VoiceCraft-X: Unifying Multilingual, Voice-Cloning Speech Synthesis and Speech Editing

Anonymous ACL submission

Abstract

We introduce VoiceCraft-X, an autoregressive neural codec language model which unifies multilingual speech editing and zero-shot Text-to-Speech (TTS) synthesis across 11 languages: English, Mandarin, Korean, Japanese, Spanish, French, German, Dutch, Italian, Portuguese, and Polish. VoiceCraft-X utilizes the Qwen3 large language model for phoneme-free cross-lingual text processing and a novel token reordering mechanism with time-aligned text and speech tokens to handle both tasks as a single sequence generation problem. The model generates high-quality, natural-sounding speech, seamlessly creating new audio or editing existing recordings within one framework. VoiceCraft-X shows robust performance in diverse linguistic settings, even with limited per-language data, underscoring the power of unified autoregressive approaches for advancing complex, real-world multilingual speech applications. Audio samples are available at https://voicecraft-x.github.io/.

1 Introduction

Highly realistic speech generation is an indispensable technology for voice assistants, content dubbing, accessibility tools, and creative media. Speech generation can be broken down into several sub-problems: *creating* new audio via Text-To-Speech synthesis (TTS) or *editing* part of an existing recording while ensuring voice consistency with the remainder of the original speech. Despite their shared goal of producing natural speech, TTS and speech editing are typically treated as *separate* problems, especially in multilingual settings, which leaves practitioners without a *single* model that can both edit and synthesize speech across languages.

Over the past several years, the quality of TTS models has improved significantly, particularly in the zero-shot setting in which a model generates speech in a new speaker's voice given a short (e.g. 3 second) audio prompt. Transformer-based neural networks have been central to this progress, leading to three broad paradigms: (i) autoregressive (AR), (ii) non-autoregressive (Non-AR), and (iii) hybrid models. AR models, such as VALL-E (Wang et al., 2023) and its successors (Zhang et al., 2023b; Han et al., 2024; Xin et al., 2024; Chen et al., 2024a; Song et al., 2025; Yang et al., 2025), generate frame-level speech tokens sequentially, where the tokens are typically derived from a neural audio codec (Défossez et al., 2022; Zeghidour et al., 2021; Zhang et al., 2023a). These models are able to perform voice-cloning TTS via Transformer language models' in-context learning ability, demonstrating high-quality speech synthesis. Non-AR models include flow-matching models such as F5-TTS (Chen et al., 2024b), as well as diffusion models such as NaturalSpeech 2/3 (Shen et al., 2023; Ju et al., 2024). These models predict all tokens representing an utterance in parallel via iterative refinement. Hybrid approaches such as Seed-TTS (Anastassiou et al., 2024), CosyVoice (Du et al., 2024b,c) and MaskGCT (Wang et al., 2024) aim to combine the strengths of both paradigms. While these models deliver impressive zero-shot quality, most of the models are either monolingual or focus on a handful of high-resource languages such as English and Chinese. This is likely due to the fact that these models are data-hungry, often requiring 10K-100K hours of training speech for SOTA performance.

The quest for broader linguistic inclusivity across the world's 7,000 spoken languages (Eberhard et al., 2024) has driven research in multilingual speech generation. Efforts include curating large corpora (e.g., VoxPopuliTTS (Liu et al., 2025), Fish-Speech (Liao et al., 2024)) and training multilingual TTS architectures like VoiceBox (Le et al., 2023), CLAM-TTS (Kim et al., 2024) and XTTS (Casanova et al., 2024). Yet even the most capable multilingual systems treat *speech editing* as a separate task—or ignore it altogether—leaving users without a unified solution.

In this paper we address this gap, by introducing **VoiceCraft-X**, a unified autoregressive neural codec language model that performs *both* speech editing and zero-shot TTS in **11 languages**: English (en), Mandarin (zh), Korean (ko), Japanese (ja), Spanish (es), French (fr), German (de), Dutch (nl), Italian (it), Portuguese (pt) and Polish (pl). Our contributions are threefold:

1. We introduce VoiceCraft-X, a single autoregressive model that unifies multilingual speech editing and zero-shot Text-to-Speech (TTS) across 11 languages.
2. Our approach leverages the Qwen3 large language model for cross-lingual text processing, without the need for phonetic pronunciation lexicons. We also propose a novel token reordering mechanism that time-aligns text and speech, enabling a unified sequence generation approach for both editing and synthesis.
3. We demonstrate VoiceCraft-X's robust generation of high-quality, natural-sounding speech across diverse languages, even with limited per-language data, and will release our code and model to the community.

2 Related Work

2.1 Speech Editing

Speech editing aims to correct mispronunciations, stutters, or recording artifacts while producing speech that is indistinguishable from natural audio. Recent approaches leverage Transformer and diffusion architectures. Borsos et al. (2022) perform audio infilling with a Transformer that maintains speaker identity and prosody, generalizing to unseen speakers. Le et al. (2023) use flow matching for versatile speech infilling, and Peng et al. (2024) show that a neural-codec language model with token infilling can concurrently handle editing and synthesis. F5-TTS (Chen et al., 2024b) and MaskGCT (Wang et al., 2024) extend this idea with flow-matching or diffusion, respectively. Despite these advances, most works are monolingual, motivating a unified multilingual solution.

2.2 Zero-Shot Speech Synthesis

The zero-shot Text-to-Speech (TTS) synthesis task entails generating speech in a new speaker's voice from a short audio prompt, without assuming that the new speaker was seen during training. Recent progress is largely driven by Transformer-based neural networks, falling into autoregressive (AR), non-autoregressive (non-AR), and hybrid.

Autoregressive (AR) models generate speech tokens sequentially. VALL-E (Wang et al., 2023) pioneered neural codec language models for high-quality zero-shot TTS via in-context learning, with subsequent works (Zhang et al., 2023b; Han et al., 2024; Chen et al., 2024a; Xin et al., 2024; Song et al., 2025; Kharitonov et al., 2023; Łajszczak et al., 2024; Peng et al., 2024; Guo et al., 2024) further refining this paradigm. Non-Autoregressive (Non-AR) models aim for faster generation by predicting tokens in parallel or using iterative refinement. Examples include flow-matching models like VoiceBox (Le et al., 2023) and diffusion-based models such as NaturalSpeech 2 (Shen et al., 2023), NaturalSpeech 3 (Ju et al., 2024), and DiTTo-TTS (Lee et al., 2024). Other notable non-AR approaches include Unicats (Du et al., 2024a), SimpleSpeech (Yang et al., 2024b,a), E2-TTS (Eskimez et al., 2024), F5-TTS (Chen et al., 2024b) and Mega-TTS 3 (Jiang et al., 2025). Hybrid systems combine aspects of both AR and non-AR methods. Seed-TTS (Anastassiou et al., 2024) uses a two-stage architecture, while CosyVoice (Du et al., 2024b,c) and MaskGCT (Wang et al., 2024) also represent efforts to balance quality, speed, and controllability. In this work, VoiceCraft-X follows the codec language modeling method of VoiceCraft (Peng et al., 2024) and enables high-quality, zero-shot multilingual speech synthesis within its unified editing and generation framework.

2.3 Multilingual Speech Generation

Prior work on multilingual speech synthesis largely pursues two complementary goals: (i) expanding language coverage and (ii) achieving zero-shot robustness to unseen speakers and languages.

On the data side, Saeki et al. (2024) show that pairing self-supervised speech representations with unsupervised text alignment scales TTS to 100+ languages, even when only scant transcriptions exist. Large curated corpora amplify these gains: VoxPopuliTTS (Liu et al., 2025) refines 30,000 hours of English, French and Spanish speech; FishSpeech (Liao et al., 2024) goes further, training on 720,000 hours while using an LLM to sidestep language-specific G2P rules. Model architectures have evolved in parallel. VoiceBox (Le et al., 2023) adopts non-autoregressive flow matching, delivering cross-lingual zero-shot TTS in six languages via in-context learning. XTTS (Casanova et al.,

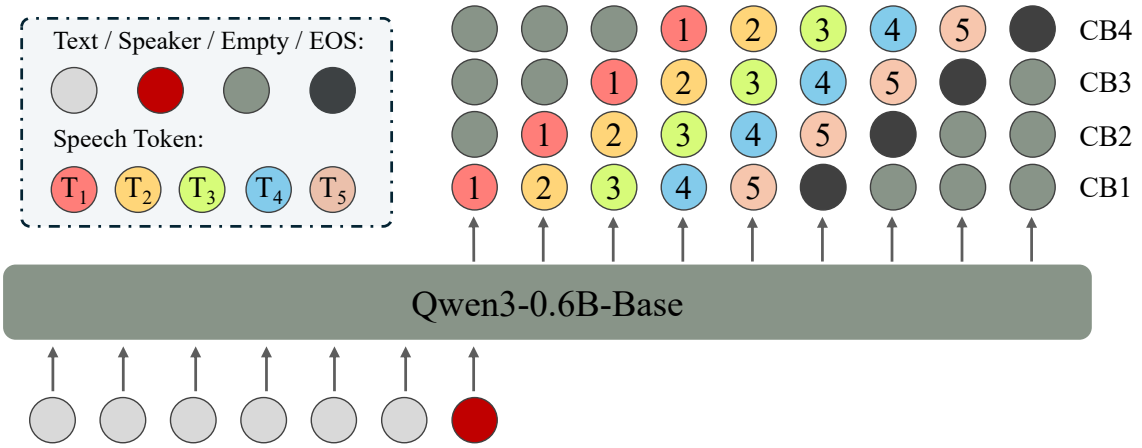

Figure 1: **Architecture Overview.** This diagram illustrates the training process for the VoiceCraft-X model. The model takes text and a speaker embedding as input and is trained to predict sequences of speech tokens. The labels CB1-CB4 represent codec tokens from different codebooks.

2024), extending Tortoise (Betker, 2023), combines a Perceiver Resampler with a speaker-consistency loss to reach 16 languages with speaker cloning. CLAM-TTS (Kim et al., 2024) improves codec language model compression with probabilistic residual vector quantization, enabling single-step multi-token generation. However, these models often treat synthesis as a distinct task from speech editing. The challenge of *unifying* high-quality, multilingual speech editing with robust multilingual speech synthesis within a single, open-source, and fully autoregressive model architecture remains largely unaddressed.

3 Method

3.1 Overview

VoiceCraft-X evolves VoiceCraft (Peng et al., 2024) into a truly multilingual speech-editing and synthesis system, treating both tasks as a single sequence-generation problem over neural codec tokens. The core of this system, as illustrated in Figure 1, is the Qwen3 (Qwen-Team, 2025) large language model. Qwen3 natively supports text input in 119 languages and dialects, which we leverage as the cross-lingual input text tokenizer for VoiceCraft-X. This eliminates the cumbersome phoneme-conversion step that was integral to the original VoiceCraft, resulting in a simplified pipeline with a shared tokenizer across languages, without the need to curate pronunciation lexicons for each language.

A further key innovation in VoiceCraft-X is its enhanced data layout: it interleaves text tokens and speech tokens in a single, time-ordered stream, whereas VoiceCraft reordered only the speech tokens. Enforcing this alignment between linguistic content and its acoustic realization yields more consistent and natural-sounding speech.

3.2 Speaker Embedding

In addition to the speech tokens representing the prompt speech, VoiceCraft-X also takes as input a speaker embedding vector extracted from this prompt speech. We follow the approach of CosyVoice (Du et al., 2024b) by using a pre-trained voiceprint model[1] to extract the speaker embedding. The resulting vector is then passed through a linear projection layer. This projection maps the speaker embedding to match Qwen3's input dimension.

3.3 Speech Tokenization

We utilize the EnCodec (Défossez et al., 2022) neural audio codec model to tokenize the input utterance. Specifically, we train a modified version of the tokenizer which outputs a sequence of four parallel token streams at a 50Hz framerate. The tokens are discretized with residual vector quantization (RVQ) with a vocabulary size of 2048 at each quantization layer.

3.4 Token Reordering

VoiceCraft-X employs several token reordering steps, illustrated in Figure 2, to unify speech editing and synthesis. We assume that our training examples consist of utterance waveforms accompanied by time-aligned word transcriptions (we use the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) in our work). During training, a text transcription is randomly segmented into prefix,

[1] https://www.modelscope.cn/models/iic/CosyVoice-300M/file/view/master/campplus.onnx

middle, and suffix portions. These are then rearranged into a "prefix-suffix-middle" sequence, where the "middle" segment serves as the prediction target. Finally, the corresponding speech tokens for each segment are reordered identically based on the alignment timings. This ensures a monotonic alignment between the text and speech tokens, even when performing speech edits which require infilling tokens in the middle of the speech sequence. This rearrangement serves to mirror the use case in which a user wishes to modify some, but not all of the words in an utterance - by using this rearrangement, the model can be trained to predict the speech tokens within the middle of an utterance, conditioned on the preceding (prefix) and following (suffix) speech tokens in addition to the desired text transcription.

3.5 Causal Masking and Delay Pattern

Following the token reordering, a learnable <*MASK*> token is inserted at two locations within the text-speech input sequence: one <*MASK*> token is inserted at the boundary between the prefix and suffix speech tokens, and a second <*MASK*> token is placed between the suffix audio tokens and the middle (target) audio tokens. These tokens serve to inform the model of the boundaries between the segments.

During training, the model is tasked with autoregressively predicting all audio tokens: encompassing those in the prefix, suffix, and the middle (target) segments. This prediction is optimized using a standard language modeling objective, where the cross-entropy loss function is applied to every token in the sequence. By training the model to predict not only the target segment but also the known prefix and suffix segments, it receives gradients for every timestep, resulting in faster training.

To model the K parallel token sequences output by the EnCodec tokenizer autoregressively, we incorporate the "Delay Pattern" proposed by MusicGen (Copet et al., 2023). Instead of predicting all K codebooks for a given audio timestep t simultaneously or flattening all codebooks across all timesteps into one long sequence, delay patterning inserts a cumulative time delay of one timestep per RVQ layer to the EnCodec token sequences. As a result, the prediction for the speech token at codebook level k at timestep t can be conditioned on the model's predictions for codebook levels 1 through $k-1$ associated with the same timestep t.

3.6 Inference

Figure 2 shows how, at inference time, VoiceCraft-X performs speech editing and zero-shot text-to-speech by preparing an input sequence based on the "prefix-suffix-middle" reordering of text and speech tokens. The system then autoregressively generates the neural codec tokens for the target audio segment.

Speech editing Let T_P, A_P be the prefix text/audio, T_S, A_S the suffix, and T_M^{new} the user-supplied replacement text for the middle segment. The model input is the concatenation

$$T_P,\ T_S,\ T_M^{\text{new}},\ \text{<}\textit{SPK}\text{>},\ A_P,\ \text{<}\textit{M}\text{>},\ A_S,\ \text{<}\textit{M}\text{>},$$

where <*SPK*> is a speaker embedding token and <*M*> is the (learnable) mask token. The decoder predicts the middle-segment audio tokens $\hat{A}_M$, which we splice between A_P and A_S before decoding the entire sequence with the EnCodec decoder network to create a seamless edit.

Zero-shot TTS If a prompt text (T_{prompt}) and its corresponding prompt speech are provided, we concatenate the prompt text and the target text (T_{target}) to form the middle text segment, and a speaker embedding is extracted from the prompt speech. If no such prompt is provided, we set the prompt text (T_{prompt}) to empty and randomly generate a speaker embedding. The final input is as follows:

$$\begin{aligned} &T_P,\ T_S,\ T_{prompt},\ T_{target},\\ &\text{<}\textit{SPK}\text{>},\ A_P,\ \text{<}\textit{M}\text{>},\ A_S,\ \text{<}\textit{M}\text{>},\ A_{prompt}, \end{aligned}$$

where $T_P = T_S = \varnothing$, $A_P = A_S = \varnothing$, and $T_{prompt} = A_{prompt} = \varnothing$ if no prompt is provided.

4 Experiments

4.1 Setup

Training Dataset. We combined speech data across public datasets over 11 languages, amounting to a total of approximately 32K hours (detailed statistics provided in Appendix §A.1). The sampling rate for all audio is 16 kHz. Audio segments longer than 25 seconds were discarded. For MLS dataset (Pratap et al., 2020), misalignment issues were particularly prominent, with approximately 20% of samples having extra or missing words in the transcript at the beginning or end. We found that this negatively impacted model performance for English, and subsequently removed utterances whose transcriptions differed significantly from

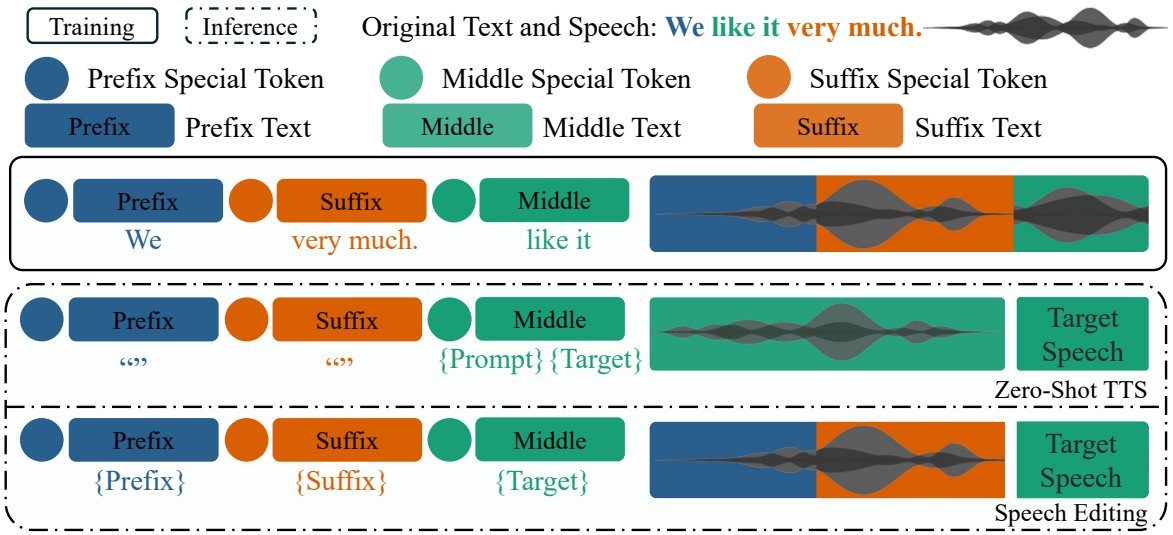

Figure 2: Illustration of Token Reordering

those produced by the Whisper (Radford et al., 2023) model. While we found similar problems with the non-English European language data in MLS, we anecdotally observed better performance on those languages without performing this filtering. We speculate that this is due to the fact that the amount of available training data for those languages is already relatively low, and the performance improvements brought by the additional training data outweigh the detriments brought by transcription noise.

Evaluation Dataset. For evaluating Text-to-Speech (TTS) performance, we curated an evaluation dataset from several established benchmarks. For English, we utilized the Seed-TTS test-en set (Anastassiou et al., 2024) (1088 samples sourced from Common Voice (Ardila et al., 2019)). For Mandarin, we employed the Seed-TTS test-zh set (2020 samples from DiDiSpeech (Guo et al., 2021)). Korean and Japanese evaluations were conducted using 200 randomly selected samples from KsponSpeech (Bang et al., 2020) and KokoroSpeech (Iida, 2021), respectively. For the remaining seven languages supported by our model (Spanish, French, German, Dutch, Italian, Portuguese, and Polish), we randomly selected 100 samples for each language from their corresponding Multilingual LibriSpeech (MLS) (Pratap et al., 2020) test sets. To evaluate speech editing, we randomly selected 100-300 samples per language from these TTS test datasets and then utilized Gemini (Team et al., 2023) to perform insertion, deletion, or substitution operations on the textual portions of these samples, with specific details available in the appendix §A.2. We conducted subjective evaluation over a subset of languages (English, Chinese, French, Italian, Portuguese, and Spanish) using a random subset of the evaluation set: 40 English samples, 50 Chinese, and 20 for others.

Training. Our model utilizes Encodec (Défossez et al., 2022) as the speech tokenizer. We retrain the model with some modifications, namely using 4 Residual Vector Quantization (RVQ) codebooks, each containing 2048 entries, and a framerate of 50Hz on audio recorded at 16 kHz. We retrain the model with our multilingual speech data. Other than those, the training process adheres to the methodology outlined in the work by (Défossez et al., 2022). Additional configuration specifics can be found in Section §B.1. To combine the parallel speech tokens when using them as input to the Transformer LM, at each timestep we sum the embeddings of the tokens across the four codebooks.

We use Qwen3-0.6B-Base as both the text tokenizer and the Transformer LM backbone (details are provided in Appendix B.2). The outputs from the final Transformer layer are then projected into four distinct linear layers, each producing the logits for one of the codec tokens. The model comprises 613 million total parameters (457 million excluding embeddings). The codebook weights $\boldsymbol{\alpha}$ are set to $(1.0, 0.8, 0.6, 0.4)$, influencing the contribution of each codebook during training (as further detailed in our loss formulation §B.3). For model training, we employ the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 4×10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, an epsilon of 1×10^{-6}, and a weight decay of 0.01. A learning rate scheduler is utilized, featuring a linear warm-up for the initial $50K$ steps, followed by a linear decay

for the remainder of the $5,000K$ total training steps. Gradient accumulation is performed over 8 micro-batches. The training of the multilingual VoiceCraft-X model took approximately one week on 16 NVIDIA A100 40GB GPUs.

Inference Figure 2 shows how, at inference time, VoiceCraft-X performs speech editing and zero-shot text-to-speech by preparing an input sequence based on the "prefix-suffix-middle" reordering of text and speech tokens; the model then autoregressively predicts the corresponding neural codec tokens for the target audio segment. Notably, the token reordering mechanism significantly enhances inference stability. This largely prevents repeating token loops, an issue in the original VoiceCraft which could cause artifacts (e.g., excessive silences) and required multi-sample filtering. Consequently, VoiceCraft-X reliably generates high-quality speech in a single pass without needing this filtering step. In all experiments, we employ nucleus sampling (Holtzman et al., 2019) with $TopK = 20, TopP = 1.0$, and a temperature of 1.

Baselines. For the English and Chinese Zero-shot TTS tasks, we compared our model with FireRedTTS (Guo et al., 2024), MaskGCT (Wang et al., 2024), F5-TTS (Chen et al., 2024b), CosyVoice (Du et al., 2024b), and CosyVoice 2 (Du et al., 2024c). For English, we also included VoiceCraft (Peng et al., 2024) in our comparison. For the remaining languages, we benchmarked our model against the multilingual XTTS (Casanova et al., 2024) model, considering both its v1 and v2 versions. For speech editing, we compared VoiceCraft-X with the original VoiceCraft (Peng et al., 2024) model on English.

Metrics. We used a combination of subjective and objective measures. Objectively, we use Word Error Rate (WER) as an automatic proxy for the intelligibility of the synthesized speech; this is calculated using Paraformer-zh (Gao et al., 2023) for Chinese and Whisper-large-v3 (Radford et al., 2023) for other languages. Additionally, speaker similarity (SIM-o) is objectively measured by computing the cosine similarity of speaker embeddings, which are extracted from both the generated and original target speech using a WavLM-based speaker verification model (Chen et al., 2022). Subjective evaluations involved human annotators (see Appendix C for details) who provide Comparative Mean Opinion Scores (CMOS) and Similarity Mean Opinion Scores (SMOS) for TTS, and Naturalness Mean Opinion Scores (NMOS) and Intelligibility Mean Opinion Scores (IMOS) for speech editing. For CMOS, evaluators assess the naturalness of the synthesized speech in comparison to the ground truth, while for SMOS, they directly score the similarity between the synthesized speech and the initial speech prompt. For NMOS and IMOS, evaluators respectively assess the naturalness and intelligibility of the synthesized and original speech.

4.2 Zero-Shot TTS

We evaluated VoiceCraft-X's zero-shot TTS performance across 11 languages, and the results are shown in Table 1. For Chinese, VoiceCraft-X was trained on a modest $5K$ hours of data, a fraction of that used by leading models (often exceeding $50K$ hours). Consequently, while its CER of 3.29 was higher than these specialized models, this was achieved with substantially less data, and its speaker similarity and subjective scores reflected this data disparity. In English, VoiceCraft-X, trained on $14K$ hours, showed marked improvements over its predecessor, VoiceCraft, reducing its WER from 5.28 to 4.37 and enhancing SIM-o from 0.51 to 0.54. Critically, its CMOS score of 0.63[2] was the highest among compared models, indicating superior perceived naturalness. While some models trained on significantly larger datasets achieved lower WERs, VoiceCraft-X's subjective quality in English was highly competitive.

For the remaining nine languages, VoiceCraft-X, compared to XTTS (versions v1 and v2), showed strong overall performance with varying focuses. VoiceCraft-X particularly excelled in European languages like German (WER significantly better than XTTS-v2 by over 50%), Spanish (WER over 40% better than XTTS-v2 and below the ground truth), and Italian (higher data efficiency), as well as in Korean (CER reduced by over 20%). However, in languages such as Japanese and Dutch, or for those where VoiceCraft-X had considerably less training data like Portuguese and Polish, XTTS-v2 achieved lower error rates. Nevertheless, VoiceCraft-X was often favored by evaluators for its better speaker similarity, naturalness, and intelligibility. (Further results are in the appendix §C).

[2]The generally higher English CMOS scores likely resulted from using Seed-TTS test set as prompts with atypical, exaggerated intonation (not standard read speech).

Table 1: Zero-Shot TTS performance across different models and languages. ‡*Training Hours* for XTTS-v2 may be an underestimation as the model is continuously updated and specific training data has not been fully disclosed. "-" indicates data not available or not applicable. *For Chinese, Korean and Japanese, figures in the WER columns represent Character Error Rate (CER). †Scores reported in baseline papers.

	Chinese*					English				
	Train (hrs)	WER	SIM-o	CMOS	SMOS	Train (hrs)	WER	SIM-o	CMOS	SMOS
Ground Truth	-	1.25	0.75	0.0	3.38	-	2.14	0.73	0.0	3.36
MaskGCT (Wang et al., 2024)	49.9K	2.27†	**0.77**†	-	-	46.8K	2.62†	**0.72**†	-	-
F5-TTS (Chen et al., 2024b)	49.9K	1.56†	0.76†	-	-	46.8K	**1.83**†	0.67†	-	-
FireRedTTS (Guo et al., 2024)	110K	**1.21**	0.65	-0.28	2.82	40K	9.08	0.45	0.27	2.97
CosyVoice (Du et al., 2024b)	130K	3.49	0.75	**0.18**	3.64	30K	3.89	0.64	0.50	3.48
CosyVoice 2 (Du et al., 2024c)	130K	1.35	0.75	-0.01	**3.86**	30K	2.69	0.65	0.59	**3.69**
VoiceCraft (Peng et al., 2024)	-	-	-	-	-	9K	5.28	0.51	0.44	3.27
VoiceCraft-X	5K	3.29	0.68	-0.39	2.94	14.5K	4.37	0.54	**0.63**	3.43

	Korean*			Japanese*			Dutch		
	Train (hrs)	WER	SIM-o	Train (hrs)	WER	SIM-o	Train (hrs)	WER	SIM-o
Ground Truth	-	8.89	-	-	9.72	0.79	-	9.54	0.65
XTTS-v1	-	-	-	-	-	-	-	78.17	0.41
XTTS-v2	539‡	40.89	**0.62**	57‡	**11.61**	0.64	74‡	**12.62**	0.59
VoiceCraft-X	832	**31.11**	0.56	3489	15.09	**0.66**	2147	16.28	**0.61**

	Italian			Portuguese			Polish		
	Train (hrs)	WER	SIM-o	Train (hrs)	WER	SIM-o	Train (hrs)	WER	SIM-o
Ground Truth	-	9.48	0.68	-	8.75	0.69	-	8.81	0.72
XTTS-v1	-	73.12	0.32	-	48.93	0.33	-	96.15	0.41
XTTS-v2	1297‡	15.52	**0.56**	2387‡	**13.48**	**0.58**	199‡	**9.47**	**0.62**
VoiceCraft-X	294	**15.46**	0.54	223	22.57	0.56	139	24.80	0.61

	French			German			Spanish		
	Train (hrs)	WER	SIM-o	Train (hrs)	WER	SIM-o	Train (hrs)	WER	SIM-o
Ground Truth	-	6.09	0.68	-	6.64	0.69	-	4.87	0.73
XTTS-v1	-	38.34	0.35	-	11.37	0.35	-	20.84	0.37
XTTS-v2	2216‡	**5.45**	0.58	3584‡	16.50	0.59	1514‡	8.11	0.58
VoiceCraft-X	1338	13.22	**0.59**	3405	**8.19**	**0.60**	1191	**4.67**	**0.63**

4.3 Transfer Learning for Multilingual TTS

To explore the benefits of multilingual training, especially for lower-resource languages, we fine-tuned *monolingual* models on individual languages starting from different pre-trained checkpoints, comparing these against training from scratch and the multilingual model (detailed in Table 2).

The universal advantage of pre-training over "from Scratch" models is paramount, especially for languages with limited data. For instance, Italian (294 hours) and Polish (139 hours) saw their WERs plummet from over 140 and 160 to under 14 and 20 respectively, demonstrating pre-training's crucial role in transferring foundational knowledge and overcoming data scarcity. Even higher-resource languages like Spanish, French and German benefited significantly. Fine-tuning from an English model initialization proved highly effective for European languages (Germanic, Romance, Slavic), leveraging linguistic similarities and robust acoustic modeling, with gains particularly vital for low-data scenarios (Italian, Portuguese, Polish). Korean showed better CER with a Japanese checkpoint (42.08) than Chinese (49.11), aligning with typological closeness. Conversely, Japanese experienced negative transfer from Chinese (CER 36.18 vs. 22.36 from scratch).

Furthermore, fine-tuning from the "multilingual checkpoint" frequently yielded superior WER/CER compared to an English-only checkpoint for a range of languages including Spanish, Dutch, Italian, Portuguese, Polish, and Japanese. This advantage held across varying data volumes (e.g., Polish 139 hours, Japanese 3489 hours), suggesting that pre-training on a diverse linguistic set fosters more generalized and transferable representations than exposure to

Table 2: Cross-lingual transfer learning performance on zero-shot TTS task. Comparison of fine-tuning from different pre-trained models versus training from scratch for various target languages. Character Error Rate (CER) for Korean and Japanese, indicated by *. "-" indicates data not available or not applicable.

Language	**#Hours**	**Multilingual**		**from Scratch**		**from English**		**from Chinese/Japanese**		**from Multilingual**	
		WER	SIM-o	WER	SIM-o	WER	SIM-o	WER	SIM-o	WER	SIM-o
Korean*	832	31.11	**0.56**	45.79	0.51	42.10	0.54	49.11/42.08	0.50/0.52	41.36	0.53
Japanese*	3489	**15.09**	0.66	22.36	0.62	-	-	36.18	0.61	19.35	**0.67**
Spanish	1191	4.67	**0.63**	7.08	0.38	4.54	0.47	-	-	**3.30**	0.52
French	1338	13.22	**0.60**	18.85	0.43	**12.50**	0.49	-	-	16.39	0.53
German	3405	8.19	**0.60**	6.43	0.43	**5.93**	0.50	-	-	7.25	0.53
Dutch	2147	16.28	**0.61**	16.85	0.37	16.02	0.35	-	-	**11.78**	0.46
Italian	294	15.46	**0.54**	142.30	0.22	13.97	0.36	-	-	**13.93**	0.46
Portuguese	223	22.57	**0.56**	91.89	0.26	15.87	0.46	-	-	**14.74**	0.55
Polish	139	24.80	**0.61**	163.08	0.25	20.73	0.46	-	-	**19.47**	0.55

English alone, capturing a broader array of phonetic and prosodic patterns.

Finally, the original multilingual model's speaker similarity is significantly higher than models fine-tuned from other checkpoints for nearly all languages. This indicates that joint training on diverse linguistic data, leveraging collective data volume, allows the model to disentangle speaker-specific characteristics from language-specific features. This robust performance across varied languages suggests it learns a more abstract, shared representation space for speech, facilitating both high-fidelity synthesis and strong cross-lingual capabilities. While fine-tuning on single language data may impact this disentanglement ability, as evidenced by SIM-o drops in many such cases.

4.4 Speech Editing

Table 3: Performance on English speech editing.

	WER	NMOS	IMOS
Original	2.42	3.78	3.79
VoiceCraft	5.99	**3.87**	**3.87**
VoiceCraft-X	**5.62**	3.68	3.79

For English speech editing (Table 3), VoiceCraft-X demonstrated a better Word Error Rate (WER) than VoiceCraft. Both models produced edited speech that listeners found to be highly natural (NMOS) and intelligible (IMOS), comparable to the original recordings. VoiceCraft's slightly higher scores in these subjective tests are not surprising, given its monolingual English focus, especially considering both models have similar parameter counts and amounts of English training data.

For multilingual speech editing in other languages—a capability where comparative baselines are notably scarce as most models do not support multilingual editing—we conducted subjective MOS evaluations. These evaluations focused on a subset of languages (French, Italian, Portuguese, and Spanish) for which MTurk annotators were available, with results presented in Table 4. The evaluations demonstrate VoiceCraft-X's effective performance in this challenging scenario. While naturalness (NMOS) scores for edited speech are, as anticipated, lower than the original recordings, intelligibility (IMOS) remains high across these languages. Particularly for Spanish and Italian, where edited NMOS and IMOS scores closely matched the original audio, these findings underscore VoiceCraft-X's significant and unique capability for coherent, comprehensible multilingual speech editing.

Table 4: Subjective performance on speech editing.

	Original		Edited	
	NMOS	IMOS	NMOS	IMOS
French	3.62	4.10	3.13	3.60
Italian	4.38	4.78	3.77	4.28
Portuguese	4.42	4.98	2.63	3.78
Spanish	3.80	3.93	3.58	3.78

5 Conclusion

We present VoiceCraft-X, an autoregressive neural codec language model that successfully unifies multilingual speech editing and Text-to-Speech (TTS) synthesis. Leveraging the Qwen3 LLM and a novel token reordering strategy, VoiceCraft-X supports eleven languages, producing high-quality, natural-sounding speech. Our model demonstrates robust performance across diverse conditions and shows that a unified framework can effectively advance both speech editing and synthesis in multilingual contexts, even with limited data for some languages. This work underscores the potential of autoregressive models for complex, real-world speech generation tasks.

Limitations

One key limitation is the scale of our training data. Although VoiceCraft-X performs well with approximately 32,578 hours across eleven languages, this is notably less than some state-of-the-art models. This comparative data scarcity, particularly for lower-resource languages in our set, may limit the model's capacity to capture the full spectrum of speech nuances as effectively as systems trained on more extensive datasets.

Secondly, while the model's multilingual support is a core feature, its current reach of eleven languages (with around 20-30 explored internally) only scratches the surface of global linguistic diversity. Expanding coverage to more languages, especially under-resourced ones, remains a significant challenge that would require substantial data curation and potential model adaptations to address varied linguistic features.

Finally, further investigation into model size scalability is also warranted. The current VoiceCraft-X utilizes the Qwen3-0.6B architecture; exploring larger model variants could unlock enhanced learning capabilities and higher fidelity in speech synthesis and editing. Systematically assessing different model sizes is crucial for optimizing the balance between performance improvements and computational demands.

References

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, and 1 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Jeong-Uk Bang, Seung Yun, Seung-Hi Kim, Mu-Yeol Choi, Min-Kyu Lee, Yeo-Jeong Kim, Dong-Hyun Kim, Jun Park, Young-Jik Lee, and Sang-Hun Kim. 2020. Ksponspeech: Korean spontaneous speech corpus for automatic speech recognition. *Applied Sciences*.

James Betker. 2023. Better speech synthesis through scaling. *arXiv preprint arXiv:2305.07243*.

Zalán Borsos, Matt Sharifi, and Marco Tagliasacchi. 2022. Speechpainter: Text-conditioned speech inpainting. *arXiv preprint arXiv:2202.07273*.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and 1 others. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, and 1 others. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. 2024a. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *In Journal JSTSP*.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024b. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. In *Proc. NeurIPS*.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Chenpeng Du, Yiwei Guo, Feiyu Shen, Zhijun Liu, Zheng Liang, Xie Chen, Shuai Wang, Hui Zhang, and Kai Yu. 2024a. Unicats: A unified context-aware text-to-speech framework with contextual vq-diffusion and vocoding. In *Proc. AAAI*.

Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. 2018. Aishell-2: Transforming mandarin asr research into industrial scale. *arXiv preprint arXiv:1808.10583*.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, and 1 others. 2024b. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024c. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2024. *Ethnologue: Languages of the World*, twenty-seventh edition. SIL International, Dallas, Texas.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, and 1 others. 2024. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *Proc. SLT*.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, and 1 others. 2023. Funasr: A fundamental end-to-end speech recognition toolkit. *arXiv preprint arXiv:2305.11013*.

Hao-Han Guo, Yao Hu, Kun Liu, Fei-Yu Shen, Xu Tang, Yi-Chen Wu, Feng-Long Xie, Kun Xie, and Kai-Tuo Xu. 2024. Fireredtts: A foundation text-to-speech framework for industry-level generative speech applications. *arXiv preprint arXiv:2409.03283*.

Tingwei Guo, Cheng Wen, Dongwei Jiang, Ne Luo, Ruixiong Zhang, Shuaijiang Zhao, Wubo Li, Cheng Gong, Wei Zou, Kun Han, and 1 others. 2021. Didispeech: A large scale mandarin speech corpus. In *Proc. ICASSP*.

Bing Han, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Yanming Qian, Yanqing Liu, Sheng Zhao, Jinyu Li, and Furu Wei. 2024. Vall-e r: Robust and efficient zero-shot text-to-speech synthesis via monotonic alignment. *arXiv preprint arXiv:2406.07855*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Katsuya Iida. 2021. Kokoro speech dataset. https://github.com/kaiidams/Kokoro-Speech-Dataset.

Ziyue Jiang, Yi Ren, Ruiqi Li, Shengpeng Ji, Boyang Zhang, Zhenhui Ye, Chen Zhang, Bai Jionghao, Xiaoda Yang, Jialong Zuo, and 1 others. 2025. Megatts 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis. *arXiv preprint arXiv:2502.18924*.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, and 1 others. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.

Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. 2023. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *In journal TACL*.

Jaehyeon Kim, Keon Lee, Seungjun Chung, and Jaewoong Cho. 2024. Clam-tts: Improving neural codec language model for zero-shot text-to-speech. In *Proc. ICLR*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. 2023. Libritts-r: A restored multi-speaker text-to-speech corpus. *arXiv preprint arXiv:2305.18802*.

Mateusz Łajszczak, Guillermo Cámbara, Yang Li, Fatih Beyhan, Arent Van Korlaar, Fan Yang, Arnaud Joly, Álvaro Martín-Cortinas, Ammar Abbas, Adam Michalski, and 1 others. 2024. Base tts: Lessons from building a billion-parameter text-to-speech model on 100k hours of data. *arXiv preprint arXiv:2402.08093*.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and 1 others. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. In *Proc. NeurIPS*.

Keon Lee, Dong Won Kim, Jaehyeon Kim, and Jaewoong Cho. 2024. Ditto-tts: Efficient and scalable zero-shot text-to-speech with diffusion transformer. *arXiv preprint arXiv:2406.11427*.

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. 2024. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis. *arXiv preprint arXiv:2411.01156*.

Wenrui Liu, Jionghao Bai, Xize Cheng, Jialong Zuo, Ziyue Jiang, Shengpeng Ji, Minghui Fang, Xiaoda Yang, Qian Yang, and Zhou Zhao. 2025. Voxpopulitts: a large-scale multilingual tts corpus for zero-shot speech generation. In *Proc. COLING*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Linhan Ma, Dake Guo, Kun Song, Yuepeng Jiang, Shuai Wang, Liumeng Xue, Weiming Xu, Huan Zhao, Binbin Zhang, and Lei Xie. 2024. Wenetspeech4tts: A 12,800-hour mandarin tts corpus for large speech generation model benchmark. *arXiv preprint arXiv:2406.05763*.

Magic Data. 2019. Magicdata mandarin chinese read speech corpus.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proc. Interspeech*.

Frederico S Oliveira, Edresson Casanova, Arnaldo Candido Junior, Anderson S Soares, and Arlindo R Galvão Filho. 2023. Cml-tts: A multilingual dataset for speech synthesis in low-resource languages. In *Proc. TSD*.

Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. 2024. Voicecraft: Zero-shot speech editing and text-to-speech in the wild. *arXiv preprint arXiv:2403.16973*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*.

Qwen-Team. 2025. Qwen3.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*.

Takaaki Saeki, Gary Wang, Nobuyuki Morioka, Isaac Elias, Kyle Kastner, Andrew Rosenberg, Bhuvana Ramabhadran, Heiga Zen, Françoise Beaufays, and Hadar Shemtov. 2024. Extending multilingual speech synthesis to 100+ languages without transcribed data. In *Proc. ICASSP*.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*.

Yakun Song, Zhuo Chen, Xiaofei Wang, Ziyang Ma, and Xie Chen. 2025. Ella-v: Stable neural codec language modeling with alignment-guided sequence reordering. In *Proc. AAAI*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2024. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*.

Detai Xin, Xu Tan, Kai Shen, Zeqian Ju, Dongchao Yang, Yuancheng Wang, Shinnosuke Takamichi, Hiroshi Saruwatari, Shujie Liu, Jinyu Li, and 1 others. 2024. Rall-e: Robust codec language modeling with chain-of-thought prompting for text-to-speech synthesis. *arXiv preprint arXiv:2404.03204*.

Dongchao Yang, Rongjie Huang, Yuanyuan Wang, Haohan Guo, Dading Chong, Songxiang Liu, Xixin Wu, and Helen Meng. 2024a. Simplespeech 2: Towards simple and efficient text-to-speech with flow-based scalar latent transformer diffusion models. *arXiv preprint arXiv:2408.13893*.

Dongchao Yang, Dingdong Wang, Haohan Guo, Xueyuan Chen, Xixin Wu, and Helen Meng. 2024b. Simplespeech: Towards simple and efficient text-to-speech with scalar latent transformer diffusion models. *arXiv preprint arXiv:2406.02328*.

Yifan Yang, Shujie Liu, Jinyu Li, Yuxuan Hu, Haibin Wu, Hui Wang, Jianwei Yu, Lingwei Meng, Haiyang Sun, Yanqing Liu, and 1 others. 2025. Pseudo-autoregressive neural codec language models for efficient zero-shot text-to-speech synthesis. *arXiv preprint arXiv:2504.10352*.

Yue Yin. 2023. Reazonspeech: A free and massive corpus for japanese asr.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *In Journal TASLPRO*.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechtokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692*.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023b. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*.

A Dataset

A.1 Training Dataset Statistics

The training datasets for each language are as shown in Table 5. For all of them, we remove all YouTube clips.

Table 5: Speech-corpus statistics used for training (**total: 32 578 h**).

Language	Dataset(s)	Hours
English	LibriTTS-R (Koizumi et al., 2023)	516
	GigaSpeech (Chen et al., 2021)	5 783
	MLS (Pratap et al., 2020)	8 235
Chinese	WenetSpeech4TTS (Ma et al., 2024)	3 282
	AISHELL-2 (Du et al., 2018)	997
	MAGICDATA (Magic Data, 2019)	707
Korean	KsponSpeech (Bang et al., 2020)	832
Japanese	ReazonSpeech (Yin, 2023)	3 489
Spanish	MLS (Pratap et al., 2020) CML-TTS (Oliveira et al., 2023)	1 191
French		1 338
German		3 405
Dutch		2 147
Italian		294
Portuguese		223
Polish		139
Total		**32 578**

A.2 Speech Editing Dataset

To create a comprehensive evaluation set for speech editing, we began by selecting a subset of samples from the Text-to-Speech (TTS) evaluation datasets described in Section 4.1. For each language, 100-300 original text samples were chosen.

Unlike RealEdit (Peng et al., 2024), which relies on manual, sentence-by-sentence human annotation and modification, a process that limits its scalability across many languages, we employed the powerful multilingual capabilities of the Gemini language model (Team et al., 2023) to systematically introduce textual modifications to the original sentences. The goal was to generate edited versions that reflect common editing scenarios. To achieve this, Gemini was instructed to perform exactly one of the following specified operations on each original sentence:

- **Insertion:** Adding a sequence of new words into the original sentence.
- **Deletion:** Removing a sequence of words from the original sentence.
- **Substitution:** Replacing a sequence of words in the original sentence with a new sequence of words.

To ensure diversity in the complexity and scope of edits, the length of the modified segments was varied. Specifically, all edits involved at least two contiguous words. The modifications ranged from short (2–3 words), to medium (4–6 words), and occasionally longer spans (7–10 words). We show examples in Table 6.

B Implementational Details

B.1 Encodec Model

The Encodec model we employ operates with a stride of 320 samples, corresponding to a codec frame rate of 50 Hz when processing audio recorded at 16 kHz. Its encoder begins with a base channel dimension of 64, which doubles at each of the five successive convolutional layers. Following (Défossez et al., 2022), we utilize the open-source audiocraft repository[3] for training. Specifically, we sample one-second speech segments from the multilingual dataset (shown in Table 5) and train for 200 epochs with a batch size of 832. Optimization is performed using the Adam algorithm (Kingma and Ba, 2014) with a base learning rate of 5e-5.

B.2 Qwen3 Base Model

The Qwen3-0.6B-Base model[4], foundational to VoiceCraft-X, is a causal language model with 0.6 billion total parameters, of which 0.44 billion are non-embedding parameters. It features 28 Transformer layers, a hidden dimension of 1024, and a feed-forward network (FFN) dimension of 3072, along with 16 attention heads. The model employs Grouped-Query Attention (16 query heads and 8 key/value heads) and supports a context length of 32,768 tokens. A key factor in its suitability for VoiceCraft-X's multilingual requirements is its pre-training on 36 trillion tokens across 119 languages. This pre-training utilized a diverse, high-quality data mix that included multilingual texts, books, and synthetic data. Furthermore, the model incorporates architectural refinements such as *qk layernorm* and benefits from a three-stage pre-training process designed for robust long-context handling.

[3] https://github.com/facebookresearch/audiocraft/blob/main/docs/ENCODEC.md

[4] https://huggingface.co/Qwen/Qwen3-0.6B-Base

Table 6: Examples of the multilingual speech editing dataset.

Language	Edit Types	Original	Edited
English	Substitution	Since I've gotten a dog, the **regular visits of the fox** have stopped.	Since I've gotten a dog, the **nightly disturbances** have stopped.
	Insertion	Increment the order quantity if you require more than one item.	Increment the order quantity **in the online form** if you require more than one item.
	Deletion	A bus shuttle took us **from the airport** to the metro.	A bus shuttle took us to the metro.
Chinese	Substitution	女主在等男主回来，**事情挺多**，不会无聊。	女主在等男主回来，**手头上的事情多得不可思议**，不会无聊。
	Insertion	那无边无际的大海啊，不会因时间的推移而变化。	那无边无际的大海啊，**其波澜壮阔的景象**不会因时间的推移而变化。
	Deletion	丈夫**又惊又怕，**再次放下了斧子，朝四周张望。	丈夫再次放下了斧子，朝四周张望。
Korean	Substitution	이렇게 안하면 니가 한번 가슴 **하면 가슴이** 지쳐서 다음 날 힘이 안 들어가는데	이렇게 안하면 니가 한번 가슴 **운동하면 가슴 근육이** 지쳐서 다음 날 힘이 안 들어가는데
	Insertion	아 뭐 계획은 거창하게 잡았는데 막상 한 건 하루라서 이제 쫌 해볼려고 하는데.	아 뭐 계획은 **아주** 거창하게 잡았는데 막상 한 건 하루라서 이제 쫌 해볼려고 하는데.
	Deletion	빼빼로 데이 **빼빼로 데이** 때 아는 동생한테 빼빼로 하나 받았다. 기프티콘으로.	빼빼로 데이 때 아는 동생한테 빼빼로 하나 받았다. 기프티콘으로.
Japanese	Substitution	一般学生よりはずっと**金持**に違いないと信じていますそうですともとＫ君はうなずいた。	一般学生よりはずっと**裕福な家庭環境**に違いないと信じていますそうですともとＫ君はうなずいた。
	Insertion	田中もそう申しておりました。それから、先生に是非お目にかかってお	田中も**全く同じように**そう申しておりました。それから、先生に是非お目にかかってお
	Deletion	私は興味にみちた眼をもって**それらの人を迎えたり送ったりした**事さえある。	私は興味にみちた眼をもって事さえある。
Spanish	Substitution	Los troyanos han **vencido** a los griegos en el llano.	Los troyanos han **derrotado completamente** a los griegos en el llano.
	Insertion	Tan esbelta y tan velera que consumió todos sus ahorros.	Tan esbelta y tan velera que **rápidamente** consumió todos sus ahorros.
	Deletion	La corrección que merodeaba **aún por allí**, y las bolsitas de cera, lo iluminaron suficientemente.	La corrección que merodeaba, y las bolsitas de cera, lo iluminaron suficientemente.
French	Substitution	Alors le malheureux navire s'enfonça **plus rapidement**.	Alors le malheureux navire s'enfonça **dans les abîmes profonds.**
	Insertion	Je m'étonne que vous m'ayez prêté de pareils sentiments.	Je m'étonne, **vraiment et très sincèrement**, que vous m'ayez prêté de pareils sentiments.
	Deletion	C'est quand elle est accomplie, qu'elle semble possible **aux êtres du commun.**	C'est quand elle est accomplie, qu'elle semble possible.
German	Substitution	Dasselbe gilt für die **so** komplizierte Entwicklung der Sexualfunktion.	Dasselbe gilt für die **außerordentlich** komplizierte Entwicklung der Sexualfunktion.
	Insertion	Aber schon hatte sich das Luftschiff fortgeschnellt.	Aber schon hatte sich das **feindliche** Luftschiff fortgeschnellt.
	Deletion	Und in des Schiffs Kielwasser schwammen **Grüngoldne Schlangen** hinterher.	Und in des Schiffs Kielwasser schwammen hinterher.
Italian	Substitution	Il professor Gori **scattò** in piedi, urlando: Lasciate!	Il professor Gori **balzò improvvisamente** in piedi, urlando: Lasciate!
	Insertion	Il terzo, che'l cibo vostro sia da bestia.	Il terzo **comandamento importante,** che'l cibo vostro sia da bestia.
	Deletion	Non era mai venuto **neppure una volta** a visitarla, è vero.	Non era mai venuto a visitarla, è vero.
Portuguese	Substitution	Astros! Qual é o mundo, **Em torno ao qual** rodais Por esse firmamento?	Astros! Qual é o mundo, **Pelo qual vocês todos** rodais Por esse firmamento?
	Insertion	Indagando com os olhos atilados o vôo do corvo.	Indagando **atentamente e curiosamente** com os olhos atilados o vôo do corvo.
	Deletion	Era preciso decidir entre os seus desejos **de vingar o sexo** e as conveniências da sua posição.	Era preciso decidir entre os seus desejos e as conveniências da sua posição.
Dutch	Substitution	Het is slechts een **zeer vage** veronderstelling.	Het is slechts een **interessante maar onbewezen** veronderstelling.
	Insertion	Wij zullen Toby bij ons houden, want hij kan ons nog van dienst zijn.	Wij zullen Toby bij ons houden **voorlopig in ieder geval,** want hij kan ons nog van dienst zijn.
	Deletion	En het oudste jongetje kwam mij vertellen, **dat ze honger en kou leden.**	En het oudste jongetje kwam mij vertellen.
Polish	Substitution	Pozostawało tylko **osnuć na nich poprzeczne drabinki.**	Pozostawało tylko **zbudować solidne rusztowanie.**
	Insertion	Jest on jedynym puklerzem niewinnej pluskwy polnej.	Jest on jedynym **skutecznym i niezawodnym** puklerzem niewinnej pluskwy polnej.
	Deletion	Podniecenie nerwów sprawiło, żem **zaraz w ciągu pierwszych minut** dostrzegł światło.	Podniecenie nerwów sprawiło, żem dostrzegł światło.

B.3 Loss Design

VoiceCraft-X is trained as an autoregressive model to predict a sequence of neural codec tokens. Given the input context, which includes text tokens, speaker embeddings, and potentially prefix/suffix audio tokens, the model predicts the target audio tokens one by one. The overall training objective is a weighted cross-entropy loss, designed to enhance learning efficiency and focus on the crucial aspects of the speech generation task.

Let the sequence of all ground truth speech tokens (encompassing prefix, suffix, and middle segments, and structured according to the delay pattern described in Section 3.5) be denoted by $Z = (z_1, z_2, \ldots, z_N)$, where N is the total number of tokens in the flattened sequence. Each token z_i in this sequence corresponds to an original codec token Y_{t_i,k_i} from timestep t_i and the k_i-th codebook of the EnCodec output (where $K = 4$ is the total number of codebooks). The model predicts the probability distribution for each token $\hat{z}_i$ conditioned on previous tokens and the input context.

The total loss $\mathcal{L}$ is a sum of individual cross-entropy losses for each token, with two layers of weighting:

1. **Codebook Weighting**: As mentioned in Section 4.1, each of the $K = 4$ parallel codebooks contributes differently to the overall perceptual quality. We assign weights $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (1.0, 0.8, 0.6, 0.4)$ to the tokens from codebook 1 to 4, respectively. So, for a token z_i corresponding to Y_{t_i,k_i}, its codebook weight is α_{k_i}.

2. **Segment Weighting**: While the model is trained to predict tokens for all three segments (prefix, middle, and suffix) to improve training efficacy and contextual understanding, the primary goal is the accurate generation of the "middle" (target) segment. To reflect this, we introduce segment-specific weights. Tokens belonging to the "prefix" and "suffix" segments are assigned a weight $w_{seg} = 1$. Tokens belonging to the "middle" segment, which is the primary target for generation or editing, are assigned a higher weight $w_{seg} = 3$. Let $w_{seg}(z_i)$ denote the segment weight for token z_i.

Combining these, the total loss $\mathcal{L}$ is formulated as:

$$\mathcal{L} = \sum_{i=1}^{N} w_{seg}(z_i) \cdot \alpha_{k_i} \cdot L_{CE}(\hat{z}_i, z_i)$$

where $L_{CE}(\hat{z}_i, z_i)$ is the cross-entropy loss for predicting token z_i. This weighted loss function guides the model to prioritize the generation of the target audio segment while still learning from the context provided by the prefix and suffix, and appropriately valuing the contribution of each codebook.

C Subjective Evaluation

C.1 Setup

To compute our subjective evaluation metrics (SMOS and CMOS for TTS, NMOS and IMOS for Speech Editing), for all languages except Chinese, we recruited Amazon Mechanical Turk workers with a minimum approval rate of 98% and at least 1000 successful HITs. We manually recruited university students for Chinese. We filtered workers by the following countries in Table 7 for each of our languages:

Language	Countries
English	United States
Chinese	China
French	Belgium, Canada, France, Luxembourg, Switzerland
Italian	Italy
Portuguese	Brazil, Portugal
Spanish	Argentina, Chile, Colombia, Mexico, Spain, United States

Table 7: Countries used to filter crowdworkers for each language

Each sample was annotated by 3 different annotators. We display annotation UIs for our metrics in Figures 4, 5, 6 and 7.

C.2 Additional Results

A scarcity of Amazon Mechanical Turk workers for less common languages prevented us from collecting subjective evaluation results for all targeted languages. Consequently, the SMOS results for French, Italian, Portuguese, and Spanish on the Zero-Shot TTS task that we were able to gather are detailed in Table 8.

Table 8: SMOS on Zero-Shot TTS.

Model	French	Italian	Portuguese	Spanish
Ground Truth	3.07	3.57	4.15	3.42
XTTS-v1	2.07	2.00	1.63	2.83
XTTS-v2	2.23	2.75	2.48	3.22
VoiceCraft-X	**3.58**	**3.30**	**2.87**	**3.58**

D Ablations

D.1 Reordering Mechanism

Table 9: Impact of token reordering in a low-resource scenario. Models were trained from scratch: one on English (585h LibriTTS-R), the other on Chinese (601h WenetSpeech4TTS Premium subset).

	English		Chinese	
	WER↓	SIM-o↑	CER↓	SIM-o↑
w/o Reordering	104.02	0.31	262.25	0.29
w/ Reordering	**11.60**	**0.32**	**19.25**	**0.46**

For this ablation study, considering the low-resource nature of most languages, we used LibriTTS-R (Koizumi et al., 2023) and the WenetSpeech4TTS Premium (Ma et al., 2024) subset as training data. LibriTTS-R contains 585 hours of speech, while the WenetSpeech4TTS Premium subset includes 601 hours[5]. Models were trained for 15 epochs, both with and without the reordering mechanism. The final epoch was then evaluated on the Seed-TTS test set. As can be seen from Table 9, the model using the reordering mechanism shows significant performance improvements across all objective evaluation metrics on both the English and Chinese datasets. Specifically, the WER for English dropped dramatically from 104.02 to 11.60, and the CER for Chinese also decreased sharply from 262.25 to 19.25. Concurrently, the SIM-o scores for both languages also showed noticeable increases, indicating an improvement in the quality and naturalness of the synthesized speech. These results strongly demonstrate that the reordering mechanism is very effective in training under low-resource scenarios.

D.2 Position of Prompt in Zero-Shot TTS Inference

The token reordering mechanism, integral to our training methodology, introduces flexibility in how prompts are structured during zero-shot Text-to-Speech (TTS) inference. To determine the optimal placement, we evaluated several configurations for incorporating the prompt text (T_{prompt}) and prompt audio (A_{prompt}) into the input sequence. These configurations are detailed in Table 10.

Our evaluation, based on WER and SIM-o, revealed that placing the prompt at the beginning of the "middle" segment yields the most favorable overall performance. Specifically, structuring the input such that the prompt text precedes the target text within the middle text segment (i.e., $T_P = \varnothing, T_S = \varnothing, T_M = (T_{prompt}, T_{target})$, with A_{prompt} appended after the mask tokens and before where A_{target} would be generated) resulted in a WER of 4.37, which is notably better than the alternative placements.

E Code-Switching

A desirable characteristic of a multilingual Text-to-Speech (TTS) model is its ability to generate code-switched speech—that is, speech that fluidly transitions between languages. Although our model was trained exclusively on monolingual data, meaning code-switched speech is an out-of-distribution phenomenon for it, the model still demonstrated a certain capacity for code-switching without needing additional language identifiers for inputs in different languages.

We also observed that the model tends to perform better when the initial language of the target text matches the language of the prompt. Conversely, if the starting language of the target text differs from the prompt, the model's performance may be significantly worse. We have made code-switched samples available on our demo page.

F Cross-lingual Finetuning Hours on Zero-Shot TTS

To further assess VoiceCraft-X's adaptability and the impact of data quantity, we extended fine-tuning experiments across diverse languages. Building on cross-lingual transfer insights (Section §4.3), we examined the correlation between per-language fine-tuning data volume and zero-shot Text-to-Speech (TTS) quality.

Figure 3 illustrates these findings, plotting per-language fine-tuning data volume (x-axis) against the relative Word Error Rate (WER) from zero-shot TTS (y-axis). This relative WER, the difference between Whisper's WER on synthesized versus ground-truth audio, offers a normalized measure of intelligibility. The figure generally shows that

[5]YouTube clips are removed.

Table 10: WER and SIM-o of different prompt positions in zero-shot TTS inference on Seed-TTS test-en set.

	WER	SIM-o
$\varnothing, \varnothing, T_{prompt}, T_{target}$, <SPK>, $\varnothing$, <M>, $\varnothing$, <M>, A_{prompt}, A_{target}	**4.37**	**0.54**
$T_{prompt}, \varnothing, T_{target}$, <SPK>, A_{prompt}, <M>, $\varnothing$, <M>, A_{target}	5.68	0.53
$\varnothing, T_{prompt}, T_{target}$, <SPK>, $\varnothing$, <M>, A_{prompt}, <M>, A_{target}	6.32	**0.54**

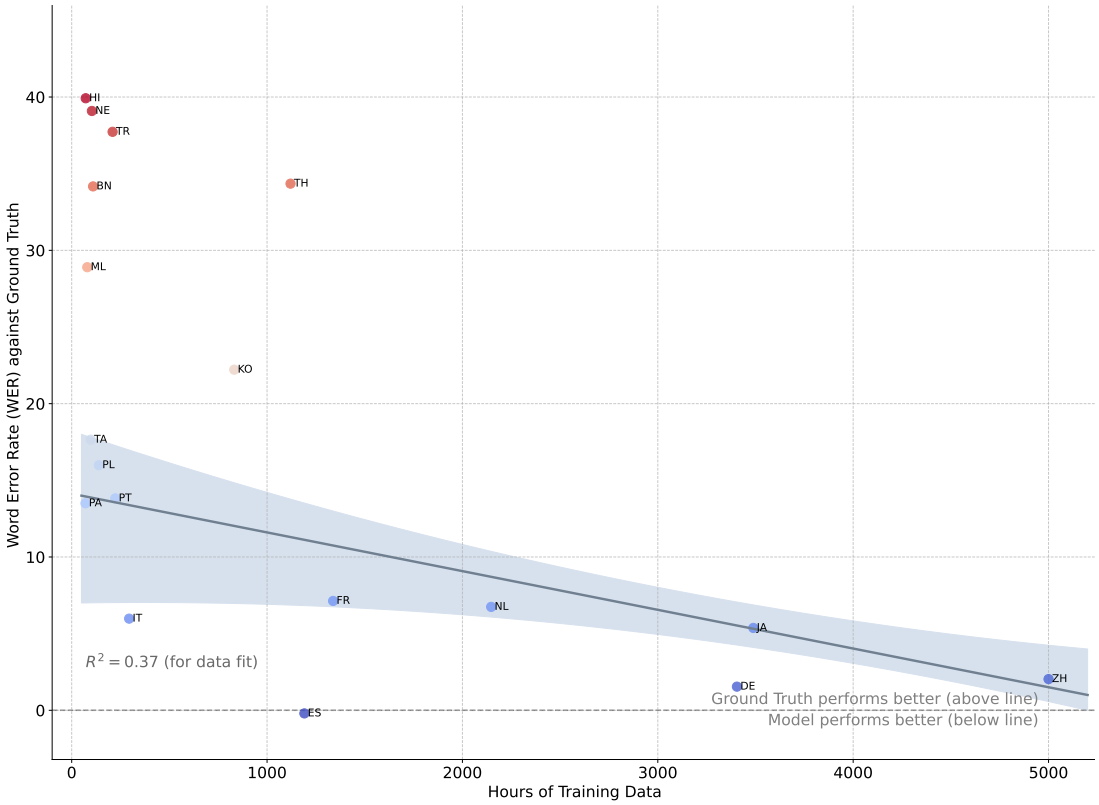

Figure 3: Relationship between per-language fine-tuning data and zero-shot TTS quality. Each point represents a target language, positioned by the number of hours used to fine-tune VoiceCraft-X (x-axis) and the relative Word Error Rate – the difference between Whisper's WER on synthesized audio and its WER on ground-truth audio.

more fine-tuning data improves pronunciation accuracy, especially for languages sharing similarities with VoiceCraft-X's initial training set. However, this correlation is not universally linear. For languages like Korean and Thai, a moderate data increase (around 1000 hours) did not yield significant WER improvements. This plateauing suggests that for such languages, substantial gains may require much larger or more diverse datasets, or different fine-tuning approaches.

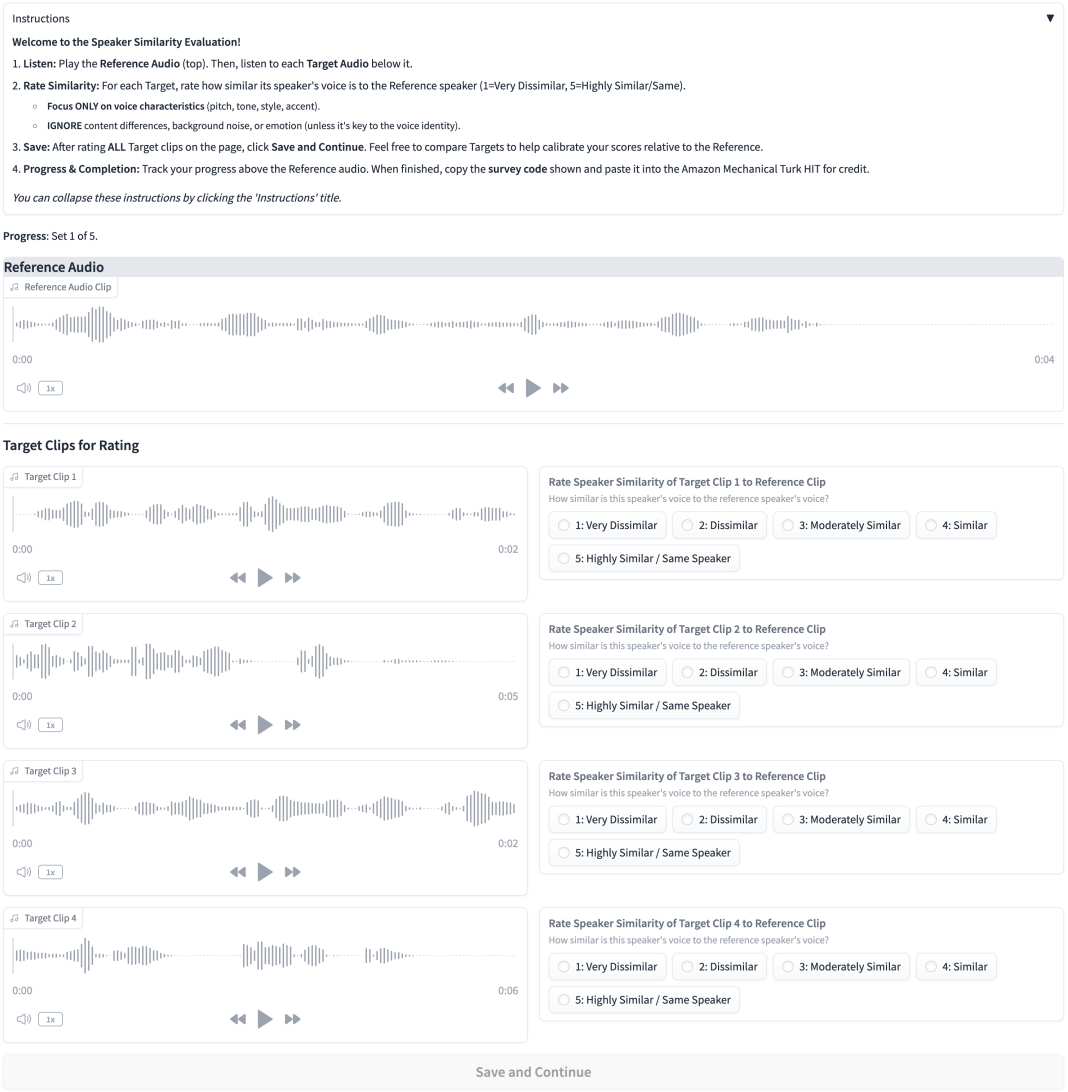

Figure 4: SMOS Annotation UI

Figure 5: CMOS Annotation UI

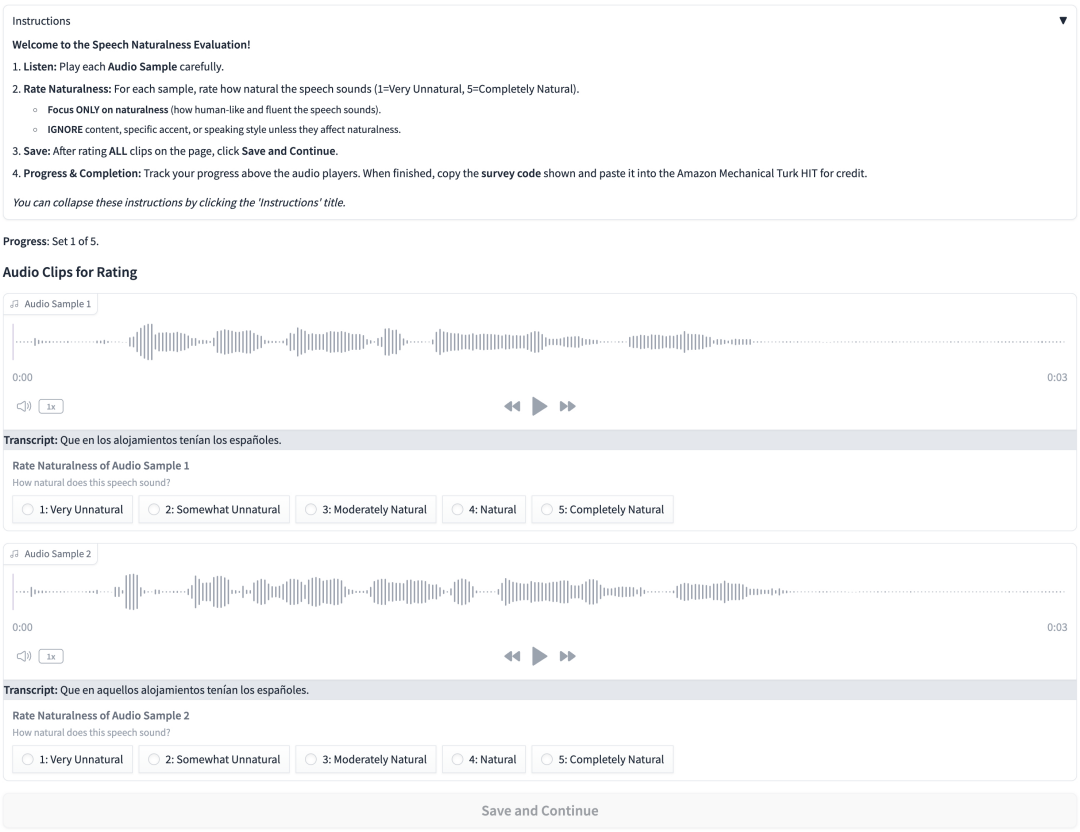

Figure 6: NMOS Annotation UI

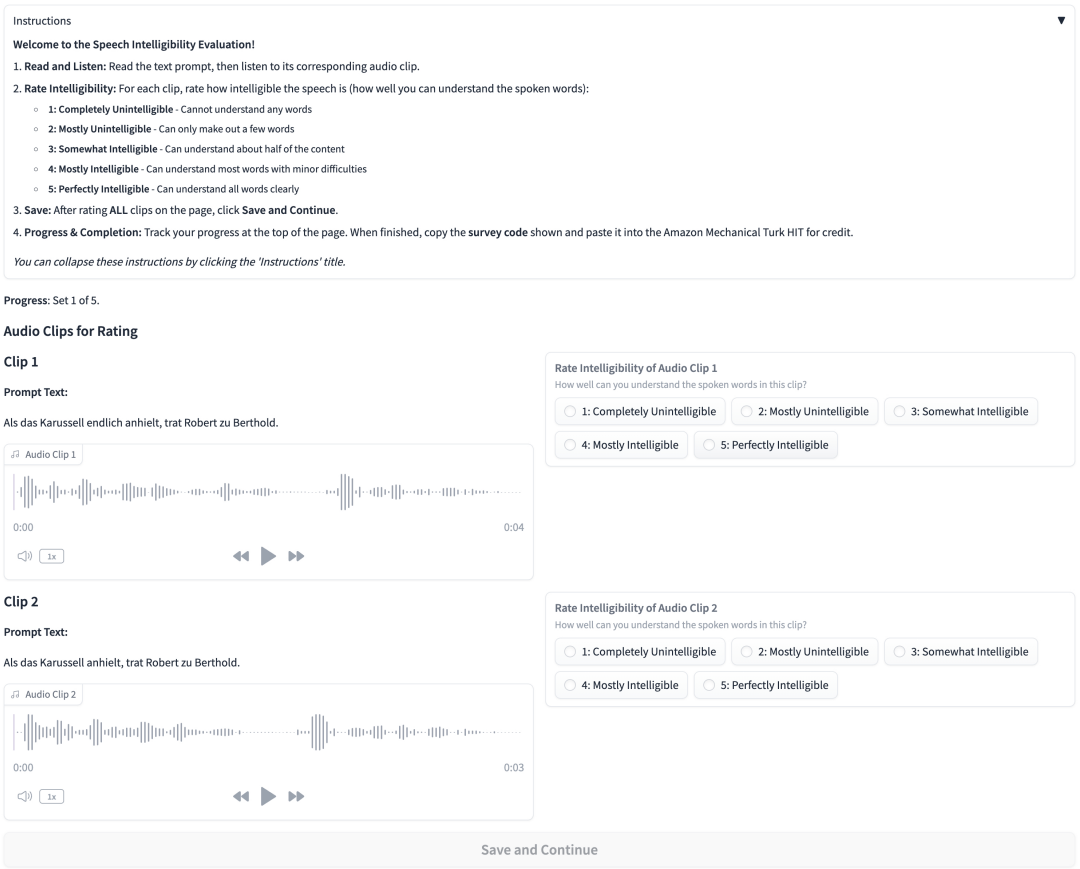

Figure 7: IMOS Annotation UI