SoundStorm

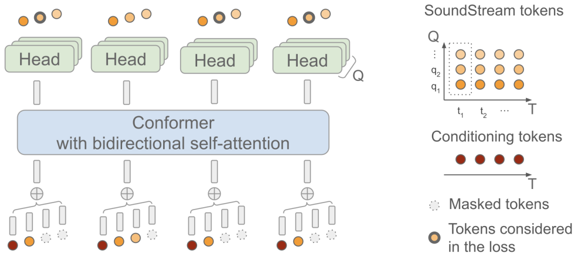


Figure 1. SoundStorm architecture and masking scheme for training (without prompting). The model reduces the input sequence length by summing up the embeddings of the tokens corresponding to the same SoundStream frame. During training, an RVQ level q is sampled ($q = 2$ out of $Q = 3$ levels in the figure), and a subset of randomly sampled tokens at level q are masked together with all tokens at RVQ levels $q + 1, \ldots, Q$. The loss is computed only on the masked tokens at level q.

- □□□□□SoundStorm: Efficient Parallel Audio Generation
- □□□□□https://arxiv.org/abs/2305.09636

□□□□□SoundStorm□□□□□□□□□□□□□□□□□□□□□□□□□token□□□□□□□□□□□□□□□□ acoustic token□□□□□□□□□□□□□□

SoundStorm□□□□□□□□□□□□□□□□□□□□□

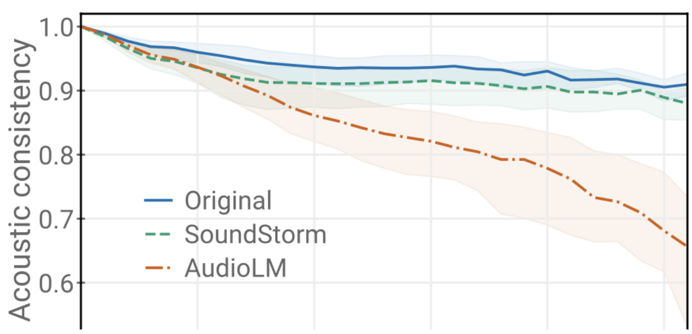
- Sample the prompt delimiter timestep $t \sim \mathcal{U}\{0, T-1\}$;
- Sample the current RVQ level $q \sim \mathcal{U}\{1, Q\}$;
- Sample the mask $M \in \{0,1\}^T$ according to a cosine schedule (Chang et al., 2022) for level q, i.e., sample the masking ratio $p = \cos(u)$ where $u \sim \mathcal{U}[0, \pi/2]$, and sample iid $M_i \sim \text{Bernoulli}(p)$.
- Mask the selected non-prompt tokens at the current RVQ level q (mask $Y_{t',q}$ if $M_{t'} = 1$ and $t' > t$) and all non-prompt tokens at finer RVQ levels ($Y_{>t,>q}$).

□□□□□□□□□□□□□□□acoustic token□□□□□□□□□□□□token□□□

□□□□□□□□□□□□□□□

Table 1. Comparing intelligibility, quality, voice preservation, and acoustic consistency of AudioLM's acoustic generator and SoundStorm. We report metric values for the 'short' (4-10 s), 'mid' (10-20 s), and 'long' (20-30 s) splits of LibriSpeech test-clean separately. SoundStorm matches AudioLM's acoustic generator in terms of audio quality, and outperforms it in terms of speech intelligibility and acoustic consistency.

	WER↓			CER↓			Audio quality↑			Voice preservation↑			Acoustic consistency↑		
	short	mid	long	short	mid	long	short	mid	long	short	mid	long	short	mid	long
Original SoundStream rec.	2.62	1.95	2.20	0.89	0.55	0.69	3.72	3.91	3.99	0.63	0.65	0.66	0.97	0.95	0.93
Without a speaker prompt															
AudioLM	4.65	3.59	4.79	2.15	1.57	2.30	3.93	4.04	4.08	–	–	–	–	–	–
SoundStorm	**3.48**	**2.55**	**3.33**	**1.39**	**0.89**	**1.29**	**4.01**	**4.16**	**4.20**	–	–	–	–	–	–
With a speaker prompt															
AudioLM	3.77	3.40	3.75	1.50	1.47	1.54	**3.91**	**4.06**	4.10	0.46	0.48	0.48	**0.96**	0.91	0.86
SoundStorm	**2.99**	**2.43**	**3.36**	**1.10**	**0.81**	**1.24**	3.81	**4.05**	**4.15**	**0.57**	**0.59**	**0.59**	**0.96**	**0.94**	**0.91**



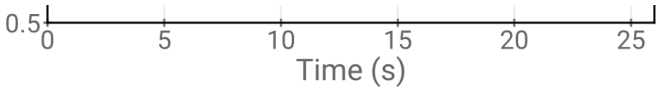


Figure 2. Acoustic consistency between the prompt and the generated audio for the samples in the 'long' split of LibriSpeech test-clean. The shaded area represents the inter-quartile range.

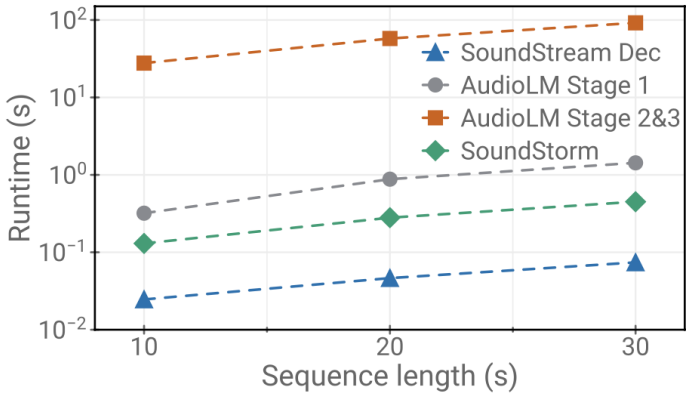


Figure 3. Runtimes of SoundStream decoding, SoundStorm and different stages of AudioLM on a TPU-v4.

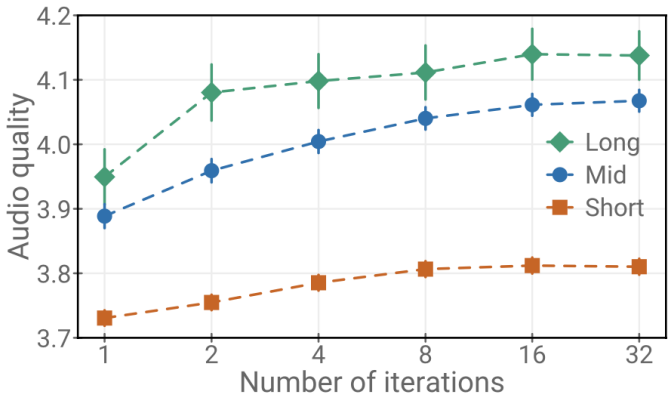


Figure 4. Audio quality with respect to the number of decoding iterations in the first RVQ level.

Newer Older

2024-07-01
SpeechGPT

2024-06-30
AnyGPT