Speaker Separation via Audio Language Modeling

Luca Lanzendörfer [*1] **Constantin Pinkl** [*1] **Florian Grötschla** [1] **Roger Wattenhofer** [1]

Abstract

We propose LlaSep, an autoregressive speaker separation model operating entirely in the discrete token domain: conditioned on a tokenized mixture and semantic features from a pretrained speech encoder, a causal language model generates per-speaker codec token streams in a single decoding pass. Trained via supervised fine-tuning on a 15k-hour synthetic multilingual dataset spanning seven languages, LlaSep matches prior baselines on LibriCSS and CallHome while delivering substantially higher audio quality.

1. Introduction

Speaker separation, the task of recovering individual speech signals from a mixture, is a core problem in multi-speaker speech processing. Real conversations are highly intermittent, alternating between silence, single-speaker regions, and overlapping speech, the latter of which remains a major failure mode for downstream tasks such as ASR (He & Whitehill, 2025). Once clean per-speaker streams are available, diarization follows directly from the non-silent regions, making high-quality separation an attractive route to robust multi-speaker understanding. Historically, separation pipelines have operated on continuous time-frequency representations using mask estimation (Luo & Mesgarani, 2018), permutation-invariant objectives (Shin et al.; Subakan et al., 2021), and clustering-based post-processing (Hershey et al., 2016a); methods such as Conv-TasNet (Luo & Mesgarani, 2018) and SepFormer (Subakan et al., 2021) advance the state-of-the-art but remain bound to task-specific continuous architectures.

Recent neural audio codecs (Zeghidour et al., 2021; Défossez et al.; Kumar et al., 2023; Siuzdak et al.) compress speech into compact discrete token sequences amenable to LLM-based modeling, enabling token-level approaches to speech synthesis (Wang et al., 2023; Du et al., 2024), music generation (Agostinelli et al., 2023; Copet et al., 2023), and audio understanding (Lanzendörfer et al., 2025b; Tang et al.). Yet applying causal language models to multi-speaker separation remains largely unexplored. We cast separation as autoregressive token generation: given a tokenized mixture of up to $N{=}4$ speakers, a causal LM generates per-speaker token streams in a single decoding pass, using a pretrained multilingual speech LM as backbone with Whisper semantic conditioning, trained via supervised fine-tuning on a large synthetic multilingual conversation dataset.

2. Related Work

2.1. Speaker Separation

Speaker separation has progressed from deep clustering (Hershey et al., 2016b) and permutation invariant training (PIT) (Yu et al., 2016) to time-domain models such as Conv-TasNet (Luo & Mesgarani, 2018) and attention-based architectures such as SepFormer (Subakan et al., 2021). Joint diarization-separation formulations such as TS-SEP (Boeddeker et al., 2024) condition mask estimation on speaker embeddings, but all of these methods operate on continuous representations with task-specific components. More recently, TokenSplit (Erdogan et al.) performs separation over discrete codec tokens within an encoder-decoder Transformer, and TSELM (Tang et al., 2025) applies a language model to target speaker extraction. However, TokenSplit trains a task-specific model from scratch rather than leveraging a pretrained language model, and TSELM requires a speaker enrollment embedding, limiting it to extracting one known speaker at a time.

2.2. Neural Audio Codecs

Neural audio codecs learn to compress waveforms into discrete token sequences via residual vector quantization (RVQ). SoundStream (Zeghidour et al., 2021) introduced a convolutional encoder-decoder with RVQ paradigm, and EnCodec (Défossez et al.) extended it with improved training objectives and a lightweight Transformer for further compression. More recently, X-Codec (Ye et al., 2025a) demonstrated that integrating semantic features from a pretrained self-supervised model such as WavLM (Chen et al., 2021) into the quantization pipeline yields tokens that better preserve content information, making them more suitable for downstream language modeling. Its successor,

[*]Equal contribution [1]ETH Zurich, Switzerland. Correspondence to: Luca A. Lanzendörfer <lanzendoerfer@ethz.ch>.

Machine Learning for Audio Workshop at the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. 2026.

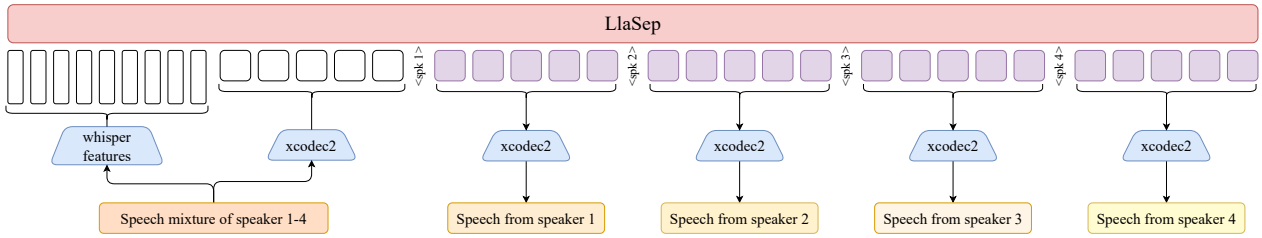

Figure 1. Architecture of the proposed token-level language model for joint diarization and separation. LlaSep supports up to 4 speakers but can be trained to support an arbitrary number of speakers. Speech snippets of up to four speakers are combined into one mixture signal during training. The mixture signal is embedded using features extracted from the Whisper encoder together with the Xcodec2 token sequence. The LLaSA-1B TTS backbone is teacher-forced on the purple output tokens during training. During inference the model generates the speech segments of each speaker autoregressively.

XCodec2 (Ye et al., 2025b), provides a single-codebook tokenization with a vocabulary of 65k entries at 50 tokens per second, balancing reconstruction fidelity with a compact token sequence ideal for language models.

2.3. Language Models for Audio

Discrete audio tokens have enabled language-model-based approaches to speech generation. VALL-E (Wang et al., 2023) first cast text-to-speech as conditional codec language modeling, demonstrating zero-shot speaker adaptation from a short prompt. LLaSA (Ye et al., 2025b) simplified this with a single-codebook codec (XCodec2) and a standard Llama (Touvron et al., 2023) architecture. Whisper (Radford et al., 2023) provides robust multilingual speech representations widely adopted as conditioning features. Despite these advances in single-stream generation, applying causal language models to blind multi-speaker separation, where the model must disentangle all overlapping sources from a single tokenized mixture without enrollment, remains unexplored. Our work bridges this gap by recasting speaker separation as autoregressive token generation from a pretrained speech language model.

3. Method

3.1. LlaSep Architecture

Overview. Our approach takes a multi-speaker audio mixture, encodes it into discrete tokens as well as continuous Whisper embeddings, and generates separated per-speaker token streams using a causal language model (cf. Figure 1). The pipeline consists of three components: (i) a discrete audio tokenizer, (ii) a semantic conditioning module using Whisper, and (iii) an autoregressive language model.

Audio Tokenization. We use XCodec2 as the audio tokenizer. Given a waveform sampled at 16 kHz, XCodec2 encodes it into a sequence of integer tokens drawn from a vocabulary of size $V = 65\,536$ at a framerate of 50 Hz. For an 8-second segment, this yields a sequence of 400 tokens per source. Both the input mixture and individual speaker sources are tokenized independently. The XCodec2 decoder reconstructs waveforms from generated token sequences at inference time.

Semantic Conditioning. To provide semantic guidance for separation, we extract embeddings from a Whisper-small encoder and project them into the hidden dimension of the audio-language model via a learned linear projection:

$$\mathbf{h}_{\text{sem}} = \mathbf{W}_{\text{proj}} \cdot \text{Whisper}(\mathbf{x}) \tag{1}$$

where $\mathbf{x}$ denotes the input audio signal. During training, the Whisper encoder receives the sum of clean source signals. During inference, it receives the mixture signal directly.

Language Model Backbone. We use LLaSA-1B-Multilingual (Ye et al., 2025b), a causal decoder-only transformer, as the backbone. Since LLaSA already operates over XCodec2 tokens, we only extend the embedding layer with special speaker delimiter tokens that mark per-speaker output regions. The Whisper projection embeddings are injected into the sequence at the corresponding positions before the speaker output regions. The model generates up to $K = 4$ speaker token streams autoregressively, conditioned on the mixture tokens and semantic embeddings as a prefix.

Training: Supervised Fine-Tuning. The model is trained with a standard cross-entropy loss on the ground-truth speaker token sequences. The loss is computed only over the speaker output tokens; mixture tokens and Whisper embeddings in the prefix are masked from the loss.

3.2. MLSEE-Conversation Dataset

To train LlaSep we construct MLSEE-Conversation, a large-scale synthetic multi-speaker conversation dataset derived from three multilingual speech corpora: MLS (Pratap et al., 2020), Emilia (He et al., 2024), and EuroSpeech (Pfisterer et al., 2025). The dataset comprises approximately 6.9 million 8-second conversation samples, each tokenized into a fixed-length sequence of 400 XCodec2 tokens.

Table 1. Performance comparison on the LibriCSS benchmark across different overlap ratios. 0% overlap with short inter-utterance silence (0S) and 0% overlap with long silence (0L). The metrics cover diarization accuracy (DER), perceptual quality (DNSMOS), and separation fidelity (ScoreQ). We find that LlaSep overall outperforms previous approaches.

Metric	Model	0S	0L	10	20	30	40	Avg
DER (%) ↓	PixIT	39.03	21.66	34.86	34.65	34.61	32.31	32.65
	SepFormer	138.98	160.03	120.07	109.43	95.46	83.98	117.99
	SepReformer	107.80	117.30	91.54	82.27	71.72	56.22	87.81
	LlaSep (ours)	**21.91**	**18.57**	**24.79**	**26.85**	**20.96**	**27.52**	**23.43**
DNSMOS-OVRL ↑	PixIT	2.59	2.50	2.49	2.57	2.65	2.58	2.56
	SepFormer	2.45	2.16	2.45	2.55	2.56	2.57	2.46
	SepReformer	2.66	2.45	2.66	2.72	2.69	2.74	2.65
	LlaSep (ours)	**3.12**	**3.06**	**3.11**	**3.14**	**3.17**	**3.19**	**3.13**
ScoreQ-NR ↑	PixIT	2.10	2.24	2.05	2.08	2.10	2.22	2.13
	SepFormer	1.90	1.74	1.83	2.07	1.87	1.84	1.87
	SepReformer	2.33	2.07	2.31	2.56	2.37	2.39	2.34
	LlaSep (ours)	**3.80**	**3.70**	**3.99**	**4.03**	**4.06**	**4.11**	**3.95**
ScoreQ-Ref ↓	PixIT	1.16	1.10	1.18	1.19	1.20	1.15	1.17
	SepFormer	1.10	1.10	1.16	1.10	1.17	1.21	1.14
	SepReformer	0.91	0.95	0.97	0.90	0.94	0.98	0.95
	LlaSep (ours)	**0.37**	**0.39**	**0.36**	**0.36**	**0.37**	**0.39**	**0.37**

Table 2. Evaluation on the held-out test split of the MLSEE-Conversation dataset, restricted to samples with up to **2 speakers**. This subset allows for direct comparison against baselines with fixed output channels (SepFormer, SepReformer). LlaSep (ours) operates on the discrete token domain, while baselines operate on continuous representations.

Metric	PixIT	SepRef.	SepF.	Ours
DER (%) ↓	63.80	136.19	99.21	**28.11**
DNSMOS-SIG ↑	2.55	2.68	2.79	**3.33**
DNSMOS-BAK ↑	3.27	3.32	3.41	**4.02**
DNSMOS-OVRL ↑	2.23	2.30	2.41	**3.04**
DNSMOS-P.808 ↑	3.07	2.83	3.41	**3.34**
ScoreQ-NR ↑	1.44	1.37	1.58	**3.41**
ScoreQ-Ref ↓	1.21	1.09	0.98	**0.42**

Table 3. Evaluation on MLSEE-Conversation test set with up to **4 speakers**. Comparison is restricted to PixIT, as the other baselines are limited to fixed two-speaker outputs.

Metric	PixIT	Ours
DER (in %) ↓	58.22	**43.79**
DNSMOS-SIG ↑	2.78	**3.27**
DNSMOS-BAK ↑	3.59	**4.03**
DNSMOS-OVRL ↑	2.45	**2.99**
DNSMOS-P.808 ↑	3.04	**3.22**
ScoreQ-NR ↑	1.36	**3.41**
ScoreQ-Ref ↓	1.18	**0.56**

Source Data and Languages. The MLS dataset contributes the large majority of source utterances (89.0%), with smaller contributions from Emilia (7.5%) and EuroSpeech (3.5%). Seven languages are represented with roughly balanced coverage: French (17.5%), English (17.0%), German (16.3%), Italian (13.4%), Portuguese (13.3%), Spanish (11.8%), and Dutch (10.8%). Conversations are predominantly monolingual (98.9%) of samples have all speakers in the same language and the remaining 1.1% form multilingual mixtures.

Conversation Generation. Speakers are randomly sampled per conversation, and each sample is synthesized according to one of four generation methods that control conversational dynamics: **Normal** (55.1%) uses sequential turn-taking where each speaker is conditioned on the previous speaker's offset, with Gaussian-distributed segment durations; **Erlang** (21.4%) samples each speaker's activity independently with Erlang-distributed segment durations, producing more frequent overlap and a more balanced distribution across up to 4 speakers; **Main+interrupts** (16.6%) features one dominant speaker with randomly sampled interruptions; and **Full overlap** (6.9%) has all speakers active simultaneously, the most challenging separation scenario.

Most conversations in the dataset feature 2 active speakers (69.3%), followed by 1 speaker (15.5%), 3 speakers (7.8%), and 4 speakers (7.4%). Individual source waveforms are normalized to a target loudness via LUFS normalization, mixed at 16 kHz with smooth fade-in/out transitions, and

peak-normalized. Sources are reordered by onset time to provide a canonical speaker ordering.

Dataset Statistics. The final combined dataset (train & test set) contains over 15k hours split into 6.8M samples. Per-sample metadata includes XCodec2 tokens for the mixture and each source, per-speaker language labels, speaker identifiers, signal-to-noise ratios, and the generation method.

4. Experiments

4.1. Setup

Evaluation Metrics. We evaluate along three axes: diarization accuracy via diarization error rate (DER) (Lanzendörfer et al., 2025a), perceptual audio quality via DNSMOS (Reddy et al., 2022), and separation fidelity via ScoreQ (Ragano et al., 2024) in both no-reference (ScoreQ-NR) and reference-based (ScoreQ-Ref) modes.

Benchmark. We compare against three baselines: PixIT (Kalda et al., 2024), a joint diarization and separation pipeline, as well as SepReformer (Shin et al.) and SepFormer (Subakan et al., 2021), both mask-based separation models. Evaluation is conducted on three benchmarks: the held-out test split of our proposed MLSEE-Conversation dataset, LibriCSS (Chen et al., 2020), and the English and German subsets of CallHome.[1] For CallHome, ground-truth source stems are not available, so we omit ScoreQ-Ref. All models operate on fixed-length audio chunks, segmented using SileroVAD (Team, 2024) (8s for LlaSep, SepFormer, SepReformer, and 5s for PixIT).

Implementation Details. The Whisper-small encoder is frozen and only a linear projection layer is trained alongside the LM parameters. We train with AdamW (Loshchilov & Hutter) on our MLSEE-Conversation dataset for 3 epochs. At inference, we generate 20 samples and compute the mean to account for nondeterministic behavior of sampling-based audio language models.

4.2. Results

Evaluation on LibriCSS. Table 1 shows results across overlap conditions. LlaSep achieves the lowest average DER (23.43% vs. 32.65% for PixIT); the mask-based baselines exceed 85% DER, as their fixed two-output design produces spurious streams when fewer speakers are active. LlaSep also leads on all perceptual quality metrics (DNSMOS, ScoreQ-NR, ScoreQ-Ref). The baselines recover signals by operating on continuous representations of the mixture, via time-frequency masking (SepFormer, SepReformer) or MixIT-based source estimation (PixIT) (Wisdom et al., 2020), which can leak residual interference and introduce processing artifacts. In contrast, LlaSep generates each stream from codec tokens, producing perceptually clean audio by construction.

[1] https://huggingface.co/datasets/talkbank/callhome

Table 4. Zero-shot evaluation on CallHome (English and German subsets), representing real-world telephone conversations. *ScoreQ-Ref* is omitted due to the unavailability of clean ground-truth source signals for this dataset. Best results in **bold**.

Metric	PixIT	SepRef.	SepF.	Ours
DER (%) ↓	30.20	65.98	79.61	**24.84**
DNSMOS-SIG ↑	2.67	2.63	2.66	**3.41**
DNSMOS-BAK ↑	3.09	3.03	2.94	**3.97**
DNSMOS-OVRL ↑	2.26	2.17	2.14	**3.09**
DNSMOS-P.808 ↑	2.80	2.76	2.73	**2.94**
ScoreQ-NR ↑	1.64	1.68	1.68	**2.80**

MLSEE-Conversations results. Tables 2 and 3 report results on held-out synthetic conversations. LlaSep leads across all metrics in both settings, though performance degrades with increasing speaker count: DER rises from 28.11% (2 speakers) to 43.79% (4 speakers). This is expected, as each additional speaker stream lengthens the output sequence and compounds error propagation during autoregressive decoding. Despite this degradation, LlaSep still outperforms PixIT in the 4-speaker setting, and unlike the mask-based baselines, does not require architectural changes to handle variable speaker counts.

CallHome results. On real telephone conversations (cf. Table 4), PixIT achieves the second best DER (30.20%) with LlaSep outperforms with (24.84%), additonally, LlaSep produces substantially higher-quality separated audio (e.g., DNSMOS-OVRL: 3.09 vs. 2.26). That LlaSep leads on both diarization accuracy and audio quality despite being trained exclusively on synthetic data suggests strong domain transfer from our MLSEE-Conversation dataset to real telephone speech. This is achieved using only synthetic training data, without exposure to any real conversational speech, further underscoring the quality of our synthetic dataset. Incorporating more naturalistic turn-taking during data generation could further widen this gap.

5. Conclusion

We presented LlaSep, an autoregressive speaker separation model capable of achieving high perceptual separation quality and competitive diarization accuracy. A key practical advantage of the autoregressive formulation is that speaker count is implicitly handled by the generation process, allowing it to easily extend to an arbitrary number of speakers and avoid the fixed-output limitation that can cause issues with mask-based methods.

Impact Statement

This work aims to advance multi-speaker speech processing by improving speaker separation and diarization, with potential benefits for meeting transcription, accessibility, and multilingual speech understanding. The generated separated signals may differ from the original speech, so the method should not be deployed without careful validation.

References

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Boeddeker, C., Subramanian, A. S., Wichern, G., Haeb-Umbach, R., and Le Roux, J. TS-SEP: Joint diarization and separation conditioned on estimated speaker embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1185–1197, February 2024.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Zeng, M., and Wei, F. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16:1505–1518, 2021. URL https://api.semanticscholar.org/CorpusID:239885872.

Chen, Z., Yoshioka, T., Lu, L., Zhou, T., Meng, Z., Luo, Y., Wu, J., Xiao, X., and Li, J. Continuous speech separation: Dataset and analysis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7284–7288. IEEE, 2020.

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., and Défossez, A. Simple and controllable music generation. *Advances in neural information processing systems*, 36:47704–47720, 2023.

Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. High fidelity neural audio compression. *Transactions on Machine Learning Research*.

Du, Z., Chen, Q., Zhang, S., Hu, K., Lu, H., Yang, Y., Hu, H., Zheng, S., Gu, Y., Ma, Z., et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

Erdogan, H., Wisdom, S., Chang, X., Borsos, Z., Tagliasacchi, M., Zeghidour, N., and Hershey, J. R. Tokensplit: Using discrete speech representations for direct, refined, and transcript-conditioned speech separation and recognition.

He, H., Shang, Z., Wang, C., Li, X., Gu, Y., Hua, H., Liu, L., Yang, C., Li, J., Shi, P., Wang, Y., Chen, K., Zhang, P., and Wu, Z. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *Proc. of SLT*, 2024.

He, X. and Whitehill, J. Survey of end-to-end multi-speaker automatic speech recognition for monaural audio. *Computer Speech & Language*, pp. 101925, 2025.

Hershey, J. R., Chen, Z., Le Roux, J., and Watanabe, S. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 31–35. IEEE, 2016a.

Hershey, J. R., Chen, Z., Roux, J. L., and Watanabe, S. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31–35, 2016b. URL https://api.semanticscholar.org/CorpusID:1058813.

Kalda, J., Pagés, C., Marxer, R., Alumäe, T., and Bredin, H. Pixit: Joint training of speaker diarization and speech separation from real-world multi-speaker recordings. In *The Speaker and Language Recognition Workshop (Odyssey 2024)*, pp. 115–122. ISCA, 2024.

Kumar, R., Seetharaman, P., Luebs, A., Kumar, I., and Kumar, K. High-fidelity audio compression with improved rvqgan. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 27980–27993, 2023.

Lanzendörfer, L. A., Grötschla, F., Blaser, C., and Wattenhofer, R. Benchmarking diarization models. *arXiv preprint arXiv:2509.26177*, 2025a.

Lanzendörfer, L. A., Pinkl, C., Perraudin, N., and Wattenhofer, R. Bootstrapping language-audio pre-training for music captioning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025b.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Luo, Y. and Mesgarani, N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27:1256–1266, 2018. URL https://api.semanticscholar.org/CorpusID:52310361.

Pfisterer, S., Grötschla, F., Lanzendörfer, L. A., Yan, F., and Wattenhofer, R. Eurospeech: A multilingual speech corpus. *arXiv preprint arXiv:2510.00514*, 2025.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., and Collobert, R. Mls: A large-scale multilingual dataset for speech research. In *Proc. Interspeech 2020*, pp. 2757–2761, 2020.

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023.

Ragano, A., Skoglund, J., and Hines, A. Scoreq: speech quality assessment with contrastive regression. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pp. 105702–105729, 2024.

Reddy, C. K., Gopal, V., and Cutler, R. Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 886–890. IEEE, 2022.

Shin, U.-H., Lee, S., Kim, T., and Park, H.-M. Separate and reconstruct: Asymmetric encoder-decoder for speech separation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Siuzdak, H., Grötschla, F., and Lanzendörfer, L. A. Snac: Multi-scale neural audio codec. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*.

Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., and Zhong, J. Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21–25. IEEE, 2021.

Tang, B., Zeng, B., and Li, M. Tselm: Target speaker extraction using discrete tokens and language models. In *National Conference on Man-Machine Speech Communication*, pp. 459–469. Springer, 2025.

Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., MA, Z., and Zhang, C. Salmonn: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Team, S. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad, 2024.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023. URL https://api.semanticscholar.org/CorpusID:257219404.

Wang, C., Chen, S., Wu, Y., Zhang, Z.-H., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., and Wei, F. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718, 2023. URL https://api.semanticscholar.org/CorpusID:255440307.

Wisdom, S., Tzinis, E., Erdogan, H., Weiss, R. J., Wilson, K., and Hershey, J. R. Unsupervised sound separation using mixture invariant training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 3846–3857, 2020.

Ye, Z., Sun, P., Lei, J., Lin, H., Tan, X., Dai, Z., Kong, Q., Chen, J., Pan, J., Liu, Q., et al. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25697–25705, 2025a.

Ye, Z., Zhu, X., Chan, C.-M., Wang, X., Tan, X., Lei, J., Peng, Y., Liu, H., Jin, Y., Dai, Z., Lin, H., Chen, J., Du, X., Xue, L., Chen, Y., Li, Z., Xie, L., Kong, Q., Guo, Y., and Xue, W. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *CoRR*, abs/2502.04128, 2025b.

Yu, D., Kolbæk, M., Tan, Z.-H., and Jensen, J. H. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 241–245, 2016. URL https://api.semanticscholar.org/CorpusID:7331600.

Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. Soundstream: An end-to-end neural audio codec, 2021.