Text vs. Phoneme Intermediates for Low-Resource Swiss German Text-to-Speech

Reza Kakooee* Vincenzo Timmel Daniel Perruchoud Michael Graber Manfred Vogel
Institute for Data Science, School of Computer Science,
University of Applied Sciences and Arts Northwestern Switzerland
reza.kakooee@fhnw.ch

Abstract

Building text-to-speech (TTS) systems for low-resource languages such as Swiss German is challenging due to limited paired data and the lack of standardized orthography. In practical Swiss settings, user input is typically written in High German, motivating pipelines that map High German text to Swiss German speech via an intermediate representation. We compare three approaches: (i) direct synthesis from High German (DE-TTS), (ii) High German → Swiss German text translation followed by synthesis (CH-TTS), and (iii) High German → automatically derived fused phoneme conversion followed by synthesis (PH-TTS). Using the SwissDial dataset, we fine-tune two TTS backbones, SpeechT5 and Orpheus, and evaluate the resulting systems with closed-loop STT metrics (WER/SacreBLEU) and human MOS. Objective transcript-overlap metrics reliably penalize PH-TTS but fail to reflect human preference between DE-TTS and CH-TTS. MOS consistently ranks CH-TTS highest for both backbones, with Orpheus achieving near-original quality and showing robustness when training data is halved; notably, under the half-data setting PH-TTS becomes close to DE-TTS, suggesting that phoneme intermediates may be more competitive in lower-resource regimes. Our analysis indicates that the current PH-TTS pipeline is limited by noisy phoneme supervision and representation mismatch, and we outline directions to make phoneme intermediates competitive in low-resource dialect TTS.

1 Introduction

Neural text-to-speech (TTS) has reached high naturalness in well-resourced settings (Le et al., 2023; Borsos et al., 2023; Wang et al., 2023; Du et al., 2024; OpenAI, 2024; Qwen Team, 2026; Google AI for Developers, 2026; ElevenLabs, 2026; Canopy AI, 2025), but building robust systems for low-resource languages and dialects remains difficult due to the limited availability of paired text–audio data (Chen et al., 2019). Prior work shows that transfer learning and multilingual training can substantially reduce the amount of target-language supervision required, yet performance can still be sensitive to representation choices and data scarcity (Lux et al., 2022).

Swiss German is a particularly challenging target: it is a dialect continuum that is used in spoken form and lacks a standardized orthography. As a result, collecting consistent text labels is difficult, and spelling variation can become a dominant source of noise for text-conditioned speech generation. The SwissDial dataset addresses part of this challenge by providing a parallel multi-dialect corpus with Swiss German audio paired with both Swiss German and Standard (High) German text references (Dogan-Schönberger et al., 2021). In many practical Swiss applications, user input is written in High German, which naturally suggests pipelines that map High German text to Swiss German speech via an intermediate representation.

A common approach is to translate High German into written Swiss German and then synthesize speech from that text. However, because written Swiss German is not standardized, the translation step can introduce variability that is hard for downstream TTS models to resolve, and "correctness" (getting the intended words and meaning) can dominate perceived quality.

This motivates the research question: *should the intermediate representation be Swiss German text at all, or can we benefit from a more acoustically grounded representation such as phonemes?*

Phoneme conditioning can reduce pronunciation ambiguity, but controlled English studies also show that imperfect phonemization introduces practical costs (Fong et al., 2019).

In this work we compare three practical Swiss German TTS pipelines that start from High German input (see Figure 1):

1. **DE-TTS (direct):** High German text → Swiss German speech

2. **CH-TTS (translation):** High German text → Swiss German text → Swiss German speech

3. **PH-TTS (phoneme):** High German text → Swiss German phoneme string → Swiss German speech

For CH-TTS and PH-TTS, the intermediate conversion is implemented using internally fine-tuned T5 models (Raffel et al., 2020).

For PH-TTS, the training data (SwissDial) already contains High German text, Swiss German text, and Swiss German recordings (Dogan-Schönberger et al., 2021). The missing modality is phoneme strings. We therefore derive phoneme supervision from audio using a wav2vec 2.0-based phonemization step (Baevski et al., 2020). The resulting *discrete* phoneme sequences are then converted into more compact *continuous* phoneme strings via a fusion procedure.

We further study the impact of the TTS backbone by fine-tuning two contrasting synthesis models. **SpeechT5** is a unified encoder–decoder model pre-trained across speech and text modalities and serves as a relatively *lightweight* baseline for data-efficient fine-tuning (Ao et al., 2022). In parallel, we evaluate **Orpheus**, a recently released Llama-based "speech-LLM" system that targets high naturalness and expressiveness, serving as a strong synthesis baseline (Canopy AI, 2025). To probe data sensitivity, we train Orpheus on the same data used for SpeechT5, and include an additional Orpheus setting trained on half of the training data.

Finally, we evaluate the synthesized audio quality subjectively. We conduct human listening tests with Mean Opinion Score (MOS), and complement them with a closed-loop objective protocol (text → TTS → audio → STT → transcript) using an internal Whisper model fine-tuned on Swiss German speech (Timmel et al., 2025a,b) that transcribes synthesized audio into High German text for consistent scoring (Radford et al., 2022).

Contributions. Our main contributions are:

- A controlled comparison of three fine-tuned Swiss German TTS pipelines (direct, translation-based, automatically derived phoneme-based) under a shared evaluation harness.
- A practical phoneme-supervision construction pipeline for Swiss German (audio-based phonemization + discrete→continuous fusion) enabling phoneme-intermediate TTS training from SwissDial-style annotations.
- An empirical comparison of a lightweight backbone (SpeechT5) versus a speech-LLM backbone (Orpheus), including a dataset scaling experiment (full vs. half data) to quantify robustness under reduced supervision.
- An evaluation setup combining MOS with closed-loop ASR scoring via Whisper-based transcription into High German.

2 Related Work

TTS for low-resource languages. Neural TTS typically requires substantial paired text–audio data, which is scarce for many languages and dialects. A common strategy is cross-lingual transfer or multilingual training, often with shared symbol spaces to improve data efficiency (**?**Lux et al., 2022). This motivates studying which intermediate representations and model choices remain robust under limited training data.

Phoneme-based TTS and phonemes vs. graphemes across languages. Phoneme conditioning is widely used in neural TTS to make pronunciation explicit, which can be beneficial when orthography is irregular, when pronunciation must be controlled, or when models are transferred across languages; the main cost is reliance on a lexicon and/or grapheme-to-phoneme (G2P) conversion, whose errors propagate to synthesis. In a controlled English study, Fong et al. (2019) compare letter-input and phone-input sequence-to-sequence TTS and show that phone inputs can reduce pronunciation errors, while also quantifying how imperfect phonemes degrade output quality. For languages with more regular spelling, the gap can shrink: Perquin et al. (2021) find that grapheme inputs can match phoneme inputs on a curated French dataset, suggesting that phoneme advantages depend on language and data conditions. In German and German varieties, phoneme (or phoneme/grapheme) tokenization remains common in practical pipelines (Govalkar et al., 2021), and dialect synthesis can combine Standard German G2P with dialect embeddings to handle regional variation under limited resources (Gutscher et al., 2023). In low-resource settings,

phoneme-based transfer learning often relies on mapping phoneme inventories across languages to reduce mismatch (Do et al., 2022), and few-shot cross-lingual TTS can be improved by learning transferable phoneme embeddings in a shared latent space (Huang et al., 2022). More recently, multilingual pretraining directly on phoneme sequences has been proposed to improve data efficiency for TTS in low-resource languages (Nguyen et al., 2023).

TTS backbone model choices. Modern TTS spans diverse architectures (autoregressive, non-autoregressive, flow/diffusion, and increasingly speech generation with large language models). We focus on two complementary backbones. **SpeechT5** is a unified encoder–decoder model pre-trained across speech and text modalities and is attractive as a relatively *lightweight* foundation for data-efficient fine-tuning (Ao et al., 2022). In contrast, **Orpheus** represents a newer *speech-LLM* direction, using a Llama-based backbone for highly natural and expressive speech generation; we include it as a strong, high-quality synthesis baseline to study trade-offs against a compact pre-trained encoder–decoder (Canopy AI, 2025).

3 Method

3.1 Problem setup and pipeline variants

Swiss German TTS is challenging because Swiss German has no standardized orthography and is primarily used in speech. In many Swiss applications, users provide **High German** text, and the system must synthesize **Swiss German** speech. Following this practical setting, we compare three pipeline variants as shown in Figure 1 that differ in the intermediate representation used before speech synthesis.

All variants share the same evaluation harness and differ only in their intermediate conversion and synthesis backbone.

3.2 Data

We use the SwissDial dataset, a parallel multi-dialect corpus of spoken Swiss German with aligned **Swiss German audio** and corresponding **Swiss German** and **Standard (High) German** text (Dogan-Schönberger et al., 2021). This means the training data already contains the modalities required for DE-TTS and CH-TTS (High German text, Swiss German text, recordings). For PH-TTS, the missing modality is the **phoneme-string** representation, which we construct from the audio as described next.

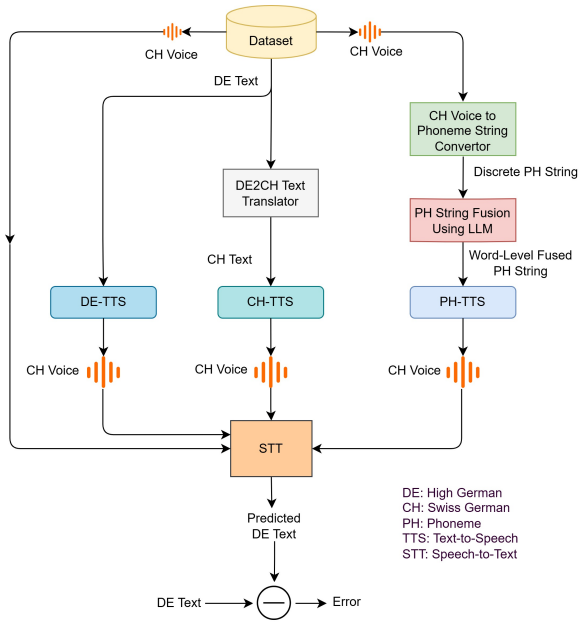

Figure 1: Overview of the three Swiss German TTS pipelines and the closed-loop evaluation. The dataset provides paired Swiss German speech (CH Voice) and High German text (DE Text). DE-TTS synthesises CH Voice directly from DE Text; CH-TTS first translates DE Text to Swiss German text (DE→CH) and then synthesises; PH-TTS derives phoneme supervision from CH Voice via an audio-to-phoneme converter and fuses discrete phonemes into word-level phoneme strings before synthesis. For objective evaluation, all synthesised CH Voice outputs are transcribed by an STT system into predicted DE text and compared against the DE reference to compute error metrics (WER and SacreBLEU).

3.3 Constructing phoneme supervision from audio

Audio-to-phoneme transcription (discrete sequences). We generate initial phoneme strings from Swiss German recordings using a pre-trained wav2vec2-based phoneme recognizer (Wav2Vec2Phoneme) (Xu et al., 2021; Baevski et al., 2020). The raw output forms **discrete phoneme sequences**, where unigraph/digraph tokens are separated by excessive whitespace. While these sequences are usable as model inputs, the redundant spacing increases sequence length and inference latency and can reduce contextual coherence for downstream models.

Discrete-to-continuous fusion (word-like units). To obtain a more compact conditioning signal,

Fusion strategy	Prec.	Rec.	Acc.
Dual wav2vec2 alignment	10.7	10.6	69.4
LLM-based fusion	**94.9**	**93.8**	**98.1**
Phoneme model alignment	30.6	30.0	76.3

Table 1: Space-boundary insertion quality on a manually fused validation samples. Scores are reported in percentages.

we convert discrete sequences into **continuous phoneme strings** by removing unnecessary spaces and re-inserting boundaries primarily at word boundaries. We compare three fusion strategies:

1. **LLM-based fusion:** an LLM is prompted to fuse the discrete phonemes into word-like units while following the Swiss German reference text as closely as possible.

2. **Phoneme alignment via eSpeak:** we derive phonemes from the Swiss German text using eSpeak/eSpeak-NG (eSpeak NG Developers, 2016), align them to the audio-derived phonemes, and transfer whitespace/boundary information from the text side.

3. **Dual wav2vec2 alignment (CTC segmentation):** we use forced alignment based on CTC segmentation to obtain timestamps for (i) text characters and (ii) audio-derived phonemes, then transfer word-boundary spacing from Swiss German text into the discrete phoneme stream (Kürzinger et al., 2020; Baevski et al., 2020).

To select the best fusion strategy, we construct a small gold set of 100 manually fused validation examples. Fusion quality is evaluated by treating *space insertion* as a binary decision between consecutive phoneme symbols and reporting precision, recall, and accuracy of boundary placement. As shown in Table 1, the LLM-based fusion strategy performs best, substantially outperforming the alignment-based alternatives. We therefore use the LLM-based strategy to generate the fused phoneme targets for PH-TTS. The same gold set is also used to tune strategy-specific parameters and prompts; while this validation set is limited in size, the performance differences between strategies are already substantial.

Tokenizer compatibility. The resulting continuous phoneme inventory contains symbols not covered by the default SpeechT5 text tokenizer (Ao et al., 2022). Therefore, we map unsupported phoneme symbols to a set of rarely used "spare" tokens in the tokenizer vocabulary. This enables end-to-end training while keeping the backbone unchanged. In addition, the default SpeechT5 text tokenizer does not cover several German-specific characters (umlauts and the Eszett/sharp s), so we apply a deterministic character mapping before tokenization:

```
ä -> æ,  Ä -> æ
ö -> é,  Ö -> é
ü -> ê,  Ü -> ê
ß-> œ  (and uppercase equivalent)
```

This normalization is applied consistently during training and inference for all text-conditioned components. For **Orpheus**, we do not face these tokenizer limitations; we therefore use the model's original tokenizer without additional character remapping or phoneme-symbol substitution.

3.4 Intermediate translation models

For branches that require intermediate conversion at inference time, we fine-tune T5-style sequence-to-sequence models (Raffel et al., 2020). During **training**, we do not rely on these translators, since the dataset already provides the aligned High German text, Swiss German text, and recordings (and phoneme strings are constructed directly from the recordings as described in Section 3.3).

DE→CH (CH-TTS). We fine-tune a T5 translator on SwissDial to map High German input into Swiss German written text, covering the eight dialects represented in the dataset. This translator provides the intermediate representation consumed by the text-conditioned TTS model in the CH-TTS branch.

DE→phoneme (PH-TTS). We fine-tune a T5 model using SwissDial paired with the **fused phoneme strings** constructed from the recordings and fused by LLM as described in Section 3.3. The model maps High German input *directly* into fused (continuous) Swiss German phoneme strings, i.e., the same format used to condition the PH-TTS models. This avoids a separate discrete-to-continuous fusion step at inference time and ensures the translator output matches the representation seen during TTS training.

3.5 Speech synthesis backbones

We evaluate two synthesis backbones under the same three-branch pipeline structure.

SpeechT5. SpeechT5 is a unified encoder–decoder model pre-trained across speech and text modalities (Ao et al., 2022). We fine-tune SpeechT5 to synthesize Swiss German speech from (i) High German text (DE-TTS), (ii) Swiss German text (CH-TTS), or (iii) fused phoneme strings (PH-TTS), using the same paired supervision from SwissDial (Dogan-Schönberger et al., 2021).

Orpheus. Orpheus is an open-source Llama-based "speech-LLM" TTS system targeting highly natural and expressive synthesis (Canopy AI, 2025). To compare robustness and data-efficiency, we train three Orpheus variants on the same dataset used for SpeechT5, and additionally train one Orpheus variant on half of the training data.

3.6 Objective and subjective evaluation

We evaluate both perceived quality and content preservation.

Closed-loop objective evaluation. We implement a closed-loop protocol: *reference text* → *TTS* → *audio* → *STT* → *transcript*, then compute **WER** and **SacreBLEU** between the transcript and the reference. For STT we use an internal Whisper model fine-tuned on Swiss German speech; synthesized audio is transcribed into **High German** to enable a unified scoring space across all branches (Timmel et al., 2025a; Radford et al., 2022).

Human listening tests (MOS). We run Mean Opinion Score (MOS) evaluations where raters judge short audio samples on a 5-point scale for overall naturalness/quality, comparing original recordings against DE-TTS, CH-TTS, and PH-TTS outputs.

4 Results

We report closed-loop STT-based metrics (WER↓, SacreBLEU↑) and human MOS for DE-TTS, CH-TTS, and PH-TTS. While the objective metrics are useful diagnostics, they do not reliably capture the perceptual ranking between DE-TTS and CH-TTS, motivating MOS as the primary signal for comparison.

4.1 Objective evaluation (closed-loop WER / SacreBLEU)

4.1.1 SpeechT5

Table 2 reports average WER and SacreBLEU for SpeechT5. First, note that the *original recordings* do not yield perfect transcript agreement with the High German reference, indicating non-trivial STT variability even on real audio. Among the synthesised conditions, PH-TTS is clearly worse (higher WER, lower SacreBLEU). In contrast, DE-TTS and CH-TTS are comparatively close under these transcript-based metrics, and DE-TTS appears best by objective scores.

4.1.2 Orpheus: full vs. half dataset

Table 3 shows the same objective metrics for Orpheus trained on the full dataset and on half of the dataset. The same pattern holds: PH-TTS is strongly penalised, while DE-TTS and CH-TTS are harder to separate and remain relatively close in score. Reducing the dataset size affects DE-TTS more noticeably than CH-TTS under these objective metrics.

Why objective metrics are insufficient. Across both backbones, the objective metrics clearly identify PH-TTS as the weakest branch, but they do not provide a clean separation between DE-TTS and CH-TTS. In particular, DE-TTS tends to score best because the evaluation reference is High German text and DE-TTS is optimised to preserve that surface form. CH-TTS, however, includes an explicit DE→CH step and aims to produce more dialect-appropriate Swiss German content before synthesis, which can change lexical choices and reduce transcript overlap even when listeners prefer the result. This motivates the subjective evaluation below.

4.2 Subjective evaluation (MOS)

4.2.1 SpeechT5

Table 4 reports MOS for SpeechT5 (5-point scale). CH-TTS is closest to the original recordings, DE-TTS is lower, and PH-TTS is lowest. Importantly, this perceptual ranking differs from the objective metrics above, which favour DE-TTS. This mismatch reinforces that listening tests are necessary when comparing pipelines with different intermediate representations.

4.2.2 Orpheus and dataset scaling

Table 5 reports aggregated MOS for Orpheus across 4 participants. With the **full dataset**, CH-TTS is the best synthesised condition (4.67), followed by DE-TTS (3.80) and PH-TTS (3.43), while original recordings remain highest overall (4.86). With **half the dataset**, CH-TTS remains comparatively stable (4.56), whereas DE-TTS drops more substantially (3.26). PH-TTS also decreases (3.25) and becomes

System	WER (synth)	sBLEU (synth)
Original (vs. DE ref)	0.235	0.607
CH-TTS	0.237	0.607
DE-TTS	**0.214**	**0.657**
PH-TTS	0.355	0.471

Table 2: SpeechT5 closed-loop objective results (synth vs. High German reference; across 316 samples). WER↓ and SacreBLEU↑ (sBLEU).

System	WER (full)	WER (half)	sBLEU (full)	sBLEU (half)
Original (vs. DE ref)	0.239		0.611	
CH-TTS	0.284	**0.285**	0.556	**0.554**
DE-TTS	**0.254**	0.316	**0.603**	0.540
PH-TTS	0.469	0.496	0.365	0.344

Table 3: Orpheus closed-loop objective results (synth vs. High German reference). Original baseline is identical across runs (across 312 samples).

Condition (SpeechT5)	MOS
Original	4.04 (119)
CH-TTS	**4.00** (124)
DE-TTS	3.53 (127)
PH-TTS	3.10 (124)

Table 4: SpeechT5 MOS results (higher is better); over 6 participants. Numbers in parentheses denote the number of rated samples per condition.

Condition (Orpheus)	Full data	Half data
Original	4.86 (66)	–
CH-TTS	**4.67** (72)	**4.56** (68)
DE-TTS	3.80 (66)	3.26 (73)
PH-TTS	3.43 (76)	3.25 (76)

Table 5: Orpheus MOS results aggregated over 4 participants. Higher is better. Numbers in parentheses denote the number of rated samples per condition.

essentially indistinguishable from DE-TTS under reduced data.

Effect of halving the dataset (Orpheus). Halving training data has the smallest effect on CH-TTS ($\Delta = -0.11$), while DE-TTS shows the largest drop ($\Delta = -0.54$). PH-TTS also degrades ($\Delta = -0.18$), and its separation from DE-TTS largely vanishes in the half-data setting (0.37 on full data vs. 0.01 on half data).

Takeaway. PH-TTS is consistently weakest across both objective and subjective evaluation. For DE-TTS vs. CH-TTS, however, the STT-based metrics tend to favor DE-TTS and provide only a limited basis for ranking systems that differ in intermediate representation. In contrast, MOS consistently prefers CH-TTS. This mismatch suggests that, in Swiss German TTS where intermediate representations can legitimately change lexical choices and surface form, transcript-overlap metrics should be treated primarily as diagnostics rather than the main selection criterion.

5 Discussion

Why subjective evaluation is necessary. Our closed-loop STT-based metrics (WER and SacreBLEU) consistently rank PH-TTS lowest, but they fail to capture the human preference difference between DE-TTS and CH-TTS. This is expected because the objective reference is *High German* text: DE-TTS is incentivised to preserve High German surface form, whereas CH-TTS intentionally alters lexical choices through the DE→CH translation step to produce more dialect-appropriate content. As a result, transcript-overlap metrics conflate "quality" with "literal matching" and provide an incomplete basis for ranking systems with different intermediate representations. This mismatch makes subjective evaluation essential for model selection in the Swiss German setting.

CH-TTS as the strongest pipeline across backbones. Across both TTS backbones, MOS consistently ranks CH-TTS as the best synthesized system, suggesting that perceived quality is dominated by dialect-appropriate wording and meaning preservation rather than strict High-German transcript match. In practice, this supports a pipeline view in which the intermediate representation is the primary driver of user-perceived quality: translating into Swiss German text before synthesis improves overall naturalness and intelligibility as judged by listeners.

Backbone effects: SpeechT5 vs. Orpheus. Compared to SpeechT5, Orpheus achieves higher MOS overall and keeps CH-TTS close to original

recordings, indicating that a strong synthesis backbone can substantially improve naturalness once a suitable intermediate representation is provided. However, the phoneme branch remains behind even with Orpheus, implying that the main bottleneck for PH-TTS is not the synthesis model itself, but the representation and supervision pipeline used to obtain phoneme conditioning.

Data scarcity and robustness. The dataset scaling experiment highlights an additional trade-off. Halving the dataset hurts DE-TTS much more than CH-TTS for Orpheus, suggesting that the translation-based intermediate provides a stronger inductive bias and better robustness under reduced supervision. This also hints that, under scarce data, conditioning on a representation that is closer to the target dialect (Swiss German text) can be more helpful than conditioning directly on High German, even if the latter yields higher transcript overlap against a High German reference.

Fairness of the phoneme comparison. Importantly, PH-TTS should not be interpreted as an oracle phoneme condition: unlike the text-based intermediates available in SwissDial, the phoneme strings are automatically derived from audio and therefore introduce additional supervision noise.

Why PH-TTS underperforms. PH-TTS likely suffers from compounded upstream noise and representation mismatch. In our training pipeline, phoneme strings are inferred from audio and then transformed via discrete-to-fused conversion. Errors introduced by audio phonemization and fusion propagate directly into the conditioning signal seen during training, turning the "phonemes as cleaner supervision" hypothesis into a noisier target in practice. In addition, phoneme symbol coverage and tokenization constraints (notably for SpeechT5) can further degrade conditioning fidelity.

Outlook for phoneme-based Swiss German TTS. The most promising next step for PH-TTS is to remove the discrete→fused phoneme noise at the source. Concretely, this could involve training a model that predicts *fused* phoneme strings directly, adopting a phoneme-native tokenizer, and scaling with more (or cleaner) supervision. These changes would test the underlying hypothesis—that phonemes provide a more stable intermediate than non-standard orthography—without the current error-compounding bottlenecks.

Implications for deployment. For product deployment, CH-TTS with Orpheus appears to offer the best trade-off: it aligns with typical user behavior (input in High German), yields near-original MOS, and remains robust when data is limited. Improving the High German→Swiss German translation component is therefore a high-leverage direction, since translation quality largely determines the content/wording that downstream TTS must realize. *In contrast*, PH-TTS is not yet competitive in perceived quality, but it remains a promising direction if the phoneme supervision and tokenization issues can be addressed.

Acknowledgements

Funding: This study was funded by grants from the Swiss Data Innovation Alliance (Project 04.04.2025-12) and Innosuisse (Project 127.910 INNO-ICT).

References

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738, Dublin, Ireland. Association for Computational Linguistics.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Zalán Borsos, Matt Sharifi, Damien Vincent, Arun Babu, Yongkeun Kim, Bowen Shi, Adam Polyak, Da-Cheng Ju, Aaron van den Oord, and Karen Simonyan. 2023. Soundstorm: Efficient parallel audio generation. *Preprint*, arXiv:2305.09636.

Canopy AI. 2025. Orpheus-tts. `https://github.com/canopyai/Orpheus-TTS`. Open-source Speech-LLM TTS; accessed 2026-02-26.

Yuan-Jui Chen, Tao Tu, Cheng-chieh Yeh, and Hung-Yi Lee. 2019. End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning. In *Interspeech 2019*, pages 2075–2079.

Phat Do, Matt Coler, Jelske Dijkstra, and Esther Klabbers. 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*, pages 16–22, Marseille, France. European Language Resources Association.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. Swissdial: Parallel multidialectal corpus of spoken swiss german. *Preprint*, arXiv:2103.11401.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *Preprint*, arXiv:2407.05407.

ElevenLabs. 2026. Eleven multilingual v2. `https://elevenlabs.io/blog/eleven-multilingual-v2`. Blog post; accessed 2026-03-04.

eSpeak NG Developers. 2016. espeak ng: Open source speech synthesizer.

Jason Fong, Jason Taylor, Korin Richmond, and Simon King. 2019. A Comparison of Letters and Phones as Input to Sequence-to-Sequence Models for Speech Synthesis. In *10th ISCA Workshop on Speech Synthesis (SSW 10)*, pages 223–227.

Google AI for Developers. 2026. Text-to-speech generation (tts) — gemini api documentation. `https://ai.google.dev/gemini-api/docs/speech-generation`. Documentation; accessed 2026-03-04.

Prachi Govalkar, Ahmed Mustafa, Nicola Pia, Judith Bauer, Metehan Yurt, Yigitcan Ozer, and Christian Dittmar. 2021. A lightweight neural tts system for high-quality german speech synthesis. In *14th ITG Conference on Speech Communication*.

Lorenz Gutscher, Michael Pucher, and Víctor Garcia. 2023. Neural speech synthesis for austrian dialects with standard german grapheme-to-phoneme conversion and dialect embeddings. In *2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 68–72.

Wei-Ping Huang, Po-Chun Chen, Sung-Feng Huang, and Hung-yi Lee. 2022. Few-shot cross-lingual tts using transferable phoneme embedding. In *Interspeech 2022*, pages 4566–4570.

Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. 2020. Ctc-segmentation of large corpora for german end-to-end speech recognition. In *Speech and Computer*, volume 12335 of *Lecture Notes in Computer Science*, pages 267–278, Cham. Springer.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. *Preprint*, arXiv:2306.15687. NeurIPS 2023.

Florian Lux, Julia Koch, and Ngoc Thang Vu. 2022. Low-resource multilingual and zero-shot multispeaker TTS. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 741–751, Online only. Association for Computational Linguistics.

Linh The Nguyen, Thinh Pham, and Dat Quoc Nguyen. 2023. Xphonebert: A pre-trained multilingual model for phoneme representations for text-to-speech. *Preprint*, arXiv:2305.19709.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Antoine Perquin, Erica Cooper, and Junichi Yamagishi. 2021. Grapheme or phoneme? an analysis of tacotron's embedded representations. *Preprint*, arXiv:2010.10694.

Qwen Team. 2026. Qwen3-tts technical report. *Preprint*, arXiv:2601.15621.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Vincenzo Timmel, Claudio Paonessa, Manfred Vogel, Daniel Perruchoud, and Reza Kakooee. 2025a. Fine-tuning whisper on low-resource languages for real-world applications. In *Proceedings of the 10th edition of the Swiss Text Analytics Conference*, pages 57–65.

Vincenzo Timmel, Manfred Vogel, Daniel Perruchoud, and Reza Kakooee. 2025b. Swiss parliaments corpus reimagined (spc_r): Enhanced transcription with rag-based correction and predicted bleu. In *Proceedings of the 10th edition of the Swiss Text Analytics Conference*, pages 149–154.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *Preprint*, arXiv:2301.02111.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2021. Simple and effective zero-shot cross-lingual phoneme recognition. *arXiv preprint arXiv:2109.11680*.