CROSS-LINGUAL INTERLEAVING FOR SPEECH LANGUAGE MODELS

Adel Moumen, Guangzhi Sun, Philip C. Woodland

Department of Engineering, University of Cambridge, UK
{am3303,gs534,pcw117}@cam.ac.uk

ABSTRACT

Spoken Language Models (SLMs) aim to learn linguistic competence directly from speech using discrete units, widening access to Natural Language Processing (NLP) technologies for languages with limited written resources. However, progress has been largely English-centric due to scarce spoken evaluation benchmarks and training data, making cross-lingual learning difficult. We present a cross-lingual interleaving method that mixes speech tokens across languages without textual supervision. We also release an EN–FR training dataset, *TinyStories* ($\sim$ 42k hours), together with EN–FR spoken *StoryCloze* and *TopicCloze* benchmarks for cross-lingual semantic evaluation, both synthetically generated using GPT-4. On 360M and 1B SLMs under matched training-token budgets, interleaving improves monolingual semantic accuracy, enables robust cross-lingual continuation, and strengthens cross-lingual hidden-state alignment. Taken together, these results indicate that cross-lingual interleaving is a simple, scalable route to building multilingual SLMs that understand and converse across languages. All resources will be made open-source to support reproducibility.

***Index Terms*—** Spoken language models, cross-lingual learning, interleaving, multilinguality

1. INTRODUCTION

Spoken Language Models (SLMs) aim to learn language directly from speech, without relying on textual supervision. This line of work emerged from *'textless' NLP* [1], which combines discrete speech representations [2] and neural language modelling [3, 4] to capture both acoustic and linguistic regularities from audio alone. Within this framework, *Generative Spoken Language Modelling* (GSLM) formalises the objective by training an autoregressive LM on sequences of discrete speech units and reconstructing waveforms with a unit vocoder, with the motivation of broadening access to NLP in settings where written resources are scarce.

Despite rapid progress, the *cross-lingual* setting, which aims to train a single SLM over multiple languages, remains comparatively under-explored [5–7]. Two practical bottlenecks persist: (i) evaluations are predominantly English-centric, hindering measurement of cross-lingual semantic competence; and (ii) training methods that encourage sharing across languages often depend on *text*, for example via speech–text *interleaving* in Text–Speech LMs (TSLMs) [5, 8–10], which is not directly compatible with *'textless'* SLMs. As a result, the original promise of *'textless'* NLP learning from audio alone at scale and across languages has not yet been fully realised.

We address these limitations by introducing a *cross-lingual interleaving* scheme that mixes speech tokens from different languages within the same training sequence, using sentence-level alignments but no text tokens. Exposure to interleaved multilingual contexts encourages a shared representational subspace and promotes transfer, while remaining compatible with pure *'textless'* pipelines. To make this practical and measurable, we curate bilingual, sentence-aligned spoken resources and new spoken semantic benchmarks. Our contributions are as follows:

(i) We propose a cross-lingual interleaving strategy that concatenates sentence-aligned speech token sequences across languages, promoting a shared representational subspace.

(ii) We release a French–English, sentence-aligned spoken corpus derived from *TinyStories* [11] (approximately 42k hours in total) together with bilingual spoken *StoryCloze* and *TopicCloze* benchmarks [12] for semantic evaluation synthetically generated using GPT-4 [13].

(iii) We show on 360M and 1B parameter SLMs that the method yields *positive transfer* to monolingual semantic tasks and strong *cross-lingual* performance, under a matched training-token budget. Additionally, representation analyses indicate stronger cross-lingual alignment in the hidden states.

Section 2 situates our approach with respect to SLMs, speech–text interleaving, and multilingual training. Section 3 formalises the modelling setup and the cross-lingual interleaving scheme; Sections 4–6 present datasets, experiments, and analyses.

2. RELATED WORK

2.1. Speech Language Models

Early SLMs instantiate the *'textless'* pipeline by modelling discrete speech units and decoding back to waveforms. GSLM [1] trains a decoder-only LM over units and synthesises speech with a unit-based vocoder; pGSLM [14] augments units with prosodic tokens to improve expressiveness; and dGSLM [15] introduces a dual-tower variant for two-channel dialogues. While these systems can produce locally fluent speech, long-form coherence is challenging. SpeechSSM [16] explores state-space models [17] for long-form generation, TWIST [18] shows that initialising from a pre-trained text Large Language Model (LLM) can boost semantics, and AudioLM [19] cascades semantic and acoustic token streams. More recently, Slamming [20] provides a competitive compute-efficient recipe, and Align-SLM [21] fine-tunes SLMs with direct preference optimisation. Our work stays within the speech-only paradigm but targets *cross-lingual* sharing through a new learning method.

2.2. Joint Speech–Text Interleaving

TSLMs extend SLMs with text to encourage cross-modal transfer. VoxtLM/SUTLM [8] jointly trains on ASR, TTS, and continuation, and identifies *interleaving* (switching between text and speech

This work was supported by the Cambridge Commonwealth, European & International Trust.

inside a sequence) as particularly effective for building a shared space. Spirit-LM [5] scales this recipe and emphasises interleaving as a key quality driver; SmolTolk [9] proposes architecture refinements for vocabulary expansion; and GLM4-Voice [6] leverages a text-to-token model [10] to generate audio tokens on-the-fly, sidestepping explicit aligners [22, 23]. However, these approaches depend on text and interleave within a single language, limiting direct applicability to textless, cross-lingual SLM training. We instead interleave *speech tokens across languages* without introducing text.

2.3. Cross-Lingual Training

Multilingual training for speech LMs is gaining traction but remains under-evaluated on spoken semantic understanding beyond English. Spirit-LM [5] includes multiple languages at scale but reports only English spoken evaluations; GLM4-Voice [6] trains on Chinese and English yet evaluates English semantics; and scheduled interleaving for speech-to-speech translation [7] interleaves *monolingually* with no cross-language interaction during continuation. In text, mixing languages can induce positive transfer [24]. Our method brings this intuition to the *'textless'* speech regime by interleaving sentence-aligned speech sequences across languages, enabling both cross-lingual abilities and improved monolingual semantics under a matched training-token budget.

3. SPOKEN LANGUAGE MODELLING

The task of modelling spoken language from raw audio typically comprises three stages: (i) quantising a waveform $\mathbf{x} \in \mathbb{R}^T$ into a sequence of L discrete speech units with $L \ll T$; (ii) training an SLM under a next-token objective on the quantised sequence; and (iii) synthesising a waveform from the discrete units via a neural decoder. We first describe speech tokenisation, then the SLM objective, and finally our cross-lingual interleaving scheme.

3.1. Speech Language Models

Speech tokenisers. Speech tokenisers transform raw waveforms into discrete representations. Let $\mathcal{X}$ denote the domain of audio samples and let a waveform be $\mathbf{x} = (x_1, \dots, x_T) \in \mathcal{X}^T$. An encoder $f : \mathbb{R}^T \to \mathbb{R}^{T' \times d_e}$ produces a sequence of continuous features $f(\mathbf{x}) = (\mathbf{v}_1, \dots, \mathbf{v}_{T'})$, sampled at a lower frame rate $T' \ll T$. A quantiser q maps these to discrete units $\mathbf{s} = (s_1, \dots, s_{T'})$ with $s_i \in \{1, \dots, K\}$, where K is the codebook size. Common choices include k-means and residual vector quantisation (RVQ).

In this work we adopt Mimi [25], a neural audio codec [26, 27] implemented as a convolutional auto-encoder with RVQ in the latent space. Following prior evidence that early codebooks capture semantics, we *only* model the first (semantic) codebook during language modelling. The coarse codebook in Mimi is semantically distilled from the WavLM representations [28], which biases the first codebook to be more semantic.

Speech language models. Given a sequence of L speech tokens $\mathbf{s} = (s_1, \dots, s_L)$ over vocabulary $\mathcal{V}_s$, an SLM parameterises

$$p_\theta(\mathbf{s}) = \prod_{i=1}^{L} p_\theta(s_i \mid s_{<i}) \,. \tag{1}$$

Decoder-only transformers [29] are typically used as SLM and trained by minimising the autoregressive negative log-likelihood

$$\mathcal{L}_{\text{LM}} = -\sum_{i=1}^{L} \log p_\theta(s_i \mid s_{<i}) \,. \tag{2}$$

Each token s_i is embedded via $E \in \mathbb{R}^{|\mathcal{V}_s| \times d}$ to $\mathbf{e}_i = E[s_i]$. A stack of m causal transformer blocks maps $(\mathbf{e}_1, \dots, \mathbf{e}_L)$ to contextual states $(\mathbf{h}_1, \dots, \mathbf{h}_L)$, with $\mathbf{h}_i$ depending only on $s_{<i}$. A projection $U \in \mathbb{R}^{d \times |\mathcal{V}_s|}$ yields logits $\boldsymbol{\ell}_i = U^\top \mathbf{h}_i$ and the predictive distribution $p_\theta(s_i \,|\, s_{<i}) = \text{softmax}(\boldsymbol{\ell}_i)$.

Table 1. Cross-Lingual *TinyStories* per-language statistics. Durations are computed over the audio stories in each language.

Split	Lang	Total (h)	Avg (s)	Min (s)	Max (s)
Train	EN	20294.90	55.52	3.12	140.48
	FR	21120.66	57.77	3.12	135.84
Validation	EN	298.04	55.22	4.64	121.44
	FR	311.64	57.74	4.08	120.64

Table 2. Cross-Lingual spoken *StoryCloze* and spoken *TopicCloze* per-language statistics. Durations are computed over the audio stories in each language.

Split	Lang	Total (h)	Avg (s)	Min (s)	Max (s)
StoryCloze	EN	15.68	15.09	6.32	24.96
	FR	16.75	16.12	5.12	26.16
TopicCloze	EN	15.64	15.05	5.92	24.08
	FR	16.76	16.13	6.48	27.20

3.2. Cross-Lingual Interleaving Training

Standard SLM training predicts the next speech token within a monolingual sequence. Interleaving has been proposed in TSLMs to bridge text–speech modality gaps by switching at word boundaries [5, 6, 8, 30], but such methods rely on text tokens and are thus incompatible with *textless* SLMs; moreover, they do not readily support mixing multiple languages [24].

We therefore formulate a *textless, cross-lingual interleaving* scheme that operates directly on speech tokens. Assuming sentence-level alignments across languages (see Section 4), let $A_j^{(l)}$ denote the jth sentence in language l, represented as a sequence of speech tokens. Given two languages l_1 and l_2, we define an interleaved sequence

$$s = \big(A_1^{(l_1)} \,\|\, A_1^{(l_2)} \,\|\, A_2^{(l_1)} \,\|\, A_2^{(l_2)} \,\|\, \dots\big),$$

where $\|$ denotes concatenation at sentence boundaries.[1] Training remains purely autoregressive with objective (2), but exposure to interleaved contexts encourages cross-lingual semantic transfer. In practice, interleaved and monolingual sequences are sampled according to a fixed ratio. Having defined the mechanism, we next describe the sentence-aligned spoken corpora that make such training feasible (see Section 4).

4. CROSS-LINGUAL SPOKEN DATASETS

Training SLMs is compute-intensive and data-hungry: effective systems require large speech corpora, while cross-lingual interleaving

[1] Word-level switching is also possible when alignments are available; we use sentence-level switching unless stated otherwise.

(see Section 3.2) additionally demands alignment. To make such training practical, we release a French–English, sentence-aligned spoken corpus derived from *TinyStories* [11], totalling approximately 42k hours of speech ($\approx$ 20.6k in English and $\approx$ 21.4k in French, including validation; Table 1). This resource enables sentence-level interleaving and provides scale sufficient to amortise SLM training costs. In addition, we introduce bilingual spoken versions of *StoryCloze* and *TopicCloze* for evaluating cross-lingual semantic understanding (Table 2). We next summarise how these datasets are constructed.

4.1. Cross-Lingual TinyStories

TinyStories is a synthetic corpus of short narratives whose vocabulary targets 3–4-year-olds, generated by large language models to probe commonsense reasoning in text LMs. Its spoken variant has proved effective for SLM training, improving both modelling and generative metrics [20, 30]. Stories are self-contained, causally structured, and fit within a typical SLM context windows.

In order to preserve cross-lingual semantic coherence, we construct sentence-level alignments that preserve order and meaning. A high-quality MT system (GPT-4) [13] translates each sentence with full story context; the entire validation set and 600k training sentences are translated into French while maintaining English alignment. We then synthesise speech with a multi-speaker TTS system [31][2] (a 1.6B model using delayed-streams modelling first introduced in Moshi [25]).

We subsample 500 cross-lingual stories from the EN-FR *sTinyStories* validation set, for each candidate voice we synthesise EN–FR pairs with the same voice. Using cosine similarity between Speaker Verification (SV) [28][3] embeddings, we retain voices scoring > 0.90 on average across paired sentences, yielding 44 voices (26 female, 18 male). We synthesise the final sets with these 44 voices, reusing the same speaker across languages to promote cross-lingual speaker consistency. The *sTinyStories* corpus statistics are given in Table 1.

4.2. Cross-Lingual StoryCloze and TopicCloze

Assessing cross-lingual semantic abilities of SLMs is challenging due to the scarcity of spoken benchmarks beyond English [5, 6]. Inspired by *StoryCloze* [12] and its spoken instantiation [18], we introduce bilingual, sentence-aligned spoken variants—*sSC* and *sTC*—in French and English.

Each benchmark comprises 4k five-sentence stories: the model observes the first four sentences and must assign higher probability (log-likelihood) to the true ending than to an adversarial alternative. *sSC* uses semantically incompatible endings, probing causal/temporal commonsense; *sTC* draws negatives from different topics, probing topical coherence.

Corpus construction mirrors that in Section 4.1: sentences are translated with GPT-4 using full story context, then synthesised with the same TTS system to preserve timing and prosody. To minimise confusion from speaker variability, we use a single female speaker (the top SV performer on validation) across both languages. Dataset statistics are given in Table 2.

5. EXPERIMENTAL SETUP

Having introduced the interleaving mechanism and the datasets that support it, we now describe our training, and evaluation protocol.

[2]https://github.com/kyutai-labs/delayed-streams-modeling

[3]HuggingFace Model: microsoft/wavlm-base-sv

5.1. Models and training

We use the Mimi tokeniser [25] at 12.5 Hz with vocabulary size K=2048 and RVQ with 32 codebooks; only the first (semantic) codebook is modelled. SLMs are initialised from a 1B-parameter Llama 3.2 family checkpoint [32] and a 360M-parameter Qwen2 family checkpoint [33], following TWIST [18], with a 2048-token context window ($\approx$ 2.73 minutes of speech). These initialisations are used for the 1B and 360M models. The textual embedding tables are replaced with audio embeddings for the new tokens.

Training uses 4 H100 (80 GiB) GPUs, per-GPU batch 153,600 tokens (total 614,400 tokens/step). The Adam [34] optimiser is used with settings of $(\beta_1, \beta_2) = (0.9, 0.98)$, gradient clipping 1.0, weight decay 0.1, and a linear warm-up of 5% to 5×10^{-4} followed by linear decay. Inputs are packed by concatenating samples until the target length is reached.

Training proceeds in 3 stages: *(1)* EN-only pre-training for 50k steps; *(2)* cross-lingual interleaving for 20k steps with a language sampling probability of 0.5 at the sentence level; *(3)* bilingual fine-tuning for monolingual generation in FR and EN for 15k steps. Unless noted, 1B and 360M models share this setup. All comparisons are made with a matched training-token budget.

5.2. Baselines

We compare to: **(i) FR-only monolingual**, trained solely on French; and **(ii) EN+FR without interleaving**, trained on mixed data but restricted to monolingual sequences. All systems are matched by training-token budget, ensuring comparable exposure to FR and EN tokens. By matching the total training-token budget and balancing EN/FR exposure across systems, we ensure that any gains arise from interleaving rather than from data volume or sampling differences.

5.3. Datasets

For EN we use *LibriHeavy* [35] (56k hours) plus our EN *sTinyStories* (see Section 4) split for a total of 76k hours. For FR we use our FR *sTinyStories* ($\sim$ 21k hours) split. Interleaving uses the full *sTinyStories* sentence-aligned EN–FR corpus described above.

5.4. Evaluation

sBLiMP [36] tests syntactic competence: for each grammatical/ungrammatical pair, the model should assign higher probability to the grammatical item.
sWUGGY [36] measures lexical discrimination by preferring a real word over a phonotactically similar nonce word.
sSC* and *sTC (see Section 4) evaluate semantic understanding by requiring preference for the true ending over an adversarial alternative.
Cross-lingual evaluation. We evaluate *cross-lingual sSC* and *sTC* (EN→FR, FR→EN): prompts are in one language and continuations in the other. A robust cross-lingual SLM should achieve comparable performance across these conditions.

6. EXPERIMENTAL RESULTS

The results are organised around three questions: (i) does interleaving yield *positive transfer* to monolingual performance? (ii) does it induce *cross-lingual abilities*? and (iii) does it help build a shared structure of cross-lingual speech representations? The findings are reported for both 360M and 1B SLMs.

Table 3. Cross-lingual results (accuracy, %). Best within each sub-block in **bold**, second-best underlined. Interleaving shows cross-lingual transfer in monolingual generation and cross-lingual performances while preserving semantic, syntactic, and lexical SLMs performances. The EN+FR baseline displays limited cross-lingual abilities, despite showing positive transfer.

Training setup	# Parameters	Token budget		*StoryCloze (sSC)*				*TopicCloze (sTC)*				*ZeroSpeech* (EN)	
		EN	FR	EN	FR	EN→FR	FR→EN	EN	FR	EN→FR	FR→EN	sBLiMP (EN)	sWUGGY (EN)
random	–	–	–	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00
Baseline EN	1B	46.08	–	56.06	–	–	–	66.43	–	–	–	61.96	**69.92**
Baseline EN	360M	46.08	–	**56.38**	–	–	–	65.20	–	–	–	61.62	69.32
Baseline FR	1B	–	15.36	–	55.31	–	–	–	67.07	–	–	–	–
Baseline FR	360M	–	15.36	–	56.44	–	–	–	69.85	–	–	–	–
Baseline EN+FR	1B	61.44	15.36	55.79	57.83	52.32	50.77	66.86	**71.24**	57.93	58.36	**62.29**	62.24
Baseline EN+FR	360M	61.44	15.36	55.26	57.93	50.56	51.25	66.00	69.48	55.58	57.34	61.17	67.71
Cross-lingual Interleaving	1B	52.22	6.14	54.40	55.47	54.56	52.64	62.26	63.17	63.28	63.44	52.73	56.74
Cross-lingual Interleaving	360M	52.22	6.14	55.90	57.08	**56.44**	**55.37**	64.00	68.67	**65.20**	**65.84**	55.35	59.56
Cross-Lingual Interleaving + EN+FR Fine-Tuning	1B	61.44	15.36	55.63	**58.31**	55.21	55.05	**67.45**	70.39	62.90	63.35	61.75	69.15
Cross-Lingual Interleaving + EN+FR Fine-Tuning	360M	61.44	15.36	55.74	57.50	55.10	53.92	67.07	70.55	59.86	62.28	61.08	68.62

6.1. Cross-lingual interleaving helps monolingual capabilities

Starting from monolingual baselines in Table 3, French-only training already acquires non-trivial semantic competence with scores 56.44% and 69.85% on *sSC* and *sTC* for the 360M model, respectively. The 1B model shows the same level of competence. Interestingly, in this data regime the 360M model often matches or exceeds the 1B model on French, a pattern consistent with a capacity–data mismatch: with limited French hours, the smaller model is easier to optimise and less prone to overfitting or optimisation instabilities. Relative to English baselines, the *sSC* task tends to be slightly more favourable to French, while *sTC* is generally easier overall and, for French, yields an average increase of 2.64 points over English.

Introducing interleaving *without* a final stabilisation phase produces a predictable trade-off. Semantic scores in English decrease by a small margin (e.g., for the 1B model, from 56.06% to 54.40%.), whereas syntactic and lexical probes (*sBLiMP*/*sWUGGY*) drop more noticeably (from 61.96% and 69.92% to 52.73% and 56.74%). An intuitive explanation is that exposure to rapidly alternating languages perturbs low-level regularities (agreement, morphology, phonotactics) more than high-level semantics. Despite this, French *sSC* improves modestly even though the model sees substantially fewer French tokens than the monolingual baseline; in this regime, the 360M model reaches 57.08% while the baseline with 3× more tokens obtains 56.44%. The gains therefore cannot be attributed to data volume and are better explained as transfer from English into French. *TopicCloze* in French may dip at this stage, suggesting that, on balance, raw interleaving trades a little monolingual polish for multilingual competence.

A brief bilingual stabilisation phase (fine-tuning on EN+FR without interleaving) restores most of the lost ground in English and consolidates the transfer into French. After stabilisation, English performance typically returns to within a hair's breadth of the English-only baseline, particularly on lexical discrimination. Indeed, for the 1B model, *sBLiMP* and *sWUGGY* recover to 61.75% and 69.15%, aligning with the EN baseline. At an equivalent French token budget, the stabilisation stage surpasses its monolingual FR baseline by a few points on both *sSC* and *sTC*, achieving 58.31% and 70.39% versus 55.31% and 67.07%. This two-stage recipe consistently benefits both model sizes.

6.2. Cross-lingual abilities emerge with interleaving

Looking again at Table 3 and the cross-lingual evaluations (EN→FR and FR→EN), mixed EN+FR training *without* interleaving underperforms when prompted in one language and continued in the other. Focussing on the 360M results, the baseline achieves 50.56% and 51.25% on *sSC* and 55.58% and 57.34% on *sTC* (EN→FR and FR→EN, respectively). These scores are above chance, especially on *sTC*. This gap likely reflects that *sTC* is simpler than *sSC*; thus, small but non-trivial gains on *sSC* translate into larger gains on *sTC*.

With interleaving, substantially larger cross-lingual performance is observed: *sTC* approaches monolingual accuracy and *sSC* narrows to within a few points. For the 360M interleaved model, *sSC* reaches 56.44% and 55.37%, and *sTC* reaches 65.20% and 65.84% (EN→FR and FR→EN). Comparing to the same model's monolingual results, *sSC* scores are 55.90% (EN) and 57.08% (FR), while *sTC* scores are 64.00% (EN) and 68.67% (FR). These results show that (i) the model develops cross-lingual abilities that approach monolingual performance, and (ii) relative to the EN+FR baseline, interleaving yields markedly stronger cross-lingual continuation. The effect is particularly strong at 360M, indicating that smaller models, under limited French, extract proportionally more benefit from multilingual context. With the stabilisation phase, we preserve these cross-lingual abilities with a small degradation: 55.10% and 53.92% on *sSC*, and 62.90% and 63.35% on *sTC* (EN→FR and FR→EN), representing average relative degradations of 1.39 points (*sSC*) and 4.45 points (*sTC*), offset by gains in monolingual performance. The same trends hold for the 1B models.

Finally, we measured layer-wise cosine similarity between hidden states for 1,000 aligned EN–FR sentences and found that interleaving leads to consistently stronger cross-lingual alignment than training on mixed data without interleaving, with average cosine similarities of 0.73 (EN+FR baseline), 0.75 (interleaving), and 0.76 (interleaving+stabilisation) for the 1B models.

7. CONCLUSIONS

We introduce a new cross-lingual interleaving scheme that mixes speech tokens across sentence-aligned languages without textual supervision, and we release an EN–FR training dataset, *TinyStories* (∼ 42k), together with spoken *StoryCloze* and *TopicCloze* benchmarks for EN–FR semantic evaluation. Under a matched training-token budget on 360M–1B SLMs, the method improves monolingual semantic accuracy, enables robust cross-lingual continuation, and strengthens cross-lingual hidden-state alignment. Because the method places minimal constraints on optimisation dynamics, it is well suited to scaling multilingual SLMs, thereby enabling models to understand and converse across languages.

8. REFERENCES

[1] Kushal Lakhotia et al., "On Generative Spoken Language Modeling from Raw Audio," *Proc. ACL*, vol. 9, pp. 1336–1354, 2021.

[2] Pooneh Mousavi et al., "Discrete Audio Tokens: More Than a Survey!," *Proc. TMLR*, 2025.

[3] Tom Brown et al., "Language models are few-shot learners," *Proc. NeurIPS*, vol. 33, pp. 1877–1901, 2020.

[4] Alec Radford et al., "Improving Language Understanding by Generative Pre-Training," 2018.

[5] Tu Anh Nguyen et al., "SpiRit-LM: Interleaved spoken and written language model," *Transactions of the Association for Computational Linguistics*, vol. 13, pp. 30–52, 2025.

[6] Aohan Zeng et al., "Scaling speech-text pre-training with synthetic interleaved data," in *Proc. ICLR*, 2025, vol. 2025, pp. 49396–49419.

[7] Hayato Futami et al., "Scheduled Interleaved Speech-Text Training for Speech-to-Speech Translation with LLMs," in *Proc. Interspeech*, 2025, pp. 36–40.

[8] Ju-Chieh Chou et al., ""Toward Joint Language Modeling for Speech Units and Text"," in *Proc. EMNLP*, Houda Bouamor, Juan Pino, and Kalika Bali, Eds., Singapore, Dec. 2023, pp. 6582–6593, Association for Computational Linguistics.

[9] Santiago Cuervo et al., "Text-speech language models with improved cross-modal transfer by aligning abstraction levels," *arXiv preprint arXiv:2503.06211*, 2025.

[10] Aohan Zeng et al., "Scaling speech-text pre-training with synthetic interleaved data," *arXiv preprint arXiv:2411.17607*, 2024.

[11] Ronen Eldan et al., "Tinystories: How small can language models be and still speak coherent english?," *arXiv preprint arXiv:2305.07759*, 2023.

[12] Nasrin Mostafazadeh et al., "A corpus and cloze evaluation for deeper understanding of commonsense stories," in *Proc. ACL*, Kevin Knight, Ani Nenkova, and Owen Rambow, Eds., San Diego, California, June 2016, pp. 839–849, Association for Computational Linguistics.

[13] Josh Achiam et al., "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[14] Eugene Kharitonov et al., "Text-free prosody-aware generative spoken language modeling," in *Proc. ACL*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, Eds., Dublin, Ireland, May 2022, pp. 8666–8681, Association for Computational Linguistics.

[15] Tu Anh Nguyen et al., "Generative spoken dialogue language modeling," *Proc. ACL*, vol. 11, pp. 250–266, 2023.

[16] Se Jin Park, Julian Salazar, Aren Jansen, Keisuke Kinoshita, Yong Man Ro, and RJ Skerry-Ryan, "Long-Form Speech Generation with Spoken Language Models," in *Proc. ICML*, 2025.

[17] Albert Gu, Karan Goel, and Christopher Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[18] Michael Hassid et al., "Textually Pretrained Speech Language Models," *Proc. NeurIPS*, vol. 36, pp. 63483–63501, 2023.

[19] Zalán Borsos et al., "AudioLM: a Language Modeling Approach to Audio Generation," *Proc. IEEE/ACM transactions on audio, speech, and language processing*, vol. 31, pp. 2523–2533, 2023.

[20] Gallil Maimon et al., "Slamming: Training a speech language model on one GPU in a day," in *Proc. ACL*, Vienna, Austria, July 2025, pp. 12201–12216, Association for Computational Linguistics.

[21] Guan-Ting Lin et al., "Align-SLM: Textless spoken language models with reinforcement learning from AI feedback," in *Proc. ACL*, Vienna, Austria, July 2025, pp. 20395–20411, Association for Computational Linguistics.

[22] Alec Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," 2022.

[23] Vineel Pratap et al., "Scaling speech technology to 1,000+ languages," *Proc. JMLR*, vol. 25, no. 97, pp. 1–52, 2024.

[24] Haneul Yoo et al., "Code-Switching Curriculum Learning for Multilingual Transfer in LLMs," in *Proc. ACL*, Wanxiang Che et al., Eds., Vienna, Austria, July 2025, pp. 7816–7836, Association for Computational Linguistics.

[25] Alexandre Défossez et al., "Moshi: a speech-text foundation model for real-time dialogue," *arXiv preprint arXiv:2410.00037*, 2024.

[26] Neil Zeghidour et al., "SoundStream: An end-to-end neural audio codec," *Proc. IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[27] Alexandre Défossez et al., "High Fidelity Neural Audio Compression," *Proc. TMLR*, 2023, Featured Certification, Reproducibility Certification.

[28] Sanyuan Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *Proc. IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[29] Ashish Vaswani et al., "Attention is All You Need," in *Proc. NeurIPS*, 2017.

[30] Santiago Cuervo et al., "Scaling Properties of Speech Language Models," in *Proc. ACL*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, Eds., Miami, Florida, USA, Nov. 2024, pp. 351–361, Association for Computational Linguistics.

[31] Neil Zeghidour et al., "Streaming sequence-to-sequence learning with delayed streams modeling," *arXiv preprint arXiv:2509.08753*, 2025.

[32] Abhimanyu Dubey et al., "The llama 3 herd of models," *arXiv e-prints*, pp. arXiv–2407, 2024.

[33] Qwen Team, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, 2024.

[34] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] Wei Kang et al., "Libriheavy: A 50,000 hours asr corpus with punctuation casing and context," in *Proc. ICASSP*. IEEE, 2024, pp. 10991–10995.

[36] Ewan Dunbar et al., "The zero resource speech challenge 2021: Spoken language modelling," in *Proc. Interspeech*, 2021, pp. 1574–1578.