LINA-SPEECH: GATED LINEAR ATTENTION IS A FAST AND PARAMETER-EFFICIENT LEARNER FOR TEXT-TO-SPEECH SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural codec language models have demonstrated state-of-the-art performance in text-to-speech (TTS) synthesis. Leveraging scalable architectures like autoregressive transformers, they capitalize on the availability of large speech datasets. When framing voice cloning as a prompt continuation task, these models excel at cloning voices from short audio samples. However this approach can't be extended to multiple speech excerpts and is limited since the concatenation of source and target speech must fall within the maximum context length which is determined during training. In this work, we propose a model that replaces transformers with emergent recurrent architecture such as Gated Linear Attention (GLA). Our model, Lina-Speech, outperforms or matches the baseline models that are up to 4x it's size. We showcase intial-state tuning as a parameter-efficient fine-tuning technique that optimizes the initial state of the recurrent layers, resulting in compact and expressive speaker embedding with fine-grained control over the speech style. Compared to prompt continuation, it allows voice cloning from multiple speech excerpts and full usage of the context window for synthesis. This approach is fast, deployable and does not rely on auxiliary modules. It also demonstrates extensive adaptation to out-of-domain data. We will release publicly our code and checkpoints. Audio samples are available at `https://anonymsubm.github.io`.

1 INTRODUCTION

Scaling text-to-speech Betker (2023) (TTS) model and data has led to dramatic improvements with regards to quality, diversity and cloning capabilities. Leveraging neural audio codec Zeghidour et al. (2021); Défossez et al. (2023) and simple problem formulation such as next-token prediction have shown state-of-the-art results in zero-shot voice cloning, extending in-context learning abilities observed primarily on natural language to the codec language. Under this setting zero-shot voice cloning is formulated as a prompt continuation task and provides state-of-the-art results starting from 3s of prompt audio. In contrast with prior works, this approach put more pressure on the pre-training stage, where large-scale speech dataset are needed in order to get sufficient in-context learning abilities and less on domain knowledge. In this direction, the transformer has been the leading architecture for scalable auto-regressive modeling of speech.

While transformer is still the dominant architecture for auto-regressive large-scale generative modeling, the attention weights learned Lemerle et al. (2024); Jiang et al. (2024) during text-to-speech synthesis suggest that self-attention might be a sub-optimal choice for this particular task. Indeed, as observed in previous works, transformers tend either to focus on local information Parcollet et al. (2024) with respect to a given time-step or learn tight cross-attention between text and audio. This potential waste of computation reflect also the lack of inductive biases towards monotonicity which result in instabilities compared to NAR TTS models Yang et al. (2024b). Moreover the quadratic complexity of self-attention and the relatively high framerate of neural audio codec prevent training on long context. Also they typically fail at extrapolating to longer lengths. As a consequence during inference, zero-shot voice cloning by prompt continuation face a trade-off between longer prompt containing more information about the target speaker and short prompt that allow the model to synthesize over the remaining of the context window. Moreover cloning a voice from short excerpts makes challenging to capture elements of the voice such as accent, thus relying primarily on pretraining or a subsequent fine-tuning.

In this work, we introduce Lina-Speech, a model that replaces transformers with Gated Linear AttentionYang et al. (2024c). Gated Linear Attention is an emergent recurrent architecture for language modeling that has shown promising result on natural language modeling while scaling linearly with the sequence length. Our contributions include the following :

- When used in a zero-shot prompt continuation, it shows competitive performance against baselines that have up to 4x times more parameters.
- We showcase a parameter-efficient fine-tuning approach (PFET) that is proper to stateful model for voice cloning. This method is fast and deployable (less than 15 seconds in average for 5-30min of speech on consumer grade GPU) and shows extensive aptitude on both in-domain and out-of-domain speech corpus. It enables voice cloning with full usage of the context length.

2 Related Work

Large-Scale TTS Large-scale TTS state-of-the-art relies massively on transformers for both autoregressive (AR) Wang et al. (2023); Betker (2023); Lyth & King (2024) and non-autoregressive Chang et al. (2022); Shen et al. (2024); Le et al. (2023) (NAR) architectures . NAR transformers for speech synthesis, such as those based on diffusion or flow-matching, traditionally require either precomputed durations or an additional generative model. While fine-grained duration annotations can be challenging to produce for noisy large-scale datasets, recent work has introduced coarse durations at the word or sentence level instead. In contrast, AR models have demonstrated strong performance when trained on in-the-wild data, without needing intermediate feature representations. Although pure NAR models tend to excel in terms of inference speed and robustness, they tend to suffer from over-smoothness Yang et al. (2024a); Ren et al. (2022), resulting in reduced diversity and less fidel prosody. Recent research has begun to blend techniques traditionally associated to NAR and AR approaches: Xin et al. (2024) explicitly models durations in an AR transformer for greater robustness, while Yang et al. (2024b) explores AR generative models for prosody and duration modeling on top of a NAR flow-matching acoustic model. While large-scale AR acoustic modeling relies heavily on neural audio codecs, diffusion and flow-matching Le et al. (2023); Betker (2023); Shen et al. (2024) methods have proven being effective at scale for both data space (e.g., mel spectrograms) and latent space modeling.

Zero-shot TTS Zero-shot TTS is the task of synthesizing speech from unseen samples at test time. Traditional approaches includes the use of speaker encoder that produces embedding. Multi-samples approaches have been introduced such as Mega-TTS2 Jiang et al. (2024) in order to close the gap with fine-tuning approach and capture aspect of the prosody that can't be contain within a single excerpt. In contrast Large-scale TTS models relies on the in-context learning abilities: prompt-continuation Wang et al. (2023); Peng et al. (2024b) and infilling strategies Le et al. (2023) have been shown succes with as few as 3 seconds of audio, including from noisy sources such as spontaneous speech Peng et al. (2024b) or podcasts.

Parameter Efficient Fine-Tuning Parameter Efficient Fine-Tuning (PEFT) Xu et al. (2023) focuses on identifying the optimal subset of parameters for fine-tuning a model, resulting in a compact variation of the original model. PEFT has gained popularity across large language models (LLMs) and other large-scale generative models, enabling adaptation on a single GPU in a fraction of the pretraining time. These methods include LoRA Hu et al. (2022); Dettmers et al. (2023) and its variants. Alternative techniques involve tuning embeddings that are not parameters during training, such as prompt tuning Lester et al. (2021); Liu et al. (2022). PEFT can outperform full fine-tuning in small data scenarios, where full fine-tuning may lead to catastrophic forgetting. Most of these techniques were initially applied to natural language processing or image generation. Qi et al. (2024) evaluates the use of LoRA for domain adaptation in emotional text-to-speech (TTS).

3 Preliminaries

Given an input $\mathbf{X} \in \mathbb{R}^{N\times d}$ self-attention for auto-regressive modeling derives three linear projections: the query matrix $\mathbf{Q} \in \mathbb{R}^{N\times d_k}$, the key matrix $\mathbf{K} \in \mathbb{R}^{N\times d_k}$, the value matrix $\mathbf{V} \in \mathbb{R}^{N\times d_v}$, and a causal mask $\mathbf{M}_{i,j} = \mathbf{1}_{i<j}$ $\mathbf{M} \in \mathbb{R}^{N\times N}$. The parrallel form of attention is defined as:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} \odot \mathbf{M}\right)\mathbf{V}, \tag{1}$$

where $\odot$ denotes element-wise multiplication, and admits the recurrent form,

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V})_t = \frac{\sum_{i=1}^{t} exp(\mathbf{q_t}\mathbf{k_i^T})\mathbf{v_i}}{\sum_{i=1}^{t} exp(\mathbf{q_t}\mathbf{k_i^T})}. \tag{2}$$

during inference.

3.1 Linear Attention

A variation of attention known as linear attention Katharopoulos et al. (2020), approximates the softmax behavior with a general similarity function k and its associated feature map ϕ. The linear attention then can be expressed as:

$$\text{LinearAttention}(\mathbf{Q},\mathbf{K},\mathbf{V})_t = \frac{\sum_{i=1}^{t} \phi(\mathbf{q_t})\phi(\mathbf{k_i})^T\mathbf{v_i}}{\sum_{i=1}^{t} \phi(\mathbf{q_t})\phi(\mathbf{k_i})^T}. \tag{3}$$

Denoting,

$$\mathbf{S_t} = \sum_{i=1}^{t} \phi(\mathbf{k_i})^T\mathbf{v_i}, \quad \mathbf{z_t} = \sum_{i=1}^{t} \phi(\mathbf{k_i})^T, \mathbf{o_t} = \frac{\phi(\mathbf{q_t})S_t}{\phi(\mathbf{q_t})\mathbf{z_t}} \tag{4}$$

it can be expressed following the updating rule:

$$\mathbf{S_t} = \mathbf{S_{t-1}} + \phi(\mathbf{k_t})^T\mathbf{v_t}, \quad \mathbf{z_t} = \mathbf{z_{t-1}} + \phi(\mathbf{k_t})^T \quad \mathbf{o_t} = \frac{\phi(\mathbf{q_t})S_t}{\phi(\mathbf{q_t})\mathbf{z_t}}, \tag{5}$$

sheding light that it boils down to a RNN with matrix-valued state.

In practice, recent works get rid of the normalization term and set ϕ as the linear kernel ($\phi = Id$). This leads to the simplified recurrent form of linear attention:

$$\mathbf{S_t} = \mathbf{S_{t-1}} + \mathbf{k_t}^T\mathbf{v_t}, \quad \mathbf{o_t} = \mathbf{q_t}\mathbf{S_t}, \tag{6}$$

where S_t acts as the constant size kv cache in traditional transformer.

3.2 Gated Linear Attention (GLA)

Despite its efficiency, basic linear attention under performs compared to standard self-attention. Recent research in linear-complexity language models (e.g., RWKV-{4,5,6}Peng et al. (2023; 2024a), GLAYang et al. (2024c), Mamba-{1,2}Gu & Dao (2024); Dao & Gu (2024)) have found that incorporating data-dependent updates into Equation 6 significantly narrows the performance gap with transformers.

For this reason Gated Linear Attention Yang et al. (2024c) (GLA) comes with a data-dependent structured decay applied to the previous state, resulting in the following update rule:

$$\mathbf{S_t} = \mathbf{G_t} \odot \mathbf{S_{t-1}} + \mathbf{k_t^T}\mathbf{v_t}, \tag{7}$$

where $\mathbf{G_t}$ is a gating mechanism that modulates the contribution of past states.

3.3 Key aspects of GLA for speech synthesis

- **Performance:** GLA achieved state-of-the-art results in linear-complexity language modeling, even matching or surpassing transformer models for some tasks at large scale.
- **Efficiency:** GLA admits hardware efficient implementationYang et al. (2024c) by imposing some structure to the gating term $\mathbf{G_t}$ and leveraging chunk wise form of **??**. Its linear scaling in sequence length makes it an attractive option for tasks like audio modeling, streaming or on device application.
- **Inductive bias of locality:** While linear language model are known to under-perform on recall-intensive tasks Arora et al. (2024), we hypothesize that they could mitigate the inefficiency of self-attention in domain like speech modeling Parcollet et al. (2024); Lemerle et al. (2024); Jiang et al. (2024), where it appears to be less critical or even unnecessary.

4 Method

Lina-Speech is an autoregressive encoder-decoder architecture that learns neural codec tokens $\mathbf{c}$ and is conditioned on text input $\mathbf{x}$ via a text encoder. When combined with a residual vector quantizer, we employ the delaying strategy introduced by MusicGen Copet et al. (2023), and if only one codebook is used, it simplifies to standard next-token prediction,

$$p(\mathbf{c}|\mathbf{x}) = \prod_{t=0}^{T} \prod_{q=1}^{Q} p(c_{q,t}|\mathbf{x}, \mathbf{c}_{<q,t}, \mathbf{c}_{:,<t}). \tag{8}$$

4.1 Model architecture and inference

Model architecture The text encoder is a non-causal transformer encoder that uses RoPE positional encoding. The acoustic model includes both an audio encoder and a decoder, featuring a transformer-like architecture (without positional encoding), where self-attention is replaced by GLA layers, and SwiGLU Shazeer (2020) is used as feed-forward network,

$$\begin{aligned} \mathbf{Y} &= \mathbf{X} + \text{GLA}(\text{LayerNorm}(\mathbf{X})), \\ \mathbf{Y}' &= \mathbf{Y} + \text{Swish}(\text{LayerNorm}(\mathbf{Y})). \end{aligned} \tag{9}$$

The decoder takes input from the audio encoder and a cross-attention layer between the text and audio encoder outputs. To improve robustness, we used specialized cross-attention from Lemerle et al. (2024), replacing sinusoidal positional encoding with convolutional positional encoding for enhanced training stability.

Inference We use top-k sampling with $k = 100$ for all models and ablations, and treats **EOS** token as an additional token to the audio codebook. If we use a RVQ as audio codec, we use greedy sampling for the residuals.

4.2 Initial State Tuning

We have seen that Gated Linear Attention achieves linear complexity by replacing the expanding key-value cache of transformers with a constant-sized memory, represented by the matrix-valued state $\mathbf{S_t}$ in 6. During training in each layer states are initialized to zeros, that is $\mathbf{S_0} = \mathbf{0}$. Recent work stemming from the RWKV community Peng et al. (2023; 2024a); Fish (2024) have demonstrated that this type of memory can be subject to PEFT for domain adaptation or instruction tuning of large language models. Since the state encodes past information without expanding on the time axis, it offers a compact alternative to prompt tuning. To the best of our knowledge, neither initial state tuning nor prompt tuning has been successfully applied to speech synthesis.

We found out that:

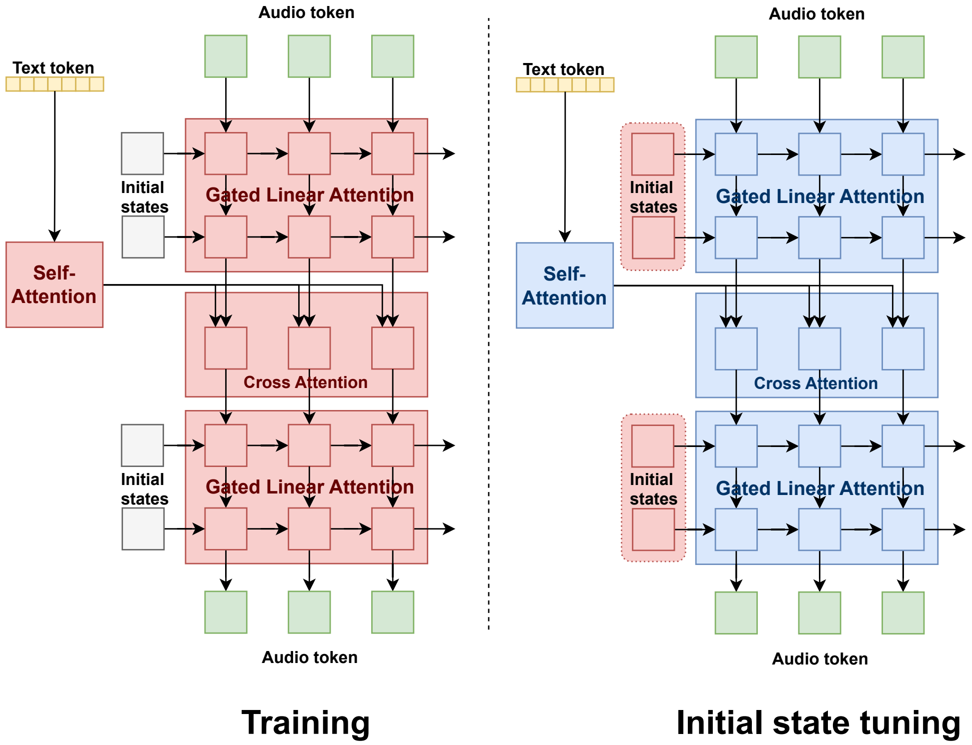

Figure 1: An overview of Lina-Speech and the Init State Tuning. Red colour denotes the modules or states considered for training. Blue colour denotes the modules that are kept frozen. Green and yellow are respectively audio and text tokens. On the left: during training an encoder-decoder architecture with cross-attention learns next-token prediction over a neural codec language. Initial states of gated linear attention are set to zero. On the right: a pre-trained model can learns a new voice or style within 20 steps by tuning a randomly initialized first state.

- This approach is robust to the choice of hyper-parameters across datasets, making it suitable for automated tasks of voice adaptation with simple tuning strategies. In practice we use the same learning rate $\lambda = 0.1$ and two pass over the target dataset with a batch size of 8 utterances for all examples.
- When restricting to 5-30 minutes of speech, the state matrix can be parameterized as a rank-1 matrix, reducing the parameters set to a pair of vectors per head and per layer, that is $\mathbf{S_0} = \mathbf{k_0^T v_0}$, without significant performance degradation.
- The tuning is fast, lasting less than 15 seconds in average on a RTX3080.

5 Experiments

5.1 Experimental setup

Datasets We trained Lina-Speech on a publicly avaible english subset of MLS[1] Lacombe et al. (2024) which consists of 10k hours of librivox recordings. We do not use the provided transcription and rather used the Automatic Speech Recognition (ASR) model Nemo [2]. We also added both LibriTTS Zen et al. (2019) and its restored version LibriTTS-R Koizumi et al. (2023) with their normalized transcripts. We used WavTokenizer Ji et al. (2024)[3] as a neural audio codec which encodes speech at a rate of 75 token/s, with a codebook size of 4096 and compared it against EnCodec Défossez et al. (2023) Koizumi et al. (2023). For text input, we learn a BPE model with vocabulary size of 256 based on the lower-cased transcripts from LibriTTS. During init state tuning, we used Expresso Nguyen et al. (2023) as an out-of-domain dataset for adaptation.

Model Configuration We provide detailed model configuration and hyper-parameters in Appendix A.

Training and Inference The main model is trained for next-token prediction with cross-entropy loss for 500k steps with a batch size of approximately 100k tokens ($\approx$ 22min of speech). We use AdamW optimizer with a learning rate of 2×10^{-4} , a cosine learning rate schedule for the first 1k steps, a weight decay of 0.1 and gradient clipping of 0.1. We group samples of similar lengths within 10 buckets in order to avoid padding. The training takes about 4 days on two RTX4090. Inference of audio speech is made by feeding predicted tokens to WavTokenizer decoder and Vocos Siuzdak (2023) for WavTokenizer and EnCodec token respectively. For long synthesis we continue training for 100k steps on librilight-medium. We rely on the official hardware efficient implementation of GLA Yang & Zhang (2024).

5.2 Objective metrics

We measure word error rate (WER) and character error rate (CER) using the same model from Nemo as for speech transcription. We also measured speaker similarity as the cosine similarity of WavLM Chen et al. (2022) embedding of target and synthesized speech using a pretrained checkpoint [4].

5.3 Subjective metrics

We conducted subjective experiment using Mean Opinion Score (MOS) to measure the naturalness and similarity to the target speaker via the platform Prolific. The complete details of the subjective are described in Appendix B.

5.4 Baselines

The baselines includes:

[1] `parler-tts/mls_eng_10k`

[2] stt_en_fastconformer_hybrid_large_pc

[3] WavTokenizer-medium-speech-75token

[4] wavlm-base-plus-sv

- The TTS enhanced version of VoiceCraft Peng et al. (2024b), a decoder-only model trained on GigaSpeech and librilight. They trained an EnCodec model on their dataset.
- StyleTTS2 Li et al. (2024) an end-to-end TTS model that leverages latent diffusion for style modeling.
- an unofficial reproduction of VALLE-X Zhang et al. (2023) that leverages an official EnCodec model. Plachtaa
- Parler-TTS Lacombe et al. (2024), is a series of reproduction of Lyth & King (2024) that allows synthesis controlled by textual description of the voice. Interestingly, this reproduction differs from the original paper by separating text and audio sequence and employing cross-attention between the two modalities instead of self-attention on the concatenation of both, making the architecture closer to Lina-Speech. They leverage DAC Kumar et al. (2024) as audio codec.

5.5 Experiments

- **Zero-shot voice cloning** We evaluate zero-shot against the baselines with the exception of Parler-TTS on the clean test split of LibriTTS.
- **Initial State Tuning** Since Parler-TTS voice adaptation is limited to a subset of speakers from the training set, we evaluate our model using an initial state tuned against the same set of speakers. For LibriTTS it consists of a list of names[5] that we give to Parler as a textual prompt. We evaluate syntheses by generating random sentences with the help of a large language model and compare them to the training set. For the Expresso dataset we employ the same strategy and tune initial states for each pair (speaker, style). We provide details on the specific guidance in the appendix.

5.6 Results and Discussion

5.6.1 Experiment #1: Zero-shot voice cloning

Table 5.6.1 presents the results of the objective and subjective evaluation conducted on the task of zero-shot voice cloning. First, results for the ground truth is presented in order to provide the lower (respectively higher) boundary which can be expected for the different measurements. Those scores generally reflect the internal noise or the non-reducible error of the measurement, which can have many causes: manual transcription errors of the ground truth, limitations of the algorithm used for automatic transcription, or internal perception noise.

Table 1: {it Zero-shot evaluation on LibriTTS test clean split. The objective evaluation includes: Word Error Rate (WER), Caracter Error Rate (CER), and cosine similarity to the reference speaker (Sim.). The subjective evaluation includes: MOS for naturalness (N-MOS) and MOS for similarity to the reference speaker (S-MOS). The number of parameters for each model is reported in #Params.

Model	Objective eval.			Subjective eval.		
	WER (↓)	CER (↓)	Sim. (↑)	N-MOS (↑)	S-MOS (↑)	#Params.
Ground Truth	4.5%	1.5%	-	4.26 ± 0.19	4.3 ± 0.22	-
StyleTTS2	**3.2**%	**0.8**%	0.89	3.93 ± 0.22	4.02 ± 0.24	148M
VALLE-X	14.1%	7.6%	0.92	3.21 ± 0.27	3.07 ± 0.29	300M
VoiceCraft	6.6%	3.4%	**0.94**	3.62 ± 0.24	3.55 ± 0.25	833M
Lina-Speech	7.5%	2.9%	0.93	$\mathbf{4.18 \pm 0.19}$	$\mathbf{4.10 \pm 0.20}$	169M

Objective evaluation Among all the TTS systems under comparison, StyleTTS2 presents by the lowest WER and CER that turns to be lower than the ground truth which could indicate that the duration prediction is biased and might be attributed to the oversmoothing Ren et al. (2022) of NAR TTS models. Among the TTS models belonging to the auto-regressive language models (LM), Lina-Speech presents WER of 7.5% and CER of 2.9% that are comparable to those of VoiceCraft

[5] speaker_ids_to_names.json

while VALLE-X presents about twice errors in proportion. Considering the cosine similarity to the speaker, the LM TTS models present the higher similarity (above 0.9 for all), compared to StyleTTS2 (below 0.9).

Subjective evaluation The right side of the Table presents the mean and standard deviation obtained in terms of naturalness (N-MOS) and the similarity (S-MOS). Lina-Speech presents the higher scores either in terms of naturalness (N-MOS=4.18) or of similarity to the reference speaker (S-MOS=4.10). A comparison with the other TTS models reveals that Lina-Speech is the only TTS models reaching MOS scores above 4.0 for both naturalness and similarity, while the others remain generally below and sometimes even closer to 3.0 as for VALLE-X. A further statistical analysis of those results was conducted between each pair of models using a Student's t-test Fisher (1925) in order to assess whether the observed differences are significant. Significant differences with p-value = 0.05 was found between Lina-Speech and all other models, except for the ground truth and StyleTTS2.

In conclusion, Lina-Speech presents comparable performance to the other LM TTS models in terms of objective metrics and is considered as significantly better in terms of subjective metrics while having much less parameters than the other models.

5.6.2 Experiment #2: Initial State Tuning

Table 5.6.2 presents the results of the objective and subjective evaluation specifically designed for a comparison of the initial state tuning in Lina-Speech against Parler-TTS. The evaluation was conducted for both in and out of domain tasks: LibriTTS was used for the in-domain evaluation, and EXPRESSO was used for the out-of-domain evaluation.

Table 2: {it Evaluation of Lina-Speech vs. Parler-TTS mini models. In-domain evaluation is reported on the LibriTTS dataset and out-of-domain evaluation is reported on the EXPRESSO dataset. The objective evaluation includes: Word Error Rate (WER), Caracter Error Rate (CER), and cosine similarity to the reference speaker (Sim.). The subjective evaluation includes: MOS for naturalness (N-MOS) and MOS for similarity to the reference speaker (S-MOS). The number of parameters for each model is reported in #Params.

Model		Objective eval.			Subjective eval.		
		WER ↓	CER ↓	Sim. ↑	N-MOS ↑	S-MOS ↑	#Params.
	Ground Truth	5.1%	1.6%		4.68 ± 0.08	4.71 ± 0.08	-
EXPRESSO	Parler Mini[6]	4.9%	4.4%	0.87	**3.69 ± 0.18**	3.39 ± 0.24	674M
	Lina-Speech	**3.6%**	**1.4%**	**0.93**	3.68 ± 0.17	**3.66 ± 0.15**	169M
	Ground Truth	4.3%	0.9%		4.22 ± 0.18	4.26 ± 0.20	-
LibriTTS	Parler Mini [7]	4.4%	2.6%	0.90	4.06 ± 0.26	3.45 ± 0.27	880M
	Lina-Speech	**2.9%**	**1.3%**	**0.94**	**4.10 ± 0.18**	**3.97 ± 0.24**	169M

Objective evaluation Lina-Speech presents systematically the best performances in terms of WER, CER, and similarity to the speaker, as compared to Parler-TTS, both for in and out of domain tasks. For the in-domain task conducted on LibritTTS, Lina-Speech has a WER of 2.9%, a CER of 1.3% and a cosine similarity to the speaker of 0.94. For the out-of-domain task conducted on EXPRESSO, the performance of Lina-Speech degrades slightly compared to the one observed for the in-domain task. Lina-Speech has a WER of 3.6%, a CER of 1.3%, and a cosine similarity to the speaker of 0.93. This indicates the consistency of Lina-Speech, and in particular its efficiency to adapt to out-of-domain speech.

Subjective evaluation For the in-domain task, Lina-Speech presents the highest score on Libri-TTS both in terms of naturalness and similarity. For the out-of-domain task, this tendency mostly remains observed: Firstly, Lina-Speech and Parler-TTS present comparable performance. The observed difference is marginal and not significant. But secondly, Lina-Speech presents a significantly higher similarity to the speaker as compared to Parler-TTS.

In conclusion, the reported results demonstrate the efficiency of the proposed Init State Tuning to condition on new speakers either in or out of domains. This strategy offers a particularly efficient and stable alternative to other strategies as the one represented by Parler-TTS. Efficient: Lina-Speech only needs 15 seconds of speech for the adaptation when Parler-TTS has been trained specifically for this task. Stable: In Lina-Speech the conditioning is effective and can generalize to out-of-domain speakers, conditions, or speaking styles), while Parler-TTS tends to not systematically or fully respect the given prompt.

6 Conclusion

In this paper, we introduced Lina-Speech a parameter-efficient model during both pretraining and fine-tuning for text-to-speech synthesis. The proposed architecture leverages on Gated Linear Attention and init state tuning for which we discussed the key properties in the context of text-to-speech synthesis: fast and compact. Lina-Speech has been compared to other existing text-to-speech models, objectively and subjectively, and both on in-domain and out-domain tasks. These evaluations revealed experimentally that Lina-Speech is **1)** a particularly efficient zero-shot learner for voice cloning with respect to its size; **2)** init state tuning is an effective PEFT method to condition effectively on in and out of domain speakers from small amounts of data. The comparison with other existing TTS models demonstrates that Lina-Speech is particularly competitive versus much larger models, possibly specifically pre-trained or fine-tuned on a given dataset while being extremely fast and compact.

References

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, Dylan Zinsley, James Zou, Atri Rudra, and Christopher Ré. Simple linear attention language models balance the recall-throughput tradeoff. In *ES-FoMo II: 2nd Workshop on Efficient Systems for Foundation Models, International Conference on Machine Learning (ICML)*, 2024.

James Betker. Better Speech Synthesis through Scaling. *arXiv preprint arXiv:2305.07243*, 2023.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. MaskGIT: Masked Generative Image Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11315–11325, 2022.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518, 2022.

Chiang Cheng-Han, Huang Wei-Ping, and Hung yi Lee. Why We Should Report the Details in Subjective Evaluation of TTS More Rigorously. In *Interspeech*, pp. 5551–5555, 2023. doi: 10.21437/Interspeech.2023-416.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and Controllable Music Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 47704 – 4772, 2023.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *International Conference on Machine Learning (ICML)*, pp. 10041–10071, 2024.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High Fidelity Neural Audio Compression. *Transactions on Machine Learning Research*, 2023.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. In *Advances in Neural Information Processing Systems*, pp. 10088–10115, 2023.

Jelly Fish. Init State Tuning repository. https://github.com/Jellyfish042/RWKV-StateTuning, 2024.

Ronald A. Fisher. Applications of ‘student’s’ distribution. *Metron*, 5:90–104, 1925.

Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In *Submitted to International Conference on Learning Representations (ICLR)*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2022.

ITU-R BS.1534-3. Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems, 2015.

ITU-T P.800.2. Methods for Objective and Subjective Assessment of Speech Quality - Mean Opinion Score Interpretation and Reporting, 2013.

Shengpeng Ji, Ziyue Jiang, Xize Cheng, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, et al. WavTokenizer: an Efficient Acoustic Discrete Codec Tokenizer for Audio Language Modeling. *arXiv preprint arXiv:2408.16532*, 2024.

Ziyue Jiang, Jinglin Liu, Yi Ren, Jinzheng He, Zhenhui Ye, Shengpeng Ji, Qian Yang, Chen Zhang, Pengfei Wei, Chunfeng Wang, Xiang Yin, Zejun Ma, and Zhou Zhao. Mega-TTS 2: Boosting Prompting Mechanisms for Zero-Shot Speech Synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attentionn. In *International Conference on Machine Learning (ICML)*, pp. 5156–5165, 2020.

Jaehyeon Kim, Keon Lee, Seungjun Chung, and Jaewoong Cho. CLaM-TTS: Improving Neural Codec Language Model for Zero-Shot Text-to-Speech. In *International Conference on Learning Representations (ICLR)*, 2024.

Ambika Kirkland, Shivam Mehta, Harm Lameris, Gustav Eje Henter, Éva Székely, and Joakim Gustafson. Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation. In *Speech Synthesis Workshop (SSW)*, pp. 41–47, 2023. doi: 10.21437/SSW.2023-7.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. LibriTTS-R: A Restored Multi-Speaker Text-to-Speech Corpus. pp. 5496–5500, 2023. doi: 10.21437/Interspeech.2023-1584.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-Fidelity Audio Compression with Improved RVQGAN. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

Yoach Lacombe, Vaibhav Srivastav, and Sanchit Gandhi. Parler-TTS. `https://github.com/huggingface/parler-tts`, 2024.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided Multilingual Universal Speech Generation at Scale. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Théodor Lemerle, Nicolas Obin, and Axel Roebel. Small-E: Small Language Model with Linear Attention for Efficient Speech Synthesis. In *Interspeech*, pp. 3420–3424, 2024. doi: 10.21437/Interspeech.2024-508.

Brian Lester, Rami Al-Rfou, and Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3045–3059, 2021.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks. In *Proceedings of the 60th Annual Meeting of the Association of Computational Linguistics (ACL)*, pp. 61–68, 2022.

Dan Lyth and Simon King. Natural Language guidance of High-Fidelity Text-To-Speech with Synthetic Annotations. *arXiv preprint arXiv:2402.01912*, 2024.

Tu Anh Nguyen, Wei-Ning Hsu, Antony d'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, et al. EXPRESSO: A Benchmark and Analysis of Discrete Expressive Speech Resynthesis. In *Interspeech*, 2023. doi: 10.21437/Interspeech.2023-1905.

Titouan Parcollet, Rogier van Dalen, Shucong Zhang, and Sourav Bhattacharya. SummaryMixing: A Linear-Complexity Alternative to Self-Attention for Speech Recognition. In *Interspeech*, pp. 3460–3464, 2024. doi: 10.21437/Interspeech.2024-40.

Bo Peng, Eric Alcaide, Quentin G. Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, G Kranthikiran, Xuming He, Haowen Hou, Przemyslaw Kazienko, Jan Kocoń, Jiaming Kong, Bartlomiej Koptyra, Hayden Lau, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Xiangru Tang, Bolun Wang, Johan Sokrates Wind, Stansilaw Wozniak, Ruichong Zhang, Zhenyuan Zhang, Qihang Zhao, Peng Zhou, Jian Zhu, and Rui Zhu. RWKV: Reinventing RNNs for the Transformer Era. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Teddy Ferdinan, Haowen Hou, Przemys l aw Kazienko, G Kranthikiran, Jan Koco'n, Bartlomiej Koptyra, Satyapriya Krishna, Ronald McClelland, Niklas Muennighoff, Fares Obeid, Atsushi Saito, Guangyu Song, Haoqin Tu, Stanislaw Wo'zniak, Ruichong Zhang, Bingchen Zhao, Qihang Zhao, Peng Zhou, Jian Zhu, and Ruijie Zhu. Eagle and Finch: RWKV with Matrix-Valued States and Dynamic Recurrence. 2024a.

Puyuan Peng, Po-Yao Huang, Abdelrahman Mohamed, and David Harwath. VoiceCraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 12442–12462, 2024b.

Plachtaa. VALL-E-X repository. URL `https://github.com/Plachtaa/VALL-E-X`.

Xin Qi, Ruibo Fu, Zhengqi Wen, Jianhua Tao, Shuchen Shi, Yi Lu, Zhiyong Wang, Xiaopeng Wang, Yuankun Xie, Yukun Liu, Guanjun Li, Xuefei Liu, and Yongwei Li. EELE: Exploring Efficient and Extensible LoRA Integration in Emotional Text-to-Speech. In *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2024.

Yi Ren, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu. Revisiting Over-Smoothness in Text to Speech. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8197–8213, 2022. doi: 10.18653/v1/2022.acl-long.564.

Noam M. Shazeer. GLU Variants Improve Transformer. *ArXiv*, abs/2002.05202, 2020. URL `https://arxiv.org/abs/2002.05202`.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. NaturalSpeech 2: Latent Diffusion Models are Natural and Zero-Shot Speech and Singing Synthesizers. In *International Conference on Learning Representations (ICLR)*, 2024.

Hubert Siuzdak. Vocos: Closing the Gap between Time-Domain and Fourier-based Neural Vocoders for High-Quality Audio Synthesis. In *International Conference on Learning Representations (ICLR)*, 2023.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

Detai Xin, Xu Tan, Kai Shen, Zeqian Ju, Dongchao Yang, Yuancheng Wang, Shinnosuke Takamichi, Hiroshi Saruwatari, Shujie Liu, Jinyu Li, and Sheng Zhao. RALL-E: Robust Codec Language Modeling with Chain-of-Thought Prompting for Text-to-Speech Synthesis, 2024. URL `https://arxiv.org/abs/2404.03204`.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. *Nature Machine Intellingence*, 5:220–235, 2023.

Dongchao Yang, Rongjie Huang, Yuanyuan Wang, Haohan Guo, Dading Chong, Songxiang Liu, Xixin Wu, and Helen Meng. SimpleSpeech 2: Towards Simple and Efficient Text-to-Speech with Flow-based Scalar Latent Transformer Diffusion Models. *Submitted to IEEE Transactions on Audio, Speech and Language (TASLP)*, 2024a.

Dongchao Yang, Dingdong Wang, Haohan Guo, Xueyuan Chen, Xixin Wu, and Helen Meng. Simplespeech: Towards simple and efficient text-to-speech with scalar latent transformer diffusion models. In *Interspeech 2024*, pp. 4398–4402, 2024b. doi: 10.21437/Interspeech.2024-1392.

Songlin Yang and Yu Zhang. FLA: A Triton-Based Library for Hardware-Efficient Implementations of Linear Attention Mechanism, January 2024. URL `https://github.com/sustcsonglin/flash-linear-attention`.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated Linear Attention Transformers with Hardware-Efficient Training. In *Proceedings of the 41st International Conference on Machine Learning (PMLR)*, 2024c. doi: 235:56501-56523.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An End-to-End Neural Audio Codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 30:495–507, 2021.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. pp. 1526–1530, 2019. doi: 10.21437/Interspeech.2019-2441.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Speak Foreign Languages with Your Own Voice: Cross-Lingual Neural Codec Language Modeling. *arXiv preprint arXiv:2303.03926*, 2023.

A Model Architecture

Lina-Speech is a 169M parameters encoder-decoder architecture with text-conditioning through cross-attention.

Text Encoder Our text encoder consists of a stack 6 non-causal transformer block with SwiGLU feed-forward network Shazeer (2020), base dimension 1024. We use dropout with rate 0.1 on the outputs of each block and RoPE positional encoding.

Audio Encoder and Decoder It consists for each of a stack of 6 causal GLA transformer with base dimension 1024. The expansion factor of key projection is set to 0.5. We do not use dropout.

Cross-Attention We use PACA cross-attention Lemerle et al. (2024) with convolutional embedding. We do not use RoPE on queries and keys.

B SUBJECTIVE EVALUATION

Subjective evaluation was also conducted in order to compare the benchmark of algorithms by using human perception. In those evaluations, the human subjects were asked to rate the naturalness of speech samples (N-MOS) and their similarity to a reference speaker (S-MOS) using Mean Opinion Score (MOS) ITU-T P.800.2 (2013) as commonly used in the literature for evaluation text-to-speech synthesis systems. In the following, we provide all details of the experimental protocol following observations reported in Kirkland et al. (2023) and recommendations presented in Cheng-Han et al. (2023). This is achieved in order to improve the transparency and reproducibility of the proposed protocol.

B.1 METHODOLOGY

Each listener was assigned a single experiment across the 3 we introduced. It consists of 12 evaluation and measured for a median time of experiment of 7 min.

B.1.1 CREATION AND PRESENTATION OF THE SPEECH STIMULI

LibriTTS clean - Zero shot We group pair of sentences from a same speaker in the test split. We first draw a prompt candidate randomly between 2 and 5 seconds, then we draw a sample so that the concatenation of both samples remains within 16s in order to account of the context window of the baselines, we discard speakers that do not contains at least 2 samples that satisfy this condition. We draw 40 pairs in this manner.

Init State Tuning In order to adapt our evaluation to Parler that does not provide test set, we synthesized random sentences with Llama3.1 8B with the following prompt.

” Please generate a diverse set of 40 random sentences designed for evaluating text-to-speech synthesis quality. Ensure that:

- Variety in sentence length: Sentences should be between 12 and 20 words long.
- Phonetic coverage: Include a wide range of phonemes, syllable structures, and sounds.
- Incorporate both simple and complex sentence structures.
- Diverse syntax and styles: Use varied syntactic forms, rhythms, and styles to capture different speech patterns and intonations.
- Natural and conversational tone: The sentences should sound natural, like everyday speech, while still offering variability.
- Descriptive and vivid: Include a mix of action-oriented, descriptive, and emotional content to test prosody and emotional intonation.
- Non-repetitive: Ensure that all sentences are distinct from each other, with no repetition in structure or wording.

„

The sentences should be suitable for use in evaluating prosody, naturalness, and phonetic diversity in a text-to-speech system. In the Expresso dataset, we omitted all excerpts where transcripts contained non-verbal instruction such as ”laugh ”, ”breathe ”.

B.1.2 INTERFACE FOR THE EXPERIMENT

The experiment was conducted online and implemented in Python. The experiment was preceded by general recommendation to the subject before starting the evaluation:

- Please use headphones or earphones in a quiet environment.
- Adjust the sound level so you can hear subtle sound differences.

Each experimental run consisted in the evaluation by the subject of 15 speech samples. The order of presentation of the speech stimuli was randomised before each experiment in order to avoid any presentation biases. Each speech sample was presented and evaluated individually and separately on a dedicated page.

On top of the page, a reminder of the instructions was presented. In the the center of the page, the speech sample to be evaluated was presented in the left side and a speech sample of the reference speaker was presented in the upper right corner. The criteria to be rated were presented right next to the speech sample to be evaluated and below the speech sample of the reference speaker The subject had to provide one and only one rating for all criteria before being allowed to proceed to the next speech sample. Subjects were not allowed to revise their rating of previous speech samples.

B.1.3 Instructions given to the subjects

In each experimental run, the subjects were asked to evaluate the naturalness of the speech sample and its similarity to a speech sample of a reference speaker. The speech sample of the reference speaker consists of a real speech sample of the reference speaker which was selected systematically different from the speech sample under evaluation. By doing so, we were intending to prevent the subject from confusing the general perception of a speaker identity with one single realisation on a particular utterance and thus trying to mitigate the linguistic biased in the perception of speaker similarity.

The following instructions were given to the subjects:

- In this experiment, you have to judge a speech sample with respect to a reference speech sample.
- The reference and the sample to judge does not pronounce the same utterance.

For each speech sample, please rate :

- The speech **NATURALNESS**: *to which extent you judge the speech sample as natural as real human speech?*
- The **SIMILARITY** to the reference speaker: *to which extent the speech sample is judged close to the reference speaker?*

Complementary recommendations were provided either to precise the definition or the task to be achieved:

1. The aspects of speech naturalness include: fluency and appropriateness of pronunciation and prosody, and diversity in the expression of styles and emotions.
2. The samples may have different recording conditions or background noise. As the scope of this experiment is only focused on speech naturalness, please try to ignore them during your evaluation.

B.1.4 Psychometric measurements and assessment methodology

The Mean Opinion Score (MOS) ITU-T P.800.2 (2013) was used to measure naturalness and similarity, using a Likert scale ranging from 1 to 5. For both criteria, we used the original scale, as suggested in Kirkland et al. (2023):

$$1(Bad), 2(Poor), 3(Fair), 2(Good), 5(Excellent)$$

Following the MUSHRA methodology ITU-R BS.1534-3 (2015), we additionally incorporated hidden references of real speech samples for each evaluation run.

B.2 Subjects Recruitment

B.2.1 Recruitment platform

Prolific crowd-sourcing platform was used for the experiment: `https://www.prolific.com/`.

B.2.2 Language background and geographic location of the evaluators

For this evaluation, we used the following filters available in this plateform to recruit subjects:

- Location: USA or UK
- Language: First language and Primary language and Fluent languages = English

The combination of these filters ended up into 91k subjects potentially skilled for the evaluation.

B.2.3 Subjects qualification

We applied several filters to assess the qualification of the subjects, in order to reject those who do not fulfill the necessary conditions to be considered qualified for the evaluation.

The list of conditions is listed as follows:

- non-native English speaker
- rate below 3 any real speech sample on the N-MOS
- time spent to complete the experiment is below 3m 30s
- the mean MOS of the subject deviates from the overall mean of all subjects by more than two standard deviations, as proposed by Kim et al. (2024)

The subject was considered not qualified if at least one of the conditions was not fulfilled.

165 subjects participated in the experiment. Applying these filters, the qualification rate of the subjects was about 76%. We rejected 39 subjects from a total of 165, so the total of qualified subjects was 126 whose ratings were further used for analysis.