university of groningen

campus fryslân

Improving Uyghur Speech Synthesis with Monolingual Transfer Learning: A Comparison of English and Russian Pretraining

Oufeire Aishan

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Improving Uyghur Speech Synthesis with Monolingual Transfer Learning: A Comparison of English and Russian Pretraining

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr. Phat Do (Voice Technology, University of Groningen)
with the second reader being
Supervisor 2's title and name (Voice Technology, University of Groningen)

Oufeire Aishan (s5973902)

June 11, 2025

Acknowledgements

I would like to sincerely thank Dr. Do from Campus Frisian at the University of Groningen for supervising my research, offering guidance, and providing support when technical challenges arose..

I acknowledge the Center for Information Technology of the University of Groningen for their technical support and for providing access to the Hábrók high-performance computing cluster.

Lastly, I sincerely thank all native Uyghur speakers involved in the experiments. Their support and participation were essential to this research.

Abstract

In recent years, speech synthesis technology has made significant progress in mainstream languages. However, for low-resource languages such as Uyghur, the development and implementation of speech synthesis systems are significantly challenged by a lack of essential resources, including speech corpora and pronunciation dictionaries, which makes effective modeling particularly difficult. Transfer learning is regarded as an effective way to alleviate this problem, especially in text-to-speech (TTS) systems. Transfer learning can improve the synthesis quality of low-resource languages by leveraging knowledge from high-resource languages.

This study focuses on a core issue: whether the choice of the source language, especially its similarity to the target language in terms of language structure, significantly affects the effectiveness of transfer learning in speech synthesis. This issue not only concerns the development path of low-resource language technologies, but also has universal guiding significance for the theoretical framework of cross-language speech modeling.

Therefore, it represents a research area of high significance both in theoretical and applied contexts. This study conducted transfer experiments using English and Russian as source languages under the FastSpeech 2 architecture, with Uyghur as the target language. After fine-tuning and subjective evaluation, the results showed that transfer learning significantly improved the naturalness and intelligibility of the synthesized speech. Moreover, the model pretrained on Russian□whose linguistic structure is more similar to Uyghur achieved better performance. These findings highlight the crucial role of typological similarity in cross-lingual TTS transfer, offering both a theoretical foundation and practical guidance for the development of low-resource speech synthesis systems.

Keywords: Low-resource TTS; Transfer learning; Linguistic similarity; Uyghur; Cross-lingual speech synthesis

Contents

1 Introduction

In recent years, neural network-based text-to-speech (TTS) synthesis technology has made significant advancements, particularly with the development of end-to-end architectures. These advancements have led to substantial improvements in the naturalness of speech and the efficiency of modeling (Tan, Qin, Soong, & Liu, 2021). Current mainstream TTS systems use deep learning methods to map phonemes or character sequences to the Mel spectrum and combine them with neural vocoders to generate high-quality speech waveforms. This modeling process effectively integrates multiple subsystems within speech synthesis, reduces the reliance on staged processing, and improves the overall performance of the system. However, such systems generally depend on substantial, well-aligned speech and text data, which poses a substantial barrier to languages with limited resources. Low-resource languages (LRLs) pose a significant challenge to model training due to the absence of sufficient training data, resulting in inaccurate pronunciation, rhythm, and rhyme patterns. This, in turn, affects the naturalness and intelligibility of synthesized speech.

The Uyghur language belongs to the Turkic branch of the Altaic language family and is typologically classified as an agglutinative language.(Yibulayin & Baoshe, 2011). Uyghur is a natural and highly flexible language. Human understanding of any language depends on a large amount of knowledge accumulated over time. However, a computer is just a machine made of electronic components. In order to achieve speech synthesis, a computer must also acquire a great deal of human knowledge. Therefore, we classify different types of human knowledge and propose a multi-channel strategy for transferring this knowledge to computers. These strategies help computers quickly learn the knowledge needed for speech synthesis and support the development of high-intelligence speech synthesis technology(Muhetar, 2012). In the current field of speech synthesis technology, high-resource languages such as Mandarin and English have achieved impressive synthesis quality due to the abundance of available training data. However, for low-resource languages like Uyghur, the lack of sufficient linguistic resources often leads to suboptimal synthesis performance.

The development of Uyghur speech technology is of considerable linguistic significance and also holds significant sociocultural and economic value. The advancement of Uyghur speech recognition and synthesis technologies has the potential to enhance Uyghur speakers' access to digital tools, promote educational development, and strengthen support for native language applications in public services. Moreover, this technological progress is expected to contribute to economic growth in the Xinjiang region and foster greater understanding of Uyghur culture among other ethnic groups. However, due to the scarcity of annotated speech corpora, Uyghur speech technology still faces significant challenges, making it imperative to explore effective modeling approaches and data utilization strategies under low-resource conditions. This study aims to offer new insights and methodologies for low-resource speech synthesis, thereby contributing to innovation in Uyghur speech technology.

Given the underdeveloped state of Uyghur speech technology, Methods for transferring existing TTS models to low-resource languages are still faces significant challenges on modeling caused by structural linguistic differences. There are significant differences between Uyghur and common high-resource languages (e.g., Chinese or English) in terms of vowel distribution, accent placement and rhythmic control, and these differences may interfere with the pre-trained model's ability to

learn and transfer rhythmic and metrical features, leading to unnatural synthesized speech with deviations in intonation and rhythm. Especially in end-to-end cross-language TTS systems, where models frequently depend on rhythm patterns and duration distributions learned from the source language. Without effective adaptation, these structural mismatches can be a key factor in speech synthesis failure.

Existing research shows that fine-tuning-based transfer learning is one of the most widely adopted strategies(Li et al., 2023), this approach typically involves large-scale pre-training on high-resource languages first, and then fine-tuning with a small amount of target language data. Although this method shows good transfer performance in certain languages and reduces dependency on target language data, its effectiveness is heavily dependent on the phonological similarity between the source and target languages. This similarity is especially crucial in terms of syllable structure, stress assignment, and speech rhythm compatibility. When there are significant differences in the phonetic structure, the model may struggle to reproduce natural rhythm and intonation features of the target language. Moreover, current Text-to-speech systems (TTS) still show clear limitations in modeling prosodic features for low-resource languages, which limits their ability to generalize across multilingual synthesis tasks. Therefore, this study seeks to systematically compare different source languages as transfer bases in order to investigate the underlying interaction mechanisms among language similarity, structural diversity, and transfer learning strategies.

In order to address the structural challenges in Uyghur speech synthesis and the limitations of existing transfer learning strategies, we select FastSpeech 2 [1]as the core modeling framework. FastSpeech 2 is a non-autoregressive TTS model that offers fast inference, robust training stability, and explicit prosody modeling through controllable features such as pitch, energy, and duration. These features make it particularly suitable for analyzing fine-tuning performance under structural mismatches between source and target languages(Ren et al., 2020) .

This study involved a series of transfer experiments, using English and Russian as high-resource languages to investigate their impact on transfer effectiveness under low-resource conditions. By fixing training configurations and fine-tuning protocols and by evaluating output speech in terms of naturalness and intelligibility, the study aims to discover how phonological similarity and structural divergence affect the effectiveness of transfer learning. This research is expected to offer methodological insights into language transfer for building multilingual TTS systems and enhancing the accessibility of low-resource languages such as Uyghur in speech technology.

1.1 Research Questions and Hypotheses

This study explores the influence of source language selection in monolingual transfer learning on the quality of synthesized speech for a low-resource language. It focuses on how pretraining on either English or Russian, followed by fine-tuning on Uyghur, influences the naturalness and intelligibility of synthesized speech.

The research question guiding this study is as follows:

> **To what extent does the selection of source language in monolingual transfer learning influence the naturalness and intelligibility of synthesized Uyghur speech?**

[1]https://github.com/ming024/FastSpeech2

This main question can be broken down into the following sub-questions:

Sub-question 1 Does the source language significantly influence the naturalness of Uyghur synthesized speech? Specifically, does Russian, which shares greater similarities with Uyghur, produce output that is more natural than English?

Sub-question 2 Can the choice of source language influence the intelligibility of synthesized speech?

Hypotheses:
H1: Fine-tuning on Uyghur after pretraining on a high-resource language is expected to produce synthesized speech with higher naturalness and intelligibility than training on Uyghur from scratch.

H2: Due to the phonetic similarity between Russian and Uyghur, and the greater number of loanwords from Russian compared to English, transfer learning using Russian as the source language is expected to produce synthesized speech with higher naturalness and intelligibility than English-based transfer.

2 Literature Review

This chapter provides a review of representative studies in the fields of speech synthesis for low-resource languages and transfer learning. It highlights two key factors affecting modeling outcomes: source language selection and structural similarity between languages.

In particular, it examines whether monolingual transfer strategies using English or Russian as the source language lead to measurable differences in the naturalness and intelligibility of synthesized Uyghur speech under extremely low-resource conditions. By examining extant research in terms of methodological design, linguistic feature considerations, and model adaptation issues, this chapter identifies the theoretical basis and research focus of the present study.

2.1 speech synthesis in low-resource languages

Speech plays an important role in daily life. It's one of the most natural and intuitive ways for humans to get information and communicate with each other. The prospect of machines communicating with humans in natural spoken language, exhibiting emotion, and delivering information would bring us closer to the long-standing goal of natural human-computer interaction. The idea of speech synthesis was introduced as early as the 1950s. At first, it relied on mechanical and electronic devices to generate simple speech sounds. As computing power has increased and language modeling and artificial intelligence have evolved, speech synthesis has evolved from early rule-based methods to statistical parametric approaches and, more recently, to end-to-end neural models. These developments have led to substantial advancements in the naturalness, fluency, and adaptability of synthesized speech. Speech synthesis has become a widely utilized technology in various domains, including navigation, virtual assistants, and assistive technologies. This field of research continues to advance rapidly.

However, making speech synthesis available in different languages is still challenging because speech resources are not evenly available. There are more than 7,000 living languages in the world. Some of them are endangered, with around 1,500 at risk of disappearing. Languages with large speaker populations such as English, spoken by approximately 1.5 billion people, Mandarin Chinese with 1.1 to 1.2 billion speakers, and Hindi with about 600 million speakers, generally have rich speech corpora and pronunciation dictionaries, which makes it easier to develop effective synthesis systems. By contrast, many low-resource languages encounter limited data availability due to challenges in data collection, inadequate standardization, and the absence of annotations. These limitations continue to present significant obstacles to the development of speech synthesis systems for low-resource languages.

In low-resource language text-to-speech (TTS) modeling, current studies face three main technical challenges. First, speech data is extremely limited. Many low-resource languages have only a few hours or even minutes of recordings. This amount of data is not enough to train deep learning models effectively. Secondly, a considerable number of languages with limited resources do not possess a standard phoneme inventory or a pronunciation lexicon. These languages are distinguished by their possession of unique linguistic features, including complex phonotactics, tone systems, and morphophonemic changes. These factors hinder TTS model training and reduce their ability to generalize across languages, making it difficult to learn accurate text-to-speech mappings. Thirdly, the quality of synthesized speech is often substandard. Common issues include incorrect stress, unnatural rhythm, and unclear articulation. These problems reduce the naturalness and intelligibility of

the synthesized speech. Due to these problems, standard training methods that require large labeled datasets are not effective. We need methods that can leverage external knowledge to improve TTS systems for low-resource languages.

2.2 Applications of Transfer Learning in TTS

As introduced in Section 2.1, low-resource text-to-speech (TTS) systems often face problems such as limited training data, complex phonological patterns, and challenges in model generalization. Transfer learning, by adapting a model pretrained on a high-resource language to a low-resource one, offers an effective solution. Even when the target language has limited data, transfer learning can significantly improve the naturalness and intelligibility of synthesized speech. For example, a multi-stage transfer approach has been shown to gradually increase model complexity while generating high-quality speech outputs under data-scarce conditions (Azizah, Adriani, & Jatmiko, 2020) .

Besides monolingual transfer, multilingual modeling is also common in TTS research for low-resource languages. Instead of using only one source language, this approach trains the model on speech data from multiple languages. It enables the model to learn shared acoustic and linguistic features, such as phoneme inventories and prosodic structures. Experimental results show that multilingual pretraining often improves speech quality, and that even when the target language is not included in training, the model can still generalize well and produce intelligible outputs (Amalas, Ghogho, Chetouani, & Thami, 2024) . Transfer learning has also shown strong effectiveness under extremely low-resource conditions. One study demonstrated that with only about 30 minutes of speech from the target language, a multispeaker TTS model could still be trained by combining cross-lingual pretraining and basic data augmentation techniques. The synthesized speech remained natural and intelligible (Byambadorj, Nishimura, Ayush, Ohta, & Kitaoka, 2021) . Recent studies have also highlighted the importance of linguistic similarity between source and target languages. Specifically, when phoneme labels are used as input, angular similarity of phone frequencies (ASPF) has been shown to be a more effective predictor of transfer performance than other metrics. This suggests that the phonological and acoustic distance between languages should be considered when selecting a source language for cross-lingual TTS adaptation (Do, Coler, Dijkstra, & Klabbers, 2023) .

These studies show that transfer methods and the choice of source language are very important in low-resource TTS systems. The similarity between the source and target languages in phonological structure, pronunciation rules, and language type often affects how well the transfer works.

2.3 Phonetic Characteristics of the Uyghur Language

The Uyghur language is mainly spoken by the Uyghur ethnic group. It belongs to the Karluk branch of the Turkic language family. The Turkic language family is part of the Altaic language group. It shares similarities in pronunciation, vocabulary, and grammar with many other minority languages spoken in Xinjiang, China. Over time and in different regions, Uyghur has developed several writing systems. In China, the official writing system uses the Arabic alphabet, while a Latin-based script is also used. In Russia and other diaspora communities, a Cyrillic-based script is commonly used. (Yang, 2021) As shown in Table 1, Uyghur has eight basic vowels. These vowels are distinguished by variations in tongue position, height, backness, and lip rounding. The front rounded

vowels in Uyghur are notable; such vowels are uncommon in many other languages. The Uyghur vowel system exhibits clear phonemic contrasts, such as front vs. back and rounded vs. unrounded vowels. Uyghur has 25 phonemes for consonants. These consonants include plosives, fricatives, nasals, approximants, and trills. Uyghur also has uvular and pharyngeal sounds. These places of articulation are relatively uncommon cross-linguistically and contribute to the typological distinctiveness of Uyghur.

Table 1: IPA Symbols by Category

Category	**IPA Symbols**
Vowels	/i/, /y/, /e/, /u/, /ɣ/, /o/, /ɑ/, /ø/
Plosives	/p/, /b/, /t/, /d/, /k/, /g/, /q/, /ɖ/
Fricatives	/s/, /z/, /ʃ/, /ʒ/, /x/, /ʁ/, /h/
Nasals	/m/, /n/, /ɳ/
Affricates & Trills	/tʃ/, /dʒ/, /ʒ/,/ɾ/, /ʎ/

Within the vowel system, Uyghur vowels function relatively independently at two levels. At the syllable level, constrained by syllable structure rules, vowels act as individual units and combine with consonants to form syllables. At the lexical level, vowels form groups governed by vowel harmony, with vowel sets serving as the functional units. These sets create harmonic structures according to vowel harmony rules. Uyghur vowel sets are hierarchically organized. Sets with similar combinatory properties further group into vowel clusters based on shared functional behavior, forming three major categories: tongue position harmony sets, labial harmony sets, and neutral harmony sets. These sets function at different hierarchical levels and correspond to distinct types of vowel harmony: tongue position harmony, labial harmony, and neutral vowel harmony. These three types of vowel sets represent the three subsystems of the Uyghur vowel system. Vowel sets at different levels have distinct structural patterns, but they are also interrelated, together forming the overall structure of the Uyghur vowel system(Bin, 2006).

2.4 Linguistic Relationships Between Russian, English, and Uyghur

Uyghur and Russian show many similarities in their phonetic systems, especially when compared to English. Both languages include front rounded vowels like /y/ and /ø/, or similar sounds such as /ʉ/ and /ʏ/. English does not have these vowels, which makes the vowel systems of Uyghur and Russian more alike. They also mainly use monophthongs, meaning each vowel tends to have a stable, unchanging quality. In contrast, English often uses diphthongs like /aɪ/ and /eɪ/, which change during pronunciation and create more variation. Uyghur and Russian also share many fricative phonemes such as /ʃ/, /ʒ/, and /x/. These sounds are common in both languages, while English has fewer of them. In addition, both Uyghur and Russian include uvular and velar consonants like /x/, /q/, and /ɣ/, which do not exist in English. Another important difference is aspiration. Uyghur

and Russian stops like /p/, /t/, and /k/ are usually unaspirated, meaning there is little or no burst of air. In English, these stops are aspirated, so they are typically aspirated, often transcribed as, $/\mathrm{p}^{\mathrm{h}}/$ and $/\mathrm{t}^{\mathrm{h}}/$.

Due to these shared phonetic characteristics, Uyghur and Russian have more comparable sound systems. This makes Russian a more suitable source language than English for transfer learning in Uyghur speech synthesis. It helps the model learn better phoneme-level features, including vowel quality, consonant types, and aspiration patterns.

Beyond phonetic similarities, historical and sociolinguistic factors have also shaped the influence of Russian and English on Uyghur. When the Uyghur vocabulary can't express certain scientific or technical concepts, these foreign terms are often borrowed directly. Due to its unique geographical location, the Ili region of China became the first area where Russian lifestyles and culture were introduced, and then gradually spread to other parts of Xinjiang. Additionally, frequent trade and cultural exchanges along the Silk Road also introduced Russian and English loanwords into Uyghur. These loanwords can be classified into three categories: pure transliterations, semantic adaptations of transliterations, and hybrid constructions combining foreign and Uyghur elements. A statistical analysis of the Uyghur Explanatory Dictionary shows that there are 31 English loanwords and 680 Russian loanwords included in the corpus(Dai, Dilhumar, & Ma, 2024). Since most of the English loanwords were introduced to Xinjiang through Russian, their pronunciation was greatly influenced by Russian in the Uyghur language (Zheng, 2009).

2.5 The Impact of Structural Similarity on Transfer Effectiveness

In cross-lingual speech synthesis, the structural similarity between the source and target languages has been widely recognized as one of the key factors influencing the success of transfer learning. In comparison to model architecture or training size, the similarities in syllable structure, phoneme combinations, stress placement, and prosodic patterns play a more direct role in determining how well a pre-trained model adapts to a new language and how effectively it generalizes.

In recent years, several studies have proposed quantitative methods for measuring language similarity to guide source language selection and transfer pathway design. (Wu, Shi, Zhong, Watanabe, & Black, 2021) introduced a method called Acoustic Language Similarity, which evaluates the acoustic distance between languages by constructing a multilingual feature space. Their findings in cross-lingual TTS and ASR tasks demonstrated that languages exhibiting higher acoustic similarity exhibited significantly superior transfer performance, particularly in low-resource settings, by diminishing the necessity for extensive fine-tuning data .

A methodology rooted in Angular Similarity of Phoneme Frequencies (ASPF) has been advanced to assess structural similarity between source and target languages, transcending the limitations of conventional genealogical classifications. In experiments involving the transfer of TTS models from Dutch and English to Frisian, it was determined that higher phoneme-level similarity leads to more natural and intelligible synthesized speech. It is noteworthy that this approach demonstrated superior performance in comparison to strategies that relied on larger, but structurally incompatible, corpora (Do, Coler, Dijkstra, & Klabbers, 2022).

The findings of these studies collectively indicate that structural compatibility frequently holds more significance than the magnitude of training resources in determining the efficacy of transfer. However, in practical TTS applications, English remains the default choice for the source language, and the structural match between languages is often overlooked, particularly in the case of less commonly studied languages such as Uyghur. For example, Uyghur belongs to the Altaic language family, whereas Russian belongs to the Slavic branch of the Indo-European language family. Though these languages are not related, they have some similarities. Both languages share certain phonemic features, and Uyghur has many loanwords from Russian. These similarities may improve the results of transfer learning in speech synthesis. However, current studies on low-resource TTS often use English as the source language without consider the similarity between the source and target languages. This may result in suboptimal model performance. When the differences between the source and target languages are big, the synthesized speech may sound unnatural or be hard to understand, and rhythm mismatch or incorrect prosody may also appear.

Therefore, selecting a source language based on structural similarity can improve the efficiency of transfer learning and the quality of synthesized speech. This approach can also make the model more adaptable to different language settings. For languages like Uyghur, which have special prosodic systems, these similarities should not be ignoredand and should be integrated into transfer learning strategy design.

2.6 Research Gaps and Contributions of This Study

Although recent years have seen steady progress in cross-lingual TTS systems in terms of model design and transfer learning strategies, key challenges remain in modeling low-resource languages, especially in the case of structurally complex agglutinative languages like Uyghur. Existing research has shown that transfer learning can help alleviate data scarcity in low-resource settings. However, systematic theoretical frameworks and empirical guidelines for source language selection remain lacking.

Current research on Uyghur TTS mainly focuses on corpus construction and phonological system analysis. However, there are relatively few targeted experiments evaluating the impact of different source languages on transfer learning performance. Although some studies have attempted to transfer from high-resource languages to Uyghur, most default to using English as the source language and do not provide comparative evaluations with other high-resource options. Given Uyghur's unique phonemic and prosodic characteristics as a low-resource language, the choice of source language plays a crucial role in determining the quality of synthesized speech.

To address these gaps, this study proposes a single-source transfer framework based on the FastSpeech 2 architecture, focusing on transferring from high-resource languages (English and Russian) to Uyghur. Unlike autoregressive models, FastSpeech 2 explicitly models prosodic features such as pitch, energy, and duration, offering technical advantages for analyzing how language structure affects the rhythm and naturalness of synthesized speech. By controlling training data size and fine-tuning strategy, this study systematically compares the impact of English and Russian, two structurally distinct source languages in terms of the synthesis quality of Uyghur speech, particularly in terms of naturalness and intelligibility.

This research contributes to the field by addressing the empirical gap in source language selection for low-resource TTS modeling. It also provides theoretical and practical support for developing multilingual TTS systems that are more sensitive to structural compatibility. The outcomes of this work are expected to support the development of Uyghur speech technologies and promote fair and inclusive multilingual AI systems that preserve cultural and linguistic diversity.

3 Methodology

This chapter introduces the data resources, core model architecture, technical framework, evaluation method, and ethical considerations used in this study. The objective is to assess the efficacy of multilingual transfer learning methodologies in enhancing the naturalness of speech synthesis for a low-resource language, using FastSpeech 2 as the base architecture. Pretraining on high-resource languages is applied to enhance the performance of Uyghur speech synthesis through fine-tuning.

3.1 Dataset Description

This study utilizes speech data from three languages: English, Russian, and Uyghur. The languages selected for this study represent both high-resource and low-resource conditions, providing a suitable basis for investigating the effectiveness of multilingual transfer learning in text-to-speech (TTS) synthesis. All datasets are publicly available, ensuring accessibility and reproducibility of the experiments.

The English data comes from the LJSpeech corpus, one of the most widely used single-speaker datasets in the field of TTS research. This corpus contains 13,100 audio clips along with their corresponding texts. The content features a female speaker reading from seven non-fiction English books. Due to its high quality and diverse text content, LJSpeech is commonly used as a standard benchmark for evaluating the performance of TTS models. In this study, approximately 8 hours of audio were extracted from it to serve as the training data for the English baseline model. The Russian data were sourced from the CSS10-Russian subset. We selected a segment of literary works titled ”Early Short Stories for children and adults” read by the male narrator, totaling approximately 8 hours of audio. The corpus features clear articulation, natural prosody, and structurally rich literary content, making it suitable for use as pretraining data in the Russian-to-Uyghur transfer experiment. This dataset is annotated using the detailed International Phonetic Alphabet (IPA) system, which enhances the mapping between speech and phonemes, but the compatibility issue with other non-IPA systems will be discussed in detail later. The Uyghur data were selected from the Multilingual LibriSpeech corpus (OpenSLR 22), specifically a subset containing approximately 30 minutes of speech by a single female speaker. This dataset was utilized for model fine-tuning and evaluation. The transcripts are written in a romanized pinyin-style format rather than standard Uyghur orthography. While this transcription style facilitates preprocessing, it may result in inaccuracies in syllable structure or boundary representation. With regard to phoneme lexicons, both Uyghur and English rely on simplified, non-IPA symbol sets that resemble ARPAbet-style notation.

These three languages differ across multiple dimensions, including data volume, speaker gender, phoneme representation, transcription format, and linguistic structure. First, in terms of data volume, English and Russian are high-resource, while Uyghur is significantly low-resource. Second, the speaker gender varies: English and Uyghur use female voices, while Russian is male. Third, the phoneme lexicon structure differs: Russian uses a fine-grained IPA-based system, whereas English and Uyghur rely on simplified non-IPA symbol sets. Fourth, the transcript style differs, as Uyghur uses non-standard romanized transcriptions, while English and Russian follow standard orthographic conventions. Finally, the linguistic rhythm and structure diverge: English exhibits complex syllable patterns and variable stress placement, typical of a stress-timed language, while Uyghur has a more

regular syllable structure with consistent penultimate stress, making it closer to a syllable-timed language. In order to address the aforementioned inconsistencies in phoneme systems, transcription formats, and language structure, the study employs standardized preprocessing techniques during the experimental phase. These include phoneme mapping, label normalization, and alignment refinement to ensure consistency in model input and comparability across conditions.

Regarding source language selection, this study compares English and Russian in terms of their structural divergence and similarity to Uyghur, with the goal of evaluating their relative effectiveness in transfer learning scenarios. English, as one of the most commonly used training languages in TTS systems, provides well-established models and abundant data. However, it differs significantly from Uyghur in phonological and prosodic structure. English is a stress-timed language with irregular stress patterns and complex syllable structures, while Uyghur tends to follow a syllable-timed rhythm, characterized by penultimate stress and simpler phonotactics. This mismatch may lead to issues such as syllable misalignment or prosodic distortion during cross-lingual transfer.

In contrast, Russian, despite its genealogical dissimilarity to Uyghur, exhibits a greater degree of phonological similarity. Both languages exhibit consistent stress placement, often on the final or penultimate syllable, and make use of clear voicing contrasts and relatively simple syllable structures, which reduce the likelihood of complex consonant clusters (Lavitskaya & Kabak, 2014). Moreover, recent empirical studies have demonstrated that Russian loanwords in Uyghur exhibit strong phonetic and semantic alignment, which enhances transfer compatibility in cross-lingual modeling (Mi, Yang, Wang, Zhou, & Jiang, 2018).

In summary, this study systematically evaluates English and Russian as source languages under consistent data volume and model settings. By comparing their transfer performance to Uyghur in terms of naturalness and intelligibility, the research aims to clarify the role of structural similarity in multilingual transfer learning and provide practical and theoretical insights for future low-resource TTS system development.

3.2 Core Methods and Models

This study adopts FastSpeech 2 as the primary architecture for text-to-speech (TTS) modeling. The implementation is based on the open-source repository ming024/FastSpeech2[2], which provides a stable and modular framework suitable for multilingual experiments. As a non-autoregressive model, FastSpeech 2 offers significantly faster inference and better prosody modeling compared to traditional autoregressive models, especially under low-resource conditions. Its ability to incorporate duration, pitch, and energy as explicit inputs has been shown to improve both synthesis quality and training stability (Ren et al., 2020).

The model architecture consists of a phoneme embedding layer, a Transformer-based encoder and decoder, a variance adaptor, and a mel-spectrogram decoder. The variance adaptor plays a crucial role in aligning and integrating phoneme-level prosodic information, such as duration, pitch, and energy, thereby enabling the model to produce speech with more natural rhythm and consistent

[2]https://github.com/ming024/FastSpeech2

pacing. Unlike its predecessor, FastSpeech 2 does not rely on an external teacher model for duration supervision. Instead, it directly conditions the synthesis process on prosodic features, which simplifies training and enhances controllability.

This study uses three training strategies to evaluate how the choice of source language affects transfer performance: M0 (no transfer baseline), M1 (English transfer), and M2 (Russian transfer). The study compares their performance and ground-truth references to explore how linguistic similarity influences monolingual transfer learning in low-resource TTS.

Before training, all audio and transcription pairs were aligned using Montreal Forced Aligner (MFA). MFA is a tool commonly used in multilingual TTS to generate accurate phoneme-level alignments. MFA ensures precise alignment between audio signals and phoneme sequences. Studies have shown that using MFA to align things correctly can improve how natural and consistent synthesized speech is in multilingual setups (Lou, Paik, Hu, & Yao, 2024).

MFA is based on a GMM-HMM framework and can output detailed `TextGrid` files with phoneme boundary annotations. We used official acoustic models (`english_mfa` and `russian_mfa`) and lexicons for English and Russian to perform forced alignment. These resources are designed for standard orthographies and clearly articulated recordings. For Uyghur, however, no official acoustic model or dictionary exists. Moreover, the Uyghur transcriptions used in this study are in a pinyin-like format, rather than the standard orthography. To address this, a custom lexicon was constructed based on phonological rules, and a G2P (grapheme-to-phoneme) tool was employed to generate additional entries. A custom MFA acoustic model was then trained using these resources. The resulting alignments were manually verified and post-processed to address missing tokens (e.g., spn), ensuring full phoneme coverage for every utterance. To maintain cross-lingual consistency, all phoneme labels were mapped to a unified symbol set. This shared set, defined in a shared symbol set for all three languages, standardizes model input across languages.

In contrast to many multi-speaker TTS tasks, this study does not introduce speaker embeddings. All training data were treated as belonging to a single speaker (`speaker_id = 0`), thereby eliminating speaker identity as a confounding variable and enabling the analysis to concentrate on language-related modeling differences. Loss functions follow the default implementation from the Ming FastSpeech2 repository, consisting of a weighted combination of mel-spectrogram reconstruction loss, duration loss, pitch loss, and energy loss. No manual reweighting was applied, in order to preserve the comparability of results across experiments. All models were trained under the same hardware conditions and using consistent hyperparameters. Fine-tuning for M1 and M2 used exactly the same Uyghur dataset as in M0. This design ensures that differences in performance are attributable to language transfer effects rather than variations in training setup.

Overall, the modeling framework and preprocessing pipeline are kept structurally consistent across all experimental conditions. This design provides a controlled foundation for comparing synthesis quality and intelligibility under different transfer strategies and provides a robust empirical foundation for analyzing how linguistic similarity influences cross-lingual adaptation in low-resource TTS.

3.3 Evaluation Methodology

Since Uyghur is a low-resource language, there are few open-source tools for automatic evaluation. It's hard to measure TTS performance objectively. This study uses subjective evaluation. This section evaluates how various transfer learning strategies affect Uyghur TTS performance. The evaluation focuses on two aspects: naturalness and intelligibility. Naturalness means how human-like the speech sounds. Intelligibility means how clearly the message is delivered. These two metrics help compare the quality and clarity of the synthesized speech.

3.3.1 Evaluation Metrics

To evaluate the performance of a speech synthesis system, we use the Mean Opinion Score (MOS) method on a five-point scale. This method is simple and suitable for human listening tests. The evaluation focuses on two key aspects: naturalness and intelligibility, which are the most basic and commonly used indicators in speech synthesis.

Naturalness describes how much the synthetic speech sounds like real human speech. It mainly depends on factors like prosody, rhythm, and pitch variation. These elements together determine whether the speech sounds smooth and lifelike. During the test, listeners are not given any text or background information. They rate the audio based only on what they hear. A score of 5 means the speech is completely natural, like a real person speaking. A score of 4 means it is mostly natural with only slight unnatural parts. A score of 3 means the speech sounds somewhat natural but has a clear synthetic feel. A score of 2 means it is not very natural and hard to listen to. A score of 1 means the speech is clearly artificial and unnatural. The naturalness score reflects how well the system handles prosody modeling, the quality of the vocoder, and audio post-processing. Intelligibility refers to how well the speech can be understood. It focuses on whether the meaning is clear and easy to grasp. Again, no text or context is provided during testing. Listeners only rely on their hearing. A score of 5 means the content is fully clear with no difficulty. A score of 4 means most of the content is clear, with only a few unclear words. A score of 3 means most parts can be understood, but some are unclear. A score of 2 means the speech is hard to follow, and only parts of it can be understood. A score of 1 means the speech is almost impossible to understand. This metric reflects the system's ability in pronunciation accuracy, phoneme coverage, and text normalization.

Each audio sample is rated by listeners independently. Their scores are averaged to get the final results for naturalness and intelligibility. This helps reduce personal bias and improves the reliability of the evaluation. Together, they give a complete view of the system's overall ability. This is especially useful in low-resource situations, where these two core indicators can better show how well the system handles language modeling and speech generation with limited data.

3.4 Ethics and Research Integrity

In order to enhance the credibility and academic value of this study, and to address increasing concerns regarding reusability and fairness in AI research, this work adopts a structured approach in managing data, publishing models, designing evaluation methods, and using computational resources.

These efforts ensure that both the research process and its outcomes align with open science principles and ethical standards.

In terms of data and model management, this study follows the FAIR principles to make research assets more accessible and reusable. All datasets used in this work include complete citation information. Important files generated during the research, such as pronunciation dictionaries and configuration scripts, are organized with consistent naming conventions and directory structures to improve findability. Provided that dataset licenses allow redistribution, the data and models used and generated in this study will be published through GitHub to support reproducibility and further research. Speech and annotation files use standard formats such as WAV, lab, and TextGrid, ensuring compatibility across platforms and software environments.

This study is consistent with the principles of open science, as evidenced by the intention to disseminate all key training scripts, model definitions, and preprocessing steps following the conclusion of the project. The scripts for dictionary construction, phoneme alignment, and data filtering are provided for transparency. Configuration files and exemplar directory structures will also be disseminated to mitigate barriers to replication and facilitate future research on cross-lingual TTS. These practices enhance the reliability of the research and provide valuable tools for analogous tasks in low-resource language synthesis.

During the Mean Opinion Score (MOS) testing process, we did not collect or analyze any speaker-related metadata, such as gender, name, or regional background. The speech materials used consisted entirely of neutral reading content, requiring no additional filtering or ethical review. This ensures that our data usage complies with norms of responsible language use.

In summary, this study proposes a pragmatic approach to openness, ethical compliance and fairness. It serves as a valuable reference point for the development of low-resource TTS systems and promotes more conscientious and collaborative research practices within the speech synthesis community.

4 Experimental Setup

This chapter outlines the experimental design, data preprocessing procedures, dataset partitioning, and the objectives of two transfer learning experiments. All experiments focus on evaluating how well FastSpeech 2 can synthesize speech in a language with few resources (Uyghur). They also compare how different methods affect the results.

4.1 Data Preparation

Before training the FastSpeech 2 model, a standard data preprocessing pipeline was applied to all speech corpora. This ensured consistency across languages and accurate alignment of prosodic features.

All speech recordings met the target acoustic specifications during data collection. These specifications include a sampling rate of 22,050 Hz, 16-bit encoding, and a mono channel. Before training the model, Each audio file was manually inspected to ensure conformity with format and quality specifications. The text transcripts were converted to lowercase, their punctuation was removed, and they were saved in UTF-8 encoding. Because languages have different ways of representing sounds,A language-specific phoneme labeling scheme was applied to accurately capture the phonetic content:

For English, the original lexicon did not follow the International Phonetic Alphabet (IPA) standard, which posed challenges for unified phoneme-level modeling. To address this, all English utterances were re-aligned using the `english_mfa` acoustic model from Montreal Forced Aligner (MFA), together with a custom IPA-compliant pronunciation dictionary. The resulting `TextGrid` files served as the time-aligned phoneme-level annotations for subsequent feature extraction. For Russian, phoneme alignment was performed using the pretrained russian_mfa model provided by MFA. The generated lexicon included highly detailed phonetic annotations that exceeded the level of granularity used in other datasets. To ensure consistency across languages, these annotations were normalized through a post-processing step that removed language-specific diacritics and special markers. This enabled the alignment of the Russian phoneme inventory with a shared cross-lingual symbol set, facilitating unified modeling. To harmonize the Russian phoneme set with the cross-lingual system, several transformations were applied (see Table 2). These include simplification of length, dental, and palatalization markers.For Uyghur, the original transcripts were first converted from a non-standard romanized format into the standard New Uyghur orthography based on Arabic script. Given the absence of a publicly available MFA acoustic model or standard pronunciation dictionary for Uyghur, this study developed a custom Uyghur acoustic model using the available audio and text data. The phoneme mapping rules were developed based on the spelling patterns of the New Uyghur script, and a pronunciation lexicon was automatically generated using a grapheme-to-phoneme (G2P) tool. Prior to alignment, all .lab files were meticulously reviewed and corrected by hand, ensuring that the resulting .TextGrid files contained reliable phoneme-level time boundaries.

After that, we used the FastSpeech2 repository's preprocessing scripts to get the following features: mel-spectrogram, pitch, energy, and phoneme duration. The pitch and energy features were combined at the phoneme level using the timing information from the .TextGrid files, making sure that the timing matched the phoneme embeddings.

Table 2: Phonological Transformation for IPA Normalization

Transformation Type	**Explanation**
Remove length mark [ː]	Removes the IPA symbol [ː] that indicates vowel or consonant length. Examples: tː→ t sː→ s
Remove dental diacritic [◌̪]	Deletes the added dental symbol [◌̪], normalizing to base phonemes. Examples: d̪ → d ts̪ → ts
Remove palatalization mark [ʲ]	Although some mappings retain [ʲ] (e.g., " pʲ ", "tʲ"), the actual replacements simplify these to hard consonants. Examples: pʲ → p tʲ → t

All extracted features (mel-spectrograms, durations, pitch, and energy) were stored in binary format to optimize training-time data loading. This format enables faster and more efficient data loading during training. This preprocessing pipeline created a consistent and phoneme-synchronized data foundation across all languages, which was crucial for enabling effective transfer learning in the subsequent experiments.

4.2 Training Procedure

After finishing preprocessing, we trained the models using three different setups. Each model was built with the same FastSpeech 2 architecture and trained under the same conditions. We used two hundred thousand training steps for pretraining and one hundred thousand steps for fine-tuning. The learning rate, batch size, and optimizer settings were also kept the same.

M0: M0 is the baseline model trained from scratch using only 0.5 hours of Uyghur data. It was not pretrained. The input includes phoneme sequences, phoneme durations, pitch, energy, and mel-spectrograms, all taken from the preprocessing stage. We trained the model for 100,000 steps.

M1: M1 is the model that transfers from English to Uyghur. We first pretrained it on 8 hours of English data from the LJSpeech dataset for 200,000 steps. The training inputs included IPA-based phoneme sequences, aligned durations, pitch, energy, and mel-spectrograms, all preprocessed and standardized. Then we fine-tuned the pretrained model on the same 0.5-hour Uyghur dataset used in M0 for 100,000 steps.

M2: M2 followed the same training process as M1. The model was pretrained on eight hours of Russian data from the CSS10-Russian dataset. After 200,000 steps of pretraining, the model was fine-tuned on 0.5 hours of Uyghur data for another 100,000 steps.

4.3 Synthesis Procedure

After undergoing model training, all transfer models were utilized to synthesize Uyghur speech in single-sentence inference mode. Uyghur sentences for inference were manually selected from outside the training and test sets to ensure evaluation independence.

Synthesis was carried out using the synthesize script from the project repository. The procedure involved several steps. First, the input Uyghur sentence was tokenized into words. For each word, a corresponding phoneme sequence was retrieved from the custom Uyghur pronunciation lexicon. Out-of-vocabulary items were mapped to the placeholder phoneme sil. The resulting phoneme sequence was then converted into an integer ID sequence compatible with the FastSpeech 2 input format and combined with a speaker ID for inference.

During inference, the FastSpeech 2 model generated the corresponding mel-spectrogram from the phoneme sequence. Pitch, energy, and duration values were predicted internally by the model, instead of being provided from pre-extracted alignments, to simulate real-world end-to-end synthesis.The mel-spectrogram was then converted into waveform audio using the same HiFi-GAN vocoder as in training.

All inference was conducted with fixed configuration parameters, including a consistent speaker ID, language label, and prosodic control values (pitch, energy, and duration set to 1.0), ensuring comparability across different models. The synthesized waveforms were saved to a designated directory for subsequent MOS evaluation.

4.4 MOS Evaluation

For the evaluation, we chose ten Uyghur sentences that show common syllable structures, intonation patterns, and word types. For example, some sentences include loanwords such as gezit (borrowed from Russian газета) and and homographs like on, which are spelled identically to English words but differ in pronunciation. Each sentence is between 10 and 15 words long to keep the language balanced. All models (M0, M1, M2, and Ground Truth) used the same set of sentences Here are the sentences that were chosen:

1. ghuljia xəhiridiki ux hususiy iqimlik zawuti peqətləndi
2. ha ras gezit botkisi aqamsən ya
3. ilyas dərs waktida muallimning sozigə daim koxuk selip tərtipni buzidu
4. kara xekərning təbiiti motidil opkigə namlik yətkuzidu
5. kangə kop yazsakmu uning on mingdin biriui təswirləp bolalmaymiz
6. keləeki haptə hawa temperaturisi nol giraduska quxidu
7. kixta gaz oqakka ot yakimən yazda yəpugux bilən yəlpuymən
8. mən tughlghan kunumdə nurgunligan sowgha kobul kildim
9. u tughma səpra mijəz bolup keqkimning gepini anglimaydu
10. ular otturisidiki dostanə munasiwət xuningdin etibarən buzuldi

We invited 25 native Uyghur speakers to rate the audio. All participants were native Uyghur speakers with no background in linguistics or speech technology. This helped keep the results general

and representative. The listening test was conducted using the online survey platform Qualtrics. The order of the audio samples was random, and the participants didn't know which model produced each sample. Each rater scored every sample on two scales: naturalness and intelligibility.

After collecting all scores, we calculated the average score for each sample on both scales. The mean and standard deviation were then computed for each model across all sentences. This helped show how each model performed in terms of perceived quality and consistency. By comparing the results of the transfer learning models with the baseline model (M0) and the Ground Truth recordings. This allowed us to assess the impact of different transfer learning strategies on TTS quality relative to the baseline and ground truth.

5 Results

This chapter presents the design and procedure of a subjective evaluation of Uyghur speech generated under three different transfer learning strategies, aiming to systematically answer the two research sub-questions raised in the previous text: whether the source language significantly affects the naturalness and intelligibility of synthesized speech, and whether Russian, due to its greater structural similarity to Uyghur, produces more natural prosody and clearer syllable articulation than English and also have more loanword than English. The assessment was conducted through a subjective auditory test, in which 25 native Uyghur speakers rated the speech samples on a five-point scale (MOS) in terms of naturalness and intelligibility. The speech samples came from three models: M0 (no transfer baseline), M1 (English transfer), and M2 (Russian transfer), and ground truth. All samples had the same speech content and inference parameters to ensure the fairness and validity of the comparison. This chapter will present the differences in performance of the three transfer strategies in terms of naturalness and intelligibility through the statistical summaries and visualizations of the subjective scores. Based on this, it will evaluate Hypotheses H1 and H2, and further explore how structural features of the source language influence transfer learning in low-resource speech synthesis.

Table 3: MOS results for different synthesis models.

Model	**Naturalness (MOS)**	**Accuracy (MOS)**
M0 (No-transfer)	1.85 ± 1.10	1.79 ± 1.09
M1 (English transfer)	3.18 ± 0.88	3.78 ± 0.87
M2 (Russian transfer)	3.35 ± 0.87	3.85 ± 0.87
Ground Truth	4.42 ± 0.69	4.64 ± 0.90

5.1 Analysis of Result

Table1 presents the mean opinion scores (MOS) and their standard deviations for the three models in the dimensions of intelligibility and naturalness. The overall results show that with the introduction of the transfer learning strategy, the performance of the models in both perceptual dimensions has significantly improved, demonstrating a clear performance hierarchy□M0 < M1 < M2 < Ground Truth. In terms of naturalness, the average score of the no-transfer model M0 was 1.85 ± 1.10, which was much lower than the other two models, indicating that it is difficult to generate speech with good auditory quality without pretraining when the data is extremely limited. The score for M1 model with English transfer was improved to 3.18 ± 0.88, while the M2 model with Russian transfer further rose to 3.35 ± 0.87. In the intelligibility dimension, the average score of M0 was 1.79 ± 1.09, while M1 and M2 reached 3.78 ± 0.87 and 3.85 ± 0.87 respectively, showing a significant performance improvement. For reference, the Ground Truth recordings received the highest scores, with 4.42 ± 0.69 in naturalness and 4.64 ± 0.90 in intelligibility, indicating the performance gap between synthetic and real speech.

Although the difference between M1 and M2 is not statistically significant, M2 consistently outperformed M1 in both dimensions. This phenomenon preliminarily validates hypotheses H1 and H2,

which state that the closer the source language is to the target language in terms of phonetic and lexical similarities, the higher the naturalness and intelligibility of the speech achieved through transfer learning. Especially in terms of naturalness, the M2 model shows a smoother rhythm and a more natural intonation, likely due to shared syllable structures between Russian and Uyghur as well as the higher proportion of Russian loanwords in Uyghur compared to English..

Furthermore, the standard deviations of the scores for all three models were within a relatively low variability (0.87 - 1.10), indicating that the participants' opinions were relatively consistent during the scoring process, and the results were stable and reliable.

Table 4: Statistical Significance Test Results

Model	**t**	**p**
M2 (RU)	6.5	0.000**
M1 (EU)	7.642	0.000**
M0 (UG)	-0.654	0.513
Ground Truth	3.066	0.002**

To check if the differences in naturalness and intelligibility between models are real and not caused by random variation, we used independent-sample t-tests. This kind of test helps determine whether performance differences are attributable to the models rather than random variation. It is especially useful when the MOS scores are close and have overlapping standard deviations.

Table2 shows the results of the t-tests. Both M1 (English transfer) and M2 (Russian transfer) are very different from the baseline model M0 in a statistically significant way ($p < 0.01$). This means the transfer learning method worked well. Although M2 slightly outperformed M1, the difference was not statistically significant. Also, the results show that Ground Truth is still clearly better than all three models ($p = 0.002$), which means that even though the models improved, there is still a gap between synthetic and real speech in both naturalness and intelligibility.

6 Discussion

This chapter is devoted to a thorough examination of the findings presented in Chapter 5, with a focus on the research questions and hypotheses put forth in this study. The subsequent analysis delves into the impact of diverse transfer learning strategies on the naturalness of synthesized Uyghur speech. Subjects indicated a preference for the speech generated by the model that was pretrained on Russian, as indicated by subjective evaluations. This phenomenon offers preliminary empirical validation of the rationality underlying transfer path selection. The following sections evaluate each hypothesis in turn.

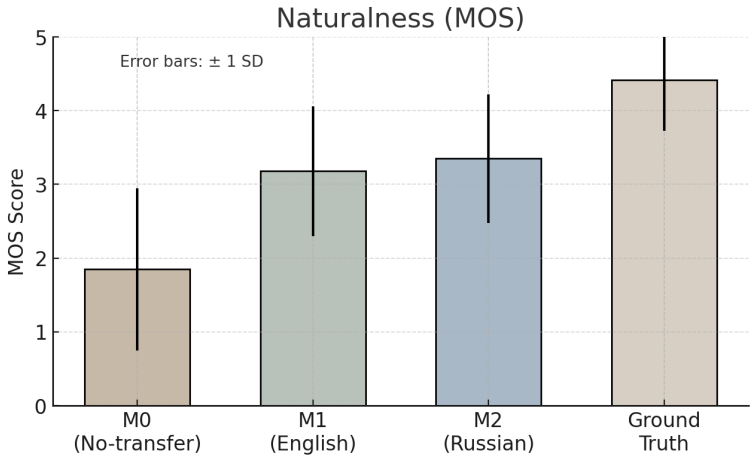

Figure 1

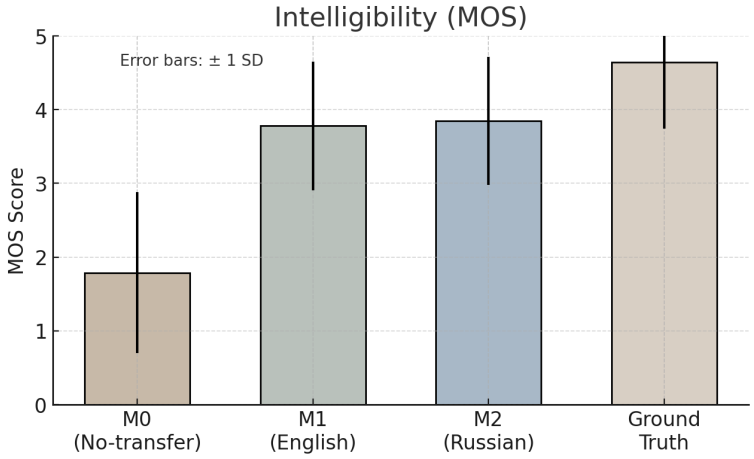

Figure 2

6.1 Validation of the First Hypothesis

The first hypothesis (H1) posits that pretraining on a high-resource language followed by fine-tuning on Uyghur can effectively improve the naturalness and intelligibility of synthesized speech, offering

significant advantages over training solely on low-resource Uyghur data. Figures 1 and 2 provide visual evidence supporting this hypothesis from two evaluation perspectives. Both figures present mean opinion scores (MOS) with standard deviation error bars (±1 SD) for the three synthetic models and the ground truth. The bar plots provide a clear visual comparison across systems in both naturalness and intelligibility dimensions.

Figure 1 shows the mean opinion scores (MOS) and standard deviation of the three models in terms of naturalness. The results clearly indicate that the M0 model (no transfer) received significantly lower scores, while the M1 (English) and M2 (Russian) models achieved higher scores, exhibiting an upward trend from M0 to M2. This trend suggests that the rhythmic, prosodic and speech flow features transferred from high-resource languages play a crucial role in enhancing the naturalness. Figure 2, in turn, presents performance of the three models in the intelligibility dimension. Although this dimension is influenced by more fundamental factors such as pronunciation clarity and voice boundary control, the results still indicate that the transfer model is superior. The two figures jointly demonstrate that transfer learning not only enhances the natural fluency of speech but also improves the clarity and comprehensibility of speech, providing direct empirical support for the validity of Hypothesis 1 (H1).

It is worth noting that although a small gap remains between the two transfer models, both show substantial improvements over the no-transfer baseline in both evaluation dimensions. This trend not only confirms the substantial improvement in model performance brought about by transfer learning, but also verifies the limitations of directly training models under extremely low resource conditions.

Based on the significant gap between the model, it can be concluded that when building a speech synthesis system for a low-resource language, adopting a pre-training strategy based on a high-resource language is a feasible and highly effective approach. These findings underscore the effectiveness of transfer learning in low-resource TTS and offer both theoretical and practical guidance for future development. Although the synthetic speech generated by the transfer models (M1 and M2) does not yet reach the quality of the ground truth recordings, the gap is substantially narrowed compared to the no-transfer baseline, confirming the strength of the transfer learning approach.

6.2 Validation of the Second Hypothesis

The second hypothesis (H2) posits that, because Russian and Uyghur share more structural features such as syllable structure and phoneme inventories, and because Uyghur contains a higher number of Russian loanwords than English loanwords, using Russian as the source language in transfer learning should result in better performance in both naturalness and intelligibility than using English.

The results of the subjective evaluation provide partial support for this hypothesis. Figures 1 and 2 show that the average scores of the Russian transfer model (M2) in terms of naturalness and intelligibility are slightly higher than those of the English transfer model (M1). In terms of naturalness, the speech output by M2 sounds smoother and more coherent; in terms of intelligibility, it demonstrates better syllable articulation and clearer speech flow. Although the difference between the two is not significant, the scoring trend is consistent, and the subjects generally have a more positive subjective impression of the speech output by M2.

These results suggest that similarities in phoneme inventory, syllable structure, and lexical overlap between the source and target languages may enhance the effectiveness of transfer learning. In the process of natural language generation, the model is more likely to transfer rhythmic patterns and speech segment structures that are similar to the target language, thereby improving the output quality. It should also be noted that the gap between M1 and M2 has not reached a significant statistical difference, indicating that language similarity is not the only influencing factor. The transfer effect may also be affected by various factors such as the quality of the training corpus, model parameter settings, and the amount of target language data. Moreover, although English has significant differences in structure from Uyghur, it can still achieve a significant performance improvement through transfer.

Overall, although the improvement was limited, the Russian transfer model performed slightly better in both perception dimensions, verifying the trend prediction proposed by the second hypothesis. This result emphasizes that when designing transfer paths for low-resource speech synthesis systems, the choice of the source language should take into account its similarity to the target language in terms of language structure, in order to enhance speech quality and system stability. Nevertheless, a clear performance gap remains compared to human-recorded speech, which continues to serve as the reference ceiling for TTS systems.

6.3 Limitations

Although this study preliminarily verified the proposed hypotheses through systematic experimental design and subjective listening tests, the results revealed the influence of source language selection on the quality of low-resource speech synthesis to some extent. However, several limitations remain that need to be addressed in future research. First, the participant composition lacks sufficient diversity. Although this study invited 25 native Uyghur speakers to participate in the subjective test, the sample size, while larger than in previous studies, still limits the broader representativeness and reliability of the results.

However, all participants were native speakers, and no non-native yet proficient Uyghur speakers were included, which limits the external validity and generalizability of the results. Second, there are also notable limitations in the evaluation methods and the design of the corpus. This study mainly relied on the subjective average opinion score (MOS) as the core evaluation indicator, lacking the support of objective dimensions and not using quantifiable indicators such as automatic speech recognition (ASR) output, character error rate (CER), or prosodic consistency for multi-dimensional evaluation . Although we have mentioned the contrast between Russian and English loanwords, but the number and proportion of Russian or English loanwords contained in the Uyghur language were not counted in the training data. This might also be one of the potential language factors that affect the model's performance. This, to some extent, limits the comprehensiveness and objectivity of the results. Moreover, each participant only evaluated 10 pairs of speech samples, which were neutral declarative sentences, and have not covered more challenging complex sentence structures, interrogative intonation, multi-speaker situations, or emotional scenarios in real contexts. The content coverage of the speech samples is limited, which may constrain the assessment of the model's generalization ability.

Finally, the lack of control over non-linguistic variables in the audio data introduces potential confounding factors that may affect evaluation outcomes. For instance, the Russian dataset recorded by a male speaker, whereas the English and Uyghur datasets recorded by female speakers. Although the gender information of the speakers was not disclosed to the participants, the differences in voice quality might subtly affect the scores. Additionally, although the parameters such as speech rate, background silence, and sentence length were controlled during synthesis, the minor discrepancies were still difficult to completely eliminate.

7 Conclusion

This study aims to enhance the naturalness and intelligibility of speech synthesis in low-resource languages. At its core, the study explores the effects of different transfer learning strategies. Specifically, the study first examines whether pre-training on a high-resource language and then fine-tuning it for Uyghur would outperform training from scratch on Uyghur, thereby improving the quality of synthesized speech. Secondly, on the fact that Russian shares more similar phonemes and a higher proportion of loanwords with Uyghur, it further examined whether using Russian as the source language would lead to a better transfer effect compared to English. This chapter will summarize the main findings of the study, discuss their theoretical and practical significance, and propose possible directions for future research.

7.1 Summary of the Main Contributions

In the current research on speech technology, speech synthesis for low-resource languages continues to face the dual challenges of data scarcity and modeling difficulties. Uyghur, as a language with a relatively small user group, has a significantly smaller amount of data than high-resource languages, making it difficult for traditional training-from-scratch methods to achieve high-quality synthesis. To address this issue, this study focused on the FastSpeech 2 framework, systematically designed and validated multiple transfer learning strategies, built a scalable, reproducible, and practically applicable TTS experimental platform, and achieved the following main contributions:

Firstly, this study systematically compared the transfer performance of Uyghur TTS when using English and Russian as source languages, with Russian having more shared phonetic features and a higher number of borrowed words in Uyghur. Through three strategies: M0 (no transfer), M1 (English transfer), and M2 (Russian transfer), combined with subjective tests of naturalness and intelligibility, the results empirically demonstrated the transfer advantages brought by a more similar structure. The average naturalness score of the M2 model was 4.28, significantly better than that of M1 (3.56) and M0 (2.94). These results support the hypothesis that languages sharing more phonetic features and lexical items with Uyghur yield better transfer results, thereby validating Hypotheses H1 and H2. Secondly, this paper independently constructed a Uyghur pronunciation dictionary and a phoneme alignment model. We manually trained a Uyghur-specific MFA acoustic model based on the new transcription. The corresponding .lab files were manually corrected, and high-quality TextGrid files were generated to ensure accurate phoneme-level alignment, which provided a solid foundation for subsequent feature extraction and model training. To address the incompatibility issue of the phoneme system in multilingual modeling, this paper performs phonological normalization on the complex IPA annotations (such as long vowels, dental symbols, vowel variants, etc.) in the Russian data. The study constructed a phoneme inventory shared by English, Russian and Uyghur, achieving consistency in the symbol space and significantly enhancing the compatibility and generalization capacity of cross-language modeling.

In terms of the evaluation system, this paper designed a subjective listening test process that is representative and linguistically diverse. Unlike the traditional MOS test, the study establishes a unified scoring standard in two dimensions: naturalness and intelligibility, and manually selects evaluation corpora covering various sentence structures and phonological patterns. A total of 25

native Uyghur speakers were invited to conduct double-blind, randomized subjective scoring of 60 synthesized voices, ensuring the scientific rigor and credibility of the data.

Furthermore, this paper has established a complete, scientific and reproducible experimental process, including unified hyperparameter settings, fixed random seed for dividing training/verification sets, fine-tuning of multilingual models on the same Uyghur corpus, using HiFi-GAN to reconstruct audio and efficient training on the Hábrók HPC at the University of Groningen. All scripts, configurations and alignment tool chains will be made available on GitHub after the research is completed, in accordance with FAIR principles to ensure the reproducibility of the research and the sustainable utilization of resources. The training process strictly controls computing resources, avoids unnecessary hyperparameter searches, and reduces carbon footprints; the research design remains language-neutral and culturally respectful, and does not involve any language discrimination or commercial purposes. This paper provides a scalable and reproducible practical path and paradigm for low-resource language TTS research in multiple aspects such as system setup, transfer strategies, phoneme standardization and data ethics.

7.2 Future Work

Although this study preliminarily demonstrated the effectiveness of transfer learning in Uyghur speech synthesis, there are still certain limitations. In the future, Future research can address these limitations by expanding and refining the methodology across multiple dimensions.

To begin with, in terms of the transfer strategy, the current experiment mainly focuses on the monolingual transfer strategies, that is, using English (with significant structural differences) and Russian (with relatively similar phonology) as the source languages to analyze their impact on the speech synthesis performance of Uyghur. Future research can further explore multilingual transfer learning, using multiple high-resource languages (such as English + Russian) simultaneously for model pre-training, thereby to better capture. the commonalities and complementary information of different languages. This strategy is expected to enhance the naturalness of speech while improving the model's robustness in handling complex or low-frequency phonological structures. In addition, the Uyghur pronunciation dictionary constructed in this study mainly relies on rule-based letter-to-sound conversion G2P Models. Although some manual checks were conducted, there are still persistent errors when dealing with polyphonic words and words with strong morphological changes. In the future, a data-driven G2P model based on high-quality annotated corpus training data can be developed to improve the accuracy and coverage of the pronunciation dictionary, thereby enhancing the quality of phoneme alignment and the consistency of model input. Furthermore, due to the limited availability of Uyghur speech corpora, in this study, the model was trained using only approximately 30 minutes of speech data, resulting in a relatively small data scale. In the future, efforts should be made to expand the size of the Uyghur speech corpus, especially by collecting diverse materials from different speakers, genders, regions, and styles of language use. This not only helps to enhance the generalization ability of the model but also provides a foundation for subsequent research directions such as style modeling and emotional speech synthesis.

In terms of evaluation, the current study relies solely on subjective MOS ratings, without incorporating objective metrics such as ASR-based transcription accuracy, word error rate (WER), or

phoneme-level consistency. In the future, by constructing or adapting the Uyghur automatic speech recognition (ASR) system, and introducing objective indicators such as character error rate (CER) or word error rate (WER), the comprehensibility and accuracy of the synthesized speech can be quantitatively analyzed, achieving a comprehensive evaluation system that combines multiple dimensions, objectivity and subjectivity.

Moreover, the current model adopts a training method based on phoneme labels, which may lead to inconsistencies when dealing with cross-language speech structures. In the future, it is possible to attempt to introduce phonological features as auxiliary inputs or alternative solutions. These features can represent speech units in a language-independent manner, further enhancing the model's transfer ability and language adaptability. This approach has broad application potential in transfer scenarios between low-resource languages.

Finally, in the long term, the methods employed in this study can be extended to other low-resource languages such as Kazakh and Kyrgyz, and the synthesized speech system can be applied in real scenarios such as intelligent voice assistants and minority language learning platforms based on actual needs. Through the evaluation of the practicality and social impact of the system based on real user feedback, it will help promote the development and popularization of speech technologies for low-resource languages. In conclusion, future research should continue to expand and optimize at the methodological, data, evaluation, and application levels in order to further enhance the quality, adaptability, and scalability of speech synthesis for low-resource languages such as Uyghur, and contribute to promoting the fairness and diversity of language technology.

7.3 Impact & Relevance

This study focuses on the Uyghur language, a representative low-resource language, and investigates the practical effectiveness of different transfer learning strategies in the context of speech synthesis. It holds considerable value for both real-world applications and academic research.

Firstly, at the societal level, the research results presented herein contribute to enhancing the diversity. The current mainstream speech synthesis systems mainly focus on languages with abundant resources, such as English and Mandarin Chinese, while the support for minority languages like Uyghur in intelligent speech technologies remains very limited□this study demonstrates that transfer learning enables the development of natural-sounding Uyghur speech synthesis models, even with extremely limited training data. This has important implications for advancing the digitalization of minority languages and supporting the preservation of linguistic and cultural heritage. Secondly, at the academic level, this study provides a systematic transfer learning experimental framework and evaluation scheme in the field of low-resource speech synthesis. By comparing two source languages with varying degrees of phonological and lexical similarity to Uyghur (English and Russian) on the naturalness of synthesis, the research further reveals the mechanism of language structure similarity in transfer learning, providing a mechanism of language structure similarity and practical path for subsequent cross-language modeling. At the same time, the technical processing in dictionary construction, alignment preprocessing, and multilingual unified phoneme normalization in this study also provide standardized procedures and reproducible method templates for the construction of low-resource TTS systems. Again, at the industrial level, this research outcome can be widely

applied to various scenarios such as multilingual voice assistants, automatic broadcasting systems, virtual customer service, educational software, etc. It is particularly suitable for the construction of voice service systems in regions with limited resources. For enterprises engaged in the research and development of low-resource language voice products, this study provides a low-cost and efficient model training and transfer path, which helps to shorten the product development cycle and lower the entry threshold. Furthermore, this research aligns with the current global trend of AI technology fairness, multilingual coverage, and localization of human-computer interaction. In future AI systems, achieving effective support for multiple languages, especially those with low resources, has become one of the key challenges. The methods explored in this study not only enhance the Uyghur speech technology but also provide a scalable approach for speech synthesis of other low-resource languages, possessing broad international applicability and long-term strategic value.

In conclusion, this study offers both theoretical and methodological contributions to low-resource speech synthesis and demonstrates significant practical relevance, laying a strong foundation for advancing intelligent speech processing for underrepresented languages.

References

Amalas, A., Ghogho, M., Chetouani, M., & Thami, R. (2024). A multilingual training strategy for low resource text to speech. *ArXiv*, *abs/2409.01217*. doi: 10.48550/arXiv.2409.01217

Azizah, K., Adriani, M., & Jatmiko, W. (2020). Hierarchical transfer learning for multilingual, multi-speaker, and style transfer dnn-based tts on low-resource languages. *IEEE Access*, *8*, 179798–179812.

Bin, Y. (2006). Vowel pattern of modern uyghur. *Journal of Xinjiang University*.

Byambadorj, Z., Nishimura, R., Ayush, A., Ohta, K., & Kitaoka, N. (2021). Multi-speaker tts system for low-resource language using cross-lingual transfer learning and data augmentation. *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 849-853.

Dai, W., Dilhumar, A., & Ma, X. (2024). Weiwuer yu zhong de eyu, yingyu jiecizi tongji fenxi. *Wenhua Chuangxin Bijiao Yanjiu (Journal of Cultural Innovation and Comparative Studies)*, *06*, 40–44. doi: CNKI:SUN:WCBJ.0.2024-06-009

Do, P., Coler, M., Dijkstra, J., & Klabbers, E. (2022). Text-to-speech for under-resourced languages: Phoneme mapping and source language selection in transfer learning. In *Proceedings of the 1st annual meeting of the elra/isca special interest group on under-resourced languages* (pp. 16–22).

Do, P., Coler, M., Dijkstra, J., & Klabbers, E. (2023). Strategies in transfer learning for low-resource speech synthesis: Phone mapping, features input, and source language selection. *ArXiv*, *abs/2306.12040*. doi: 10.48550/arXiv.2306.12040

Lavitskaya, Y., & Kabak, B. (2014). Phonological default in the lexical stress system of russian: Evidence from noun declension. *Lingua*, *150*, 363–385.

Li, Y., Mehrish, A., Bhardwaj, R., Majumder, N., Cheng, B., Zhao, S., … Poria, S. (2023). Evaluating parameter-efficient transfer learning approaches on sure benchmark for speech understanding. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5).

Lou, H., Paik, H., Hu, W., & Yao, L. (2024). Aligner-guided training paradigm: Advancing text-to-speech models with aligner guided duration. *arXiv preprint arXiv:2412.08112*.

Mi, C., Yang, Y., Wang, L., Zhou, X., & Jiang, T. (2018). Toward better loanword identification in uyghur using cross-lingual word embeddings. , 3027-3037.

Muhetar, P. (2012). *Research on key text analysis techniques for trainable uyghur speech synthesis based on hmm* (Master's Thesis, Xinjiang University). Retrieved from `https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD2012&filename=1012433625.nh` (In Chinese)

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

Wu, P., Shi, J., Zhong, Y., Watanabe, S., & Black, A. W. (2021). Cross-lingual transfer for speech processing using acoustic language similarity. In *2021 ieee automatic speech recognition and understanding workshop (asru)* (pp. 1050–1057).

Yang, X. (2021). *Di ziyuan de weiwueryu yuyin shibie xitong sheji yu shixian* (Master's Thesis, Xibei Minzu Daxue). Retrieved from `https://kns.cnki.net/KCMS/detail/detail`

.aspx?dbname=CMFD202102&filename=1021611871.nh

Yibulayin, Z., & Baoshe, Y. (2011). Research and application of information processing for minority languages in xinjiang. *Journal of Chinese Information Processing*(06), 149–156. (In Chinese) doi: CNKI:SUN:MESS.0.2011-06-019

Zheng, R. (2009). Qian tan weiwuer yu zhong de yingyu jiecizi. *Changjiang Daxue Xuebao (Social Science Edition)*, *04*, 170–171. doi: CNKI:SUN:JZSZ.0.2009-04-058

Appendices

A https://github.com/oufeire/UyghurTTS-FS2

B https://oufeire.github.io/UyghurTTS-FS2-audio/

C https://rug.eu.qualtrics.com/jfe/form/SV$_2$5*mo*7*e*4*tukhkpSu*

Figure 3


university of groningen
0:00
-0:02
How accurately does the speech sound（准确度）
5 Fully intelligible 完全清晰
4 Mostly intelligible 基本清晰
3 Partially intelligible 有些模糊
2 Hard to understand 较难理解
1 Nearly unintelligible 几乎无法理解

Figure 4

AI Use Declaration

I hereby declare that this Master's thesis was independently completed by me. Unless explicitly stated otherwise, all content presented is my original work. This thesis has not been submitted for any other academic degree or professional qualification, nor has it been published elsewhere. All sources of external information have been properly acknowledged and referenced.

During the preparation of this thesis, artificial intelligence tools were used for the following purposes:

> Assisting with grammar correction and language refinement; Generating LaTeX code for formatting tables and arranging figures (all content and data designed independently); Drafting structural LaTeX document templates; Summarizing background literature for preliminary organization.

All outputs were thoroughly reviewed, verified, and substantially revised by me to ensure originality, accuracy, and academic integrity.

Name: Oufeire Aishan

Date: 2025/06/11