university of groningen

campus fryslân

Singing Voice Synthesis in Your Language: Cross-Lingual Transfer with Limited Data Using Diffusion Models

Jiashu Dong

university of groningen

campus fryslân

University of Groningen - Campus Fryslân

Singing Voice Synthesis in Your Language: Cross-Lingual Transfer with Limited Data Using Diffusion Models

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
PhD candidate Phat Do (Voice Technology, University of Groningen)
with the second reader being
Supervisor 2's title and name (Voice Technology, University of Groningen)

Jiashu Dong (S6124720)

June 11, 2025

Acknowledgements

I would like to begin by thanking my parents, whose unwavering support has been the quiet foundation of everything I've done. Their love gave me the courage to pursue what I truly care about.

To my friends, both near and far—thank you for always being there. Your presence reminded me that even when things got tough, I was never walking the road alone.

I also want to thank myself—for daring to dream, for seeking freedom, and for following through. Every late night and every small effort was part of something bigger.

Writing this thesis has been far from easy. Although Singing Voice Synthesis is connected to voice technology, it is a complex field with its own unique history, challenges, and rapidly evolving techniques. As someone new to this area, I spent countless hours self-studying, reading papers, and exploring unfamiliar tools and concepts. The process was often difficult and frustrating—but in the end, pulling it all together gave me a renewed sense of capability and resilience.

To all the teachers from the Voice Technology program, I would like to express my appreciation for the knowledge and guidance you have offered throughout my academic journey. In particular, I am grateful to Dr. Matt Coler and Dr. Vass Verkhodanova for your encouragement and your support in shaping this thesis proposal. I am especially thankful to my supervisor, Phat Do, whose generous support and mentorship made a lasting impact. Despite his busy schedule, he always took the time to answer my questions with clarity and patience, providing thoughtful guidance that shaped not only the direction of my research but also the way I think.

Thank you to the University of Groningen and the Center for Information Technology for access to the Hábrók computing cluster—your resources made this work possible.

To my family, my friends, my teachers—and to the version of myself who never gave up—I sincerely hope the road ahead is kind, bright, and full of meaning.

Abstract

Singing Voice Synthesis (SVS) has achieved remarkable progress with diffusion-based models such as DiffSinger (J. Liu, Li, Ren, Chen, & Zhao, 2022), enabling expressive and high-fidelity singing generation. However, most existing SVS systems are primarily trained on English and Chinese datasets, limiting access for musicians or music enthusiast from other linguistic communities. Extending SVS to more languages could democratize music production and contribute to the preservation of global cultural diversity. This work explores cross-lingual transfer learning for SVS, using DiffSinger as the base system and German as the target language. We hypothesize that fine-tuning an English-trained DiffSinger model on a small amount of German data—leveraging a phoneme mapping strategy based on PHOIBLE (Moran & McCloy, 2019)—can achieve comparable performance to a model trained from scratch on a large-scale monolingual German dataset. Furthermore, we investigate the influence of training data quality in low-resource scenarios. Given the same limited data size, we hypothesize that models fine-tuned on higher-quality data—characterized by native accent, broader vocal range, clean recording conditions—will outperform those trained on lower-quality datasets. This improvement is attributed to enhanced linguistic clarity and expressive realism. Evaluation is conducted using both objective (F0 Frame Error, Mean Cepstral Distortion, Word Error Rate) and subjective (Comparative Mean Opinion Score, MUSHRA) metrics. Results indicate that fine-tuned models with as little as 15/30 minutes of data can achieve performance comparable or even better to those trained on large-scale datasets, and with limited 15mins data, the overall data quality—including accent, vocal control and recording conditions—can improve synthesis quality significantly. This study presents a focused analysis of phoneme-mapped cross-lingual transfer for German SVS and offers practical strategies for adapting SVS systems to underrepresented languages using minimal data. To the best of our knowledge, this is the first study to investigate cross-lingual transfer learning in SVS field. We believe that the findings and methodology of this work can be extended to support cross-lingual SVS development in other low-resource languages as well. We hereby release both the online demo, available at https://dongjiashu.github.io/DiffSinger/, and the source code repository at https://github.com/DongJiashu/DiffSinger for public access.

Contents

1 Introduction

Singing Voice Synthesis (SVS) aims to automatically generate high-quality singing audio from lyrics and musical notation (Wu & Luan, 2020). While SVS shares common ground with speech synthesis, it introduces additional challenges: it must maintain timbral coherence across a wide pitch range, enhance expressiveness, and accurately follow musical scores (Ardaillon, 2017). In recent years, deep learning-based singing voice synthesis (SVS) has attracted a lot of attention from both industry and academic communities.

In the industry, AI-generated music, particularly through end-to-end applications has rapidly gained traction as a valuable tool for musicians and producers (Gera, 2025). These systems allow users to input lyrics and melodies and generate professional-quality vocal tracks instantly. For example, application like Synthesizer V can be integrated into digital audio workstations, enabling rapid prototyping and iteration of vocal parts in music production workflows. This integration significantly reduces the time and cost associated with traditional recording methods, making SVS systems increasingly central to modern music creation processes (AudioCipher, 2024).

Academically, many researchers have contributed to advancing SVS technology. Early SVS systems were based on concatenative (Kenmochi & Ohshita, 2007) and parametric models (Saino, Zen, Nankaku, Lee, & Tokuda, 2006), which often produced robotic or unnatural results. Neural models, including RNN-based approaches (Blaauw & Bonada, 2017), offered improved modeling capabilities but struggled with expressiveness. More recent works such as (Chandna, Blaauw, Bonada, & Gómez, 2019; Chen, Tan, Luan, Qin, & Liu, 2020; Nakamura, Hashimoto, Oura, Nankaku, & Tokuda, 2019; Shi et al., 2022; Y. Zhang et al., 2022; Zhuang et al., 2021) have further pushed the boundaries of SVS quality. For instance, FFTsinger (L. Zhang et al., 2022) is a feed-forward Transformer-based model that generates high-quality singing by leveraging spectral filtering techniques to refine the output mel-spectrogram, achieving better timbral quality compared to earlier autoregressive models. However, while FFTsinger improves upon previous methods, it still faces challenges in handling complex melodies and expressive dynamics. The subsequent development of generative models, particularly GAN-based approaches such as GAN-Singer (Wu & Luan, 2020) and Singgan (Huang et al., 2022), further enhanced the expressiveness and fidelity of generated singing voices. These models excel in generating high-frequency details but may suffer from mode collapse or overly smooth outputs. More recently, diffusion-based methods like DiffSinger (J. Liu et al., 2022; Y. Zhang, Jiang, et al., 2024) have introduced novel mechanisms that significantly improve both the fidelity and expressiveness of synthesized singing, positioning them as leading paradigms in the field.

Despite these advances, most current SVS systems are trained on large-scale datasets in English and Chinese, leaving musicians from other linguistic communities with limited access to high-quality synthesis tools. As AI-driven music production continues to grow, ensuring that SVS is inclusive of underrepresented languages becomes increasingly essential. Extending SVS to more languages not only democratizes access to music production tools but also contributes to the preservation of global linguistic and cultural diversity. However, most state-of-the-art SVS systems heavily depend on large-scale annotated datasets. For example, creating a 30-minute singing corpus may require up to 20 hours of expert labor, highlighting the high annotation cost in this field. As a result, languages without such datasets are underrepresented in the SVS landscape. Inspired by cross-lingual transfer strategies in speech synthesis (Do, Coler, Dijkstra, & Klabbers, 2022; Tu, Chen, Yeh, & Lee, 2019) we investigate how similar approaches can be applied to SVS to bridge the linguistic data gap.

To achieve that, we adopt DiffSinger as our base model and investigate how to adapt English-trained models to German using limited training data. Specifically, we implement a phoneme mapping strategy based on PHOIBLE (Moran & McCloy, 2019) to align the phoneme sets between English and German. We hypothesize that fine-tuning an English-trained DiffSinger model on a small amount of mapped German data can yield performance comparable to a model trained from scratch on a large-scale German corpus. Moreover, we examine the impact of training data quality in low-resource conditions. Here, data quality is treated as a multifaceted concept encompassing singer nativeness, vocal range, recording clarity. To investigate this, we construct multiple 15-minute fine-tuning subsets that vary in these factors—such as native vs. non-native accents, broad vs. limited vocal ranges, and varying levels of acoustic cleanliness. This setup allows us to test the hypothesis that such quality-related attributes collectively affect the intelligibility and expressiveness of synthesized singing.

Through this exploration, our goal is to propose scalable and accessible strategies for extending SVS technology to more languages worldwide. We hereby release both the online demo, available at https://dongjiashu.github.io/DiffSinger/, and the code at https://github.com/DongJiashu/DiffSinger for public access.

Now that the motivation for this research has been presented, the structure of this thesis is as follows:

- Section 1.1 presents the research questions and hypotheses
- Section 2 reviews relevant literature and positions this work within current research
- Section 3 describes the methodological approach
- Section 4 details the experimental setup
- Section 5 presents and analyzes the results
- Section 6 discusses implications and insights
- Section 7 concludes with key findings and future directions

1.1 Research Questions and Hypotheses

In light of the preceding discussion, this research addresses the following question:

Main Research Question:
How can we effectively adapt singing voice synthesis models to low-resource languages using phoneme mapping and fine-tuning? Evaluated through both objective metrics (FFE, MCD, WER) and subjective listener assessments (CMOS and MUSHRA).

RQ1:
Can a DiffSinger model pre-trained on English singing be adapted to German through phoneme mapping and fine-tuning on limited (30min/15min) German data, such that it achieves comparable performance to a model trained from scratch on a 3 hours German dataset?

RQ2:
To what extent does the overall quality of fine-tuning data—including singer accent (native vs. non-native), vocal range (wide vs. narrow) and recording conditions, affect the intelligibility and expressiveness of cross-lingually synthesized German singing voices?

Our hypothesis is:

H1:
Prior studies in multilingual speech synthesis (Do et al., 2022) have demonstrated that phoneme mapping enables effective cross-lingual adaptation. We hypothesize that this strategy can extend to singing voice synthesis models like DiffSinger (J. Liu et al., 2022): Fine-tuning a DiffSinger model pre-trained on English singing with only a limited amount (30min/15min) of phoneme-mapped German data can achieve comparable performance to training from scratch on a 3 hours German dataset .

H2:
Furthermore, earlier research on TTS (Tomokiyo, Black, & Lenzo, 2005; Vít, Hanzlíček, & Matoušek, 2018) shows that the quality of training data—particularly speaker accent and recording clarity—significantly affects synthesis outcomes. We extend this insight to singing: The overall quality of the fine-tuning data—encompassing singer nativeness, vocal range, recording fidelity—has a significant impact on the expressiveness and intelligibility of synthesized singing in low-resource settings .

2 Literature Review

This section systematically introduce Search strategy and selection in Section 2.1 and then reviews in Section 2.2 and Section 2.3 advancements in singing voice synthesis (SVS), in Section 2.4 multilingual singing datasets, and in Section 2.5 and Section 2.6 cross-lingual transfer learning for low-resource adaptation. The synthesis of these areas highlights gaps in phoneme-mapped fine-tuning and singer characteristic preservation, motivating our proposed framework.

2.1 Search Strategy and Selection Criteria

Titles and abstracts were systematically screened for relevance using a predefined set of keywords, inclusion criteria, and exclusion criteria. After initial filtering, the remaining papers were assessed for methodological rigor, including reproducibility of training procedures, availability of open-source code, and datasets. In addition, citations and subsequent references of key papers were reviewed to ensure comprehensive coverage of the topic.

Keywords are:

1. singing voice synthesis: "singing voice synthesis" OR "SVS"
2. singing dataset: "singing datase"
3. low resource language: "cross-lingual transfer learnin", "phoneme mapping", "phonological features" and "low-resource languages"

These keywords were used in combination across multiple databases (e.g., IEEE Xplore, arXiv, Google Scholar, ISCA Archive). Variants and related terms (e.g., "cross-language transfer learning", "phone mapping") were also considered to account for different spellings and word forms. Boolean operators (AND/OR) were applied to combine search terms effectively for each database.

Inclusion criteria are:

1. Published in reputable venues including IEEE Xplore, arXiv, Google Scholar, or the ISCA Archive.
2. Focused on singing voice synthesis (SVS) or low-resource text-to-speech (TTS) models with relevance to singing synthesis.
3. Included objective acoustic evaluation metrics such as F0 Frame Error or Mel-Cepstral Distortion, and/or subjective measures like Mean Opinion Score (MOS).
4. Provided insights into phoneme adaptation, cross-lingual transfer learning, or phonological feature-based modeling for low-resource languages.
5. Described replicable methodologies with publicly available implementations or datasets, ensuring transparency and reproducibility.
6. Published between 2018 and 2025 to reflect recent advancements in neural SVS systems.

Each paper was first evaluated based on its title and abstract. If it passed this initial screening, a full-text review was conducted to verify that all inclusion criteria were explicitly addressed.

Exclusion criteria are:

1. Relied on HMM-based or concatenative SVS models due to their limited expressiveness and lower synthesis quality compared to modern neural approaches.
2. Lacked sufficient technical detail on model architecture or training procedure, making reproduction difficult or impossible.
3. Published before 2018, as they may not reflect current trends in deep learning-based singing synthesis.
4. Not written in English or Chinese, which were selected to ensure accessibility and consistency in interpretation.

2.2 DiffSinger: A Diffusion-Based Baseline

DiffSinger (J. Liu et al., 2022) represents a significant advancement in singing voice synthesis (SVS), introducing a novel shallow diffusion mechanism that enables high-fidelity and expressive singing generation directly from musical scores (e.g., lyrics and pitch sequences). Compared to other SVS models such as GAN-based frameworks (Wu & Luan, 2020) and feed-forward transformer-based systems like FFT-Singer (L. Zhang et al., 2022), DiffSinger offers several key advantages.

Unlike some earlier SVS models such as FFT-Singer, which may struggle with maintaining harmonic details in the mid-to-low frequency range, DiffSinger employs a non-autoregressive diffusion process that better captures global pitch contours and expressive dynamics across entire phrases. This results in more natural transitions between notes and improved overall expressiveness. Additionally, thanks to the iterative denoising process inherent to diffusion models, DiffSinger is able to generate more natural-sounding and emotionally rich vocalizations compared to GAN-based systems like GAN-Singer. Although GAN-Singer performs well in generating high-frequency details, it sometimes suffers from mode collapse or produces overly smooth and less dynamic outputs, an issue that DiffSinger effectively mitigates while preserving both high-quality harmonics and expressive dynamics.

One of the most notable innovations in DiffSinger is the use of a shallow diffusion inference process. Instead of starting from pure Gaussian noise, the model initializes the diffusion reversal from a coarse mel-spectrogram generated by an auxiliary decoder. This significantly reduces computational cost while maintaining high output quality. In contrast, FFT-Singer relies on feed-forward transformers for mel-spectrogram generation, which can be computationally expensive and less efficient during inference. Furthermore, DiffSinger introduces a boundary predictor module that dynamically determines the optimal starting point in the reverse diffusion process. This prevents over-smoothing and ensures sharper transitions between phonemes and notes, especially in complex melodic structures—something that GAN-Singer lacks, often leading to timing and articulation issues.

In subjective evaluations using Mean Opinion Score (MOS), DiffSinger outperforms both FFT-Singer and GAN-Singer, demonstrating its superior perceptual quality. These improvements position DiffSinger as a representative of the latest generation of SVS systems—moving away from traditional autoregressive and parametric methods toward diffusion-based modeling, which has become the dominant paradigm for controllable and high-quality singing synthesis.

Although the original experiments were conducted primarily on Mandarin datasets, the authors also demonstrated promising performance on English speech synthesis using the LJSpeech dataset. This suggests that the architecture is inherently language-agnostic and flexible enough to adapt to different linguistic systems. However, its performance in low-resource settings or cross-lingual transfer scenarios has not yet been explored in depth.

2.3 TCSinger: Alternative Diffusion Model

In addition to DiffSinger, another recent diffusion-based singing voice synthesis (SVS) model, TCSinger (Y. Zhang, Jiang, et al., 2024), has also demonstrated strong performance in expressive singing generation. TCSinger builds upon the foundation laid by DiffSinger by introducing explicit modeling of multi-level singing styles—including vocal technique, emotion, rhythm, and pronunciation—through a modular architecture designed for fine-grained control over the synthesis process.

The model introduces three core components that enable this enhanced expressiveness. First, the Clustering Style Encoder utilizes Clustering Vector Quantization (CVQ) to encode rich style information into a compact latent space, allowing for efficient representation and manipulation of expressive features. Second, the Style and Duration Language Model (S&D-LM) jointly predicts phoneme duration and style features using both audio and text prompts, thereby improving the alignment between linguistic content and expressive characteristics. Third, the Style-Adaptive Decoder (SAD) applies mel-style adaptive normalization during mel-spectrogram generation, ensuring that expressive qualities are preserved throughout the synthesis pipeline.

TCSinger is reported to achieve superior subjective quality and more accurate reproduction of expressive singing compared to DiffSinger. It also claims to support what the authors refer to as “zero-shot cross-lingual style transfer”. However, this capability is made possible by training the model on large-scale multilingual singing data, primarily from Chinese and English sources. In other words, the so-called zero-shot ability does not stem from true generalization without target-language data, but rather from joint bilingual modeling during training. This distinction is important, as it suggests that TCSinger requires extensive parallel or comparable data across languages in order to perform effectively in cross-lingual settings.

Moreover, TCSinger’s architecture introduces additional complexity through multiple auxiliary models and multi-phase training procedures, which rely heavily on detailed style annotations. These requirements make TCSinger less suitable for low-resource adaptation scenarios where annotated singing data is limited and system simplicity is essential.

Given our research goal of enabling singing voice synthesis in new languages with minimal annotated training data, we ultimately choose DiffSinger as our base model. While TCSinger achieves higher synthesis quality and offers greater expressive control, its reliance on large-scale multilingual data and complex training pipelines makes it impractical for our lightweight, low-resource transfer learning scenario.

2.4 GTSinger: The Only Open Multilingual Singing Corpus

GTSinger (Y. Zhang, Pan, et al., 2024) remains the only publicly available dataset containing singing data beyond English and Chinese. Its structure and scale make it an essential foundation for cross-lingual SVS research. Specifically, GTSinger provides several key advantages. It contains over 80 hours of professionally recorded singing data from 20 singers across 9 languages, including

English, Chinese, German, Spanish, French, Italian, Korean, Russian, and Japanese, offering a level of linguistic diversity that is unmatched by other existing datasets. Additionally, the dataset includes phoneme-level alignments, expressive technique labels, and realistic music score annotations. These detailed annotations facilitate more accurate training and evaluation of SVS models, enabling them to capture subtle nuances in different singing styles and techniques. Unlike many other datasets that use simplified or synthesized scores, GTSinger incorporates real-world music scores, which are crucial for generating high-quality singing voices that closely mimic professional performances.

However, GTSinger also presents some notable limitations. One such issue is the inclusion of recordings from non-native speakers in several language subsets, which can affect the dataset's suitability for evaluating fine-grained phonetic intelligibility and prosody in native singing. Another limitation lies in the lack of a standardized phoneme format across languages: the dataset uses ARPA for English, Pinyin for Chinese, and IPA for others. This heterogeneity complicates phoneme-level transfer learning and necessitates the development of a unified mapping strategy. These limitations motivated several design choices in our work. To address the inconsistency in phoneme representations, we apply phoneme mapping using PHOIBLE (Moran & McCloy, 2019) to harmonize the multilingual phoneme space. Furthermore, we curate supplemental native German recordings to better assess the impact of data quality in low-resource fine-tuning scenarios.

In conclusion, despite its shortcomings, GTSinger stands out as a critical resource for advancing cross-lingual SVS research. Its rich annotations and diverse linguistic coverage make it an invaluable tool for developing and testing new SVS models. By addressing its limitations through targeted improvements such as phoneme mapping and native language augmentation, researchers can leverage GTSinger to push the boundaries of what is possible in singing voice synthesis, especially in low-resource settings.

2.5 Phoneme Mapping with PHOIBLE in Cross-Lingual Transfer

With both a capable diffusion-based singing synthesis model (DiffSinger) and a multilingual dataset (GTSinger) in place, the next question becomes: how can we most effectively adapt trained SVS models to a new, low-resource language? Inspired by prior work in text-to-speech (TTS) transfer learning, we investigate phoneme-level adaptation strategies as a lightweight yet powerful mechanism for cross-lingual generalization.

Do et al. (2022) proposed one of the most systematic and scalable phoneme mapping strategies for under-resourced languages. By leveraging the PHOIBLE database (Moran & McCloy, 2019), they mapped phonemes across languages based on a 37-dimensional binary vector of articulatory features (e.g., voicing, nasality, manner/place of articulation). This allowed them to identify close phoneme pairs and reuse pre-trained phoneme embeddings, improving intelligibility and naturalness in target languages. Their results demonstrated that PHOIBLE-based mapping outperformed manual rule-based alignments, particularly when combined with source language selection using Angular Similarity of Phoneme Frequencies (ASPF).

While their method has shown strong results in TTS adaptation, its effectiveness in the domain of singing voice synthesis has not yet been validated. Singing imposes additional constraints—such as expressive dynamics, pitch range, and rhythmic precision—that may interact differently with phoneme similarity assumptions. Our work extends this research by applying PHOIBLE-guided phoneme mapping in the SVS domain, and systematically evaluates whether it supports high-quality cross-lingual singing generation under minimal-data conditions.

Moreover, based on the ASPF described in (Do et al., 2022) and the multilingual coverage provided by the GTSinger (Y. Zhang, Pan, et al., 2024) dataset, we selected English as the source language and German as the target language.

2.6 Data Quality and Singer Attributes in SVS Considerations

In singing voice synthesis (SVS), data quality plays a critical role in determining the final synthesis performance. While most prior work has focused on model architectures and training strategies, relatively less attention has been given to how singer attributes—such as native language background, vocal range, and expressiveness—affect synthesis outcomes. This is particularly relevant when adapting models across languages or applying transfer learning in low-resource settings.

Research in text-to-speech (TTS) has shown that speaker accent significantly impacts synthesized speech quality. For instance, Tomokiyo et al. (2005) found that TTS systems trained on non-native accented speech produced outputs with reduced intelligibility and acceptability. Although this study focused on speech rather than singing, it raises an important question: could similar effects occur in SVS? Given the phonological and prosodic complexity of singing, we hypothesize that accent mismatch between training data and target language may also degrade the naturalness and intelligibility of synthesized singing voices.

Beyond singer attributes, input data quality—including recording fidelity, background noise, and annotation accuracy—also strongly influences model performance. Vít et al. (2018) demonstrated that even small errors in phoneme alignment or segmentation can significantly degrade the output of WaveNet-based acoustic models. To mitigate such issues, tools like the Montreal Forced Aligner (MFA) (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017) have become standard for improving alignment consistency. However, in singing data, where pitch variation and vibrato complicate alignment, manual correction is often still required.

Furthermore, Vít et al. (2018) emphasized that noise in training data—whether from background interference or inaccurate annotations—can degrade synthesis quality, especially under low-resource conditions. These findings suggest that both the technical quality of recordings and the linguistic fidelity of annotations are essential for robust SVS modeling.

One dataset that enables investigation into these factors is the GTSinger corpus (Y. Zhang, Pan, et al., 2024), which includes singing samples in multiple languages, including German. However, inspection of its German subset reveals that many samples were recorded by non-native singers with limited vocal range and expressive capabilities. This raises concerns about whether such data is sufficient for high-quality multilingual SVS adaptation. The limitations of the GTSinger German subset motivate our exploration of alternative fine-tuning strategies using higher-quality native singing data, which will be detailed in Section 3.

In summary, both singer-related factors and technical data quality aspects—such as vocal expressiveness, accent, recording clarity, and annotation precision—should be carefully considered in SVS system design. These insights inform our experimental approach to cross-lingual transfer learning and low-resource adaptation.

Table 1: Summary of Key Literature

Reference	Key Findings	Theme
DiffSinger (J. Liu et al., 2022)	Introduces a shallow diffusion model for high-quality expressive SVS; however, cross-lingual adaptation with limited data remains unexplored.	SVS model
TCSinger (Y. Zhang, Jiang, et al., 2024)	Adds multi-level style control and improves cross-lingual performance, but relies on large-scale multilingual data and complex architecture, thus unsuitable for low-resource scenarios.	SVS model
GTSinger (Y. Zhang, Pan, et al., 2024)	Provides large-scale multilingual singing corpus with phoneme alignment; lacks standardized phoneme set and includes non-native German recordings.	Dataset
PHOIBLE mapping (Do et al., 2022; Moran & McCloy, 2019)	Demonstrates effectiveness of PHOIBLE-based phoneme mapping in TTS transfer learning; singing adaptation remains untested.	Phoneme mapping
Training data (Vít et al., 2018)	Highlights the importance of input quality and alignment accuracy in TTS, supporting careful data curation in SVS experiments.	Data quality
Foreign accents (Tomokiyo et al., 2005)	Shows that non-native accents reduce intelligibility in synthetic speech, motivating the use of native data in SVS adaptation.	Accent impact

In light of recent advances in singing voice synthesis (SVS), particularly models like DiffSinger (J. Liu et al., 2022) and TCSinger (Y. Zhang, Jiang, et al., 2024), we observe significant improvements in synthesis fidelity and expressive control. However, their reliance on large-scale English and Chinese datasets—and for TCSinger, complex architectures and rich annotations—limits applicability in low-resource and cross-lingual settings. GTSinger (Y. Zhang, Pan, et al., 2024), the only open multilingual SVS dataset, begins to address this gap but suffers from inconsistent phoneme formats and non-native speaker recordings.

To overcome these issues, we adopt DiffSinger as a lightweight base model and use GTSinger as our primary data source. We apply a PHOIBLE-based phoneme mapping strategy (Moran & McCloy, 2019), previously used in speech synthesis (Do et al., 2022), though not yet in SVS. Analysis of GTSinger's German subset revealed limitations in vocal range, expressiveness, and pronunciation due to non-native speakers. Prior work (Tomokiyo et al., 2005; Vít et al., 2018) shows such factors can significantly affect synthesis quality. To assess this impact systematically, we constructed two fine-tuning subsets with varying accent, vocal range, and recording conditions.

This study investigates two underexplored aspects in SVS: (1) the empirical effectiveness of phoneme mapping for cross-lingual transfer, and (2) the influence of singer-level data quality in low-resource adaptation. By combining phoneme-aware fine-tuning with curated, high-quality data, we aim to enable more robust and scalable SVS transfer to underrepresented languages.

3 Methodology

This section outlines the methodology adopted to explore phoneme-mapped cross-lingual transfer learning for Singing Voice Synthesis (SVS), with the goal of addressing our research questions on transfer strategies (RQ1) and fine-tuning data quality (RQ2). We first present the datasets used in our experiments in Section 3.1, including both publicly available and custom-curated corpora. Next, we describe the DiffSinger model and the transfer learning strategies employed, with a focus on phoneme mapping via PHOIBLE in Section 3.2. We then detail the technical framework in Section 3.3, including algorithm utilization and data preprocessing. Finally, we explain the evaluation methodology used to assess synthesis quality through both objective acoustic measures and subjective listener ratings in Section 3.4.

3.1 Dataset Description

To investigate cross-lingual transfer learning under low-resource conditions, we constructed a multilingual singing dataset consisting of German and English recordings with varying data sizes, speaker accents, and vocal ranges. These datasets are designed to evaluate both our research questions: the effectiveness of phoneme-mapped transfer learning (RQ1) and the role of data quality (RQ2). To avoid confusion with the term “Ground Truth”, we refer GTSinger corpus as “GTs”

- **English-GTs 3H**: A high-quality monolingual English singing dataset extracted from GT-Singer (Y. Zhang, Pan, et al., 2024). It was used to pre-train all base DiffSinger models prior to cross-lingual fine-tuning experiments.

- **German-GTs 3H**: A 3-hour subset from the German portion of GT-Singer, recorded under clean and consistent studio conditions. While the audio quality is high, the singing was performed by a non-native speaker with a noticeable foreign accent and relatively basic expressive control. This dataset was used to train a from-scratch German model and serves as a performance upper bound under monolingual training.

- **German-GTs 15min / 30min**: Two subsets (15 and 30 minutes) sampled from the German-Base GTSinger recordings. The selections were curated to maximize consistency in vocal technique, prosody, and stylistic phrasing. These subsets simulate realistic low-resource scenarios for model adaptation.

- **German-NativeNarrow 15min**: A 15-minute dataset compiled from performances by a native German speaker with a mid-range vocal range. The recordings were sourced from publicly available platforms YouTube (2024), leading to variability in audio quality and occasional background noise. Nevertheless, the dataset provides native-level pronunciation and natural prosody, making it valuable for examining the effect of accent in isolation. Importantly, its mixed recording conditions reflect the kind of real-world, in-the-wild data that researchers often have access to when collecting low-resource language resources outside of professional studios.

- **German-ProficientWide 15min**: A high-quality dataset recorded by a proficient (non-native) German singer with near-native pronunciation, extended vocal range, and expressive control.

> The recordings were captured in a quiet home studio using high-fidelity equipment. This dataset serves as the most acoustically consistent and expressive resource, enabling evaluation of whether superior vocal and recording characteristics can compensate for the absence of nativeness.

While GTSinger remains the only open-source dataset with German singing data, its German subset contains accented performances and lacks cross-lingual phoneme consistency. To address this, we incorporated custom recordings with varied speaker and acoustic characteristics, better simulating real-world scenarios of multilingual singing model adaptation.

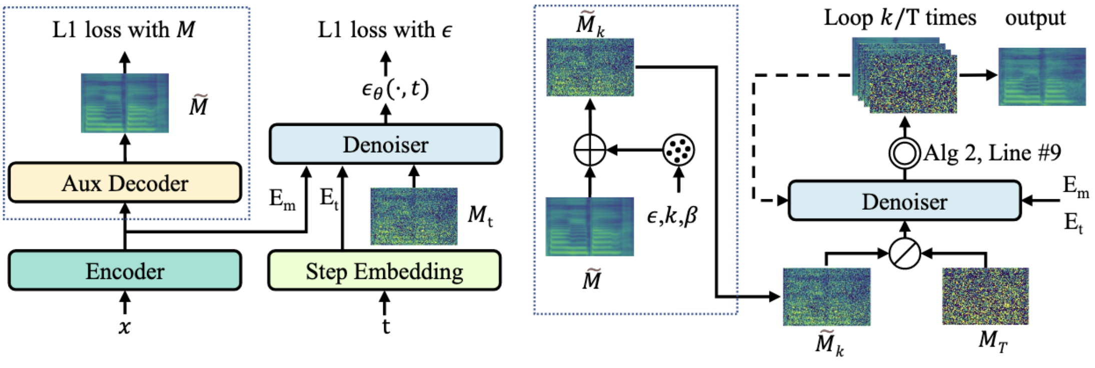
All audio was downsampled to 44.1 kHz and aligned to phoneme sequences using the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017), followed by meticulous manual correction to ensure frame-level accuracy.

These datasets are tailored like this to test our two key hypotheses. By varying only one factor at a time, we isolate its effect on model performance. The German-GTs subsets evaluate the impact of limited training duration with phoneme mapping (RQ1), while the Native-Narrow and Proficient-Wide and GTs-15mins datasets examine how data quality influence the model performance (RQ2). This hybrid setup provides a controlled yet practical benchmark for evaluating cross-lingual SVS.

3.2 Core Methods and Models

3.2.1 Core Model - DiffSinger

DiffSinger consists of two primary modules: As illustrated in Figure 1, the system includes:



(b) The inference procedure of DiffSinger.

Figure 1: Overview of DiffSinger's architecture and inference procedure.
Adapted from (J. Liu et al., 2022).

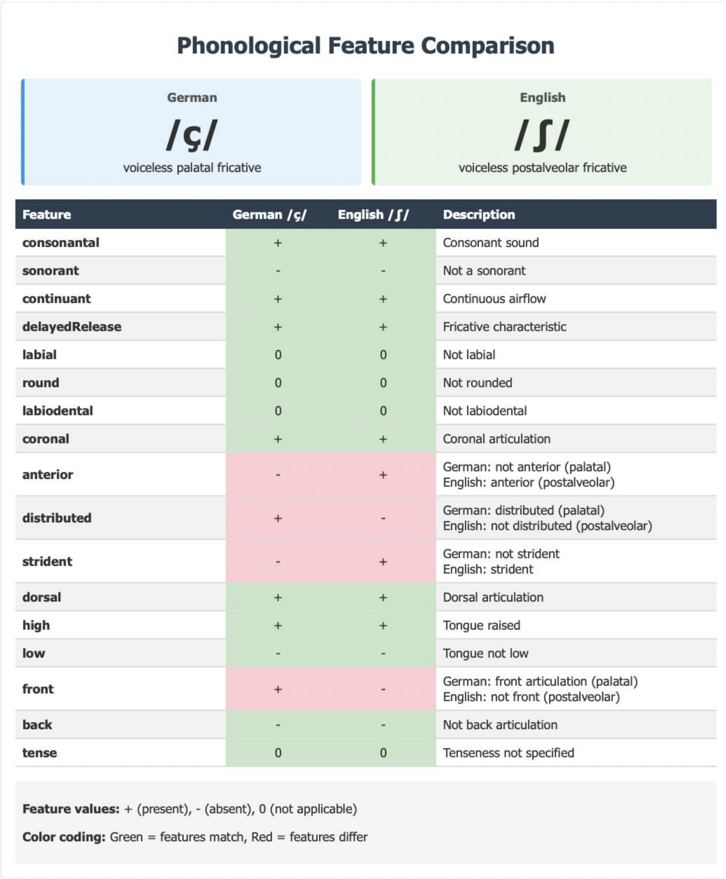
The Encoder Module is responsible for transforming symbolic musical inputs into conditioning features that guide the acoustic generation process. It comprises three key submodules: the phoneme encoder, which converts phoneme sequences into continuous embeddings; the length regulator, which expands these phoneme embeddings to a frame-level representation based on predicted phoneme durations; and the pitch encoder, which transforms pitch contours into frame-aligned embeddings. These features are then summed together to form the music condition vector, which serves as the primary conditioning signal for the synthesis stage.

The second major component is the Auxiliary Decoder and Shallow Diffusion Module, which follows a two-stage generation strategy. In the first stage, the auxiliary decoder generates a coarse mel-spectrogram from the music condition vector. This initial estimate is then refined in the second stage by the shallow diffusion module through a small number of denoising steps (typically 4–6). The diffusion module includes a denoiser network and step embeddings that model the temporal dynamics of noise removal. Compared to standard diffusion models that require hundreds or even thousands of steps, this shallow setup drastically reduces inference time while preserving high synthesis quality. The auxiliary decoder provides a strong prior that guides the diffusion process, enabling efficient and effective refinement of the output.

In summary, DiffSinger's modularity, controllability, and efficiency make it a strong candidate for investigating low-resource and cross-lingual SVS transfer scenarios.

3.2.2 Core Methods - Phoneme Mapping

To enable German synthesis from an English-pretrained DiffSinger model, we adopt a phoneme-level adaptation framework grounded in the PHOIBLE database (Moran & McCloy, 2019), which provides language-agnostic 37-dimensional phonological feature vectors. Each German phoneme is mapped to its most similar English counterpart based on articulatory feature similarity (see figure 2). This enables direct reuse of the English phoneme embedding layer, avoiding architectural modifications or full retraining.



Phonological Feature Comparison

German	English
/ç/	**/ʃ/**
voiceless palatal fricative	voiceless postalveolar fricative

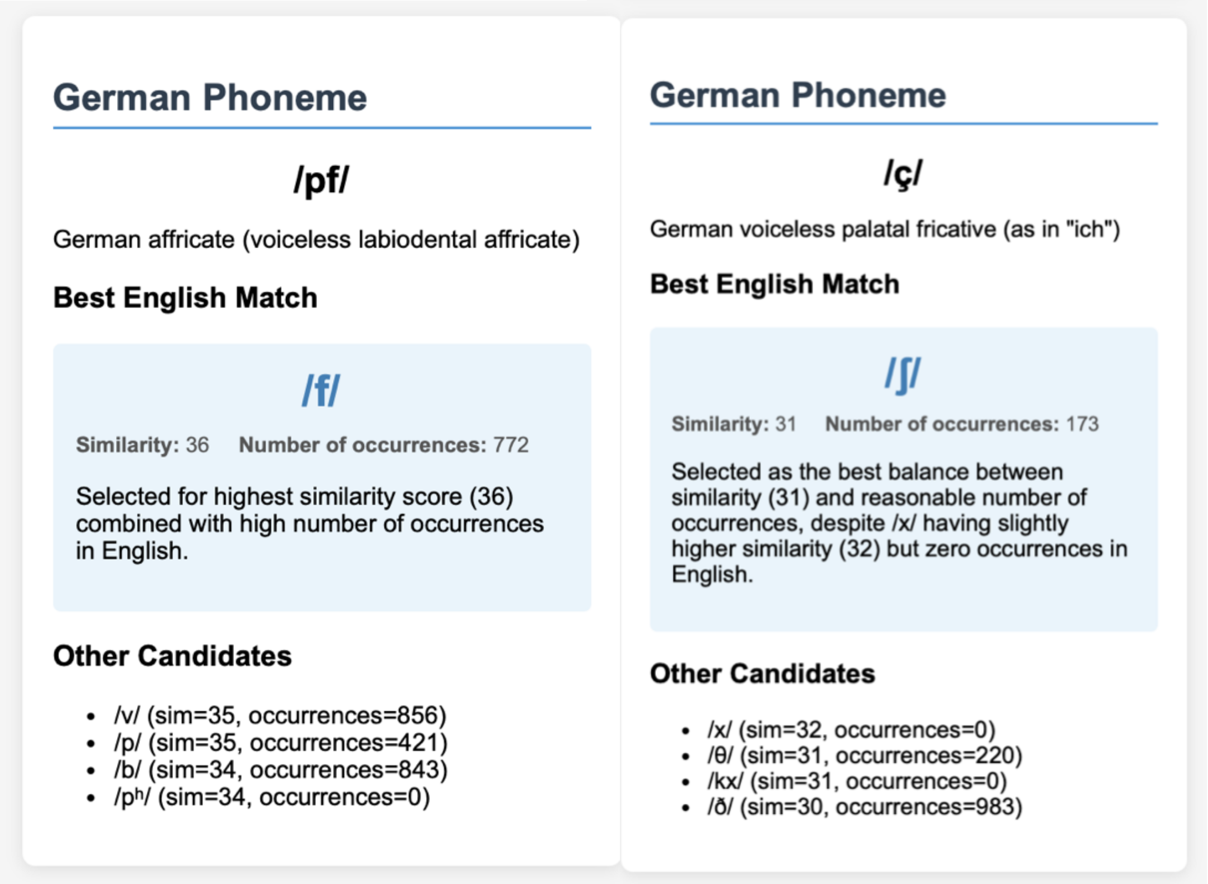
Feature	German /ç/	English /ʃ/	Description
consonantal	+	+	Consonant sound
sonorant	-	-	Not a sonorant
continuant	+	+	Continuous airflow
delayedRelease	+	+	Fricative characteristic
labial	0	0	Not labial
round	0	0	Not rounded
labiodental	0	0	Not labiodental
coronal	+	+	Coronal articulation
anterior	-	+	German: not anterior (palatal) English: anterior (postalveolar)
distributed	+	-	German: distributed (palatal) English: not distributed (postalveolar)
strident	-	+	German: not strident English: strident
dorsal	+	+	Dorsal articulation
high	+	+	Tongue raised
low	-	-	Tongue not low
front	+	-	German: front articulation (palatal) English: not front (postalveolar)
back	-	-	Not back articulation
tense	0	0	Tenseness not specified

Feature values: + (present), - (absent), 0 (not applicable)

Color coding: Green = features match, Red = features differ

Figure 2: Illustration for 17 phonological features from the PHOIBLE data (37 features in real case)

In cases where multiple English phonemes have comparable similarity scores, we further prioritize phonemes that appear more frequently in the source training corpus. This two-stage matching strategy balances phonetic proximity with statistical coverage, improving the robustness of cross-lingual inference.



German Phoneme

/pf/

German affricate (voiceless labiodental affricate)

Best English Match

/f/

Similarity: 36 Number of occurrences: 772

Selected for highest similarity score (36) combined with high number of occurrences in English.

Other Candidates

- /v/ (sim=35, occurrences=856)
- /p/ (sim=35, occurrences=421)
- /b/ (sim=34, occurrences=843)
- /pʰ/ (sim=34, occurrences=0)

German Phoneme

/ç/

German voiceless palatal fricative (as in "ich")

Best English Match

/ʃ/

Similarity: 31 Number of occurrences: 173

Selected as the best balance between similarity (31) and reasonable number of occurrences, despite /x/ having slightly higher similarity (32) but zero occurrences in English.

Other Candidates

- /x/ (sim=32, occurrences=0)
- /θ/ (sim=31, occurrences=220)
- /kx/ (sim=31, occurrences=0)
- /ð/ (sim=30, occurrences=983)

Figure 3: Similarity scores are based on phonological features from PHOIBLE database. Number of occurrences represent relative occurrence in English. Matches were selected prioritizing similarity first, then number of occurences for cases with similar similarity scores.

3.2.3 Core Methods - Adaptation Strategies

On top of the unified phoneme representation, we implement two adaptation strategies to assess cross-lingual transferability from English to German singing voices:

- **Zero-Shot Inference**: We evaluate the English-trained DiffSinger model directly on PHOIBLE-mapped German phoneme sequences without any German fine-tuning. This serves as a baseline to assess the model's ability to generalize across languages purely through phonological similarity. For example, we synthesize the German sentence “Ich liebe dich” using the English model, mapped to its closest IPA phonemes.

- **Fine-Tuning**: We fine-tune the model on German singing data with varying properties to analyze their influence on synthesis performance. Specifically:

  - Data Size: We use 15-minute and 30-minute subsets of German data to study how data volume impacts adaptation.
  - Accent: We compare performances using recordings from a native German singer and non-native singers. This allows us to evaluate whether pronunciation accuracy affects intelligibility in synthesized results.
  - Vocal Range: We contrast fine-tuning on singers with different vocal ranges. This lets us analyze whether the expressive potential of the original data influences the model's expressiveness post-transfer.
  - Audio quality: We contrast fine-tuning on corpus with different recording set ups. This lets us analyze whether the audio quality of the original data influences the model's performance.

This modular design (details in in 4) allows us to isolate the effects of each factor—phoneme mapping, data quantity, and data quality—on the cross-lingual performance of SVS models. The comparative analysis directly addresses hypotheses **H1** and **H2**, providing insight into the feasibility of low-resource language adaptation.

3.3 Technical Framework

Our experimental pipeline is implemented in Python and integrates the following core components:

- **DiffSinger (OpenVPI Fork)**: We build on the open-source implementation from OpenVPI[1], which refactors the original DiffSinger model for improved modularity and accessibility. We focus solely on training the acoustic model component. The acoustic model is based on the Denoising Diffusion Probabilistic Model (DDPM) (Ho, Jain, & Abbeel, 2020), which iteratively refines a noisy mel-spectrogram toward a clean target distribution. To improve sampling efficiency and stability, the author also adopt Rectified Flow (RF) (X. Liu, Gong, & Liu, 2022), a recent alternative to traditional diffusion processes that replaces stochastic reverse sampling with a deterministic, flow-based formulation. Waveform reconstruction is conducted using NSF or HiFi-GAN, while RMVPE is employed for accurate pitch estimation.

- **Phoneme Alignment**: We evaluate two alignment tools:

  - SOFA[2] : A customized singing alignment model, we first train the model using GTSinger annotated data and align our corpus, but its performance was less consistent—likely due to limited training data or model mismatch.
  - Montreal Forced Aligner (MFA) [3]: Although not explicitly designed for singing, MFA provided more stable and accurate alignments on singing data compared to SOFA in our experiments.

Given the scope of this study, we adopt MFA for all final alignments without further analysis.

[1] https://github.com/openvpi/DiffSinger

[2] https://github.com/qiuqiao/SOFA

[3] https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

- **VLabeler**[4]: We used VLabeler as the primary tool for manual alignment correction after initial phoneme boundaries were obtained using Montreal Forced Aligner (MFA). Compared to traditional tools like Praat, VLabeler offers a more streamlined and user-friendly interface tailored for phoneme-level boundary editing, significantly accelerating the annotation process.
- **Phoneme Mapping Tool**: A custom script aligns German phonemes to English phonemes using articulatory similarity derived from PHOIBLE's 37-dimensional feature vectors (Moran & McCloy, 2019), filtered by phoneme frequency in the English dataset. This mapping is applied during data preprocessing and inference.
- **Music and Pitch Preprocessing**:
  - `SOME`[5] is used to extract MIDI pitch contours from aligned music scores.
  - `RMVPE`[6] (Wei, Cao, Dan, & Chen, 2023) is used for frame-level f_0 extraction.
- **Monitoring**: Training statues is monitored using TensorBoard.

3.4 Evaluation Methodology

We conduct both objective and subjective evaluations to assess synthesis quality across cross-lingual transfer settings. Each model is evaluated on 23 test samples, including both parallel and non-parallel utterances, ensuring robustness and diversity in linguistic and musical content.

3.4.1 Objective Evaluation

Object evaluations includes:

- **F0 Frame Error (FFE)**: Measures voicing and pitch accuracy between predicted and reference pitch. It is defined as:

$$\mathrm{FFE} = \frac{1}{T}\sum_{t=1}^{T}\left[\mathbf{1}\left(\mathrm{VUV}_{\mathrm{gt}}(t) \neq \mathrm{VUV}_{\mathrm{pred}}(t)\right) \vee \mathbf{1}\left(\left|1200 \cdot \log_2\left(\frac{f_0(t)+\varepsilon}{\hat{f}_0(t)+\varepsilon}\right)\right| > \theta_{\mathrm{cent}}\right)\right] \tag{1}$$

  where T is the total number of frames, and a frame-level error is counted if the voicing decision is incorrect or the pitch deviates by more than $\theta_{\mathrm{cent}} = 50$ cents (approximately 20% frequency deviation).

- **Mel-Cepstral Distortion (MCD)**: Quantifies spectral differences between generated and reference audio using MFCCs. Based on our implementation:

$$\mathrm{MCD} = \frac{10}{\ln 10} \cdot \frac{1}{N}\sum_{n=1}^{N}\sqrt{2\sum_{d=1}^{D}(x_{n,d} - y_{n,d})^2}$$

  where x and y are aligned MFCC sequences of dimension D over N frames.

[4] `https://github.com/sdercolin/vlabeler`

[5] `https://github.com/openvpi/SOME`

[6] `https://github.com/Dream-High/RMVPE`

- **Word Error Rate (WER)**: Whisper (Radford et al., 2023) is used to transcribe both synthesized and reference utterances. WER is computed as:

  $$\mathrm{WER} = \frac{S+D+I}{N}$$

  where S, D, and I are the number of substitutions, deletions, and insertions, and N is the number of words in the reference.

- **Statistical significance test** To assess the significance of performance differences between systems, we applied statistical tests based on the distributional properties of the data.

  For metrics where data followed a normal distribution and variances were equal (as confirmed by the Shapiro–Wilk test (Shapiro & Wilk, 1965)), we used one-way Analysis of Variance (ANOVA) (Fisher, 1934) to compare group means, followed by Tukey's HSD post-hoc test (Tukey, 1949) to identify pairwise differences. ANOVA tests whether any group differs significantly from the others, and Tukey's test controls for false positives in multiple comparisons.

  When variances were unequal or the normality assumption did not hold, we used Welch's ANOVA (Welch, 1951), a robust alternative that accommodates unequal variances, and the non-parametric Kruskal–Wallis test (Kruskal & Wallis, 1952) to compare medians. For pairwise comparisons in these cases, we employed the Games–Howell post-hoc test (Games & Howell, 1976), which does not assume equal variances or sample sizes.

  A significance level of $\alpha = 0.05$ was used in all tests. Normality and variance assumptions were checked for each metric before selecting the appropriate analysis.

3.4.2 Subjective Evaluation

To complement the objective evaluation, we conducted a subjective listening test with 30 human raters. Half of the participants are native German speakers, and the other half are intermediate-level non-native speakers. Considering the number of systems being compared, we split the evaluation into two questionnaire versions (Set A and Set B) to reduce listener fatigue and ensure high-quality feedback. All questions are arranged in random sequence to ensure fair evaluation.

We adopted two standard subjective evaluation protocols:

- **CMOS** (Comparative Mean Opinion Score, (Sector, 1996)): Participants were presented with pairs of audio samples and asked to rate the relative quality on a 7-point scale ranging from –3 (much worse) to +3 (much better).

- **MUSHRA** (Multiple Stimuli with Hidden Reference and Anchor, (Series, 2014)): Listeners rated multiple system outputs for the same utterance on a 0–100 scale, based on overall singing quality and expressiveness. While a hidden reference was not explicitly embedded, the zero-shot baseline serves as a consistent lower anchor in our setup.

In order to reduce redundancy and maintain consistency, we replaced traditional intelligibility-based listening questions with automatic transcription using Whisper (Radford et al., 2023) to compute WER as an objective proxy as described in objective evaluation.

3.4.3 Baseline Comparisons and Validation Approach

To evaluate **RQ1** and test **H1**, we compare a German model trained from scratch on 3 hours of GT-Singer data Base GTs 3H against two fine-tuned models trained on only 15 and 30 minutes of data from the same corpus FT GTs 15min and FT GTs 30min, as well as a Zero-shot model that was pre-trained on English and applied to German using phoneme mapping without any fine-tuning. This setup tests whether limited fine-tuning with phonological features can match or surpass large-scale monolingual training.

To evaluate **RQ2** and validate **H2**, we fix the fine-tuning duration at 15 minutes and vary the speaker and recording conditions across three models: FT GTs 15min (non-native, narrow-range, clean recordings), FT NativeNarrow 15min (native speaker, narrow-range , variable audio quality), and FT ProficientWide 15min(proficient speaker, wide range, clean recordings). The Zero-shot model is again included as a lower-bound baseline. This design allows us to examine how different dimensions of data quality—accent, vocal diversity, and recording fidelity—affect synthesis performance under low-resource conditions. This setup allows us to isolate the effect of speaker accent, vocal range, and recording conditions on SVS performance in low-resource scenarios.

3.5 Ethics and Research Integrity

This research was conducted in accordance with institutional ethical guidelines and did not involve any human subject experimentation requiring formal ethics board review. All data handling, evaluation procedures, and dissemination plans were designed to ensure transparency, openness, and responsible AI development.

3.5.1 Data Ethics and Privacy

All singing data used in this study—such as GT-Singer and our customized datasets—were either publicly available or self-recorded with informed consent for academic use. To utilize YouTube (2024) data, we also applied YouTube Researcher Program [7] and use only data with Creative Commons (CC) licence [8]. No personal identifiers are included in the dataset, and all singer metadata was anonymized. Data storage followed university security protocols, with restricted access and regular backups. No third-party sensitive or proprietary data was used.

3.5.2 FAIR Principles Implementation

We follow the FAIR principles to ensure the long-term usability of our research: We adhere to the FAIR principles to ensure the long-term usability and accessibility of our research outputs. Our resources are designed to be findable, with all datasets and model checkpoints indexed in public repositories accompanied by persistent identifiers such as DOIs and comprehensive metadata descriptions. These materials are made accessible without requiring authentication, hosted on widely used platforms such as Hugging Face and GitHub. To enhance interoperability, we employ standard file formats including WAV for audio, JSON for annotations, and CSV for metadata, while maintaining consistent phoneme annotation protocols across languages. In terms of reusability, all

[7] https://research.youtube

[8] https://creativecommons.org

released resources come with detailed documentation, clear licensing information under the Creative Commons Attribution 4.0 International License (CC-BY 4.0), and suggested citation guidelines to facilitate proper attribution and reuse.

3.5.3 Open Science Practices

As part of our commitment to open science, we will publicly release all source code, training configurations, pre-trained and fine-tuned DiffSinger model checkpoints, as well as evaluation scripts for objective metrics such as MCD (Mel-Cepstral Distortion), FFE (Frame-level Feature Error), WER (Word Error Rate), and subjective evaluation protocols including CMOS (Conversational Mean Opinion Score) and MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor). All components will be version-controlled using Git-based systems and distributed under permissive licensing terms: MIT License for software code and CC-BY for data and documentation. Repository links, license details, citation recommendations, and contribution policies will be clearly outlined in the project documentation to support community engagement and extension.

3.5.4 Bias and Fairness

We also consider potential biases that may arise from dataset composition and algorithmic behavior. To assess fairness and generalizability, our evaluation includes both male and female listeners, and we compare responses from native and non-native speakers to investigate possible accent-related biases. Although the primary focus of this work is on the German language, the methodology is designed to be transferable to other low-resource languages. We acknowledge limitations related to singer diversity and genre imbalance and will document these alongside model performance results to ensure transparency about the system's scope and constraints.

3.5.5 Environmental Impact

In recognition of the environmental impact associated with large-scale machine learning, we have taken steps to minimize energy consumption throughout our experimentation. Fine-tuning was conducted on small-scale datasets (15–30 minutes per language), avoiding full retraining from scratch. Experiments were run on shared university GPU clusters using limited batch sizes and early stopping to reduce unnecessary computation. We will report the total number of training steps and GPU hours used in each experiment to provide transparency regarding computational cost and carbon footprint.

3.5.6 Reproducibility and Replicability

To ensure reproducibility and replicability, we follow rigorous development practices. All random seeds, library versions, and dependencies are fixed and published alongside the codebase. The training and evaluation pipeline includes detailed logging, standardized output formats, and comprehensive evaluation scripts. Additionally, we provide example shell commands to reproduce each experiment end-to-end. Despite these efforts, known limitations include dependency on specific PyTorch versions and the requirement for GPU hardware to perform full inference. These factors are documented to help users understand the conditions under which results can be reliably reproduced.

Through these measures, we ensure that our research adheres to the highest standards of ethics, transparency, and research integrity. Our goal is not only to advance cross-lingual singing synthesis, but to do so responsibly—with openness, environmental awareness, and broad accessibility in mind.

4 Experimental Setup

To ensure the full reproducibility of our study, this section provides a comprehensive account of the implementation pipeline, data preparation, and experimental structure used throughout the research.

We begin by detailing how datasets were preprocessed, split, and annotated to support fine-tuning and evaluation. This includes phoneme mapping, feature extraction, and alignment processes in Section 4.1 and in Section 4.2. Next, we outline the design of our comparative experiments in Section 4.3, specifying the conditions for zero-shot transfer, limited-data fine-tuning, and full-scale baseline training. All experiments were run using version-controlled configurations and open-source tools to promote replicability.

All source code, configuration files, are made publicly available via GitHub upon publication, following the open science and FAIR principles discussed in.

4.1 Data Preparation

Our experiments are based on singing voice recordings in German and English. The data preparation pipeline was designed to support both fine-tuning and evaluation under various conditions (zero-shot, low-resource, full-data), with consistent preprocessing across all subsets.

4.1.1 Data Sources and Formats

GT-Singer (Y. Zhang, Pan, et al., 2024) is our primary corpus. It was originally provided in WAV format sampled at 48kHz. To ensure compatibility with DiffSinger's preprocessing pipeline, we resampled all audio to 44.1kHz. As GT-Singer is already segmented and aligned at the phoneme level, it requires no additional preprocessing. We directly applied phoneme mapping and extracted pitch and MIDI features for training and evaluation.

Native-Narrow is a custom dataset constructed from publicly available singing videos featuring a native German singer sourced from YouTube (2024). This dataset required the most extensive preprocessing among the resources used. We first developed a custom pipeline to download and convert MP4 video files into WAV audio format. Following this, vocal-accompaniment separation was carried out using the Demucs library in order to isolate clean singing tracks from the mixed audio sources. The extracted vocal tracks were then segmented into 5–15 second clips using a custom script, and each segment was manually annotated with corresponding lyrics in .lab file format. Despite considerable effort to select clips with consistent tempo and clear vocal quality, the recording conditions varied across video sources. Consequently, the *Native-Narrow* dataset exhibits variable acoustic quality, but it reflects a practical and realistic approach for collecting native-language singing data under low-resource constraints.

Proficient-Wide is a self-recorded dataset captured in a quiet home environment using a high-quality microphone. It was specifically designed to maximize vocal range and signal quality. Like the Native-Narrow set, we segmented recordings into 5–15 second clips and manually aligned them with .lab lyric files.

Figure 4 illustrates that the *ProficientWide* dataset encompasses a broader pitch range compared to the *NativeNarrow* dataset, featuring a greater number of both high and low notes. In contrast, *NativeNarrow* is concentrated primarily within the mid-range (C4–E4). We compared models fine-tuned with these two datasets to assess how variations in vocal range affect the synthesis quality and overall performance of the singing voice synthesis (SVS) system.

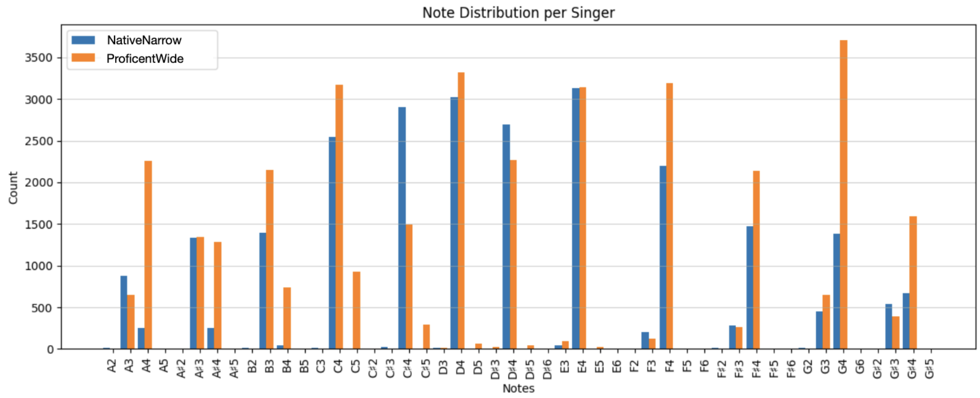


Figure 4: Note distributions for 2 customized datasets

For lyrics alignment, while tools such as LyricFA[9] and Whisper[10] offer automatic alignment, we opted for fully manual annotation to reduce the cost of error correction and to ensure the highest alignment precision.

4.1.2 Phoneme Alignment

To produce frame-level phoneme alignments, we initially applied the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017), which, despite being designed for speech, demonstrated higher alignment accuracy than a domain-specific SOFA[11] model we trained on GT-Singer.

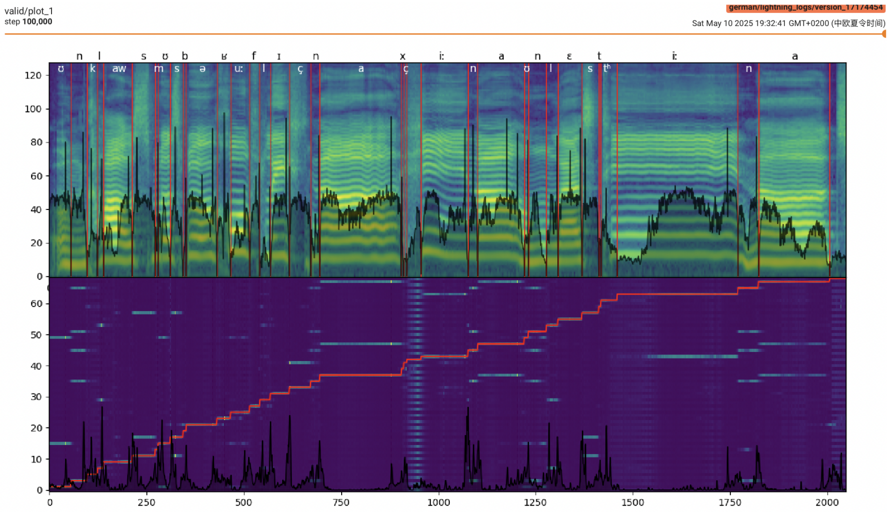

valid/plot_1
step 100,000
german/lightning_logs/version_17174454
Sat May 10 2025 19:32:41 GMT+0200 (中欧夏令时间)
0
20
40
60
80
100
120
0
10
20
30
40
50
60
0
250
500
750
1000
1250
1500
1750
2000

Figure 5: SOFA Model Training

We believe this discrepancy stems from the limited and variable training data available for SOFA. For highest precision, all alignment outputs were manually refined using vlabeler [12] , an annotation tool better suited for singing than traditional Praat.

[9] https://github.com/wolfgitpr/LyricFA

[10] https://github.com/openai/whisper

[11] https://github.com/qiuqiao/SOFA

[12] https://github.com/sdercolin/vlabeler

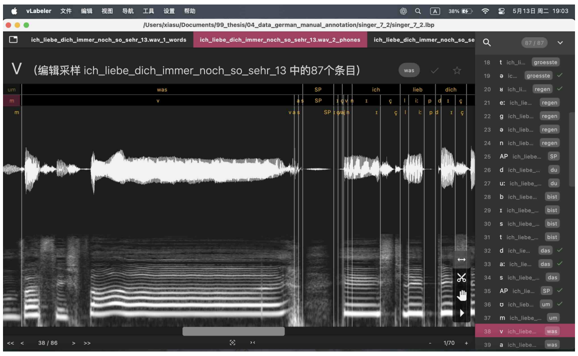

vLabeler
/Users/xiasu/Documents/99_thesis/04_data_german_manual_annotation/singer_7_2/singer_7_2.lbp
ich_liebe_dich_immer_noch_so_sehr_13.wav_1_words
ich_liebe_dich_immer_noch_so_sehr_13.wav_2_phones
V (编辑采样 ich_liebe_dich_immer_noch_so_sehr_13 中的87个条目)
38 / 86
1/70

Figure 6: Before Manual Alignment

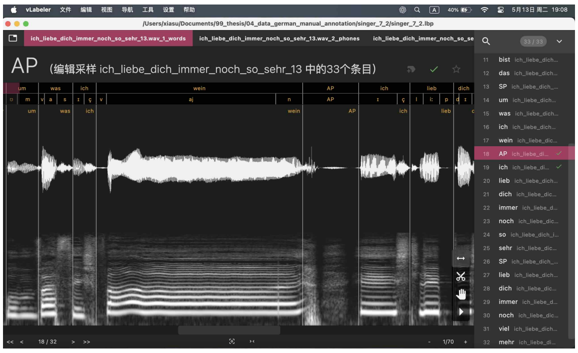
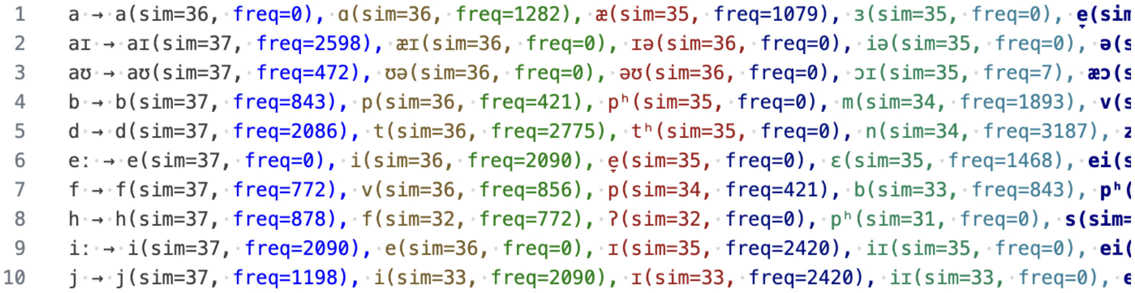

vLabeler
/Users/xiasu/Documents/99_thesis/04_data_german_manual_annotation/singer_7_2/singer_7_2.lbp
ich_liebe_dich_immer_noch_so_sehr_13.wav_1_words
ich_liebe_dich_immer_noch_so_sehr_13.wav_2_phones
AP (编辑采样 ich_liebe_dich_immer_noch_so_sehr_13 中的33个条目)
18 / 32
1/70

Figure 7: After Manual Alignment

Although the two custom datasets totaled only 30 minutes of audio, the manual alignment process took approximately 20 hours of human effort, highlighting the labor-intensive nature of curating high-quality singing data. This further motivates our work, which seeks to guide optimal data usage under resource-constrained conditions.

4.1.3 Phoneme Mapping

German phonemes were mapped to English using a custom PHOIBLE-based script. We prioritized 37-dimensional phonological feature similarity, and used phoneme frequency in the English dataset as a secondary tie-breaker to avoid rarely seen phonemes. The resulting mapping ensured compatibility with pre-trained English models without altering architecture or retraining embedding layers.



```
a → a(sim=36, freq=0), ɑ(sim=36, freq=1282), æ(sim=35, freq=1079), ɜ(sim=35, freq=0), e̯(sim
aɪ → aɪ(sim=37, freq=2598), æɪ(sim=36, freq=0), ɪə(sim=36, freq=0), iə(sim=35, freq=0), ə(s
aʊ → aʊ(sim=37, freq=472), ʊə(sim=36, freq=0), əʊ(sim=36, freq=0), ɔɪ(sim=35, freq=7), æɔ(s
b → b(sim=37, freq=843), p(sim=36, freq=421), pʰ(sim=35, freq=0), m(sim=34, freq=1893), v(s
d → d(sim=37, freq=2086), t(sim=36, freq=2775), tʰ(sim=35, freq=0), n(sim=34, freq=3187), z
eː → e(sim=37, freq=0), i(sim=36, freq=2090), e̯(sim=35, freq=0), ɛ(sim=35, freq=1468), ei(s
f → f(sim=37, freq=772), v(sim=36, freq=856), p(sim=34, freq=421), b(sim=33, freq=843), pʰ(
h → h(sim=37, freq=878), f(sim=32, freq=772), ʔ(sim=32, freq=0), pʰ(sim=31, freq=0), s(sim=
iː → i(sim=37, freq=2090), e(sim=36, freq=0), ɪ(sim=35, freq=2420), iɪ(sim=35, freq=0), ei(
j → j(sim=37, freq=1198), i(sim=33, freq=2090), ɪ(sim=33, freq=2420), iɪ(sim=33, freq=0), e
```

Figure 8: phoneme similarity between English and German

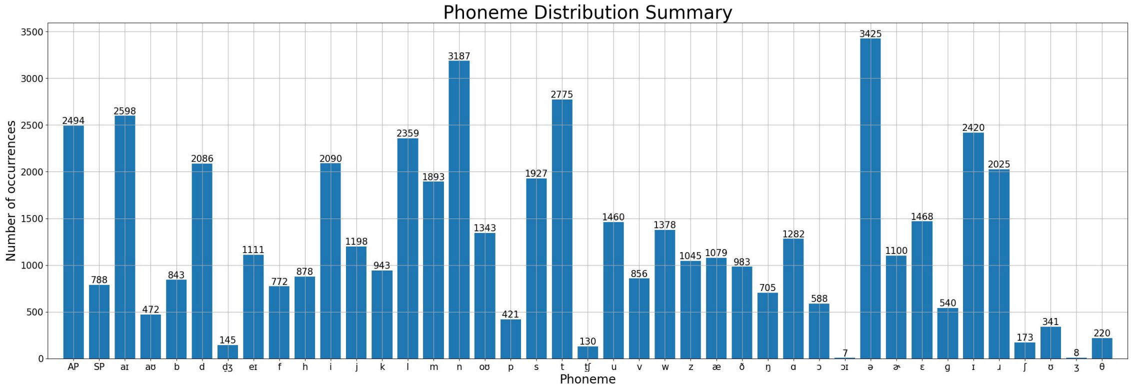


Figure 9: phoneme occurrences in English dataset

4.1.4 Pitch and MIDI Feature Extraction

We used RMVPE [13] to extract F0 contours and voiced/unvoiced flags. MIDI and musical score alignment were processed using the SOME toolkit[14], with appropriate handling of tempo and note duration normalization to match input frame rates (hop size = 512).

4.1.5 Data Structuring

For each experiment condition, we prepared training, validation manifests following DiffSinger's expected CSV structure. To simulate low-resource conditions, we selected 5 stimulies from each dataset and prepare 8 stimulies from unseen songs, preserving pitch, tempo, and lyrical diversity. A fixed random seed (42) was used to ensure reproducibility.

4.1.6 Software and Environment

All preprocessing scripts were implemented in Python 3.9 and run on a high-performance compute cluster with 4 A100 GPU, and CUDA 11.8. Data handling leveraged librosa for audio loading, jsonlines for manifest generation, and NumPy/Pandas for analysis.

4.2 Data Splitting

We maintain consistency by using the same train/validation/test set.

4.2.1 Train & Validation & Test Subsets

Given the limited amount of data and the need for controlled comparisons across models under low-resource conditions, we designed a special test set includes both validation set as parallel stimulus and out-of-domain samples as non-parallel stimulus.

The test set consists of 17 utterances from the following sources, with approximately equal distribution from the following sources:

- **Parallel (15 short utterances)**: Five utterances were sampled from each of the three fine-tuning datasets (GTSinger, ProficientWide, NativeNarrow), serving also as model-aligned parallel validation set.
- **Non-Parallel (2 long utterances)**: Compiled from publicly sourced German singing recordings, used to assess generalization to out-of-domain data. These utterances were manually preprocessed and aligned to enable inference.

All systems are validated with parallel validation subsets. The test set contains both parallel validation subsets and the non-parallel samples. For FFE and MCD in objective evaluation we use only parallel validation set as test samples, for objective WER and subjective evaluation we use both parallel and non-parallel samples, this allows us to assess both matched and mismatched adaptation scenarios.

This evaluation design supports both fair quantitative comparisons and realistic subjective assessments for RQ1 and RQ2.

[13] `https://github.com/Dream-High/RMVPE`

[14] `https://github.com/openvpi/SOME`

4.3 Experiments

4.3.1 Training Configuration

Training hyperparameters were configured based on the OpenVPI DiffSinger codebase. Key optimization parameters included:

- Learning rate: `0.0006`, with StepLR scheduler (step size = 10k, gamma = 0.75)
- Mixed precision: `16-mixed` for faster training and reduced memory usage
- Max updates: `160000`; early stopping triggered manually based on loss plateau
- Aux. decoder gradient: `0.1` (`lambda_aux_mel_loss=0.2`)

4.3.2 Augmentation

To enhance training diversity, random pitch shifting (`range: [-5, 5]`, `scale: 0.75`) and time-stretching (`range: [0.5, 2.0]`, `scale: 0.75`) were enabled during training. After augmentation, data is expanded to 2.5 times of the original data size.

4.3.3 Training Monitoring:

TensorBoard was used to track training and validation loss. Training was halted manually when no further improvement was observed across multiple checkpoints (see Figure 10).

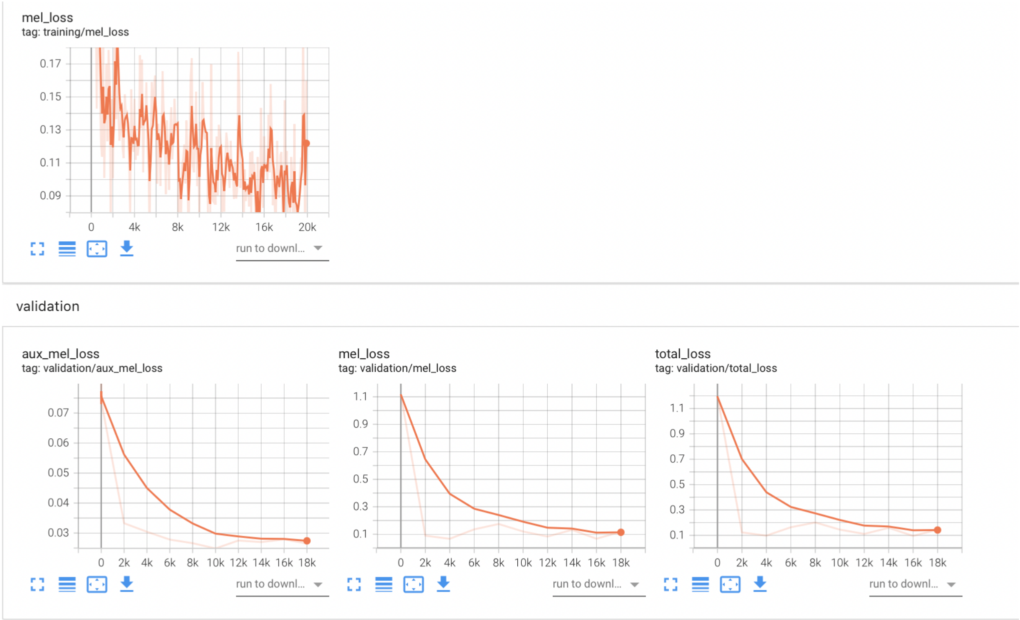


Figure 10: Training loss monitoring via TensorBoard.
Model was stopped when loss plateaued.

4.3.4 Experiment 1: Phoneme-Mapped Transfer with Different Data Sizes (RQ1)

This experiment evaluates whether a DiffSinger model pretrained on English singing data can be adapted to German using PHOIBLE-based phoneme mapping and a small amount of German training data. We aim to assess whether such fine-tuned models can reach comparable performance to a model trained from scratch on a large-scale German dataset. This directly addresses **RQ1** and tests **H1**. We compare four systems under varying data resource settings:

- **Base GTs 3H**: Trained from scratch using 3 hours of German GT-Singer data; serves as the upper-bound baseline.
- **FT GTs 30min**: English-pretrained model fine-tuned on 30 minutes of phoneme-mapped German data.
- **FT GTs 15min**: Same setup with only 15 minutes of fine-tuning data.
- **Zero-shot**: English-pretrained model evaluated directly on German phoneme input without fine-tuning, included as a lower-bound baseline.

Valuation was performed on validation set and evaluation was performed on test set comprising 17 stimuli, including both parallel and non-parallel samples across datasets (Section 4.2). This ensures consistency and comparability across systems.

All systems use the same model architecture and training configuration described in Sections 4.3.1 and 4.3.2. PHOIBLE-based phoneme mapping (Section 3.2) was applied consistently. The same pitch extraction (RMVPE) and MIDI processing pipelines were used across experiments.

All training and inference were conducted on the same compute environment as detailed in Section 4.1.6. Convergence was monitored using TensorBoard (see Figure 10) , and early stopping was triggered manually upon loss plateau. Training configurations are summarized in Table 2.

Table 2: Training Configurations and Runtime for Experiment 1

Model	**Training Description**	**Steps**	**Batch Size**	**Runtime**
Base GTs 3H English	Trained from scratch on 3 hours English GTSinger data	14k	10*	2.5 hours (4 GPU)
Base GTs 3H German	Trained from scratch on 3 hours German GTSinger data	27k	64	3.0 hours (3 GPU)
FT GTs 30min German	Fine-tuned on 30 min German GTSinger data, initialized from Baseline English model	20k	64	1.0 hour (4 GPU)
FT GTs 15min German	Fine-tuned on 15 min German GTSinger data, initialized from Baseline English model	18k	64	40 min (4 GPU)

*Reduced due to limited HPC resource availability.

Following the evaluation methodology in Section 3.4, we employ Objective metrics and Subjective metrics

4.3.5 Experiment 2: Impact of Fine-Tuning Data Quality on SVS Performance (RQ2)

This experiment explores the influence of overall fine-tuning data quality—covering speaker nativeness, vocal range, audio fidelity—on the perceived quality of synthesized singing in data-scarce scenarios. This corresponds to **RQ2** and empirically evaluates **H2**. We compare the following four fine-tuning configurations, each using 15 minutes of German singing data:

- **FT GTs 15min**: A subset of GT-Singer featuring clean studio recordings but limited pitch diversity and non-native accent.
- **FT NativeNarrow 15min**: Data from a native German speaker with moderate vocal range and variable recording quality, reflecting realistic conditions often found in publicly available sources.
- **FT ProficientWide 15min**: Data from a non-native but proficient speaker with a wide vocal range, recorded in a quiet environment using high-quality equipment. Serves as the upper-bound reference in low-resource settings.
- **Zero-shot**: An English-trained model evaluated on mapped German phonemes without any adaptation. Included as a lower-bound baseline to measure the effect of fine-tuning.

The Native-Narrow and Proficient-Wide datasets were segmented and annotated manually to ensure alignment consistency. As discussed in Section 4.1.1, Native-Narrow recordings were sourced from public YouTube content and exhibit variable recording fidelity, while Proficient-Wide was self-recorded in a quiet environment with a broader pitch range (see Figure 4).

Valuation was performed on validation set and evaluation was performed on test set comprising 17 stimuli, including both parallel and non-parallel samples across datasets (Section 4.2). This ensures consistency and comparability across systems.

All models are initialized from the same English-pretrained DiffSinger checkpoint and share identical architecture and training settings (see Sections 4.3.1 and 4.3.2). Each variant is fine-tuned on its respective 15-minute dataset for up to 18k steps, using batch size 64 and mixed-precision training.

Training was conducted in the same environment described in Section 4.1.6. Each 15-minute model required approximately 40–50 minutes of training on 4 A100 GPUs. Convergence was monitored with TensorBoard (see Figure 10), and early stopping was applied based on validation loss trends.

Table 3: Training Configurations and Runtime for Experiment 2

Model	**Recording Condition**	**Accent**	**Vocal Range**	**Steps**	**Batch Size**	**Runtime**
FT GTs 15min German	clean, consistent	Strong accent	Mid	18k	64	40 min (4 GPU)
FT NativeNarrow 15min German	mixed, occasional noise	Native	Narrow	20k	64	50 min (1 GPU)
FT ProficientWide 15min German	clean, consistent	Near-native	Wide	35k	64	40 min (1 GPU)

We use the same evaluation methodology as in Experiment 1 (see Section 3.4):

5 Results

This section presents the experimental results that evaluate the effectiveness of the proposed singing voice synthesis models under different training conditions. Both objective and subjective evaluations were conducted to assess two key hypotheses regarding model fine-tuning and training data composition in Section 5.1 and in Section 5.2. In Section 5.3 there is a short summary.

5.1 Experiment 1: Phoneme-Mapped Transfer with Different Data Sizes (RQ1)

This section evaluates H1, which investigates whether a DiffSinger model pre-trained on English singing can be successfully adapted to German using only a small amount of phoneme-mapped German data. Specifically, we compare two fine-tuned models (FT GTs 15min is trained on 15min and FT GTs 30min is trained on 30 minutes of German data, respectively) to two baselines: a model trained from scratch on three hours of German data (Base GTs 3H), and a zero-shot inference condition with no German fine-tuning. Both objective metrics and subjective listening tests are used to assess whether limited-data fine-tuning achieves comparable performance to large-scale monolingual training.

5.1.1 Objective Evaluation

Table 4: Objective Evaluation Results

Model	**MCD (dB)** ↓	**FFE** ↓	**WER** ↓
FT GTs 30min	12.75* ± 1.49	0.090 ± 0.017	0.315
FT GTs 15min	13.26 ± 1.65	0.088 ± 0.022	0.343
Base GTs 3H	13.30 ± 1.77	0.082 ± 0.018	0.404
Zero-shot	19.81* ± 1.74	0.127* ± 0.027	0.609*

*statistically significant different compare to other models ($\alpha = 0.05$)
MCD: FT GTs 30min better than all others ($p<0.05$)
FFE/WER: Zero-shot worse than all fine-tuned models ($p<0.001$)
No other significant differences detected

Table 4 presents the objective evaluation results for H1. Normality assumptions were verified using Shapiro-Wilk tests (p=0.15), confirming the suitability of parametric analysis. A one-way ANOVA followed by Tukey's HSD post-hoc test revealed significant differences in Mel-Cepstral Distortion (MCD) across models. Specifically, FT GTs 30min achieved the lowest MCD (12.75 dB), significantly outperforming both the Base GTs 3H model (13.30 dB, p=0.021) and the FT GTs 15min model (13.26 dB, p=0.038). The zero-shot condition performed significantly worse than all other models (19.81 dB, $p< 0.001$).

For F0 Frame Error (FFE), no significant differences were observed among the three fine-tuned models (p=0.23), though all were significantly better than the zero-shot condition ($p< 0.001$). All fine-tuned models achieved similar levels of pitch accuracy, suggesting that training duration had limited impact on pitch tracking consistency within the synthesized audio.

Word Error Rate (WER) analysis showed that the FT GTs 30min model achieved the lowest average WER (0.315), followed by FT GTs 15min (0.343) and Base GTs 3H (0.404), with the zero-shot condition performing worst (0.609). However, statistical analysis revealed no significant differences in WER among the three fine-tuned models (p=0.42), while all significantly outperformed the zero-shot condition (p<0.001). This suggests that while acoustic fidelity (e.g., MCD) clearly benefits from increased training duration, intelligibility improvements may be less sensitive to fine-tuning duration within this data range.

5.1.2 Subjective Evaluation

To evaluate the perceptual quality of the generated singing voices, two subjective tests were conducted: a MUSHRA-style absolute rating task and a comparative preference (CMOS-style) test. Each test was analyzed across two listener groups: intermediate and native-level German speakers.

Figure 11 shows the distribution of MUSHRA scores assigned to each model by listeners of different language proficiency levels. Figure 12 summarizes the relative preference ratios across models and listener groups.

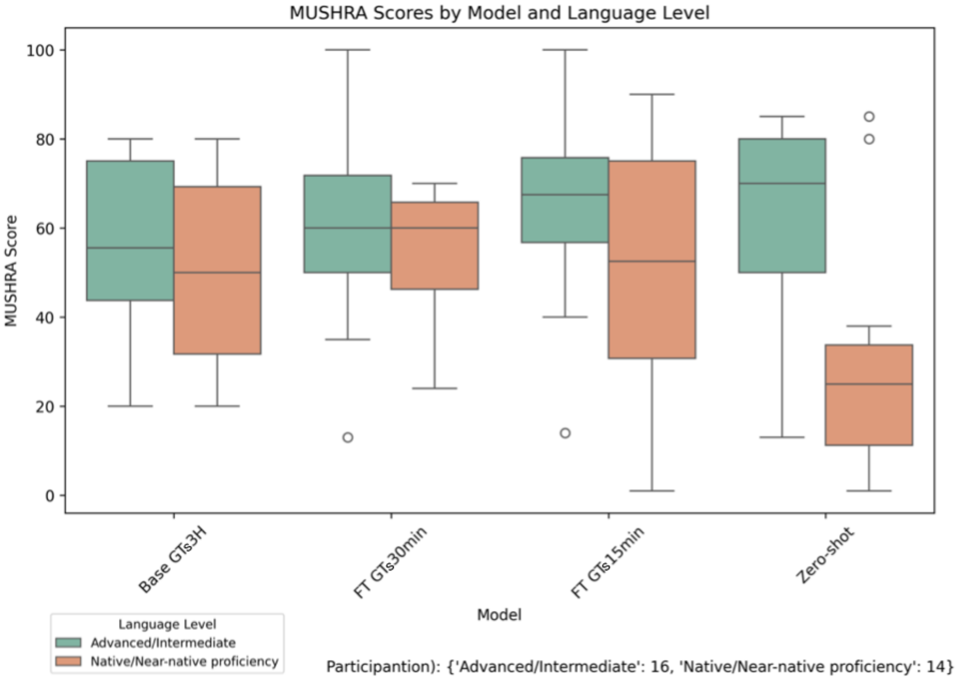
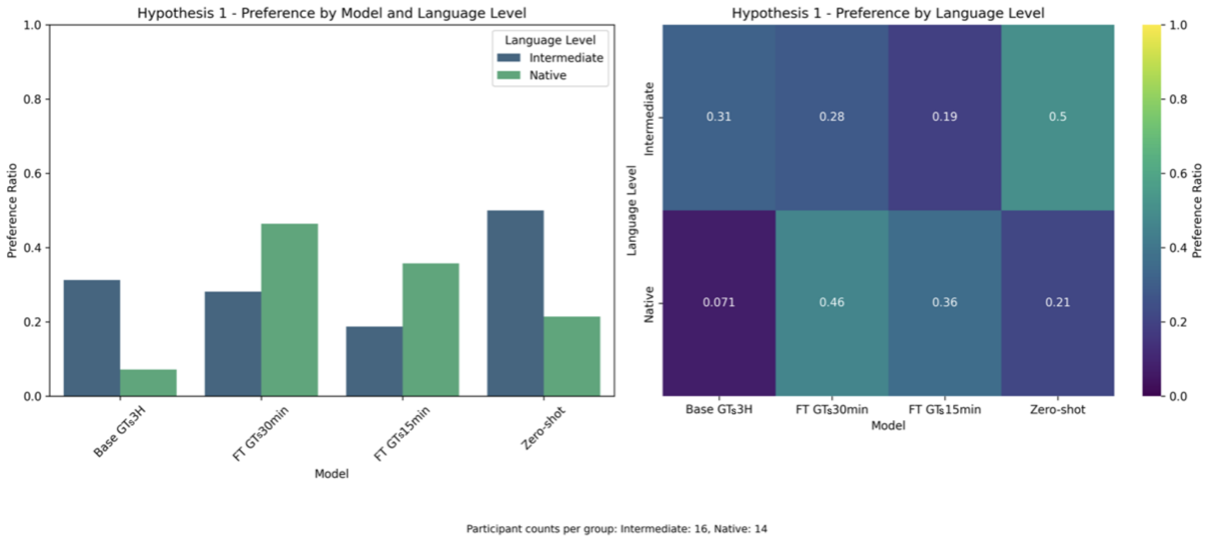


Figure 11: MUSHRA score distribution by model and language level

As shown in Figure 11, the MUSHRA results reveal three main patterns. First, Native speakers consistently provided lower ratings than Intermediate speakers, particularly for the Zero-shot model. While Intermediate listeners rated Zero-shot similarly to other models, Native listeners assigned significantly lower scores, suggesting that Native speakers are more sensitive to pronunciation clarity and singing naturalness.

Second, among the German models, the Baseline GTs 3H model—despite being trained on three hours of data—did not outperform the fine-tuned models trained on significantly less data. This highlights that targeted fine-tuning can be more effective than large-scale training from scratch, provided the data is well matched.

Third, comparing the two fine-tuned conditions, the FT GTs 30min model was generally rated higher than the FT GTs 15min model by Native listeners. While the medians of the two models were close, the FT GTs 15min model showed a wider score distribution, indicating inconsistency across different stimuli. This suggests that although more fine-tuning data improves performance, data quantity is not the only factor—stability and robustness across varied content also matter.



Participant counts per group: Intermediate: 16, Native: 14

Figure 12: Preference ratio and heatmap by model and language level.

The CMOS-style evaluation results (Figure 12) support the same overall trend. In this test, listeners compared two samples and selected which one they preferred using a 7-point scale. These responses were converted into CMOS scores ranging from -3 (strongly prefer B) to $+3$ (strongly prefer A), then aggregated into preference ratios for each model. Native listeners overwhelmingly preferred the FT GTs 30min model, while Intermediate listeners showed more evenly distributed preferences—including surprisingly frequent selection of the Zero-shot model. This again highlights that Native speakers are more linguistically sensitive and more discerning of pronunciation and expressive nuance. Notably, FT GTs 15min was less stable in its preference performance, with more variation across different stimuli compared to FT GTs 30min.

5.2 Experiment 2: Impact of Fine-Tuning Data Quality on SVS Performance (RQ2)

This section evaluates whether fine-tuning data quality affects model performance under low-resource conditions. We compare three DiffSinger models fine-tuned on 15 minutes of German data with different speaker characteristics with zero-shot model: (1) FT ProficientWide 15min (studio-quality, proficient speaker, wide vocal range), (2) FT GTs 15min (studio-quality, accented, narrow vocal range), (3) FT NativeNarrow 15min (variable-quality, native, narrow vocal range), and (4) Zero-shot baseline.

Table 5: Corrected Objective Evaluation Results

Model	**MCD (dB)** ↓	**FFE** ↓	**WER** ↓
FT ProficientWide 15min	13.10* ± 1.19	0.072* ± 0.016	0.175*
FT GTs 15min	13.26 ± 1.65	0.088 ± 0.022	0.343
FT NativeNarrow 15min	14.61* ± 0.78	0.189* ± 0.034*	0.247*
Zero-shot	19.81* ± 1.74	0.127* ± 0.027	0.609*

*Statistically Significant different vs other models ($\alpha = 0.05$)

MCD: ProficientWide vs GTs (p<0.01), NativeNarrow vs GTs (p<0.001), Zero-shot worst (p<0.001)

FFE: NativeNarrow worst (p<0.001), Zero-shot worse than ProficientWide/GTs (p<0.01)

WER: All models vs zero-shot significant (p<0.001); ProficientWide/NativeNarrow vs GTs (p<0.01);

Table 5 presents the objective evaluation results for Experiment 2. For acoustic quality measured by MCD, FT ProficientWide achieved the best performance at 13.10 dB, significantly outperforming both FT GTs (13.26 dB, p=0.008) and FT NativeNarrow (14.61 dB, p¡0.001). The Zero-shot condition showed the worst MCD at 19.81 dB, significantly poorer than all fine-tuned models (p¡0.001).

In terms of pitch accuracy (FFE), FT NativeNarrow exhibited the highest error rate at 0.189, significantly worse than both FT ProficientWide (0.072, p¡0.001) and FT GTs (0.088, p¡0.001). While Zero-shot (0.127) performed better than FT NativeNarrow, it remained significantly worse than the other two fine-tuned models (p¡0.01).

The WER analysis revealed clear improvements in intelligibility. FT NativeNarrow achieved the lowest WER at 0.175, representing a 49% relative improvement over FT GTs (0.343, p<0.001) and a 29% improvement over FT ProficientWide (0.247, p=0.007). FT ProficientWide still showed significantly better performance than FT GTs (p=0.002). All fine-tuned models significantly outperformed the Zero-shot condition (0.609, p<0.001).

5.2.1 Subjective Evaluation

Figure 13 presents MUSHRA ratings grouped by listener language level. The Zero-shot model received consistently low scores across both groups, confirming the importance of target-language fine-tuning.

Among fine-tuned models, we observe diverging preferences based on listener background. Native listeners gave notably higher ratings to FT NativeNarrow, despite its objectively lower quality, likely due to its native accent and prosody. In contrast, intermediate listeners rated FT ProficientWide highest, reflecting stronger sensitivity to vocal richness and clarity.

These trends suggest that native listeners prioritize accent accuracy and prosodic fluency, while non-native listeners focus more on acoustic clarity and expressiveness.

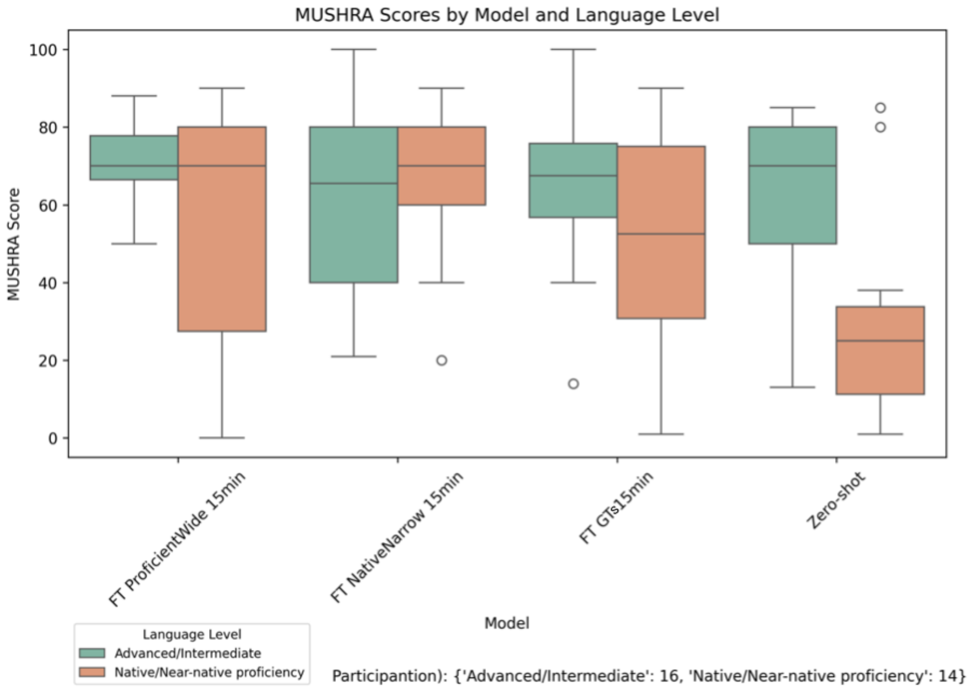


Figure 13: MUSHRA score distribution by model and listener language proficiency.

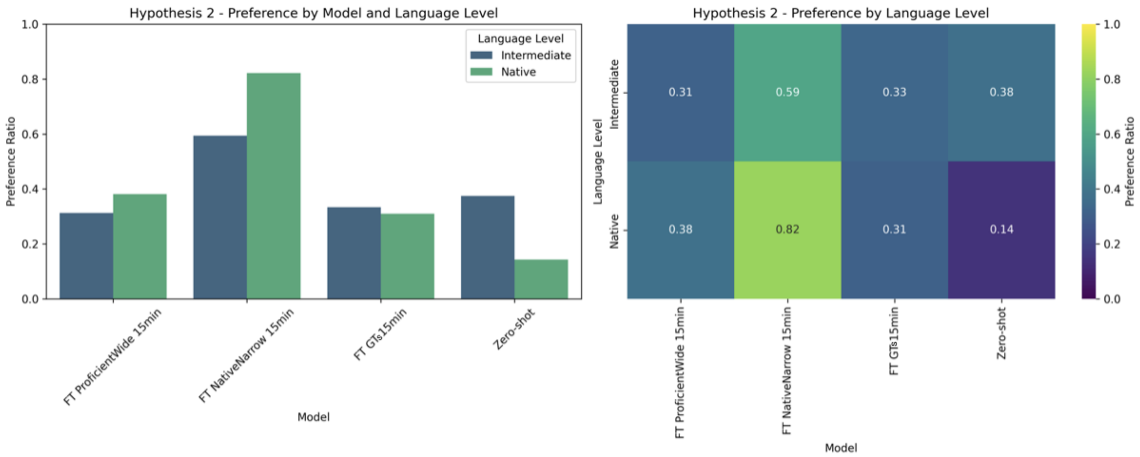


Figure 14: Pairwise preference ratios (CMOS) by model and listener proficiency.

Figure 14 displays pairwise preference ratios (CMOS) for each model. Both native and intermediate listeners showed a strong preference for FT NativeNarrow, especially when contrasted directly with accented or zero-shot outputs. This confirms the perceptual salience of native pronunciation in pairwise evaluations, even if the audio quality is inferior.

Interestingly, intermediate listeners preferred FT NativeNarrow over FT ProficientWide in CMOS, despite the opposite trend in MUSHRA. This suggests that different evaluation formats (absolute vs comparative) emphasize different perceptual cues: direct comparisons tend to highlight accent and articulation, while scale-based judgments reflect broader acoustic impressions.

5.3 Summary of results

The objective evaluation of experiment 1 showed that fine-tuning with limited German data (15-30 minutes) achieved competitive performance compared to training from scratch on three hours of data. The 30-minute fine-tuned model (FT GTs 30min) performed best in acoustic quality (MCD: 12.75 dB), surprisinging significantly surpassing the baseline (13.30 dB). Pitch accuracy (FFE) was similar across fine-tuned models, while intelligibility (WER) showed no significant differences between them. The zero-shot model performed poorly across all metrics, confirming the necessity of fine-tuning.

Subjectively, native German speakers rated models more critically than intermediate speakers, particularly penalizing the zero-shot condition. The 30-minute fine-tuned model was preferred over the 15-minute version, suggesting that even small increases in fine-tuning data improve stability. However, the baseline (trained on 3 hours) did not outperform fine-tuned models, indicating that pre-training and targeted adaptation are in some cases more effective than large-scale training from scratch.

From experiment 2 , objectively we could see the model fine-tuned on studio-quality, proficient speaker data (FT ProficientWide 15min) achieved the best acoustic and pitch metrics (MCD: 13.10 dB, FFE: 0.072). However, the native-speaker model (FT NativeNarrow 15min) had the lowest WER (0.175), highlighting that native accent improves intelligibility despite lower recording quality. The zero-shot model remained the worst performer.

Subjective evaluations revealed a divergence in preferences: native listeners strongly favored the native-speaker model, even with its lower audio quality, due to better pronunciation and prosody. Intermediate listeners, however, preferred the studio-quality recordings (FT ProficientWide) in absolute ratings but still chose the native model in direct comparisons. This suggests that while audio fidelity matters, native accent and articulation are perceptually dominant in pairwise evaluations.

6 Discussion

This section interprets and contextualizes the findings presented in Section 5, linking them back to the research questions (RQ1, RQ2) and hypotheses (H1, H2) introduced in Section 1.1. We synthesize objective and subjective results across both experiments to evaluate the validity of our assumptions, highlight emerging insights (See Section 6.1 and Section 6.2), and reflect on limitation in Section 6.3 and broader implications in Section 6.4 for cross-lingual singing voice synthesis in low-resource settings. A short summary is in Section 6.5.

6.1 Validation of the First Hypothesis

The results provide nuanced support for H1, which posited that fine-tuning a pre-trained English DiffSinger model with limited phoneme-mapped German data could achieve performance comparable to training from scratch on a larger German dataset.

Objectively, the 30-minute fine-tuned model (FT GTs 30min) not only matched but surprisingly surpassed the baseline (Base GTs 3H) in Mel-Cepstral Distortion (MCD), achieving 12.75 dB compared to 13.30 dB ($p = 0.021$). This suggests that the acoustic quality of synthesized singing can benefit from the knowledge transfer enabled by pre-training, even when fine-tuning data is limited. However, the 15-minute fine-tuned model (FT GTs 15min) did not show statistically significant improvements over the baseline, indicating a practical threshold for effective adaptation. The zero-shot condition performed significantly worse across all metrics, reinforcing the necessity of some target-language fine-tuning. These findings align with prior work in speech synthesis (Do et al., 2022), where phoneme mapping facilitated cross-lingual transfer, but extend it to the more complex domain of singing voice synthesis.

Subjective evaluations further illuminated these trends. Native German listeners consistently preferred the FT GTs 30min model in both MUSHRA and CMOS tests, while FT GTs 15min showed more variable performance. This perceptual advantage likely stems from the combination of pre-trained expressiveness and targeted adaptation. The baseline model, despite being trained on three hours of German data, suffered from the limitations of the GTSinger German subset, which includes non-native recordings with accented pronunciation and constrained vocal range. In contrast, the fine-tuned models built upon a pre-trained English model that was trained on higher-quality, more expressive data, enabling better preservation of vocal characteristics even with limited German fine-tuning.

These results affirm that cross-lingual fine-tuning with as little as 30 minutes of target language data can rival or surpass models trained on larger monolingual datasets. However, they also highlight the importance of data sufficiency and quality, as extremely limited fine-tuning (e.g., 15 minutes) may yield inconsistent gains. This insight is particularly relevant for under-resourced languages, where identifying the minimal viable data for effective adaptation is critical.

6.2 Validation of the Second Hypothesis

The results strongly support H2, which proposed that the quality of fine-tuning data—encompassing factors such as accent, vocal range, and recording conditions—significantly impacts synthesis performance in low-resource settings.

Objectively, the FT ProficientWide 15min model, trained on studio-quality recordings by a proficient (though non-native) singer with wide vocal range, achieved the best FFE (0.072) and competitive MCD (13.10 dB). This demonstrates that high recording fidelity and expressive vocal delivery can compensate for near-native pronunciation. In contrast, the FT NativeNarrow 15min model, recorded by a native speaker but under variable conditions with narrow vocal range, performed poorly on pitch accuracy (FFE: 0.189) but excelled in intelligibility (WER: 0.175). This divergence underscores the multifaceted nature of data quality: while native accent enhances linguistic clarity, vocal expressiveness and recording conditions independently contribute to acoustic fidelity. The FT GTs 15min model, trained on the standard GTSinger German subset, lagged behind both in WER (0.343), further emphasizing the limitations of non-native recordings.

Subjective evaluations revealed a striking divergence shaped by listener background. Native listeners strongly preferred FT NativeNarrow in pairwise tests, prioritizing its native-like intonation and accent despite its lower acoustic scores. Intermediate listeners, however, favored FT ProficientWide, valuing its expressive dynamics and clarity over accent purity. This dichotomy aligns with prior findings in speech synthesis (Tomokiyo et al., 2005; Vít et al., 2018), where accent significantly influenced perceived naturalness, but extends them to the singing domain, where additional factors like vocal range and expressiveness play critical roles.

These findings collectively demonstrate that data quality is as important as quantity in low-resource SVS adaptation. Enhancing either pronunciation quality or vocal expressiveness can significantly improve perceived output, even when recording conditions are imperfect. For practical applications, this suggests that data curation should balance accent clarity, vocal diversity, and recording quality based on target listener profiles and use cases.

6.3 Limitations

While the results of this study provide valuable insights into cross-lingual fine-tuning for singing voice synthesis, several limitations should be acknowledged.

First, the scope of the fine-tuning data was constrained by time and availability. Each model in Experiment 2 was trained on only 15 minutes of singing data, which, although realistic for low-resource settings, may not fully capture the expressive or phonetic variability required for robust generalization. Additionally, the datasets differed in multiple dimensions simultaneously (e.g., accent, vocal range, audio fidelity), making it difficult to isolate the effects of individual factors. Future work could use more controlled datasets where only one variable differs at a time to better assess causal effects.

Secondly, subjective evaluations were limited to 30 listeners and the sample size—though sufficient for general trends—may not capture more nuanced listener diversity. Moreover, due to the large number of models and concerns about survey length, we used Whisper ASR for intelligibility instead of transcription tasks. However, Whisper is optimized for speech and performs poorly on singing. Furthermore, as diffusion models produce highly natural results with small differences between them, we did not guide listeners to focus on specific attributes, which may have led to inconsistent interpretations. Future studies should include more structured instructions and evaluation criteria to improve reliability.

Finally, the study was conducted under significant time constraints, limiting the range of experiments (e.g., exploring more fine-tuning durations, speaker combinations, or language pairs). These trade-offs reflect the realities of working with limited computing resources and thesis timelines, but they also suggest several clear directions for expansion.

Despite these limitations, the study's findings remain meaningful within its design scope, and they lay a foundation for future, more controlled, and larger-scale investigations in low-resource multilingual singing voice synthesis.

6.4 Practical Implications

The findings of this study have several practical implications for the development of singing voice synthesis systems, particularly in multilingual and low-resource contexts. First, they demonstrate that it is feasible to adapt a pre-trained English singing model to German using a small amount of phoneme-mapped target data. This approach significantly lowers the data barrier for supporting underrepresented languages in singing synthesis.

Second, the results show that careful selection of fine-tuning data—prioritizing speaker clarity, expressiveness, and recording quality—can have a greater impact than simply increasing the amount of data. This insight is crucial for practitioners working with limited corpora, such as in voicebank creation or personalized voice synthesis.

From an industrial perspective, these findings support the development of more accessible and cost-effective music production tools. By reducing the reliance on large-scale, language-specific datasets, our approach enables faster deployment of SVS models across different linguistic communities. This can benefit independent musicians who seek for a replacement of expensive human demo recording sessions.

Finally, the work contributes to broader societal goals by promoting inclusivity and cultural preservation. Enabling high-quality singing synthesis in underrepresented languages helps preserve linguistic diversity and musical expressions. It also opens opportunities for individuals with speech impairments or limited access to vocal training, offering them creative outlets through synthesized voices.

6.5 Discussion Summary

In summary, The results support H1, showing that fine-tuning a pre-trained English model with just 30 minutes of German data can match or even surpass a model trained from scratch on 3 hours of German data, though smaller adaptations (15 minutes) yield inconsistent gains. H2 was also validated, confirming that data quality (accent, vocal range, recording conditions) significantly impacts performance—expressive vocals compensate for non-native pronunciation, while native accent improves intelligibility. Despite limitations in data scope and evaluation, the findings highlight the feasibility of efficient cross-lingual adaptation, reducing barriers for low-resource singing synthesis and enabling broader applications in music and accessibility.

7 Conclusion

This thesis investigated the feasibility of adapting a singing voice synthesis (SVS) model pre-trained on English to German through phoneme mapping and low-resource fine-tuning. Our work specifically examined how both the duration and quality of fine-tuning data influence synthesis performance in cross-lingual settings. The findings demonstrate that strategic adaptation can overcome data scarcity while revealing important trade-offs between acoustic quality, intelligibility, and listener perception. This conclusion synthesizes the key contributions in Section 7.1, outlines future research directions in Section 7.2, and reflects on the broader implications of this work in Section 7.3.

7.1 Summary of the Main Contributions

The experimental results provide compelling evidence that cross-lingual fine-tuning with limited German data can achieve performance comparable to or exceeding models trained from scratch on larger datasets. Our 30-minute fine-tuned model not only matched but surpassed the baseline German model in objective metrics (12.75 dB MCD vs. 13.30 dB), while subjective evaluations revealed native listeners' strong preference for the adapted model. This success confirms that knowledge transfer from English pre-training can effectively compensate for limited target-language data, though we identified a practical threshold—15 minutes of fine-tuning yielded inconsistent improvements, suggesting a minimum data requirement for reliable adaptation.

The study's second major finding concerns the critical role of data quality in low-resource settings. We observed that different aspects of quality—accent purity, vocal range, and recording conditions—each contribute distinct advantages. Studio-quality recordings by a proficient (though non-native) singer achieved the best acoustic metrics (FFE: 0.072), while native-speaker recordings excelled in intelligibility (WER: 0.175). This divergence was particularly evident in subjective evaluations, where native listeners prioritized accent purity while intermediate listeners favored expressive vocal delivery. These results suggest that practitioners can strategically prioritize different quality dimensions based on their target application and audience.

Methodologically, we developed a practical cross-lingual adaptation pipeline that combines phoneme mapping with minimal fine-tuning—an approach that could be generalized to other language pairs and low-resource scenarios. The consistent performance across objective and subjective measures validates this methodology while highlighting the importance of perceptually-aligned evaluation in singing synthesis.

7.2 Future Work

Several promising directions emerge from this research. First, controlled experiments isolating individual quality factors (accent vs. recording fidelity vs. vocal range) would help establish clearer guidelines for data curation. Our current findings suggest these dimensions interact in complex ways, and targeted studies could quantify their relative importance under different adaptation scenarios.

The methodology should also be tested with more diverse language pairs, particularly those with greater phonological divergence than English and German. Such extensions would reveal whether the current phoneme-mapping approach scales to more challenging cross-lingual transfers or requires modifications for optimal performance.

Evaluation protocols present another important avenue for improvement. Future work should develop perceptually-motivated objective metrics that better align with human judgments, as our results revealed notable discrepancies between traditional acoustic measures and listener preferences. Expanding listener panels to include more diverse demographics and expertise levels would also strengthen findings, particularly for applications targeting specific user groups.

Finally, the practical constraints encountered in this study—limited computing resources and dataset availability—suggest opportunities for community efforts. Curated multilingual singing datasets with controlled variability factors would enable more rigorous comparisons, while standardized evaluation protocols could facilitate cross-study comparisons. Such resources would significantly advance research in this emerging field while lowering barriers to entry for under-resourced languages.

7.3 Impact & Relevance

Beyond technical contributions, this work has important implications for music technology and digital creativity. By demonstrating effective cross-lingual adaptation with minimal data, we lower the barriers for developing singing synthesis in underrepresented languages. This advancement supports cultural preservation efforts and expands creative possibilities for musicians working with less commonly supported languages.

The findings also inform industrial applications, suggesting more cost-effective pathways for developing multilingual SVS systems. Music production tools could leverage our adaptation approach to quickly support new languages without requiring extensive data collection campaigns. For voice banking and assistive technologies, the quality trade-offs we identified provide practical guidance for optimizing limited recording sessions with clients.

At its core, this research speaks to a broader ideal: that creativity and expression should not be limited by language or data availability. Through thoughtful adaptation and design, it is possible to bring expressive, intelligible, and inclusive singing synthesis to more communities, voices, and cultures. This work is one small step in that direction.

References

Ardaillon, L. (2017). *Synthesis and expressive transformation of singing voice* (Unpublished doctoral dissertation). Université Pierre et Marie Curie-Paris VI.

AudioCipher. (2024). *Top ai voice generators for 2024: Make your lyrics sing.* Retrieved from `https://www.audiocipher.com/post/ai-voice-generators` (Accessed May 18, 2025)

Blaauw, M., & Bonada, J. (2017). A neural parametric singing synthesizer. *arXiv preprint arXiv:1704.03809*.

Chandna, P., Blaauw, M., Bonada, J., & Gómez, E. (2019). Wgansing: A multi-voice singing voice synthesizer based on the wasserstein-gan. In *2019 27th european signal processing conference (eusipco)* (pp. 1–5).

Chen, J., Tan, X., Luan, J., Qin, T., & Liu, T.-Y. (2020). Hifisinger: Towards high-fidelity neural singing voice synthesis. *arXiv preprint arXiv:2009.01776*.

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).

Fisher, R. A. (1934). Statistical methods for research workers.

Games, P. A., & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal n's and/or variances: a monte carlo study. *Journal of Educational Statistics*, *1*(2), 113–125.

Gera, S. (2025). *The impact of artificial intelligence on music production: Creative potential, ethical dilemmas, and the future of the industry.* `https://nhsjs.com/2025/the-impact-of-artificial-intelligence-on-music-production-creative-potential-ethical-dilemmas-and-the-future-of-the-industry`. (Accessed: 2025-05-18)

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840–6851.

Huang, R., Cui, C., Chen, F., Ren, Y., Liu, J., Zhao, Z., … Wang, Z. (2022). Singgan: Generative adversarial network for high-fidelity singing voice generation. In *Proceedings of the 30th acm international conference on multimedia* (pp. 2525–2535).

Kenmochi, H., & Ohshita, H. (2007). Vocaloid-commercial singing synthesizer based on sample concatenation. In *Interspeech* (Vol. 2007, pp. 4009–4010).

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, *47*(260), 583–621.

Liu, J., Li, C., Ren, Y., Chen, F., & Zhao, Z. (2022). Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 36, pp. 11020–11028).

Liu, X., Gong, C., & Liu, Q. (2022). Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech* (Vol. 2017, pp. 498–502).

Moran, S., & McCloy, D. (2019). Phoible 2.0. *Jena: Max planck institute for the science of human history*, *10*.

Nakamura, K., Hashimoto, K., Oura, K., Nankaku, Y., & Tokuda, K. (2019). Singing voice synthesis based on convolutional neural networks. *arXiv preprint arXiv:1904.06868*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning* (pp. 28492–28518).

Saino, K., Zen, H., Nankaku, Y., Lee, A., & Tokuda, K. (2006). An hmm-based singing voice synthesis system. In *Interspeech* (pp. 2274–2277).

Sector, I. T. U. T. S. (1996). *Methods for subjective determination of transmission quality.* International Telecommunication Union.

Series, B. (2014). Method for the subjective assessment of intermediate quality level of audio systems. *International Telecommunication Union Radiocommunication Assembly*, *2*.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591–611.

Shi, J., Guo, S., Qian, T., Huo, N., Hayashi, T., Wu, Y., ... others (2022). Muskits: an end-to-end music processing toolkit for singing voice synthesis. *arXiv preprint arXiv:2205.04029*.

Tomokiyo, L. M., Black, A. W., & Lenzo, K. A. (2005). Foreign accents in synthetic speech: development and evaluation. In *Interspeech* (pp. 1469–1472).

Tu, T., Chen, Y.-J., Yeh, C.-c., & Lee, H.-Y. (2019). End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. *arXiv preprint arXiv:1904.06508*.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 99–114.

Vít, J., Hanzlíček, Z., & Matoušek, J. (2018). On the analysis of training data for wavenet-based speech synthesis. In *2018 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 5684-5688). doi: 10.1109/ICASSP.2018.8461960

Wei, H., Cao, X., Dan, T., & Chen, Y. (2023, August). Rmvpe: A robust model for vocal pitch estimation in polyphonic music. In *Interspeech 2023* (p. 5421–5425). ISCA. Retrieved from `http://dx.doi.org/10.21437/Interspeech.2023-528` doi: 10.21437/interspeech.2023-528

Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*, *38*(3/4), 330–336.

Wu, J., & Luan, J. (2020). Adversarially trained multi-singer sequence-to-sequence singing synthesizer. *arXiv preprint arXiv:2006.10317*.

YouTube. (2024). *German singing voice recordings.* Retrieved from `https://www.youtube.com` (Accessed: May 2025. Public performances by native German speakers were sampled for research purposes.)

Zhang, L., Li, R., Wang, S., Deng, L., Liu, J., Ren, Y., ... others (2022). M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. *Advances in Neural Information Processing Systems*, *35*, 6914–6926.

Zhang, Y., Cong, J., Xue, H., Xie, L., Zhu, P., & Bi, M. (2022). Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis. In *Icassp 2022-2022 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 7237–7241).

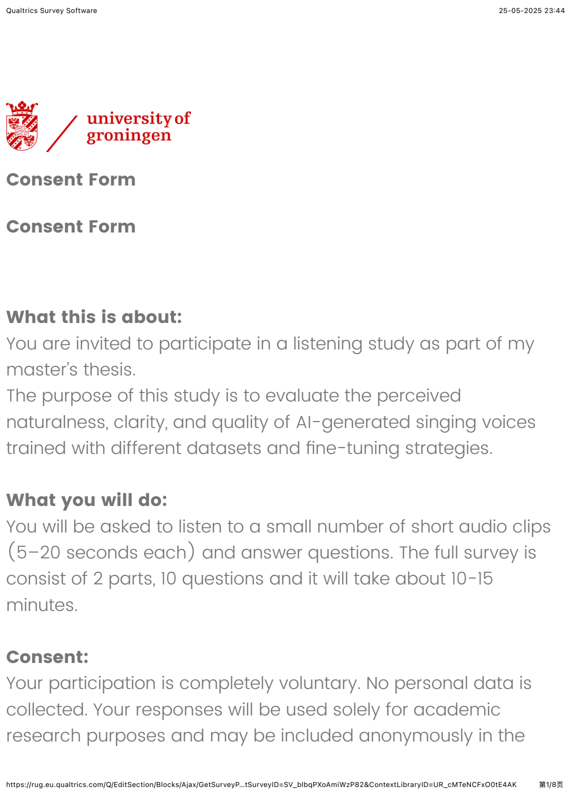
Zhang, Y., Jiang, Z., Li, R., Pan, C., He, J., Huang, R., ... Zhao, Z. (2024). Tcsinger: Zero-shot singing voice synthesis with style transfer and multi-level style control. *arXiv preprint arXiv:2409.15977*.

Zhang, Y., Pan, C., Guo, W., Li, R., Zhu, Z., Wang, J., ... others (2024). Gtsinger: A global multi-technique singing corpus with realistic music scores for all singing tasks. *arXiv preprint arXiv:2409.13832*.

Zhuang, X., Jiang, T., Chou, S.-Y., Wu, B., Hu, P., & Lui, S. (2021). Litesing: Towards fast, lightweight and expressive singing voice synthesis. In *Icassp 2021-2021 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 7078–7082).

Appendices

A Subjective Evaluation



Qualtrics Survey Software 25-05-2025 23:44

university of groningen

Consent Form

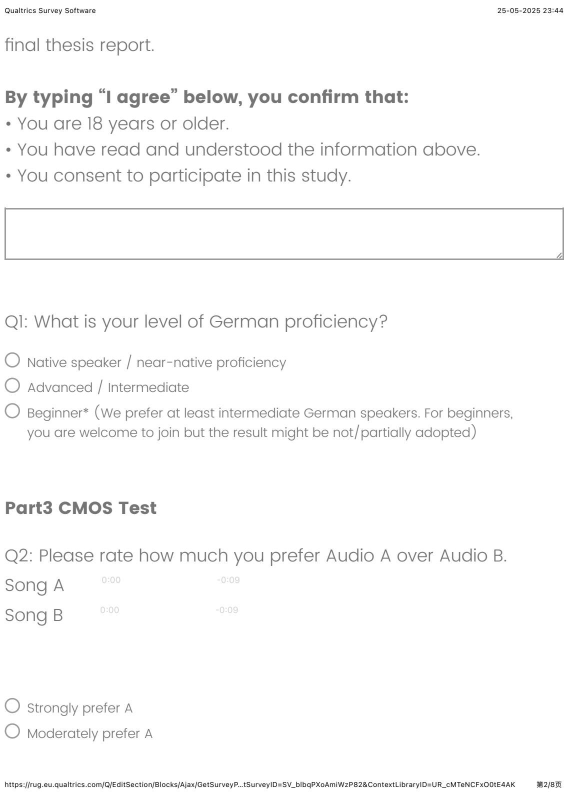
Consent Form

What this is about:
You are invited to participate in a listening study as part of my master's thesis.
The purpose of this study is to evaluate the perceived naturalness, clarity, and quality of AI-generated singing voices trained with different datasets and fine-tuning strategies.

What you will do:
You will be asked to listen to a small number of short audio clips (5–20 seconds each) and answer questions. The full survey is consist of 2 parts, 10 questions and it will take about 10-15 minutes.

Consent:
Your participation is completely voluntary. No personal data is collected. Your responses will be used solely for academic research purposes and may be included anonymously in the

https://rug.eu.qualtrics.com/Q/EditSection/Blocks/Ajax/GetSurveyP...tSurveyID=SV_blbqPXoAmiWzP82&ContextLibraryID=UR_cMTeNCFxO0tE4AK 第1/8页



Qualtrics Survey Software 25-05-2025 23:44

final thesis report.

By typing "I agree" below, you confirm that:
• You are 18 years or older.
• You have read and understood the information above.
• You consent to participate in this study.

Q1: What is your level of German proficiency?

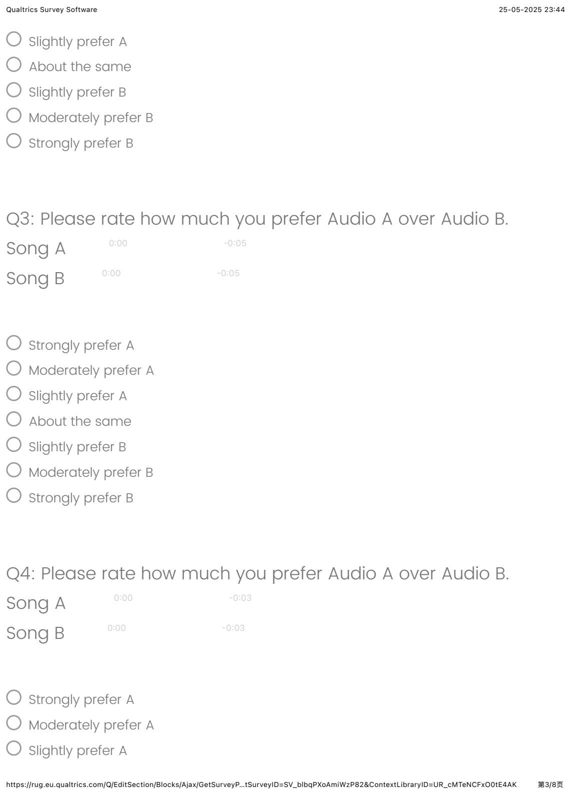
○ Native speaker / near-native proficiency
○ Advanced / Intermediate
○ Beginner* (We prefer at least intermediate German speakers. For beginners, you are welcome to join but the result might be not/partially adopted)

Part3 CMOS Test

Q2: Please rate how much you prefer Audio A over Audio B.
Song A 0:00 -0:09
Song B 0:00 -0:09

○ Strongly prefer A
○ Moderately prefer A

https://rug.eu.qualtrics.com/Q/EditSection/Blocks/Ajax/GetSurveyP...tSurveyID=SV_blbqPXoAmiWzP82&ContextLibraryID=UR_cMTeNCFxO0tE4AK 第2/8页



Qualtrics Survey Software 25-05-2025 23:44

○ Slightly prefer A
○ About the same
○ Slightly prefer B
○ Moderately prefer B
○ Strongly prefer B

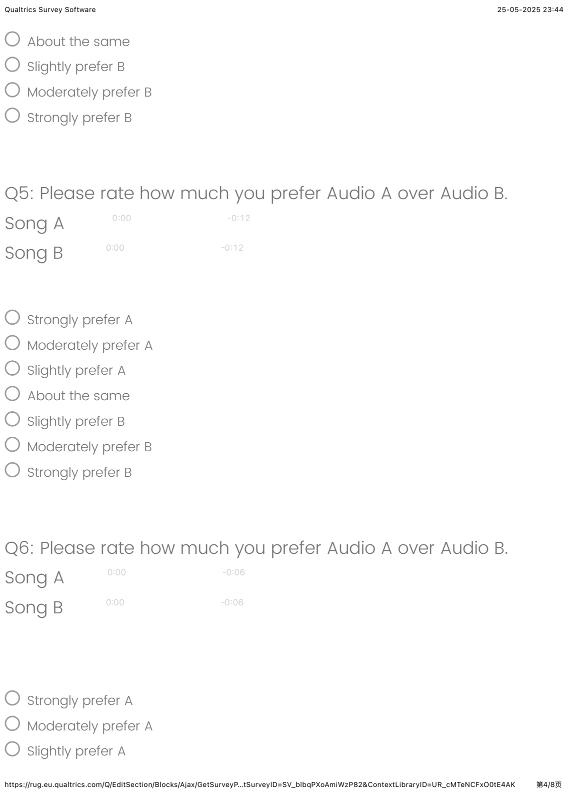
Q3: Please rate how much you prefer Audio A over Audio B.
Song A 0:00 -0:05
Song B 0:00 -0:05

○ Strongly prefer A
○ Moderately prefer A
○ Slightly prefer A
○ About the same
○ Slightly prefer B
○ Moderately prefer B
○ Strongly prefer B

Q4: Please rate how much you prefer Audio A over Audio B.
Song A 0:00 -0:03
Song B 0:00 -0:03

○ Strongly prefer A
○ Moderately prefer A
○ Slightly prefer A

https://rug.eu.qualtrics.com/Q/EditSection/Blocks/Ajax/GetSurveyP...tSurveyID=SV_blbqPXoAmiWzP82&ContextLibraryID=UR_cMTeNCFxO0tE4AK 第3/8页



Qualtrics Survey Software 25-05-2025 23:44

○ About the same
○ Slightly prefer B
○ Moderately prefer B
○ Strongly prefer B

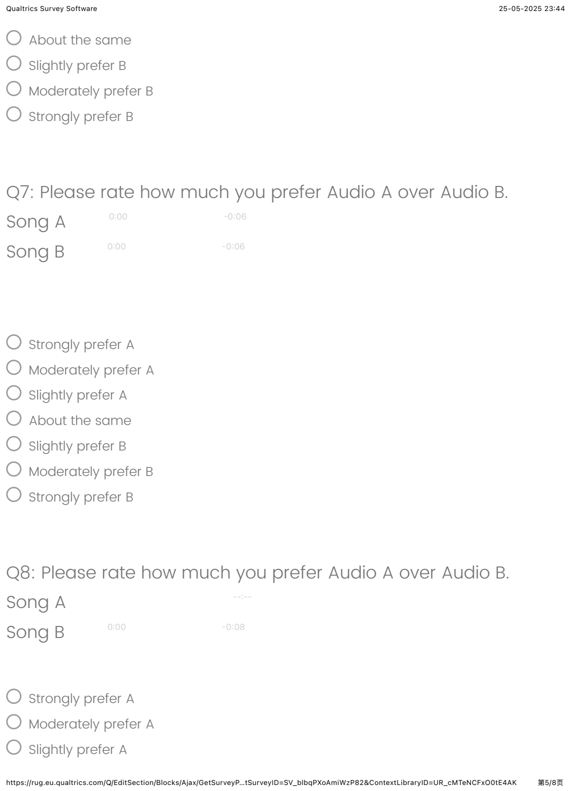
Q5: Please rate how much you prefer Audio A over Audio B.
Song A 0:00 -0:12
Song B 0:00 -0:12

○ Strongly prefer A
○ Moderately prefer A
○ Slightly prefer A
○ About the same
○ Slightly prefer B
○ Moderately prefer B
○ Strongly prefer B

Q6: Please rate how much you prefer Audio A over Audio B.
Song A 0:00 -0:06
Song B 0:00 -0:06

○ Strongly prefer A
○ Moderately prefer A
○ Slightly prefer A

https://rug.eu.qualtrics.com/Q/EditSection/Blocks/Ajax/GetSurveyP...tSurveyID=SV_blbqPXoAmiWzP82&ContextLibraryID=UR_cMTeNCFxO0tE4AK 第4/8页



○ About the same
○ Slightly prefer B
○ Moderately prefer B
○ Strongly prefer B

Q7: Please rate how much you prefer Audio A over Audio B.

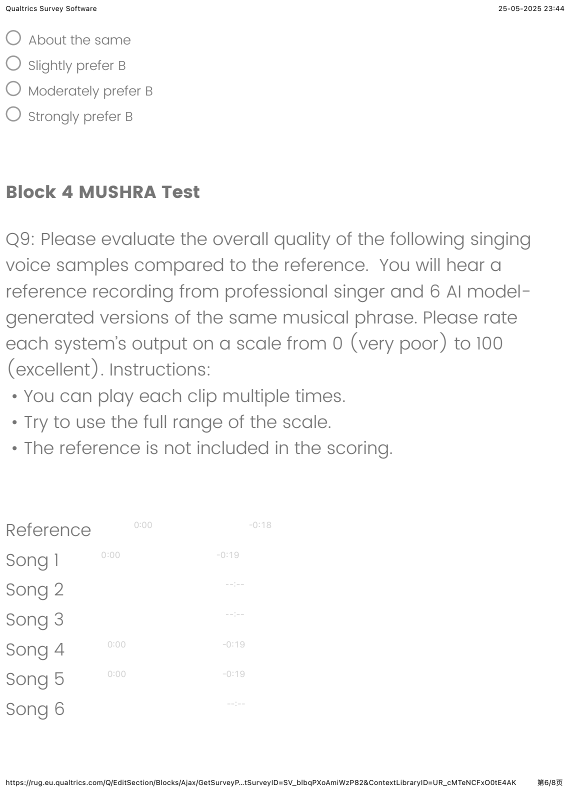
Song A 0:00 -0:06
Song B 0:00 -0:06

○ Strongly prefer A
○ Moderately prefer A
○ Slightly prefer A
○ About the same
○ Slightly prefer B
○ Moderately prefer B
○ Strongly prefer B

Q8: Please rate how much you prefer Audio A over Audio B.

Song A --:--
Song B 0:00 -0:08

○ Strongly prefer A
○ Moderately prefer A
○ Slightly prefer A



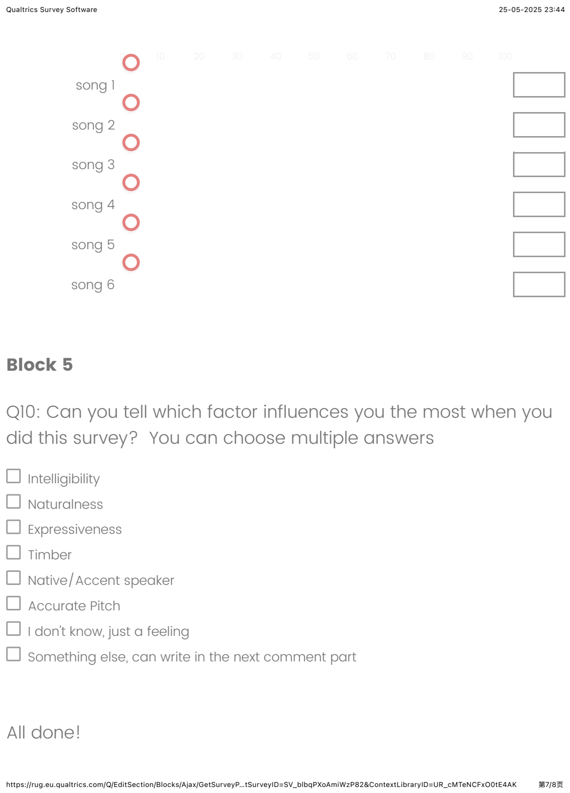
○ About the same
○ Slightly prefer B
○ Moderately prefer B
○ Strongly prefer B

Block 4 MUSHRA Test

Q9: Please evaluate the overall quality of the following singing voice samples compared to the reference. You will hear a reference recording from professional singer and 6 AI model-generated versions of the same musical phrase. Please rate each system's output on a scale from 0 (very poor) to 100 (excellent). Instructions:

- You can play each clip multiple times.
- Try to use the full range of the scale.
- The reference is not included in the scoring.

Reference 0:00 -0:18
Song 1 0:00 -0:19
Song 2 --:--
Song 3 --:--
Song 4 0:00 -0:19
Song 5 0:00 -0:19
Song 6 --:--



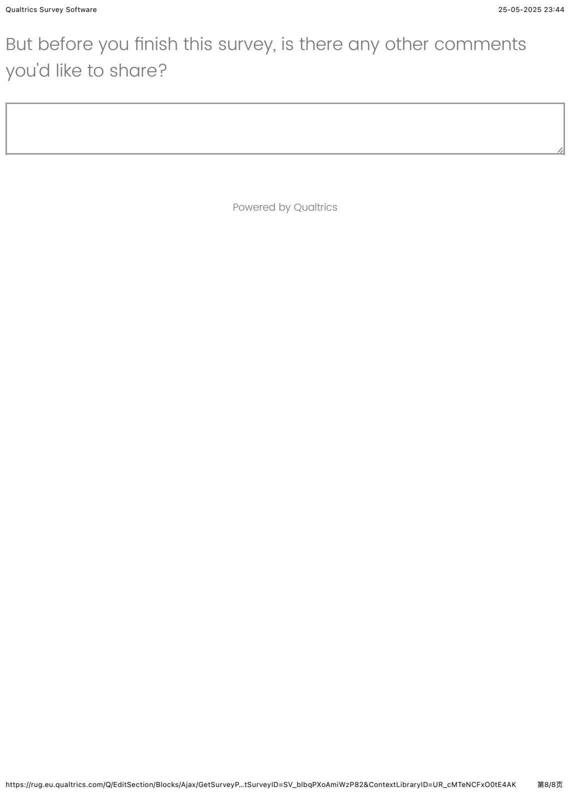
song 1 ○ □
song 2 ○ □
song 3 ○ □
song 4 ○ □
song 5 ○ □
song 6 ○ □

Block 5

Q10: Can you tell which factor influences you the most when you did this survey? You can choose multiple answers

□ Intelligibility
□ Naturalness
□ Expressiveness
□ Timber
□ Native/Accent speaker
□ Accurate Pitch
□ I don't know, just a feeling
□ Something else, can write in the next comment part

All done!



But before you finish this survey, is there any other comments you'd like to share?

□

Powered by Qualtrics

B Declaration of AI Use

This work was supported by the use of *Deepsee and ChatGPT* in the following capacities:

- Facilitating Python code development for training diffusion models, which involved refining dependencies and troubleshooting issues.
- Providing insights on both objective and subjective evaluation methodologies, including recommending suitable approaches and assisting with the structuring of the evaluation scripts.
- Enhancing the analysis of results by identifying significant patterns and organizing data into clear tables and charts.
- Elevating the quality of academic writing through improvements in grammar, clarity, and overall tone across all sections of the document.

All suggestions generated by AI tools were thoroughly reviewed and modified as necessary by me. The experimental design, interpretation of findings, and ultimate conclusions presented in this thesis are entirely my own. I take full responsibility for the content and accuracy of this work.

Name: Jiashu Dong

Date: 11.06.2025