TokenMapper: Towards a Universal Interoperability Layer for Speech Tokens

Anonymous ACL submission

Abstract

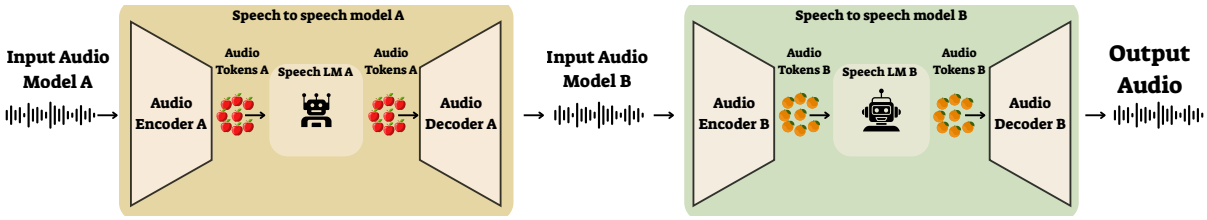
Neural audio codecs discretize speech into token sequences, but the resulting token spaces differ in vocabulary and codebook structure, preventing direct communication across models. This limitation affects applications such as conversational voice agents and speech to speech translation systems where multiple speech models must interact. As a result, transferring information between speech systems typically requires decoding to waveform audio and re-encoding with a second tokenizer, increasing latency and introducing potential information loss. To address this limitation, we present **TokenMapper**, a direction aware framework for direct token to token translation between heterogeneous speech tokenizers in the discrete domain. TokenMapper supports structurally mismatched token spaces, including mappings between single codebook and multi codebook representations. Experiments on GLM-4-Voice, MiMi, and DualCodec show consistent cross model performance. Specifically, translation WER approaches native reconstructions within 2.5-6.8% absolute WER, with UTMOS between 1.95-3.38 and end to end latency reductions of 4.8-94.5%, reaching up to 972 ms per utterance relative to waveform bridging. These results demonstrate practical cross model speech interoperability without intermediate waveform reconstruction.

1 Introduction

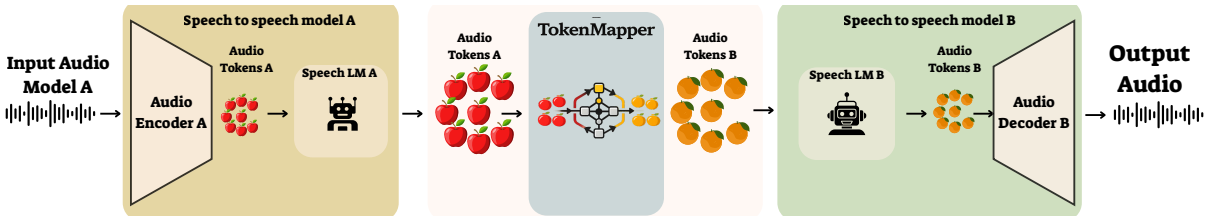
Discrete audio tokenization has become a central component of modern neural audio frameworks, enabling efficient compression and compatibility with language modeling. Early works such as VQ-VAE (van den Oord et al., 2018), vq-wav2vec (Baevski et al., 2020), and HuBERT (Hsu et al., 2021) demonstrated the effectiveness of discrete representations for speech and audio modeling. A major advance followed with SoundStream (Zeghidour et al., 2021), which introduced Residual Vector Quantization (RVQ) to hierarchically encode audio using multiple residual codebooks, a paradigm later refined by EnCodec (Défossez et al., 2022). RVQ based codecs typically expose multiple token streams operating at different information scales, enabling flexible conditioning and efficient generation. This design has been adopted or extended by a wide range of models and tasks, including GSLM (Lakhotia et al., 2021), AudioLM (Borsos et al., 2023), VioLA (Wang et al., 2023), SpeechTokenizer (Zhang et al., 2024), NanoCodec (Casanova et al., 2025), as well as recent Speech LLM codecs such as Moshi's Mimi (Défossez et al., 2024), DualCodec (Li et al., 2025a), and XY Tokenizer (Gong et al., 2025).

Neural audio codecs have consequently diverged substantially in their tokenization strategies. Variants modify the quantization mechanism itself e.g., Group RVQ or complex spectral modeling in HiFi-Codec (Yang et al., 2023a), AudioDec (Wu et al., 2023), and ComplexDec (Wu et al., 2025b), explicitly disentangle semantic and acoustic representations, SpeechTokenizer (Zhang et al., 2024), RepCodec (Huang et al., 2024b), SAC (Chen et al., 2025), or introduce adaptive temporal resolutions ,SNAC (Siuzdak et al., 2024), FlexiCodec (Li et al., 2025b). Beyond RVQ based designs, alternative tokenization schemes have also emerged, such as the unified tokenizer used in CosyVoice (Du et al., 2024a,b) and GLM-4-Voice (Zeng et al., 2024). While these advances enable applications ranging from real time conversational agents (Wang et al., 2024) to speech to speech translation pipelines (Tran et al., 2022), the resulting fragmentation of discrete token spaces prevents direct interaction across models.

Today, information transfer between models typically requires decoding tokens to waveforms and re-encoding them, introducing additional latency and potential information loss. In other words, tokens produced by one model cannot be directly in-



(a) Conventional waveform bridged communication



(b) TokenMapper: Direct Token to Token Translation

Figure 1: Comparison between conventional waveform bridging and the proposed TokenMapper framework for cross model speech communication. (a) Conventional waveform bridging: tokens produced by model A (illustrated as apples) must first be decoded to waveform audio and then re-encoded by model B (whose tokens are illustrated as oranges). (b) TokenMapper: direct translation from apples to oranges, i.e., from the token space of model A to the token space of model B, eliminating intermediate waveform reconstruction.

terpreted by another, even when both operate over the same underlying speech signal. As illustrated in Fig. 1, this mismatch resembles an "apples to oranges" problem: tokens produced by one tokenizer must first be converted back to waveform audio before they can be re-encoded by another model. To address this interoperability problem, we propose **TokenMapper**, a framework for direct token to token translation between heterogeneous audio tokenizers. TokenMapper operates entirely in the discrete domain and is designed to bridge structurally mismatched token spaces, including mappings between single codebook and multi codebook RVQ representations. We evaluate TokenMapper on three representative and structurally distinct tokenizers: Moshi (Mimi), GLM-4-Voice, and DualCodec, covering a broad spectrum of design choices in contemporary audio models. Our results demonstrate that direct token level translation is feasible and preserves intelligibility to a meaningful degree, establishing cross model speech interoperability as a practical capability rather than a theoretical abstraction. Together, these findings suggest that discrete token spaces contain transferable structure that can be aligned across independently trained speech models. In summary, our main contributions are:

- **TokenMapper Framework:** Direct token to token translation that eliminates waveform decoding and re-encoding, reducing latency and computational overhead, achieving 4.8-94.5% end to end latency reduction, up to 972 ms per utterance.
- **Structural Mismatch Handling:** A direction aware mapping strategy that bridges incompatible token spaces, including single codebook and multi codebook (RVQ) representations with differing vocabulary sizes and depths.
- **Cross tokenizer transfer:** Empirical evidence that direct token space translation preserves intelligibility across GLM-4-Voice, MiMi, and DualCodec, achieving 5.85-9.98% WER while remaining within 2.5-6.8% absolute WER of native reconstructions.

1.1 Related work

Discrete audio representations have become central to speech enabled and multimodal language models, supporting universal audio understanding (Chu et al., 2023), end to end spoken interaction (Li et al., 2025c; Zhang et al., 2023a; Mitsui et al., 2024), interleaved speech text modeling (Nguyen et al., 2024), and real time spoken agents (Fang et al., 2025a,b). In parallel, extensive work on neural audio codecs explores diverse discretization strategies, including diffusion and transformer based models (Yang et al., 2023b; Shen et al., 2023; Ju et al., 2024; Ziv et al., 2024), low bitrate and high fidelity designs (Jenrungrot et al., 2023; Niu et al., 2024; Ren et al., 2024; Jiang et al., 2025; Shi

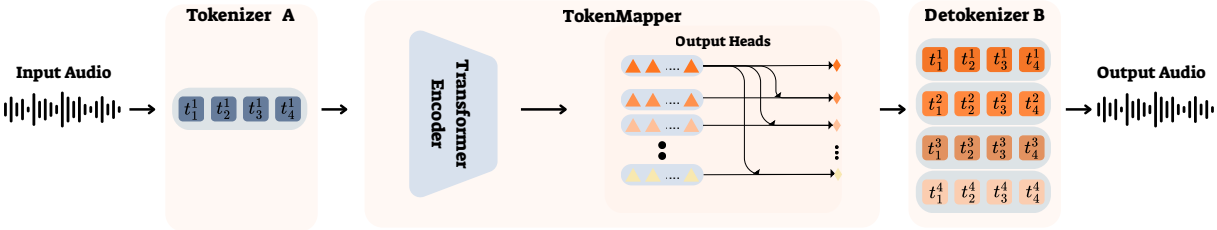


Figure 2: Overview of the TokenMapper pipeline. Input audio is tokenized by source tokenizer A, encoded by a Transformer encoder with routing to codebook specific output heads, and projected into the token space of tokenizer B. Tokens are denoted t_i^j, where j indexes the codebook and i indexes the timestep.

et al., 2025; Ahn et al., 2024; Wu et al., 2024; Ai et al., 2024; Xin et al., 2024; Parker et al., 2024; Wu et al., 2025a), semantic and disentangled representations (Liu et al., 2024; Zhang et al., 2025b; Jiang et al., 2022; Pan et al., 2024; Huang et al., 2024a), variable rate and structured token streams (Chae et al., 2025; Zhang et al., 2025a; Li et al., 2024; Tang et al., 2024; Guo et al., 2024; Shi et al., 2024; Har-Tuv et al., 2025), and practical toolkits (Du et al., 2023; Jiang et al., 2024). Discrete tokens are further used for downstream speech and language modeling, including self supervised learning (SSL) based tokenization and automatic speech recognition (ASR) (Cui et al., 2024; Bai et al., 2024), large scale multilingual modeling (Zhang et al., 2023b; Chen et al., 2024), codec based audio language models (Ye et al., 2024; Casanova et al., 2024; Ji et al., 2025b,a; Yang et al., 2024), and high quality reconstruction (Siuzdak, 2024). While discrete tokenization has proven effective within individual models, from early music generation (Dhariwal et al., 2020) to modern speech codecs, these representations remain incompatible across models, TokenMapper addresses this gap by translating directly between heterogeneous discrete token spaces without waveform reconstruction.

2 Method

TokenMapper performs directionaware translation between heterogeneous discrete token spaces, as illustrated in Fig. 2, enabling mappings between single codebook and multi codebook tokenizers without leaving the token domain.

2.1 Problem Formulation

We consider supervised token to token translation between heterogeneous speech tokenizers. Let A and B denote a source and target tokenizer, respectively. Let $X = \{x_c[t]\}_{c=1,t=1}^{C_A,T}$ denote the discrete token sequence produced by tokenizer A, and let $Y = \{y_{c'}[t]\}_{c'=1,t=1}^{C_B,T}$ denote the corresponding token sequence produced by tokenizer B for the same utterance. Here, C_A and C_B denote the number of codebooks of tokenizers A and B, respectively, and T is the sequence length. We assume a shared token rate.

Our objective is to learn a conditional mapping

$$f_\theta : (\mathbb{N}^{C_A \times T}, \omega_{A\to B}) \to \mathbb{N}^{C_B \times T}. \quad (1)$$

where $\omega_{A\to B}$ is a learned direction embedding that specifies the source to target tokenizer pair.

TokenMapper is trained such that the predicted sequence $\hat{Y} = f_\theta(X, \omega_{A\to B})$ matches the distribution of token sequences produced by the original tokenizer of the target model. Fig. 3 illustrates TokenMapper's direction aware routing mechanism.

By defining the translation directly between discrete spaces with potentially different number of codebooks ($C_A \neq C_B$) and vocabulary structures, this formulation naturally supports mappings between single codebook and multi codebook tokenizers.

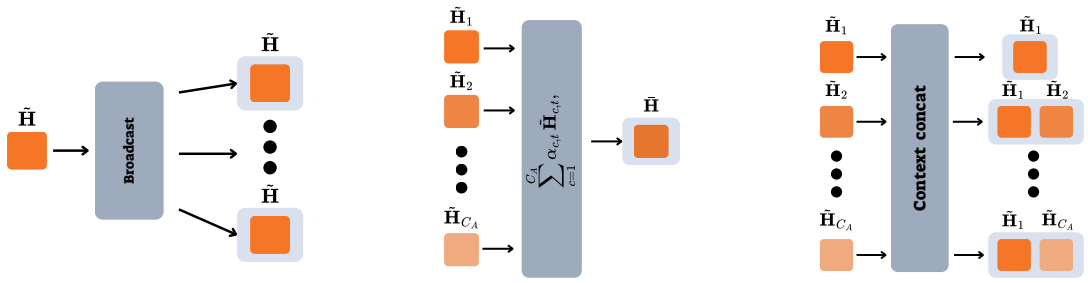
2.2 Model Architecture

2.2.1 Input Representation

Each discrete token $x_c[t]$ is first mapped to a learned token embedding $\mathbf{e}_{\text{tok}}(x_c[t]) \in \mathbb{R}^D$, where D denotes the embedding dimensionality of the encoder backbone and is shared across all token, positional, codebook, and direction embeddings. This embedding is then summed with three additional embeddings: a positional embedding $\mathbf{e}_{\text{pos}}(t)$, a codebook embedding $\mathbf{e}_{\text{cb}}(c)$, and a direction embedding $\mathbf{e}_{\text{dir}}(\omega_{A\to B})$:

$$\mathbf{h}_{c,t} = \mathbf{e}_{\text{tok}}(x_c[t]) + \mathbf{e}_{\text{pos}}(t) + \mathbf{e}_{\text{cb}}(c) + \mathbf{e}_{\text{dir}}(\omega_{A\to B}) \quad (2)$$

Stacking all token embeddings yields $\mathbf{H} \in \mathbb{R}^{C_A \times T \times D}$



(a) Single→Multi routing (b) Multi→Single routing (c) Multi→Multi routing

Figure 3: Direction conditioned encoder to codebook routing schemes. (a) Single→Multi: a single encoder stream is broadcast to all target codebooks. (b) Multi→Single: multiple encoder streams are aggregated via attention pooling. (c) Multi→Multi: each target codebook receives a concatenation of a shared base stream and its corresponding encoder stream.

2.2.2 Encoder Backbone

Encoding is performed independently for each codebook channel, with self attention restricted to the temporal axis.

$$\tilde{\mathbf{H}} = ENC(\mathbf{H}), \qquad \tilde{\mathbf{H}} \in \mathbb{R}^{C_A \times T \times D} \tag{3}$$

2.2.3 Direction Specific Output Heads

We define the post encoder processing as a routing function $\phi_{A\to B}(\omega_{A\to B}, \tilde{\mathbf{H}})$, which maps the encoder output $\tilde{\mathbf{H}}$ to the inputs of the direction specific output heads $\mathbf{Z}$:

$$\mathbf{Z} = \phi_{A\to B}(\omega_{A\to B}, \tilde{\mathbf{H}}), \qquad \mathbf{Z} \in \mathbb{R}^{C_B \times T \times D_Z} \tag{4}$$

where D_Z depends on the routing case below. The resulting stream is passed to a one layer Transformer output head followed by a linear projection to the target vocabulary. The following subsections describe the different possible routing configurations for realizing this mapping.

2.2.4 Single to Single Output Head

When both the source and target tokenizers consist of a single codebook ($C_A = C_B = 1$), the routing function is the identity:

$$\mathbf{Z} = \tilde{\mathbf{H}}, \qquad \mathbf{Z} \in \mathbb{R}^{T \times D} \tag{5}$$

2.2.5 Multi to Single Output Head

When the source tokenizer has multiple codebooks and the target tokenizer has a single codebook, the routing function $\phi_{A\to B}$ performs learned attention pooling across the codebook axis Fig. 3(b), as described in (dos Santos et al., 2016):

$$\mathbf{Z}[t] = \sum_{c=1}^{C_A} \alpha_{c,t}(\theta)\, \tilde{\mathbf{H}}_{c,t}, \qquad \sum_{c=1}^{C_A} \alpha_{c,t}(\theta) = 1 \tag{6}$$

where $\alpha_{c,t}(\theta)$ are learned attention weights.

2.2.6 Multi to Multi Output Head

When both the source and target tokenizers have multiple codebooks, the routing function $\phi_{A\to B}$ constructs a separate input stream for each target codebook Fig. 3(c) by concatenating a shared base stream, meaning the first source codebook channel with the corresponding encoded channel:

$$\mathbf{Z}_{c'} = \mathrm{concat}\left(\tilde{\mathbf{H}}_1, \tilde{\mathbf{H}}_{c'}\right), \qquad \mathbf{Z}_{c'} \in \mathbb{R}^{T \times 2D} \tag{7}$$

This design choice is motivated by hierarchical neural audio codecs, where the first codebook is typically associated with higher level semantic structure.

2.2.7 Single to Multi Output Head

For single codebook sources, $\phi_{A\to B}$ serves as a broadcast by duplication to all C_B target codebooks Fig. 3(a):

$$\mathbf{Z} = (\tilde{\mathbf{H}}, \tilde{\mathbf{H}}, \ldots, \tilde{\mathbf{H}}), \qquad \mathbf{Z} \in \mathbb{R}^{C_B \times T \times D} \tag{8}$$

2.3 Training Objective

Training uses token level cross-entropy over all target codebooks and timesteps. Given a target tokenizer with C_B codebooks, the loss is defined as

$$\mathcal{L}_{\text{CE}} = \sum_{c'=1}^{C_B} \sum_{t=1}^{T} \text{CE}\big(f_{\theta,c'}(X, \omega_{A\to B})[t], y_{c'}[t]\big) \tag{9}$$

where $y_{c'}[t]$ is the ground truth token, and $f_{\theta,c'}(X, \omega_{A\to B})[t]$ denotes the predicted logits for that codebook. All directions share a conditioned encoder, with direction specific output heads.

3 Experiments

3.1 Experimental Setup

We evaluate cross-model token translation across heterogeneous audio tokenizers. Given an utterance, audio is encoded using a source tokenizer, translated into the target token space by the proposed model, and decoded using the target tokenizer's native decoder. We evaluate all ordered source→target pairs among GLM-4-Voice, which uses a single codebook ($C = 1$), and Moshi (MiMi) and DualCodec, both of which use eight codebooks ($C = 8$). For brevity, we refer to GLM-4-Voice as GLM and to Moshi (MiMi) as Moshi throughout. Audio examples of native reconstructions and cross model translations are provided in the supplementary material.

3.2 Datasets

3.2.1 LibriSpeech

The first dataset is LibriSpeech, a large-scale English speech corpus derived from audiobook recordings (Panayotov et al., 2015). Training is performed on the `train` subset, while evaluation is conducted on the `test-clean` subset. LibriSpeech provides high-quality, well-segmented speech with relatively consistent recording conditions, making it a standard benchmark for evaluating speech representation learning and reconstruction quality.

3.2.2 VCTK

To assess generalization beyond audiobook speech, we additionally evaluate on VCTK, a multi-speaker English corpus containing speech from 110 speakers with diverse accents (Yamagishi et al., 2019). Each speaker reads approximately 400 utterances recorded under controlled conditions. Compared to LibriSpeech, VCTK introduces greater speaker diversity and accent variability, making it a useful complementary benchmark for evaluating the robustness of cross-tokenizer translation across different speaking styles and speaker characteristics.

3.3 Evaluation Metrics

We use complementary metrics to assess intelligibility and perceptual speech quality.

3.3.1 Word Error Rate (WER)

WER is computed from ASR transcripts to better reflect the contextual quality of the reconstructed audio, and is compared against the corresponding ground truth transcripts. ASR decoding is performed using Whisper large-v3 (Radford et al., 2022), a strong multilingual speech recognition model, allowing WER to serve as a proxy for semantic and linguistic preservation under token translation. Before computing WER, both hypotheses and references are normalized by removing punctuation and standardizing whitespace to match the LibriSpeech transcription format.

3.3.2 UTMOS

UTMOS is a neural mean-opinion-score predictor used to estimate perceptual speech quality and naturalness (Saeki et al., 2022). We report scores produced by the UTMOS-v2 model, which is trained to correlate with human MOS judgments and provides a reference free assessment of speech quality.

These metrics capture complementary aspects of translation quality: WER reflects semantic and linguistic preservation, while UTMOS evaluates perceptual speech naturalness independent of transcription accuracy.

3.4 Implementation Details

TokenMapper is trained for 600 epochs on a single NVIDIA RTX 6000 GPU. The model consists of a shared Transformer encoder and direction specific output heads. The encoder is a four layer Transformer with $D = 256$ and $n_{\text{heads}} = 8$. The output heads are implemented as single layer Transformers with $n_{\text{heads}} = 8$, followed by linear projections to the target codebook vocabularies. Training is performed using the Adam optimizer with a batch size of 32 and a learning rate of 5×10^{-4}.

4 Results

We now present the empirical performance of TokenMapper on cross model token translation. Results are reported on LibriSpeech test-clean and VCTK to evaluate robustness across datasets with different speaker distributions and recording conditions.

Table 1: Cross model WER (%) on LibriSpeech (lower is better). Diagonal entries correspond to native reconstructions, off diagonal entries denote source (rows)→target (columns) token translation.

WER (%) (↓)	GLM	Moshi	DualCodec
GLM	4.75	9.31	5.85
Moshi	8.55	4.02	7.91
DualCodec	8.34	9.98	3.29

Table 2: Cross model WER (%) on VCTK (lower is better). Diagonal entries correspond to native reconstructions, off diagonal entries denote source (rows)→target (columns) token translation.

WER (%) (↓)	GLM	Moshi	DualCodec
GLM	5.32	9.49	7.95
Moshi	8.79	4.39	9.18
DualCodec	8.28	9.92	2.35

Table 3: Cross model UTMOS on LibriSpeech (higher is better). Diagonal entries correspond to native reconstructions, off diagonal entries denote source→target token translation.

UTMOS (↑)	GLM	Moshi	DualCodec
GLM	3.27	2.10	2.70
Moshi	3.24	2.93	2.85
DualCodec	3.16	1.98	3.32

Table 4: Cross model UTMOS on VCTK (higher is better). Diagonal entries correspond to native reconstructions, off diagonal entries denote source→target token translation.

UTMOS (↑)	GLM	Moshi	DualCodec
GLM	3.39	2.01	2.80
Moshi	3.38	3.01	2.85
DualCodec	3.33	1.95	3.38

4.1 Semantic preservation (WER)

Semantic metric results shown in Table 1 suggest that cross model WER remains close to native reconstruction performance on LibriSpeech. While native WER ranges between 3.29-4.75%, cross-model translation remains within 5.85-9.98%, corresponding to a modest degradation of only 2.56-5.96%. The strongest direction is GLM→DualCodec with 5.85%, representing only a 2.56% increase over native DualCodec reconstruction. The most challenging direction is DualCodec→Moshi with WER of 9.98%, indicating increased difficulty when translating between multi codebook, RVQ based, tokenizers.

Table 2 shows a similar trend on VCTK, where cross model WER remains within 2.96-6.83% of native performance despite increased speaker and accent variability. This consistency across datasets suggests that TokenMapper captures tokenizer level linguistic structure rather than dataset specific acoustic properties.

Together, these results indicate that TokenMapper preserves semantic content with only limited degradation relative to native reconstruction. Performance remains stable across both single stream and multi codebook tokenizers, suggesting that the linguistic information encoded in these representations is largely transferable.

4.2 Acoustic quality (UTMOS)

Table 3 shows that perceptual quality remains reasonably close to native reconstruction on LibriSpeech. Native scores range from 2.93-3.32, while cross model translation achieves 1.98-3.24, corresponding to a degradation of 0.03-1.06. Translations into GLM remain particularly strong, reaching 3.24 and 3.16 compared to a native score of 3.27.

For RVQ based targets, perceptual quality is more variable, with the lowest score observed for DualCodec→Moshi of 1.98. This again highlights the increased difficulty of reconstructing hierarchical acoustic detail across heterogeneous multi codebook token spaces.

Table 4 shows a similar pattern on VCTK. Translations into GLM achieve the highest perceptual quality with UTMOS of 3.38 and 3.33, closely approaching the native GLM score of 3.39, while DualCodec→Moshi remains the most challenging direction with a score of 1.95.

These results suggest that while semantic content transfers reliably, reconstructing fine acoustic structure remains more sensitive to tokenizer mismatch.

4.3 Latency reduction

We evaluate end to end latency under two transfer pipelines. The waveform bridging baseline performs full reconstruction in both domains: $\text{encode}_{src} \rightarrow \text{decode}_{src} \rightarrow \text{encode}_{tgt} \rightarrow \text{decode}_{tgt}$. In contrast, TokenMapper operates entirely in the discrete domain, requiring only $\text{encode}_{src} \rightarrow$ TokenMapper $\rightarrow \text{decode}_{tgt}$. As shown in Table 5, latency reduction is direction dependent. When

Table 5: End to end latency comparison between waveform bridging (Base) and TokenMapper (TM). Time saved is reported in milliseconds and relative percentage.

Direction	Base	TM	Δ ms	Δ%
Moshi→GLM	800.9	755.3	45.6	5.7
GLM→Moshi	802.4	52.0	750.4	93.5
Moshi→DualCodec	280.6	35.3	245.3	87.4
GLM→DualCodec	1029.2	56.9	972.3	94.5
DualCodec→Moshi	283.4	253.9	29.5	10.4
DualCodec→GLM	1027.6	978.8	48.8	4.8

baseline runtime is dominated by neural decoding e.g., GLM→DualCodec, TokenMapper reduces latency by up to 94.5%, while directions with more balanced runtimes yield modest gains of 4.8-10.4%. These improvements arise since TokenMapper bypasses two computationally expensive stages in the conventional pipeline: waveform synthesis by the source decoder and re-tokenization by the target encoder. The largest gains occur when decoding dominates the baseline pipeline, highlighting the practical benefit of eliminating intermediate waveform reconstruction.

5 Ablations

To better understand the contribution of individual architectural components and assess the effect of different design choices, we perform a set of targeted ablation studies on the proposed model. Specifically, we evaluate the following modifications:

- **No direction conditioning.** We remove the direction embedding that informs the model about the source→target mapping and instead use a shared representation across all translation directions.
- **Single shared output head.** We replace direction specific output heads with a single shared projection head used for all translation directions.
- **MLP projection head.** We replace the Transformer based output heads in the proposed model with a simple fully connected MLP projection heads (4 layers).
- **Cross codebook attention in the encoder.** We ablate the original design by enabling self attention across both time and codebook dimensions, allowing interactions between codebook channels within the encoder.
- **Full context concatenation.** For multi→multi mappings, the proposed model concatenates a shared base stream (codebook 1) with the corresponding target codebook channel. We ablate this design by concatenating all source codebook representations and feeding the full context to each target projection head.

These ablations are evaluated across representative translation directions covering the main structural cases: single→multi (GLM→Moshi), multi→single (Moshi→GLM), and multi→multi (Moshi→DualCodec).

The ablation results in Table 6 confirm that direction conditioning and direction specific projection heads are critical for accurate token translation. Removing direction conditioning increases WER from 9.31% to 57.62% for GLM→Moshi and from 7.91% to 77.21% for Moshi→DualCodec. Using a single shared output head leads to even larger degradation from WER 8.55% to 83.01% for Moshi→GLM.

Replacing the Transformer based output head with a simple MLP also degrades performance e.g., WER 9.31% to 74.23%, showing the importance of sequence modeling in the output stage. Preserving codebook structure is similarly important: cross codebook attention increases WER from 8.55% to 18.86% and from 7.91% to 22.83% while also adding significant computational overhead. Full context concatenation further degrades multi→multi translation, WER from 7.91% to 26.79%.

Overall, these results validate the proposed architectural choices and highlight the importance of direction awareness, structural alignment, and sequence aware output modeling.

6 Conclusion

We introduced TokenMapper, a framework for direct token to token translation between heterogeneous speech tokenizers that enables cross model communication without intermediate waveform reconstruction.

Across both LibriSpeech and VCTK, results show that TokenMapper preserves semantic content while remaining close to the native reconstruction upper bound of each target tokenizer. Rather than absolute WER alone, a more informative measure is the performance gap relative to native decoding. On LibriSpeech, cross model translation increases

Table 6: Ablation results for representative translation directions. Lower WER is better, higher UTMOS is better.

Direction	Model variant	WER (%) (↓)	UTMOS (↑)
GLM→Moshi	**Full model**	9.31	2.10
	No direction conditioning	57.62	1.68
	Single shared output head	88.82	1.71
	MLP output head (4 layers)	74.23	1.66
Moshi→GLM	**Full model**	8.55	3.24
	No direction conditioning	33.52	3.11
	Single shared output head	83.01	2.47
	MLP output head (4 layers)	67.52	2.52
	Cross codebook attention	18.86	3.26
Moshi→DualCodec	**Full model**	7.91	2.85
	No direction conditioning	77.21	1.61
	Single shared output head	96.75	1.96
	MLP output head (4 layers)	99.46	1.53
	Cross codebook attention	22.83	2.36
	Full context concatenation	26.79	2.12

WER by only 2.56 to 5.96% compared to native reconstruction, while on VCTK the gap remains similarly limited at 2.96 to 6.83% despite increased speaker variability. These results indicate that TokenMapper operates near the intrinsic performance limits imposed by the target tokenizer.

Perceptual quality follows a similar pattern. While acoustic quality varies depending on tokenizer structure, degradation relative to native reconstruction remains limited. On LibriSpeech, UTMOS decreases by as little as 0.03 and at most 1.06 compared to native decoding, while on VCTK translations into GLM reach 3.38 and 3.33 compared to a native score of 3.39. These findings suggest that while reconstructing fine acoustic structure remains sensitive to tokenizer mismatch, TokenMapper is able to preserve most perceptual quality achievable by the target model.

Analysis across tokenizers further reveals that translation difficulty is strongly influenced by tokenizer structure. Translations into RVQ based tokenizers remain more challenging due to the need to reconstruct hierarchical residual acoustic detail, whereas translations into single stream tokenizers such as GLM remain consistently closer to native performance. This suggests that semantic information transfers more easily across token spaces than fine acoustic structure encoded in multi stage residual quantization.

Beyond reconstruction quality, TokenMapper also provides substantial efficiency benefits. By removing redundant decoding and re-encoding stages, TokenMapper reduces end to end latency by up to 94.5 percent, with the largest improvements observed in pipelines dominated by neural waveform synthesis. These results demonstrate that token level interoperability can simultaneously approach native reconstruction quality while significantly improving system efficiency.

Taken together, these results suggest that discrete speech token spaces learned by independently developed models may share sufficient semantic structure to enable translation that approaches native decoding performance. This supports the view that speech tokens may function as a model agnostic intermediate representation and suggests a path toward modular speech systems where interoperability can be achieved directly in the token domain.

Future work may explore scaling TokenMapper to additional tokenizers, reducing the remaining gap to native reconstruction, improving acoustic detail recovery for structurally mismatched RVQ codecs, and extending the approach to streaming and real time speech systems.

7 Limitations

While TokenMapper achieves strong semantic preservation, perceptual quality can remain below native reconstructions for structurally mismatched multi codebook targets. This behavior is observed across both LibriSpeech and VCTK, for example in the DualCodec→Moshi direction, where UTMOS remains lower than the native Moshi reconstruction of 1.98 vs. 2.93 on LibriSpeech and 1.95 vs. 3.01 on VCTK. These results suggest that accurately reconstructing fine grained acoustic detail across heterogeneous RVQ codec hierarchies remains challenging.

8 Generative AI Use Disclosure

Generative AI tools were utilized to edit and polish the phrasing of this manuscript.

References

Sunghwan Ahn, Beom Jun Woo, Min Hyun Han, Chanyeong Moon, and Nam Soo Kim. 2024. Hilcodec: High-fidelity and lightweight neural audio codec. *Preprint*, arXiv:2405.04752.

Yang Ai, Xiao-Hang Jiang, Ye-Xin Lu, Hui-Peng Du, and Zhen-Hua Ling. 2024. Apcodec: A neural audio codec with parallel amplitude and phase spectrum encoding and decoding. *Preprint*, arXiv:2402.10533.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2020. vq-wav2vec: Self-supervised learning of discrete speech representations. *Preprint*, arXiv:1910.05453.

He Bai, Tatiana Likhomanenko, Ruixiang Zhang, Zijin Gu, Zakaria Aldeneh, and Navdeep Jaitly. 2024. dmel: Speech tokenization made simple. arXiv:2407.15835v2.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. Audiolm: a language modeling approach to audio generation. *Preprint*, arXiv:2209.03143.

Edresson Casanova, Ryan Langman, Paarth Neekhara, Shehzeen Hussain, Jason Li, Subhankar Ghosh, Ante Jukić, and Sang gil Lee. 2024. Low frame-rate speech codec: a codec designed for fast high-quality speech llm training and inference. *Preprint*, arXiv:2409.12117.

Edresson Casanova, Paarth Neekhara, Ryan Langman, Shehzeen Hussain, Subhankar Ghosh, Xuesong Yang, Ante Jukić, Jason Li, and Boris Ginsburg. 2025. Nanocodec: Towards high-quality ultra fast speech llm inference. *Preprint*, arXiv:2508.05835.

Yunkee Chae, Woosung Choi, Yuhta Takida, Junghyun Koo, Yukara Ikemiya, Zhi Zhong, Kin Wai Cheuk, Marco A. Martínez-Ramírez, Kyogu Lee, Wei-Hsiang Liao, and Yuki Mitsufuji. 2025. Variable bitrate residual vector quantization for audio coding. *Preprint*, arXiv:2410.06016.

Wenxi Chen, Xinsheng Wang, Ruiqi Yan, Yushen Chen, Zhikang Niu, Ziyang Ma, Xiquan Li, Yuzhe Liang, Hanlin Wen, Shunshun Yin, Ming Tao, and Xie Chen. 2025. Sac: Neural speech codec with semantic-acoustic dual-stream quantization. *Preprint*, arXiv:2510.16841.

William Chen, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiatong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu, and Shinji Watanabe. 2024. Towards robust speech representation learning for thousands of languages. *Preprint*, arXiv:2407.00837.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *Preprint*, arXiv:2311.07919.

Mingyu Cui, Daxin Tan, Yifan Yang, Dingdong Wang, Huimeng Wang, Xiao Chen, Xie Chen, and Xunying Liu. 2024. Exploring ssl discrete tokens for multilingual asr. *Preprint*, arXiv:2409.08805.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A generative model for music. *Preprint*, arXiv:2005.00341.

Cicero dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. 2016. Attentive pooling networks. *Preprint*, arXiv:1602.03609.

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. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *Preprint*, arXiv:2407.05407.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *Preprint*, arXiv:2412.10117.

Zhihao Du, Shiliang Zhang, Kai Hu, and Siqi Zheng. 2023. Funcodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec. arXiv:2309.07405.

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

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *Preprint*, arXiv:2410.00037.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025a. Llama-omni: Seamless speech interaction with large language models. *Preprint*, arXiv:2409.06666.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. 2025b. Llama-omni2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. *Preprint*, arXiv:2505.02625.

Yitian Gong, Luozhijie Jin, Ruifan Deng, Dong Zhang, Xin Zhang, Qinyuan Cheng, Zhaoye Fei, Shimin Li, and Xipeng Qiu. 2025. Xy-tokenizer: Mitigating the semantic-acoustic conflict in low-bitrate speech codecs. *Preprint*, arXiv:2506.23325.

Haohan Guo, Fenglong Xie, Kun Xie, Dongchao Yang, Dake Guo, Xixin Wu, and Helen Meng. 2024. Socodec: A semantic-ordered multi-stream speech codec for efficient language model based text-to-speech synthesis. *Preprint*, arXiv:2409.00933.

Nadav Har-Tuv, Or Tal, and Yossi Adi. 2025. Past: Phonetic-acoustic speech tokenizer. arXiv:2505.14470v1.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *Preprint*, arXiv:2106.07447.

Lun Huang, Qiang Qiu, and Guillermo Sapiro. 2024a. Pq-vae: Learning hierarchical discrete representations with progressive quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Zhichao Huang, Chutong Meng, and Tom Ko. 2024b. Repcodec: A speech representation codec for speech tokenization. *Preprint*, arXiv:2309.00169.

Teerapat Jenrungrot, Michael Chinen, W. Bastiaan Kleijn, Jan Skoglund, Zalán Borsos, Neil Zeghidour, and Marco Tagliasacchi. 2023. Lmcodec: A low bitrate speech codec with causal transformer models. *Preprint*, arXiv:2303.12984.

Shengpeng Ji, Minghui Fang, Jialong Zuo, Ziyue Jiang, Dingdong Wang, Hanting Wang, Hai Huang, and Zhou Zhao. 2025a. Language-codec: Bridging discrete codec representations and speech language models. *Preprint*, arXiv:2402.12208.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. 2025b. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *Preprint*, arXiv:2408.16532.

Jiawei Jiang, Linping Xu, Dejun Zhang, Qingbo Huang, Xianjun Xia, and Yijian Xiao. 2025. Ldcodec: A high quality neural audio codec with low-complexity decoder. *Preprint*, arXiv:2510.15364.

Xiao-Hang Jiang, Yang Ai, Rui-Chen Zheng, Hui-Peng Du, Ye-Xin Lu, and Zhen-Hua Ling. 2024. Mdct-codec: A lightweight mdct-based neural audio codec towards high sampling rate and low bitrate scenarios. *Preprint*, arXiv:2411.00464.

Xue Jiang, Xiulian Peng, Yuan Zhang, and Yan Lu. 2022. Disentangled feature learning for real-time neural speech coding. arXiv:2211.11960.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *Preprint*, arXiv:2403.03100.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Hanzhao Li, Liumeng Xue, Haohan Guo, Xinfa Zhu, Yuanjun Lv, Lei Xie, Yunlin Chen, Hao Yin, and Zhifei Li. 2024. Single-codec: Single-codebook speech codec towards high-performance speech generation. *Preprint*, arXiv:2406.07422.

Jiaqi Li, Xiaolong Lin, Zhekai Li, Shixi Huang, Yuancheng Wang, Chaoren Wang, Zhenpeng Zhan, and Zhizheng Wu. 2025a. Dualcodec: A low-frame-rate, semantically-enhanced neural audio codec for speech generation. *Preprint*, arXiv:2505.13000.

Jiaqi Li, Yao Qian, Yuxuan Hu, Leying Zhang, Xiaofei Wang, Heng Lu, Manthan Thakker, Jinyu Li, Sheng Zhao, and Zhizheng Wu. 2025b. Flexicodec: A dynamic neural audio codec for low frame rates. *Preprint*, arXiv:2510.00981.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, Jianhua Xu, Haoze Sun, Zenan Zhou, and Weipeng Chen. 2025c. Baichuan-audio: A unified framework for end-to-end speech interaction. *Preprint*, arXiv:2502.17239.

Haohe Liu, Xuenan Xu, Yi Yuan, Mengyue Wu, Wenwu Wang, and Mark D. Plumbley. 2024. Semanticodec: An ultra low bitrate semantic audio codec for general sound. *IEEE Journal of Selected Topics in Signal Processing*, 18(8):1448–1461.

Kentaro Mitsui, Koh Mitsuda, Toshiaki Wakatsuki, Yukiya Hono, and Kei Sawada. 2024. Pslm: Parallel generation of text and speech with llms for low-latency spoken dialogue systems. *Preprint*, arXiv:2406.12428.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoit Sagot, and Emmanuel Dupoux. 2024. Spirit lm: Interleaved spoken and written language model. *Preprint*, arXiv:2402.05755.

Zhikang Niu, Sanyuan Chen, Long Zhou, Ziyang Ma, Xie Chen, and Shujie Liu. 2024. Ndvq: Robust neural audio codec with normal distribution-based vector quantization. *Preprint*, arXiv:2409.12717.

Yu Pan, Lei Ma, and Jianjun Zhao. 2024. Promptcodec: High-fidelity neural speech codec using disentangled representation learning based adaptive feature-aware prompt encoders. arXiv:2404.02702v2.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Julian D. Parker, Anton Smirnov, Jordi Pons, C. J. Carr, Zack Zukowski, Zach Evans, and Xubo Liu. 2024. Scaling transformers for low-bitrate high-quality speech coding. arXiv:2411.19842.

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.

Yong Ren, Tao Wang, Jiangyan Yi, Le Xu, Jianhua Tao, Chuyuan Zhang, and Junzuo Zhou. 2024. Fewer-token neural speech codec with time-invariant codes. *Preprint*, arXiv:2310.00014.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *Preprint*, arXiv:2204.02152.

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. *Preprint*, arXiv:2304.09116.

Jiatong Shi, Xutai Ma, Hirofumi Inaguma, Anna Sun, and Shinji Watanabe. 2024. Mmm: Multi-layer multi-residual multi-stream discrete speech representation from self-supervised learning model. arXiv:2406.09869v1.

Jiatong Shi, Haoran Wang, William Chen, Chenda Li, Wangyou Zhang, Jinchuan Tian, and Shinji Watanabe. 2025. Pure codec: Progressive unfolding of residual entropy for speech codec learning. arXiv:2511.22687.

Hubert Siuzdak. 2024. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. *Preprint*, arXiv:2306.00814.

Hubert Siuzdak, Florian Grötschla, and Luca A. Lanzendörfer. 2024. Snac: Multi-scale neural audio codec. *Preprint*, arXiv:2410.14411.

Yuxun Tang, Yuning Wu, Jiatong Shi, and Qin Jin. 2024. Singomd: Singing oriented multi-resolution discrete representation construction from speech models. *Preprint*, arXiv:2406.08905.

Viet Anh Khoa Tran, David Thulke, Yingbo Gao, Christian Herold, and Hermann Ney. 2022. Does joint training really help cascaded speech translation? *Preprint*, arXiv:2210.13700.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2018. Neural discrete representation learning. *Preprint*, arXiv:1711.00937.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6).

Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. 2023. Viola: Unified codec language models for speech recognition, synthesis, and translation. *Preprint*, arXiv:2305.16107.

Haibin Wu, Naoyuki Kanda, Sefik Emre Eskimez, and Jinyu Li. 2025a. Ts3-codec: Transformer-based simple streaming single codec. *Preprint*, arXiv:2411.18803.

Yi-Chiao Wu, Israel D. Gebru, Dejan Marković, and Alexander Richard. 2023. Audiodec: An open-source streaming high-fidelity neural audio codec. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 1–5. IEEE.

Yi-Chiao Wu, Dejan Marković, Steven Krenn, Israel D. Gebru, and Alexander Richard. 2024. Scoredec: A phase-preserving high-fidelity audio codec with a generalized score-based diffusion post-filter. *Preprint*, arXiv:2401.12160.

Yi-Chiao Wu, Dejan Marković, Steven Krenn, Israel D. Gebru, and Alexander Richard. 2025b. Complexdec: A domain-robust high-fidelity neural audio codec with complex spectrum modeling. *Preprint*, arXiv:2502.02019.

Detai Xin, Xu Tan, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2024. Bigcodec: Pushing the limits of low-bitrate neural speech codec. *Preprint*, arXiv:2409.05377.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92).

Dongchao Yang, Haohan Guo, Yuanyuan Wang, Rongjie Huang, Xiang Li, Xu Tan, Xixin Wu, and Helen Meng. 2024. Uniaudio 1.5: Large language model-driven audio codec is a few-shot audio task learner. *Preprint*, arXiv:2406.10056.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. 2023a. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *Preprint*, arXiv:2305.02765.

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. 2023b. Diffsound: Discrete diffusion model for text-to-sound generation. *Preprint*, arXiv:2207.09983.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, Yike Guo, and Wei Xue. 2024. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. *Preprint*, arXiv:2408.17175.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *Preprint*, arXiv:2107.03312.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *Preprint*, arXiv:2412.02612.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *Preprint*, arXiv:2305.11000.

Hanglei Zhang, Yiwei Guo, Zhihan Li, Xiang Hao, Xie Chen, and Kai Yu. 2025a. Unlocking temporal flexibility: Neural speech codec with variable frame rate. arXiv:2505.16845.

Ruonan Zhang, Xiaoyang Hao, Yichen Han, Junjie Cao, Yue Liu, and Kai Zhang. 2025b. Mbcodec:thorough disentangle for high-fidelity audio compression. *Preprint*, arXiv:2509.17006.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024. Speechtokenizer: Unified speech tokenizer for speech large language models. *Preprint*, arXiv:2308.16692.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, and 8 others. 2023b. Google usm: Scaling automatic speech recognition beyond 100 languages. *Preprint*, arXiv:2303.01037.

Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Jade Copet, Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi. 2024. Masked audio generation using a single non-autoregressive transformer. In *International Conference on Representation Learning*, volume 2024, pages 43249–43268.