ADATS: ADAPTIVE TOKEN SAMPLING FOR EFFICIENT SPEECH LANGUAGE MODELS

Sonal Sannigrahi[♣♠] **Giuseppe Attanasio**[♣], **André F.T. Martins**[♣♠◇]
[♣]Instituto de Telecomunicações, Lisbon, Portugal
[♠]Instituto Superior Técnico, Universidade de Lisboa, Portugal
[◇]TransPerfect, Lisbon, Portugal
sonal.sannigrahi@tecnico.ulisboa.pt

ABSTRACT

Speech Language Models (SLM) have demonstrated strong capabilities in end-to-end speech understanding and reasoning tasks by incorporating speech tokens into a Large Language Model (LLM). However, most common designs are i) token-intensive, since a large part of the LLM context is allotted to audio tokens, and ii) inefficient, as audio representation is often redundant, hindering SLMs' capabilities to handle long-form tasks. To address token inefficiency, we propose a dynamic sampling method that adaptively groups and merges speech tokens where the signal is less information-dense. Our approach reduces speech length by $2\times$ on average while yielding performance comparable to or better than standard convolutional downsampling across Speech Recognition (ASR), Speech Question-Answering (SQA), and Speech Translation (ST). Through extensive empirical analysis, we demonstrate the effectiveness of this strategy in preserving speech content and exhibiting general speech understanding capabilities, while substantially reducing token redundancy and inference cost by 40%. We release all of our code to the community[1].

1 INTRODUCTION

Recent advances in Large Language Models (LLMs) have encouraged substantial progress in extending language understanding and reasoning capabilities beyond text to other modalities Xu et al. (2025a); Fang et al. (2024); Dash et al. (2025). Among these, Speech Language Models (SLMs) have emerged as a promising direction, aiming to equip LLMs with the ability to process, reason over, and generate speech signals directly, rather than relying solely on text-based representations Shi et al. (2026); Cui et al. (2025); Tang et al. (2024). These models primarily achieve their strength in reasoning by processing long contexts, such as video or long-form audio, using a transformer architecture Shao et al. (2025b).

Despite this progress, current models are severely bottlenecked by the quadratic complexity of the self-attention mechanism. More so than text, speech and audio are *token expensive*, and as the number of tokens increases, the additional complexity leads to a significant computational and memory footprint. For instance, modern audio encoders typically operate at frame rates of 20ms or finer, producing thousands of tokens per minute of audio Hsu et al. (2021); Hu et al. (2024). As a result, even a 10-minute clip may exceed 30K tokens, which is at the upper limit of the effective context length for models that support extended context windows Hsieh et al. (2024), making real-world scenarios with long-form audio entirely impractical. Despite the token-intensive encoding of speech, the resulting representations are surprisingly redundant, with less than 50% of tokens actually being attended to by the models Shang et al. (2025); Shao et al. (2025a;b). Current SLMs implicitly shift the burden of filtering this redundancy to the backbone LM itself, thereby limiting the effective content provided to the model and restricting reasoning over long-from speech. Therefore, addressing this computational roadblock is crucial for allowing the real-world adoption of SLMs.

[1] https://github.com/sonalsannigrahi/AdaTS

One approach to overcome the challenges imposed by the longer contexts of multimodal tokens is *token sampling*. This approach is not only highly effective but also practical to implement as both a test-time technique or as part of the training framework itself. Existing approaches subsample from the input sequence through temporal pooling or convolutional downsampling Attanasio et al. (2025); Xu et al. (2025a); Umberto et al. (2025). While they can reduce sequence length, these methods are *static and indiscriminate*, treating all tokens uniformly regardless of their informational density. This creates a fundamental trade-off between the compression rate and semantic preservation, often discarding content-critical regions together with low-information segments. Other methods introduce learned or iterative fusion strategies for long-speech processing Guo et al. (2025); Li et al. (2023), but they require additional parameters, more complex training pipelines, or large backbone models ($>$7B parameters). Consequently, efficient speech modelling remains restricted not only by model size but also by how speech tokens are represented and allocated.

In this work, we propose ADATS, a dynamic sampling framework for efficient SLMs that adapts to each speech signal. ADATS operates directly on the output of a pretrained speech encoder, following which we apply a *score-and-merge* mechanism based on pairwise token similarity. Unlike uniform downsampling, speech tokens are adaptively split into local subgroups and merged according to their information density, producing coarse representations in redundant regions while preserving high-resolution detail in content-bearing segments. We build upon the success of small-scale LLMs Microsoft et al. (2025); Xu et al. (2025b); Martins et al. (2025) and train ADATS with a standard two-stage SLM training pipeline on models under 2B parametres. We demonstrate through evaluations across several tasks that ADATS preserves temporal consistency and fine-grained phonetic content while reducing overall redundancy. On automatic speech recognition (ASR), speech translation (ST), and speech question answering (SQA), including long-form SQA, we achieve **up to** $4\times$ **compression and** $2\times$ **on average** while consistently improving upon downstream performance. In addition, ADATS reduces inference cost by **approximately 40%** compared to convolutional downsampling at smaller model scales for better performance. Our findings open the pathways of efficient SLM by showing that scaling is not only about larger backbone LLMs or longer contexts, but also about how speech tokens are utilised.

2 RELATED WORK

Multimodal Token Compression The challenge of learning efficient multimodal token representations is an active area of research. Across image, video, and speech domains, token compression approaches can be split into four categories: transformation-based, similarity-based, query-based, and attention-based Shao et al. (2025b). Focusing on transformation and similarity-based approaches, the former leverages redundancy in multimodal tokens to effectively apply convolutional-pooling as a downsampling strategy while preserving the original structural representation. Qwen models Bai et al. (2025) leverage pooling layers for parameter-free down sampling. Furthering this strategy, applying 1-D convolutional pooling across the temporal dimension can also reduce the number of speech tokens Attanasio et al. (2025). Recent works such as the InternVL models Chen et al. (2024), Qwen models Bai et al. (2025), and NVLM Dai et al. (2024) utilize a feature-map transformation[2] to reduce the number of visual tokens by a factor of 0.25. Analogous to images, for the speech modality, models utilize *token-stacking* wherein consecutive tokens are stacked along the hidden dimension as seen in LLaMA-Omni Fang et al. (2025). Shifting focus to similarity-based approaches, several works across both the image and speech modality make use of such strategies in different flavors. DynTok Zhang et al. (2025) employs a threshold cut-off and merge strategy on visual tokens to reduce the spatial redundancy while retaining crucial, information-dense patches. A-ToME Li et al. (2023) applies adjacent token merging within the Transformer module to combine tokens with high similarity scores between their key values. More recently, FastLongSpeech Guo et al. (2025) incorporates an iterative fusion strategy with dynamic compression learned via a CTC-decoder in a SLM. Differing from previous approaches in SLMs, our method is parameter-free, with limited training overhead, and preserves the temporal consistency of the speech input despite more aggressive compression as shown by our compression rates.

[2]We refer to pixel unshuffle here where a transformation consists of moving from high spatial resolution with less channels to low spatial resolution with more channels.

Speech Language Modelling Following the adoption of LLMs for the text domain, a similar trend of moving from task-specific to general purpose models is observed for speech processing as well Fang et al. (2025); Cui et al. (2025). Of these, we can generally categorise SLMs into three categories: 1) Pure Speech: where the model learns a next token prediction objective on unlabelled speech data Hsu et al. (2021) 2) Speech and Text LMs: where the model jointly learns the distribution of speech and corresponding text through text-aligned speech data such as ASR Nguyen et al. (2025); Maiti et al. (2024) and 3) Speech-aware LMs: where models combine pre-trained LMs with speech encoders to maintain instruction-following general purpose capabilities of LMs but extend their processing capabilities to the speech domain Arora et al. (2025). Focusing on 2) and 3), we can split approaches by pre and post training methods. A common approach for pre-training SLMs is continually pre-training a text-only LM with a next-token prediction task using general speech-text data (e.g., ASR) Ambilduke et al. (2025). To obtain a Speech-aware LM, instead of continued pre-training, several works adopt an alignment phase where the speech and text embedding spaces are aligned either via explicit or implicit signals. The pre-training phases enables the modelling of joint speech-text data distributions, but still lacks the capability to solve downstream tasks. The post-training phase is used to generally bias the SLM towards specific tasks, usually specified via an instruction template. Following the modality and length adapters to align the output space of a pre-trained speech encoder and the input space of a text decoder, a full fine-tuning stage is typically used as the post-training strategy Verdini et al. (2024). Another parameter-efficient approach is to not fully train the entire LM but only update the parameters of the adapter Microsoft et al. (2025) and adding to the LM a set of parameter-efficient training modules.

3 ADAPTIVE SPEECH SAMPLING

ADATS follows a two-stage training strategy where we first align the output space of the speech encoder and the input space of the text decoder keeping the components frozen, then we learn a full fine-tuning stage where the decoder is trainable. We incorporate the sampling module in the second stage of the training following the audio encoder, prior to passing the input into the text decoder.

3.1 SAMPLING MODULE

ADATS including the sampling module is depicted in Figure 1. From a pre-trained audio encoder, we first generate a sequence of high-dimensional feature vectors $A = a_1, a_2, \ldots, a_n$. Next, we compute pair-wise token similarity and for a chosen threshold t, we then create a subgroup of tokens with $\mathrm{sim}(a_i, a_j) > t$ which are merged as follows:

$$\mathrm{sim}(a_i, a_j) = \frac{a_i a_j}{|a_i||a_j|}, a_{\mathrm{merged}} = \frac{\sum_{i=1}^{N}(1 - \mathrm{sim}(a_i, a_{i-1}))a_i}{N},$$

where $0 < i$ and $N < n$ is the total number of tokens with similarity below t. We then replace tokens $a_i, \ldots, a_N$ with a_{merged} in output vector A' resulting in fewer tokens overall. We also note that for our choice of speech encoder, which is trained with contrastive cosine loss, to produce A', cosine similarity is an appropriate choice to measure token similarity Zhang et al. (2025); Zhou et al. (2022).

3.2 MODEL ARCHITECTURE

The three primary components are the **audio encoder**, **projector**, and the **LM decoder**. In line with recent multimodal encoder-decoder models Attanasio et al. (2025); Bai et al. (2025); Grattafiori et al. (2024), we follow a two-stage training approach to first learn a linear projector between the frozen audio encoder and frozen LLM decoder then in the second stage, we unfreeze the LLM decoder and fully finetune the model with a task-specific data mix.

Modality Alignment (MA) In the first training stage, we align the embedding spaces of the audio encoder and text decoder with linear multi-layer perceptron (MLP) layers. We extract 80-dimensional Mel-filterbank audio representations using our audio encoder. These output representations are then further processed by a **modality and length adapter** to project them to the input embedding space of the text decoder. During this stage, all components are kept frozen and only the pre-encoder layers are trainable using ASR data. We follow all hyperparmetres suggested in Attanasio et al. (2025).

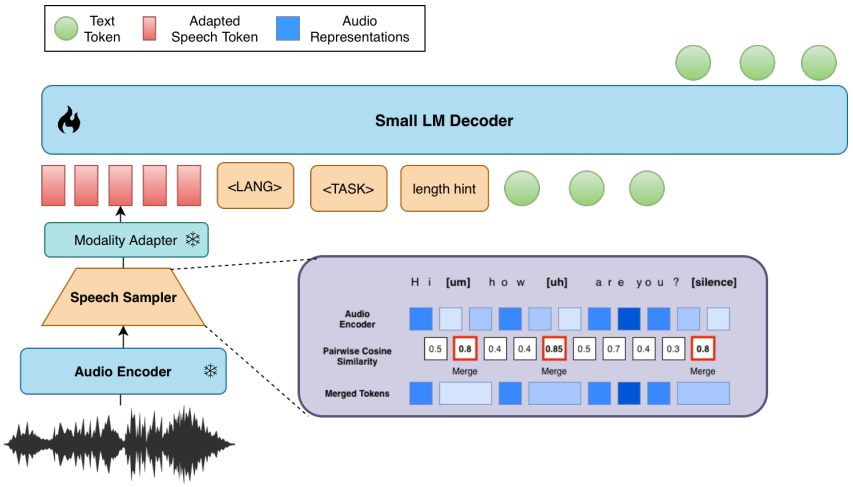

Figure 1: Illustration of ADATS including the sampling module which consists of three steps: obtaining representations from the audio encoder, computing pairwise cosine similarity, and merging tokens.

Instruction Fine-tuning (IFT) Following the modality alignment, we concatenate the audio embeddings to the text input embeddings and perform IFT using a suite of speech-to-text tasks for English. Specifically, we train on ASR in English, SQA in English, and ST from English to 10 other languages [3]. In this stage, all components are trained end-to-end jointly.

Training Setup We train the model using a standard cross-entropy loss with a cosine scheduler on reference transcripts on 4 H100 GPUs for 4 days. For both stages, we use 20 warmup steps with 128 gradient accumulation steps. We use AdamW Kingma & Ba (2015) as our optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. We also use bfloat16 mixed precision. We set our learning rate as 6×10^{-6} for the encoder, 2×10^{-5} for the decoder, and 2×10^{-4} for the adapter.

3.3 DATA

Table 1 lists the data used per task and training stage category. We also complement natural datasets with synthetically generated datasets as well to expand upon under-resourced tasks such as ST. Additionally, we filter SQA data to better match real-world use cases. We restrict our mix to open-license data only. In total, we use 80k hours of labelled speech data.

ASR We use LibriSpeech Panayotov et al. (2015), VoxPopuli Wang et al. (2021), CommonVoice 16.1 Ardila et al. (2020) during the less data-intensive modality alignment phase. Notably, we only use ASR data during the alignment phase as previous works have shown ASR data promotes a general understanding of speech-text alignment prior to task-specific training Arora et al. (2025). In the IFT phase, we use all the data from the previous stage and add Peoples Speech, Giga Speech, and the English split on Multilingual LibriSpeech.

ST We use gold-standard CoVoST-2 for En-De and a pseudo-labeled speech translation corpora, Spite [4], to augment our ST data mix for remaining language pairs. Previous works have shown this to be an effective strategy when coupled with quality filtering Attanasio et al. (2025); Ambilduke et al. (2025). Spite was generated with a two-step process of translating ASR data and filtering the result with COMETKiwi Rei et al. (2022), a quality estimation (QE) metric for text-only machine translation Attanasio et al. (2025). We use the version of Spite translated using TowerPlus-9B and with transcripts from CommonVoice 16.1 as the translation quality of TowerPlus-9B is the highest among the models released Rei et al. (2025).

[3] French, German, Italian, Dutch, Korean, Portuguese, Russian, Chinese, Spanish

[4] https://huggingface.co/datasets/bpop/spite-CV16-TP9B

Table 1: Data statistics with license and hours of speech data across all languages and task splits considered.

Task	Data	License	Hours
	Modality Alignment (MA)		
ASR	LibriSpeech (LS)	CC-BY 4.0	1K
	VoxPopuli	CC-BY 4.0	1.8K
	CommonVoice 16.1	CC-BY 4.0	4K
	Instruction Fine-Tuning (IFT)		
ASR	*All MA data*		
	Peoples Speech	CC-BY 4.0	12K
	CV 16.1 PL	-	30K
	GigaSpeech	Apache-2.0	10K
	Multilingual LS	CC-BY 4.0	2K
ST	CoVoST-2	CC-BY NC 4.0	3K
	Spite-TP	CC-BY 4.0	15K
SQA	SpokenSQuAD	CC BY-SA 4.0	-
	SLUE SQA-5	CC BY-SA 4.0	244
	LibriSQA	CC-BY 4.0	360

SQA We use the SpokenSQuAD Lee et al. (2018), SLUE SQA-5 Shon et al. (2023), and LibriSQAZhao et al. (2024) datasets for the task of English SQA which contain a mix of synthesized speech as well as natural human speech sources paired with text targets. Following insights from Lee et al. (2025), we further clean the QA pairs obtained from SpokenSQuAD and SLUE SQA-5 using Qwen2.5-7B to reformulate the text-based answers into stand alone sentences.

Audio Chunking While most of the audio samples in the training data are less than 30 seconds in nature, we include an audio chunking module to process longer audio sequences, which is applied on the input speech prior to passing the audio to the encoder. We use a chunk length of 100s with a 1s overlap. We process audio chunks in parallel and concatenate the encoded chunks. While concatenating, we blend overlapping regions of two chunks by computing a weighted average. For audios smaller than the chunk length, we do not apply any chunking. This allows us to evaluate our model on long-form audio tasks despite training only on shorter sequences Guo et al. (2025).

4 EXPERIMENTS

We train and evaluate our models on three tasks: ASR, ST, SQA, and long-form SQA. We first discuss the decoding method using followed by the different system settings we considered for our method. Next, we detail comparable baseline approaches. Lastly, we list all the evaluation datasets and metrics used for evaluation.

Evaluation Set-up We decode with zero shot prompting using task-specific tags unless specified otherwise. For all tasks, we generate using beam search with beam size 3, a repetition penalty of 1.6, up to 1024 tokens. We constrain the text generation using a target language and task token. The task tag is limited to `<|transcribe|>`, `<|translate|>`, or `<|reply|>` for ASR, ST, SQA respectively. We use ISO codes for the languages considered as their language tags.[5]

4.1 SYSTEM SETTINGS

Encoder-Decoder We report results with `Wav2Vec2Bert` as our audio encoder. For the language decoder, we restrict our approach to small-scale language models and report results with `Qwen2.5 1.5B`, `Llama3.2 1B`, and `EuroLLM 1.7B`.

Method Settings Within the sampling module, we report results with the following variations:

[5] `<|en|>,<|de|>,<|es|>,<|fr|>,<|nl|>,<|pt|>,<|ko|>,<|zh|>,<|ru|>,<|it|>.`

- **Threshold:** We modulate the similarity threshold between 0.7 and 0.9 in 0.05 increments.
- **Pooling method:** We report three approaches: 1) average uniformly across the entire token subgroup obtained for a certain threshold t, 2) weight each token in the subgroup by $(1 - s)$ where s is the previously computed pairwise cosine similarity per (1), 3) prune all but the first token in the subgroup (top-1).
- **Training stage:** Within the two-stage training set-up, we include the sampling module during the MA phase, the IT phase, or in both.

4.2 BASELINES

We consider two approaches: static downsampling (conv-based and WLQF) and content-dependent sampling (CTC-based) Verdini et al. (2024). We fix the speech encoder to Wav2Vec2Bert and use Qwen2.5 1.5B as the backbone decoder for all cases.

- **Conv-based:** We use 3 1D-convolutional layers with stride 2 and kernel size 3 after the modality adapter.
- **WLQF:** We use a Window-level Q-former (WLQF) Tang et al. (2024) to replace the modality and length adapter. The module takes in variable-length outputs from the speech encoder and splits them into non-overlapping windows of size N which is then passed to a Q-former Li et al. (2022). We set $N = 0.33$ seconds and the number of queries to 1 as per Tang et al. (2024). We set the adapter kernel size to 8 and stride to 8 per the best setting in Verdini et al. (2024).
- **CTC-based:** FastLongSpeech (FLS) Guo et al. (2025) uses an iterative fusion strategy trained with a CTC decoder to merge speech tokens. FLS is trained with Qwen2 Audio 7B as the base LLM, to be comparable we report results with the 7B and 1.5B variant.

4.3 DATASETS

We report Word Error Rate (WER) on the LibriSpeech test split (clean and other) Panayotov et al. (2015), VoxPopuli Wang et al. (2021), and FLEURS Conneau et al. (2023) for English ASR. We report results on FLEURS ST for ST across all language pairs considered. We report average COMET-22 Rei et al. (2020) across the en→xx directions. Lastly, for SQA, we report F1 accuracy on the test split of Spoken SQuAD Lee et al. (2018) and frame-F1, defined as F1 accuracy on the answer spans, on SLUE SQA-5 Zhao et al. (2024). We further report results on LongSpeechEval (LS Eval) [6] which is a long form SQA evaluation dataset where the average duration of audios are 2 minutes. We evaluate LongSpeechEval using a LLM-as-a-judge protocol as described in Guo et al. (2025).

5 MAIN RESULTS

Table 2 presents our main results on ASR, SQA, and ST. First, we demonstrate that on the ASR task, our sampling strategy is able to retain the temporal dependencies of the input speech and preserve the content. Next, on SQA and ST tasks we show the improved speech-understanding capabilities of our model. Lastly, we report results on long-form spoken QA tasks to show the applicability of our approach beyond short audio sequences.

Speech Content Preservation We observe in Table 2, that across all ASR datasets, ADATS consistently outperforms other methods. Considering both static downsampling as well as content-dependent sampling, we find that ADATS remains the least sensitive to the subsampling of the speech sequence and scale of the LLMs used. This finding suggests that we are able to effectively preserve fine-grained details required to transcribe content accurately while reducing the overall redundancy of the input. We discuss exact compression rates achieved in Section 6. Among content-dependent methods, the FastLongSpeech (FLS) variants trained with 1.5B and 7B backbone LLMs exhibit a clear performance gap, showing that the size of the decoder is crucial to final performance for this approach– a dependency ADATS does not share, as it achieves competitive results even with smaller backbones. Considering variants with several small-scale backbone LMs, we find that the choice of

[6] https://huggingface.co/datasets/ICTNLP/LongSpeech-Eval

Table 2: Main results on ASR, SQA, and ST. We report baselines above and highlight models with ADATS.

Models	ASR (↓)				SQA (↑)			ST(↑)
	Clean	Other	VoxPop	FLEURS	S-SQuAD	SLUE	LSEval	FLEURS
Pooling	5.3	8.2	9.8	11.2	45.1	27.3	2.7	80.1
WLQF	3.6	4.5	10.2	10.1	53.1	30.4	3.0	81.3
FLS-1.5B	5.2	7.8	10.1	12.1	50.4	32.3	3.2	80.2
FLS-7B	4.1	7.2	9.8	10.3	-	-	**3.6**	-
LLaMA 3.2 1B	4.3	10.2	15.5	12.3	53.5	32.6	3.2	81.1
EuroLLM 1.7B	5.6	9.8	14.4	10.9	52.3	31.4	3.1	78.6
Qwen 2.5 1.5B	**2.9**	**6.2**	**9.5**	**8.5**	**58.9**	**37.5**	3.5	**82.0**

backbone strongly determines overall performance, with Qwen 2.5 1.5B emerging as a consistently strong foundation that yields the best results across ASR and ST tasks alike. However, despite the same backbone, ADATS improves upon uniform pooling, even surpasses FLS-7B, demonstrating that effective compression can compensate for reduced model size.

Speech Understanding In Table 2, for both tasks of short-form SQA and AST, ADATS outperforms other approaches. Notably, for SQA, pooling has the largest performance gap with other approaches showing the limitations of static downsampling. WLQF improves considerably upon vanilla convolutional downsampling and achieves comparable results to CTC-based FLS. Compared to ASR, FLS is a stronger model for SQA tasks showing that CTC-based decoding is able to better retain the overall information in the speech input but is worse at preserving the required phonetic characteristics for accurate ASR. ADATS is able to improve upon the task-specific performance of FLS as well the overall phonetic content preservation of WLQF. On long-form QA, LSEval, we find that ADATS is able to perform competitively with FLS-7B showing that despite a small proportion of audio longer than 30 seconds, training on only 30 second segments is sufficient to endow the model with capabilities to process longer-form audio.

6 ANALYSIS

Following our main results, we report results from ablations in the model architecture and merge strategy. Lastly, we share statistics regarding inference efficiency in terms of the compression rate and TFLOPs, which measures the average number of floating-point operations (FLOPs) for a given sample.

Training Stage Ablation Table 3 shows our results on applying the sampling module in modality alignment phase of our training pipeline, IFT, or both. Across all tasks we find that the best strategy is to perform modality alignment with the full sequences and then perform IFT with the compressed tokens. This is in line with previous literature where the sampling strategy is often learnt during the IFT phase of training. We further find that while compressing sequences during IFT is an advantageous approach, the reverse significantly degrades the performance. We observe that while applying the compression strategy in both stages retains good performance, we are able to improve upon these results by only fine-tuning using compressed representations in Stage 2. However, applying compression only in Stage 1 of the training leads to a subpar model that is unable to perform speech-text tasks. Therefore, the alignment learnt during the MA phase is crucial for the downstream performance of the model. Our findings empirically motivates the approach of subsampling the audio representation further in the IFT phase as the MA phase allows the model to recover even from heavily compressed representations in the IFT phase.

Table 3: Training Stage Ablations. Results on LS Clean, SLUE, and FLEURS.

MA	IFT	ASR (↓)	SQA (↑)	ST (↑)
✗	✗	5.3	25.6	80.1
✗	✓	**3.2**	**35.4**	**81.2**
✓	✗	20.3	20.4	58.8
✓	✓	3.6	29.7	80.3

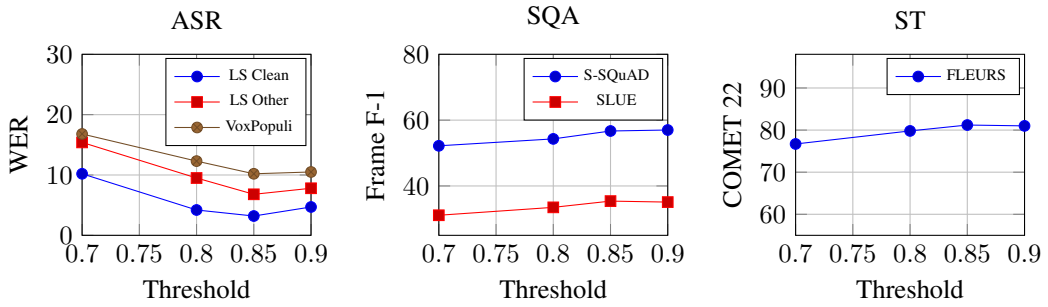

Figure 2: Performance vs. Threshold for ASR, SQA, and ST.

Threshold Sensitivity Figure 2 illustrates the performance of our evaluation suite against different threshold values t. Across all tasks, we observe a similar trend with respect to t, where lower values degrade performance. Comparing the three tasks, we note that the ASR and ST are much more sensitive to t with significant performance drops with changes in the value. However, in the SQA task the effect of the threshold is much less pronounced. This suggests that tasks which require high content preservation also in turn require a more fine grained choice of t, whereas tasks requiring general speech understanding are more robust to larger downsampling. Fixing $t = 0.85$, we now investigate the optimal merge strategy. Table 4 reports results with three pooling methods: averaging, averaging using weights as expressed in Section 3.1, and pruning all tokens but the first in the token subgroup (Top-1). While simple averaging provides a strong baseline, we are able to further improve upon it by weighted average which we find to be the best performing approach. Pruning as a strategy is surprisingly effective in downstream tasks as well suggesting that the token subgroups obtained by sampling are indeed low in information density.

Table 4: Merge Strategies on LS Clean, SLUE, and FLEURS.

Strategy	ASR (↓)	SQA (↑)	ST (↑)
Avg	3.2	35.4	81.2
Weighted Avg	**2.9**	**37.5**	**82.0**
Top-1	3.0	37.2	81.0

Efficiency After having found the best settings, we show the compression rates achieved by our sampling strategy. Table 5 reports the compression rates achieved by ADATS, averaging across datasets in each evaluation task. We report compression statistics at $t = 0.85$, which we found work best among the tasks considered. On average, we obtained a 2x compression rate across all tasks. In particular, for ASR and ST tasks, we are able to compress more aggressively for the same threshold when applied to SQA showing that the actual compression rate achieved is inherently dataset dependent. Lastly, ADATS requires 2.14 TFLOPS for inference[7] on a 10s audio while convolutional downsampling requires 3.23 TFLOPs, yielding a 34% reduction in compute. For a comparable performance, FLS-7B requires 8.54 TFLOPs representing 4 times the computational effort.

Table 5: Compression Ratio Across Test Tasks with $t = 0.85$. We report averages over all datasets in each task for brevity.

Dataset	**ASR**	**SQA**	**ST**
Min	1.23	1.16	1.25
Max	4.16	3.21	4.24
Avg	1.76	1.65	1.85

7 CONCLUSION

In this work, we presented ADATS, a parameter-free, dynamic sampling method based on pairwise token similarity. Our experimental results show that ADATS achieves competitive performance on several tasks across benchmarks while reducing the token footprint up to 2x on average, and 4x at maximum while obtaining significant gains in terms on inference efficiency. In particular, we demonstrate that by an improved encoding of an incoming speech signal, we can bypass the additional

[7]We computed FLOPS using `https://github.com/MrYxJ/calculate-flops.pytorch`

effort required by the text decoder to process longer contexts. Lastly, we report the best settings to use similarity-based sampling to further guide research in token compression strategies. For future work, we would like to further include a feature to measure content relevance as well depending on the task. This is relevant when extending this framework to tasks such as summarisation wherein in addition to mitigating temporal redundancies as measured on smaller segments of the entire audio, we would also like to judge which parts of the audio are at all relevant to process further.

8 Ethics Statement

Our work produces a Speech Language Model trained on 11 language pairs with model efficiency as a key element. We train with open license data and filter for quality carefully when possible, however we acknowledge possible biased outputs and unintended misuse stemming from diversity in the data or the backbone model itself. While we use a diverse range of sources for our training data and test with various models, an effort to curate better data covering more aspects such as accent or gender would be a welcome contribution. We do not specifically test for harmful biases in our model but we also welcome future research in this direction. We do not work with any critical use cases or use any personally-sensitive data. All content presented was authored by humans without the assistance of LLMs.

9 Reproducibility Statement

In our work, we include details regarding the datasets and metrics used, training hyperparameters, architectural details, design choices as well synthetic data generation pipelines to ensure completely reproducible results. We also release all relevant code for the work to facilitate further research. In addition, we only use publicly available datasets with CC-BY 4.0 or more permissive licenses and open-weight backbone models allowing for easy replication of our work.

10 Acknowledgements

We gratefully acknowledge feedback in early stages of this work from Marco Gaido. This work was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for ResponsibleAI), by the project DECOLLAGE (ERC-2022-CoG 101088763), and by FCT/MECI through national funds and when applicable co-funded EU funds under UID/50008: Instituto de Telecomunicações.

References

Kshitij Ambilduke, Ben Peters, Sonal Sannigrahi, Anil Keshwani, Tsz Kin Lam, Bruno Martins, André FT Martins, and Marcely Zanon Boito. From tower to spire: Adding the speech modality to a translation-specialist llm. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 19658–19673, 2025.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pp. 4218–4222, 2020.

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. On the landscape of spoken language models: A comprehensive survey. *arXiv preprint arXiv:2504.08528*, 2025.

Giuseppe Attanasio, Sonal Sannigrahi, Ben Peters, and André Filipe Torres Martins. Instituto de telecomunicações at IWSLT 2025: Aligning small-scale speech and language models for speech-to-text learning. In Elizabeth Salesky, Marcello Federico, and Antonis Anastasopoulos (eds.), *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pp. 347–353, Vienna, Austria (in-person and online), July 2025. Association for Computational Linguistics. ISBN 979-8-89176-272-5. doi: 10.18653/v1/2025.iwslt-1.36. URL https://aclanthology.org/2025.iwslt-1.36/.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 798–805. IEEE, 2023.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y Guo, and Irwin King. Recent advances in speech language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13943–13970, 2025.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*, 2024.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, et al. Aya vision: Advancing the frontier of multilingual multimodality. *arXiv preprint arXiv:2505.08751*, 2025.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*, 2024.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Shoutao Guo, Shaolei Zhang, Qingkai Fang, Zhengrui Ma, Yang Feng, et al. Fastlongspeech: Enhancing large speech-language models for efficient long-speech processing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, et al. Wavllm: Towards robust and adaptive speech large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4552–4572, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *ICLR (Poster)*, 2015. URL `http://dblp.uni-trier.de/db/conf/iclr/iclr2015.html#KingmaB14`.

Beomseok Lee, Marcely Zanon-Boito, Laurent Besacier, and Ioan Calapodescu. Naver labs europe submission to the instruction-following track. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pp. 186–200, 2025.

Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. In *Proc. Interspeech 2018*, pp. 3459–3463, 2018.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Yuang Li, Yu Wu, Jinyu Li, and Shujie Liu. Accelerating transducers through adjacent token merging. In *Proc. Interspeech 2023*, pp. 1379–1383, 2023.

Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-weon Jung, Xuankai Chang, and Shinji Watanabe. Voxtlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13326–13330. IEEE, 2024.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Nicolas Boizard, et al. Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*, 2025.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL `https://arxiv.org/abs/2503.01743`.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R Costa-Jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, et al. Spirit-lm: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13:30–52, 2025.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702, 2020.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, et al. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 634–645, 2022.

Ricardo Rei, Nuno M Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André FT Martins. Tower+: Bridging generality and translation specialization in multilingual llms. *arXiv preprint arXiv:2506.17080*, 2025.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *ICCV*, 2025.

Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Holitom: Holistic token merging for fast video large language models. In *NeurIPS*, 2025a.

Kele Shao, Keda Tao, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. When tokens talk too much: A survey of multimodal long-context token compression across images, videos, and audios. *arXiv preprint arXiv:2507.20198*, 2025b.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan Sharma, Wei-Lun Wu, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. Slue phase-2: A benchmark suite of diverse spoken language understanding tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8906–8937, 2023.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Cappellazzo Umberto, Liu Xubo, Ma Pingchuan, Petridis Stavros, and Pantic Maja. Omni-avsr: Towards unified multimodal speech recognition with large language models. *arxiv 2511.07253*, 2025.

Francesco Verdini, Pierfrancesco Melucci, Stefano Perna, Francesco Cariaggi, Marco Gaido, Sara Papi, Szymon Mazurek, Marek Kasztelnik, Luisa Bentivogli, Sébastien Bratières, et al. How to connect speech foundation models and large language models? what matters and what does not. *arXiv preprint arXiv:2409.17044*, 2024.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 993–1003, 2021.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025b. URL https://arxiv.org/abs/2509.17765.

Hongzhi Zhang, Jingyuan Zhang, Xingguang Ji, Qi Wang, and Fuzheng Zhang. Dyntok: Dynamic compression of visual tokens for efficient and effective video understanding. *arXiv preprint arXiv:2506.03990*, 2025.

Zihan Zhao, Yiyang Jiang, Heyang Liu, Yu Wang, and Yanfeng Wang. Librisqa: A novel dataset and framework for spoken question answering with large language models. *IEEE Transactions on Artificial Intelligence*, 2024.

Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. Problems with cosine as a measure of embedding similarity for high frequency words. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 401–423, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.45. URL https://aclanthology.org/2022.acl-short.45/.