contain vowel patterns that are penalized by the simplified HVR rule, while a candidate with a lower harmony penalty may still have a higher WER.

The third example is particularly informative: re-ranking replaces the reference form ‘жаздыгүні’ (“in summer”) with ‘жаздың күні’, resulting in additional word errors despite the lower harmony penalty assigned to the selected candidate. This case demonstrates that the present rule-based metric can prefer a phonologically simpler candidate even when it is less consistent with the reference transcription.

6.4.3. Summary Comparison

A summary comparison of both approaches is provided in Table 19.

Table 19. Summary comparison of both harmony integration approaches.

System	Val WER (%)	Test WER (%)	Test HVR (%)	Additional Harmony Component
Baseline—fine-tuning evaluation	9.62	9.32	3.153	None
HarmonyConstrainedWhisper	9.25	9.30	3.168	Auxiliary harmony loss
Baseline—N-best evaluation	9.298	9.31	3.150	None
Re-ranking ($\alpha = 0.05$)	9.288	9.33	3.111	Sentence-level harmony penalty

Table 19 summarizes the two harmony-integration strategies together with their corresponding evaluation baselines. HarmonyConstrainedWhisper reduces test WER from 9.32% to 9.30%, while HVR changes slightly from 3.153% to 3.168%. The N-best re-ranking experiment reduces HVR from 3.150% to 3.111%, but test WER increases from 9.31% to 9.33%. Thus, neither strategy produces a test-set improvement large enough to support a claim of improved recognition accuracy under the present experimental setup. Paired significance testing reported in Section 5.4.5 shows that neither strategy produces a statistically significant change in test WER. The HVR reduction after re-ranking is statistically significant ($p < 0.001$), although its absolute magnitude remains small and should be interpreted as a diagnostic effect rather than an improvement in ASR accuracy.

The main outcome of these experiments is therefore diagnostic rather than performance-oriented. The two strategies do not provide a measurable advantage in overall WER, but they identify three specific limitations of straightforward harmony integration: the training-time regularizer operates locally rather than over the stem–suffix relation, byte-level fragmentation does not itself explain vowel recognition errors, and N-best candidate sets contain substantially less harmony-relevant diversity than lexical diversity. These findings narrow the design space for future phonology-aware ASR methods and constitute the principal empirical contribution of the study.

6.5. Summary

Harmony re-ranking provides a computationally efficient mechanism for incorporating phonological information at inference time. Under the present setup, it produces a small reduction in HVR but does not improve test WER. The diagnostic analysis of the five-best hypotheses shows that the candidate sets contain substantial lexical variation, but harmony-relevant alternatives are available only for a subset of utterances, and the validation-selected penalty weight changes the original top-ranked hypothesis in relatively few cases. Moreover, a lower harmony penalty is not consistently associated with a lower reference-conditioned word error. These findings suggest that the limited effect of re-ranking is better explained by the restricted availability of useful harmony-sensitive alternatives and by the imperfect alignment between the current rule-based harmony penalty and transcription accuracy, rather than by a general lack of N-best diversity.

7. Discussion

7.1. Byte-Level BPE as a Representational Constraint

The fine-tuning and re-ranking experiments lead to a similar practical outcome: neither strategy produces a meaningful improvement in test WER, and the changes in HVR remain small. These results do not demonstrate that Whisper is unable to model Kazakh vowel harmony. Rather, they indicate that the straightforward integration strategies evaluated here have limited effect under the present representation and decoding setup.

Whisper's byte-level BPE tokenization is one factor that may influence explicit harmony-aware modeling. As shown in Section 2.3, several Kazakh harmony-bearing vowels, including ɣ, ə, ө, and ұ, are represented as multi-token byte sequences rather than as single atomic vocabulary entries. However, the direct analysis in Section 5.4.3 shows that this fragmentation is not associated with lower vowel recognition accuracy in the present baseline outputs. The aggregate recognition accuracy of the byte-fragmented vowels is comparable to that of the atomic vowels, and the bootstrap confidence interval does not indicate a systematic disadvantage for the fragmented group.

This result narrows the interpretation of tokenization in the present study. Byte-level fragmentation should not be treated as a direct cause of Kazakh vowel recognition errors. Its relevance is instead specific to explicit token-level constraint integration. The harmony regularizer used in HarmonyConstrainedWhisper relies on fixed front- and back-harmony vocabulary masks constructed from decoded tokenizer entries. Individual byte fragments that do not independently expose the identity or harmony class of a vowel cannot contribute to these masks in the same way as atomic vowel tokens.

The cross-class confusion analysis in Section 5.4.4 further supports this interpretation. HarmonyConstrainedWhisper does not systematically reduce front/back vowel substitutions relative to the baseline, indicating that the local token-class regularizer does not translate into a general reduction in harmony-relevant vowel confusions. These findings are consistent with the methodological limitation that the auxiliary objective operates at individual teacher-forced token positions rather than explicitly modeling the contextual relation between a stem and its suffix.

The situation is different for N-best re-ranking. Re-ranking is performed after complete hypotheses have been decoded into Unicode text, so the full vowel and word forms are available to the sentence-level harmony-penalty calculation. The diagnostic analysis in Section 6.4.1 also shows that the five-best lists contain substantial lexical variation. The limited effect of re-ranking is therefore better associated with the availability of useful harmony-sensitive alternatives and with the imperfect correspondence between the rule-based harmony penalty and reference-conditioned transcription accuracy, rather than with a general lack of candidate diversity.

Overall, the present results identify byte-level tokenization as a potential implementation constraint for explicit token-level harmony regularization, but not as a demonstrated source of recognition errors. The direct atomic-versus-fragmented comparison provides no evidence that fragmented vowels are inherently recognized less accurately. A controlled comparison with character-level, Kazakh-specific, or morphology-aware tokenization would still be required to determine whether alternative representations improve the effectiveness of explicit phonological-constraint modeling.

7.2. HVR as a Diagnostic Metric

The results support the use of HVR as a complementary diagnostic measure rather than as an independent indicator of ASR quality. HVR captures mixed front/back vowel patterns that are not distinguished by conventional WER or CER, and can therefore help identify a linguistically motivated subset of system outputs for further analysis. However, a detected HVR violation should not automatically be interpreted as a phonological error.

Sensitivity to mixed-vowel patterns. WER and CER measure transcription differences without distinguishing their linguistic type. HVR provides an additional view by identifying words that contain both front- and back-harmony vowels according to the rule defined in Section 4.3.3. This makes it useful for comparing changes in harmony-related output patterns across systems, provided that the results are interpreted together with lexical accuracy measures.

Reference-level interpretation. The non-zero HVR observed in the reference transcriptions shows that the current rule also identifies mixed-vowel patterns that occur in valid corpus text. These may include loanwords, proper names, compounds, abbreviations, and code-switched expressions. Reference HVR should therefore be treated as an empirical reference level rather than as a calibrated lower bound on phonological correctness. Likewise, a system HVR below the reference value does not by itself demonstrate either improved phonological modeling or over-correction.

Limitations. The current HVR implementation does not distinguish native Kazakh words from lexical items for which mixed front/back vowel patterns are legitimate. It also does not model morphological structure or determine whether a detected pattern represents an actual stem–suffix harmony violation. A future lexicon-aware or morphology-aware version of HVR could restrict the analysis to linguistically verified contexts and thereby provide a more specific diagnostic signal.

For these reasons, HVR should be reported and interpreted together with WER, CER, and, where possible, qualitative error analysis. Its main role in the present study is to characterize changes in harmony-related output patterns rather than to replace standard measures of transcription accuracy.

7.3. Per-Domain Analysis

The per-domain results in Table 16 show a consistent pattern across all five domains. Harmony-based re-ranking produces only small WER changes, ranging from +0.01 percentage points for the crowdsourced and podcast subsets to +0.04 percentage points for parliament and radio speech. None of the domains shows a WER improvement under the selected re-ranking setting.

The magnitude of the change is small relative to the substantial differences in baseline WER across domains. Baseline WER ranges from 4.75% for crowdsourced speech to 15.45% for podcasts, reflecting the different acoustic and linguistic characteristics of the test subsets. However, the present experiment does not provide sufficient evidence to attribute the small domain-specific re-ranking differences to any single factor such as loanword frequency, speaking style, or morphological complexity.

These results indicate that the limited effect of harmony-based re-ranking is not restricted to a single domain. Across all five subsets, the sentence-level harmony penalty does not provide a consistent lexical-accuracy advantage over the baseline decoder. A more detailed domain-level analysis of lexical composition, loanwords, compounds, and harmony-sensitive error types would be required to explain the observed differences more precisely.

7.4. Implications for Kazakh ASR

The present findings suggest three practical directions for future work on Kazakh ASR.

First, tokenization should be treated as an important design factor in explicit phonology-aware modeling, but the present results do not support interpreting byte-level fragmentation as a direct source of recognition errors. The direct comparison in Section 5.4.3 shows that byte-fragmented harmony-bearing vowels are recognized at rates comparable to atomic vowels in the baseline system. Nevertheless, fragmentation remains relevant to the implementation of the current training-time regularizer because non-atomic byte fragments may not expose front/back vowel identity to the fixed vocabulary masks. Controlled comparisons with character-level, Kazakh-specific, or morphology-aware tokenization are therefore needed to determine whether alternative representations improve explicit harmony integration rather than general vowel recognition.

Second, HVR can be useful as a complementary diagnostic measure when reported together with WER and CER. Rather than treating HVR as a replacement for standard ASR

metrics or as a universal benchmark criterion, future evaluations may use it to examine specific changes in mixed-vowel output patterns. Such analyses should account for the influence of loanwords, proper names, compounds, code-switching, and other legitimate mixed-vowel forms. Lexicon-aware or morphology-aware extensions of HVR would provide a more linguistically specific evaluation in future work.

Third, the limited improvements obtained with both training-time regularization and inference-time re-ranking provide a reproducible baseline for more advanced phonology-aware approaches. The experiments indicate that simply introducing an auxiliary harmony term or a sentence-level harmony penalty is not sufficient to produce reliable improvements in test WER. Future work should therefore investigate representation-aware training objectives, more linguistically precise constraint functions, and decoding methods that can exploit harmony-sensitive alternatives without overriding well-supported acoustic evidence.

7.5. Broader Applicability

The observations reported in this study may also be relevant to other Turkic languages with vowel harmony or related phonological regularities, including Turkish, Uzbek, Kyrgyz, Azerbaijani, Uyghur, Tatar, and Bashkir. In multilingual models, similar representational issues may arise when phonologically important characters or morphological units are not represented as atomic vocabulary entries. However, the tokenizer analysis conducted in the present study is specific to Kazakh, and comparable analyses would be required before extending these conclusions to other languages.

More generally, the N-best re-ranking framework used in Section 6 can be adapted to linguistic properties that can be evaluated over fully decoded hypotheses. In principle, language-specific penalty functions could be developed for other phonological phenomena, provided that the relevant linguistic property can be defined reliably and its interaction with lexical accuracy is evaluated separately.

Taken together, the findings suggest that phonology-aware ASR depends on the interaction between linguistic structure, model representation, training objectives, decoding, and evaluation. The present Kazakh results therefore provide a case study rather than direct evidence for equivalent behavior across other languages.

8. Conclusions

Kazakh vowel harmony provides a useful case for examining how language-specific phonological information can be incorporated into a large multilingual ASR model. In this study, we evaluated two complementary strategies using Whisper large-v3 and the ISSAI KSC2 corpus: differentiable harmony-aware regularization during fine-tuning and harmony-penalty-based N-best re-ranking during inference.

HarmonyConstrainedWhisper reduced validation WER from 9.62% to 9.25%, whereas the corresponding test-set change was small, from 9.32% to 9.30%. Paired significance testing showed that this WER difference was not statistically significant. Test HVR also remained essentially unchanged, and the additional cross-class vowel confusion analysis showed no systematic reduction in front/back vowel substitutions. These findings indicate that the differentiable auxiliary regularizer can influence model training, but under the present configuration it does not provide evidence of improved test-set recognition accuracy or more consistent harmony-related vowel recognition.

The inference-time re-ranking experiment produced a small but statistically significant reduction in HVR while slightly increasing test WER; the WER difference itself was not statistically significant. Additional analysis of the five-best hypotheses showed that the candidate lists contain substantial lexical variation, including a considerable oracle WER gap, but useful harmony-sensitive alternatives are available only for a subset of utterances.

With the validation-selected penalty weight, re-ranking changed the original top-ranked hypothesis in only a small fraction of test cases, and a lower harmony penalty was not consistently associated with lower word error.

The tokenizer analysis provides an additional perspective on these results. Several Kazakh harmony-bearing vowels are represented by Whisper as multi-token byte sequences rather than as atomic vocabulary entries. However, the direct test-set comparison of atomic and byte-fragmented vowels provides no evidence that fragmented vowels are recognized less accurately in the baseline system. Byte-level fragmentation should therefore not be interpreted as a demonstrated cause of recognition errors. Its relevance in the present study is narrower: non-atomic byte fragments may reduce the coverage of the fixed front/back vocabulary masks used by the training-time regularizer, making explicit token-level harmony integration more difficult.

The study also introduces HVR as a complementary diagnostic measure for mixed front/back vowel patterns in ASR output. HVR should not be interpreted as a standalone measure of phonological correctness, because valid loanwords, proper names, compounds, abbreviations, and code-switched expressions may also contain mixed-vowel patterns. Its main value is therefore diagnostic when used together with WER, CER, and qualitative error analysis.

Future work should evaluate character-level, Kazakh-specific, and morphology-aware tokenization under controlled experimental conditions; investigate alternative formulations and weights for differentiable harmony-aware objectives; and develop lexicon- or morphology-aware variants of HVR. More advanced decoding strategies could also examine whether linguistically useful alternatives can be generated more consistently before re-ranking is applied.

Overall, the findings do not indicate that Whisper is unable to represent Kazakh vowel harmony. Rather, they show that the straightforward training-time regularization and inference-time re-ranking strategies evaluated here are insufficient to produce reliable improvements in test WER. Effective phonology-aware ASR is likely to require closer alignment between linguistic structure, model representation, training objectives, candidate generation, and evaluation.

Author Contributions: Conceptualization, methodology, project administration, and supervision, M.T. and A.Y.; software and investigation, A.B., D.O., B.S. and T.T.; data curation, B.S., B.A., T.T. and N.B.; writing—original draft preparation, M.T., A.Y., B.A., N.B. and E.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24992875).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The ISSAI Kazakh Speech Corpus 2 (KSC2) is available at https://doi.org/10.21437/Interspeech.2022-421 (accessed on 22 March 2026).

Acknowledgments: The authors would like to thank colleagues and collaborators for their valuable feedback and support throughout this research. We also acknowledge the open research community for providing publicly available datasets and tools, in particular the ISSAI Kazakh Speech Corpus 2 (KSC2), which enabled reproducible experiments in Kazakh automatic speech recognition [2]. During the preparation of this manuscript, the authors used ChatGPT (OpenAI, GPT-based language model, version GPT-5.3) for language editing and improving the clarity of the text. The authors have carefully reviewed and edited the generated content and take full responsibility for the final version of this publication.

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

Abbreviations

The following abbreviations are used in this manuscript:

ASR	Automatic Speech Recognition
BPE	Byte-Pair Encoding
WER	Word Error Rate
CER	Character Error Rate
HVR	Harmony Violation Rate
KSC2	Kazakh Speech Corpus 2
GPU	Graphics Processing Unit
CTC	Connectionist Temporal Classification
HMM	Hidden Markov Model
DNN	Deep Neural Network
MFCC	Mel-Frequency Cepstral Coefficients
NLP	Natural Language Processing

References

1. 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 ICML 2023, Honolulu, HI, USA, 23–29 July 2023.
2. Mussakhojayeva, S.; Khassanov, Y.; Atakan Varol, H. KSC2: An industrial-scale open-source Kazakh speech corpus. In Proceedings of the Interspeech 2022, Incheon, Republic of Korea, 18–21 September 2022. [CrossRef]
3. Mamyrbayev, O.; Turdalyuly, M.; Mekebayev, N.; Alimhan, K.; Kydyrbekova, A.; Turdalykyzy, T. Automatic recognition of Kazakh speech using deep neural networks. In Proceedings of the Asian Conference on Intelligent Information and Database Systems, Yogyakarta, Indonesia, 8–19 April 2019. [CrossRef]
4. Mamyrbayev, O.; Turdalyuly, M.; Mekebayev, N.; Mukshina, K.; Keylan, A.; BabaAli, B.; Nabieva, G.; Duisenbayeava, A.; Akhmetov, B. Continuous speech recognition of Kazakh language. In Proceedings of the ITM Web of Conferences, Rome, Italy, 23—-25 November 2018. [CrossRef]
5. BabaAli, B. Speech Recognizer-Based Non-Uniform Spectral Compression for Robust MFCC Feature Extraction. *PRZEGLĄD ELEKTROTECHNICZNY* **2018**, *94*, 90–93. [CrossRef]
6. Orken, M.; Dina, O.; Keylan, A.; Tolganay, T.; Mohamed, O. A study of transformer-based end-to-end speech recognition system for Kazakh language. *Sci. Rep.* **2022**, *12*, 8337. [CrossRef] [PubMed]
7. Bekarystankyzy, A.; Mamyrbayev, O.; Mendes, M.; Fazylzhanova, A.; Assam, M. Multilingual end-to-end ASR for low-resource Turkic languages with common alphabets. *Sci. Rep.* **2024**, *14*, 13835. [CrossRef] [PubMed]
8. Oralbekova, D.; Mamyrbayev, O.; Yerimbetova, A.; Bekarystankyzy, A.; Turdalyuly, M. Application of Non-Autoregressive Decoding for Kazakh Speech Recognition. *Her. Kazakh-Br. Tech. Univ.* **2025**, *22*, 23–30. [CrossRef]
9. Kozhirbayev, Z. Kazakh Speech Recognition: Wav2vec2.0 vs. Whisper. *J. Adv. Inf. Technol.* **2023**, *14*, 1382–1389. [CrossRef]
10. Li, J.; Pu, Y.; Sun, Q.; Zhang, W.-Q. Improving Whisper's recognition performance for under-represented language Kazakh leveraging unpaired speech and text. *arXiv* **2024**, arXiv:2408.05554.
11. Karibayeva, A.; Karyukin, V.; Rakhimova, D.; Amirova, D.; Aliyev, R.; Karybayeva, A. Collecting, processing, and evaluating the performance of Kazakh automatic speech recognition. In Proceedings of the ACIIDS 2025, Kitakyushu, Japan, 23–25 April 2025. [CrossRef]
12. Du, W.; Maimaitiyiming, Y.; Nijat, M.; Li, L.; Hamdulla, A.; Wang, D. Automatic Speech Recognition for Uyghur, Kazakh, and Kyrgyz: An Overview. *Appl. Sci.* **2023**, *13*, 326. [CrossRef]
13. Karabaliyev, Y.; Kolesnikova, K. Kazakh speech and recognition methods: Error analysis and improvement prospects. *Sci. J. Astana IT Univ.* **2024**, *20*, 62–75. [CrossRef]
14. Karibayeva, A.; Karyukin, V.; Abduali, B.; Amirova, D. Speech Recognition and Synthesis Models and Platforms for the Kazakh Language. *Information* **2025**, *16*, 879. [CrossRef]
15. Liao, D.; Shi, F. How Tokenization Limits Phonological Knowledge Representation in Language Models and How to Improve Them. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics, San Diego, CA, USA, 2–7 July 2026.
16. McCollum, A.G. Gradience and locality in Kazakh vowel harmony. *Phonology* **2019**, *36*, 287–345. [CrossRef]
17. McCollum, A.G.; Chen, S. Kazakh. *J. Int. Phon. Assoc.* **2021**, *51*, 276–298. [CrossRef]

18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *NeurIPS* **2017**, *30*, 6000-6010.
19. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. In Proceedings of the ACL 2016, Berlin, Germany, 7–12 August 2016. [CrossRef]
20. Toraman, C.; Yilmaz, E.H.; Şahïnuç, F.; Ozcelik, O. Impact of Tokenization on Language Models: An Analysis for Turkish. *ACM Trans. Asian Low.-Resour. Lang. Inf. Process.* **2023**, *22*, 1–21. [CrossRef]
21. Kaya, Y.B.; Tantuğ, A.C. Effect of tokenization granularity for Turkish large language models. *Intell. Syst. Appl.* **2024**, *21*, 200335. [CrossRef]
22. Tukeyev, U.; Rysbek, B. Morphology-Aware Segmentation and Tokenization for Turkic Languages: A CSE-Guided Framework (The Kazakh Case). *Information* **2026**, *17*, 128. [CrossRef]
23. Slam, W.; Li, Y.; Urouvas, N. Frontier Research on Low-Resource Speech Recognition Technology. *Sensors* **2023**, *23*, 9096. [CrossRef] [PubMed]
24. Mukhamadiyev, A.; Mukhiddinov, M.; Khujayarov, I.; Ochilov, M.; Cho, J. Development of Language Models for Continuous Uzbek Speech Recognition System. *Sensors* **2023**, *23*, 1145. [CrossRef] [PubMed]
25. Yeshpanov, R.; Mussakhojayeva, S.; Varol, H.A. Using songs to improve Kazakh automatic speech recognition. *arXiv* **2026**, arXiv:2603.00961.
26. Rust, P.; Pfeiffer, J.; Vulić, I.; Ruder, S. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, Virtual Event, 1–6 August 2021.
27. Xue, L.; Barua, A.; Constant, N.; Al-Rfou, R.; Narang, S.; Kale, M.; Roberts, A.; Raffel, C. ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 291–306. [CrossRef]

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.