AdCraft: AI-Powered Radio Ad Generation Platform

How much does it cost to make a radio ad?

£500 ~ £5000

Radio Advertising Costs 2024 | RadioAdvertising.co.uk | UK.

Ad creation timeline

This may not be easily accessible for smaller businesses due to:

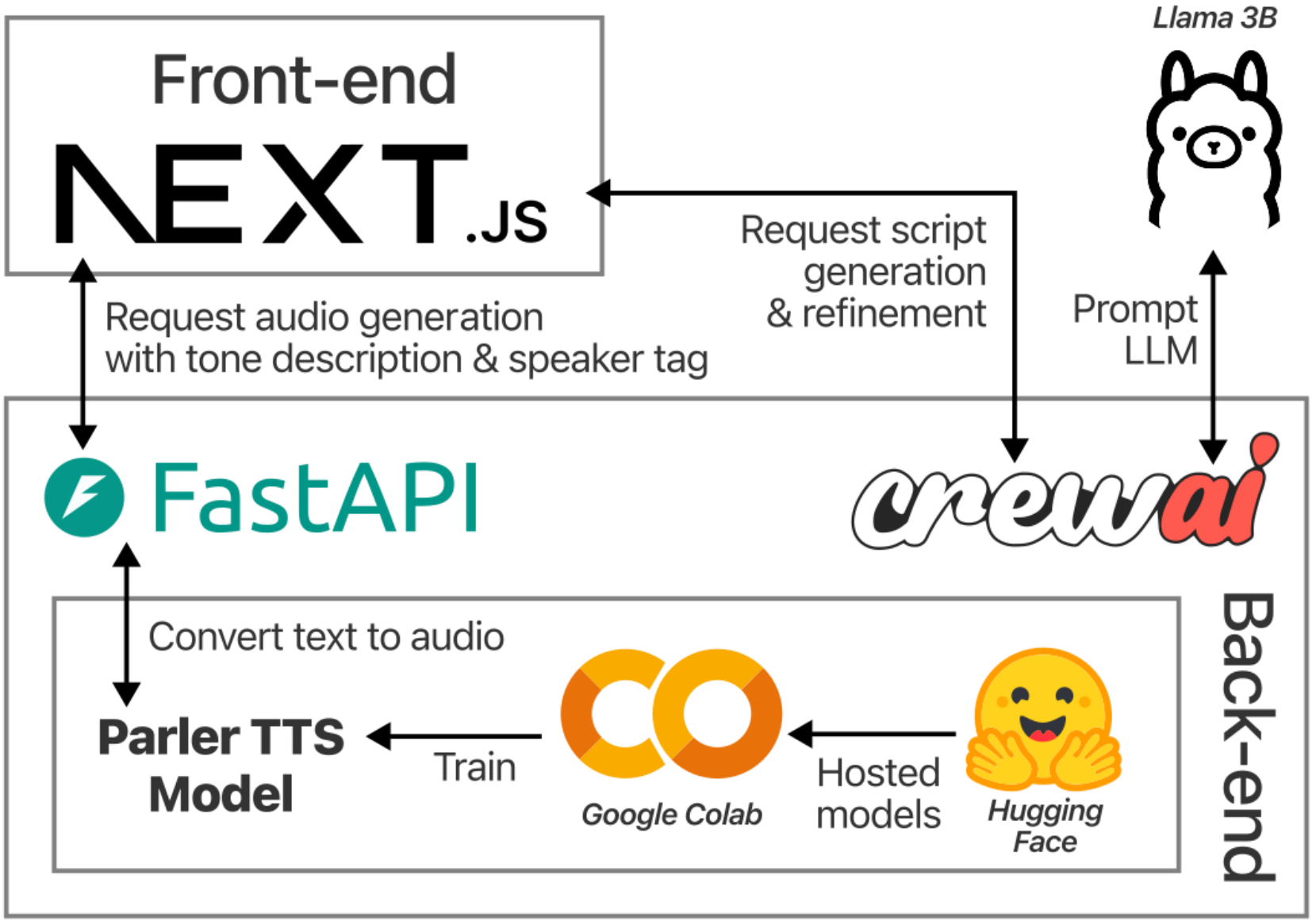
- High cost
- Long timeframe (may be up to several weeks)
- Little flexibility once production starts

Our product

AI-powered webapp that produces quality audio ad in a few minutes.

- Generate customised ad scripts from user prompts
- Refine script with AI
- Generate emotive audio based on art direction
- Track version history of scripts and audio outputs
- Cost a fraction of the traditional workflow

Technical architecture



Parler-TTS paper: https://arxiv.org/abs/2402.01912

Demo

Fine-tuned:

Line-by-line art directions and script:

Base model:

Unified art directions and script:

We’ve come a long way :)

I just graduated from college, I’m so happy!

Voice direction: “A young man speaks excitedly with fast pace. His voice is very expressive and there is minimal noise.”

He’s gonna kill us all! Run!

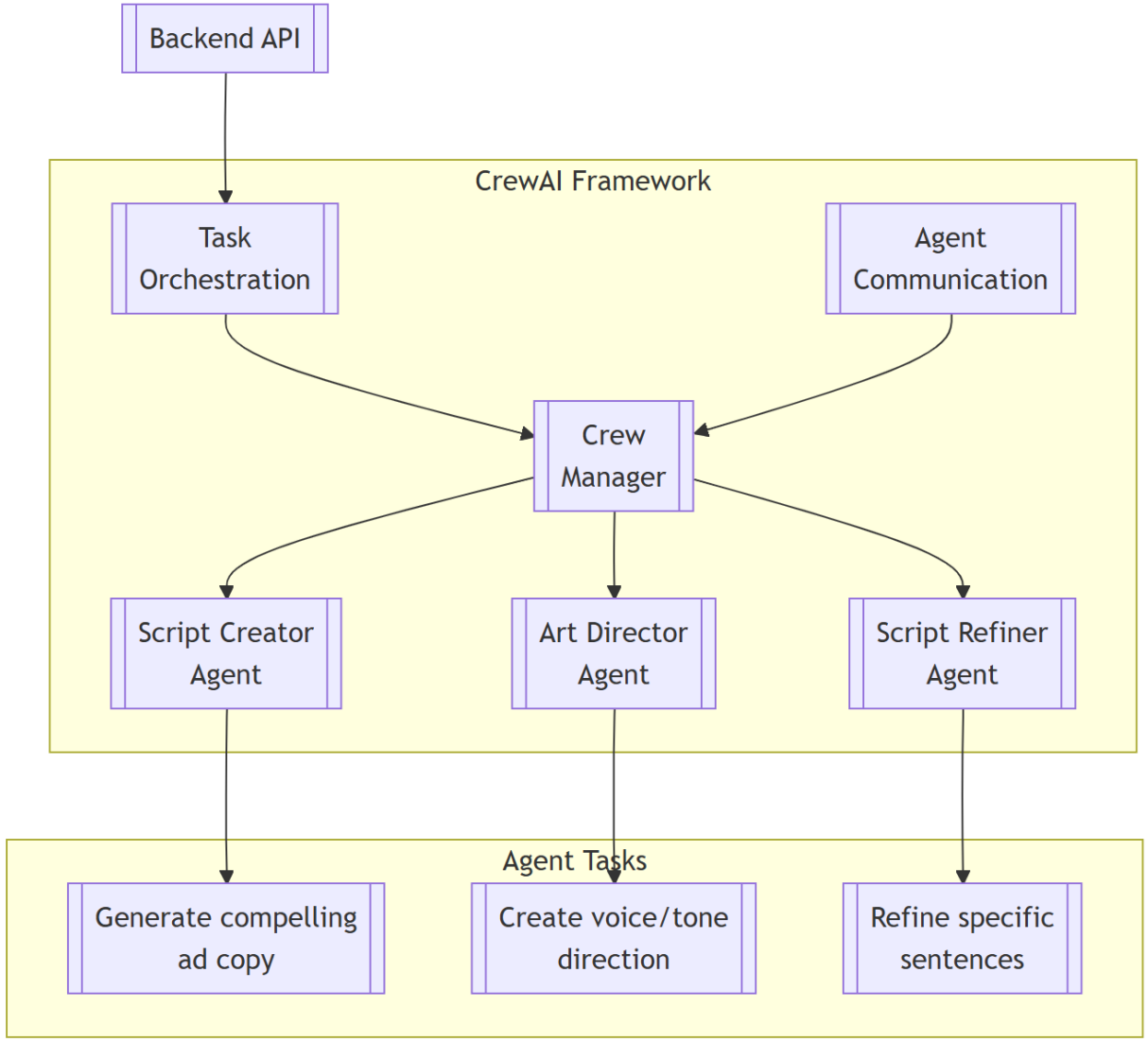
Voice direction: “A female voice in a state of chaos. There is lots of background noise as she is running away.”

Eww, the apple is so rotten, how could you even pick it up?

(Note: the last word has been improvised by the model)

Voice direction: “A young woman with a high-pitched voice speaks slowly and filled with disgust in a confined space. Her voice is very expressive and there is minimal noise.”

The CrewAI framework



Creating audio for engaging ads

Problem: A lot of existing open-source TTS models produce monotone speech with no emotion at all.

Finetuning a TTS model involves:

1) Collecting and preparing datasets with emotional labels for speech
2) Extracting audio features from datasets
3) Generating prompts that include emotional context.
4) Train the model on this set (additionally to its original training set)

Data collection

-Compiled 4 different open-source emotional audio datasets.

-Used DataSpeech to curate natural language prompts using bin values of each feature.

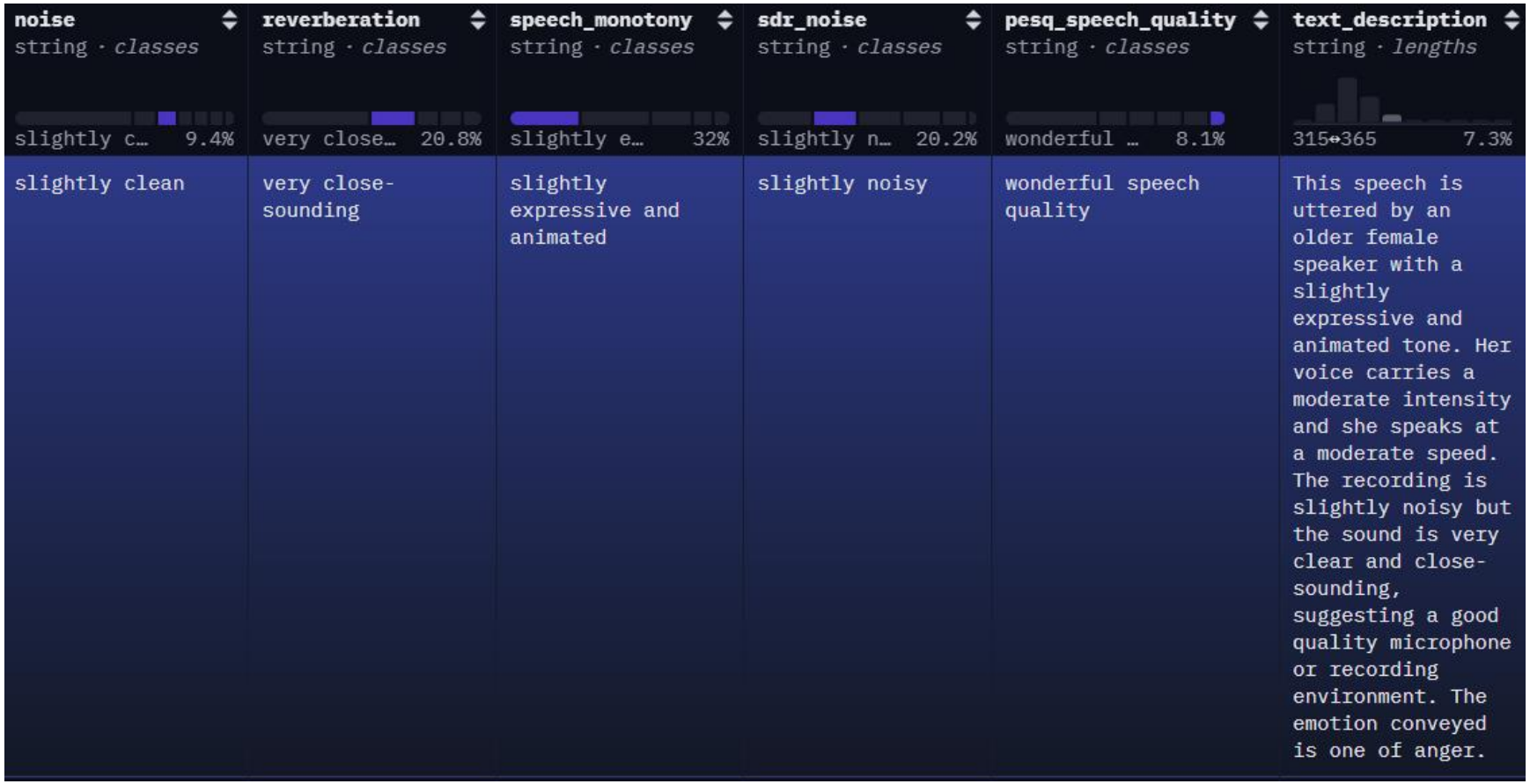
-Added speaker IDs for each distinct speaker.

Data collection

text string · *classes*	emotion string · *classes*	intensity string · *classes*	gender string · *classes*	age string · *classes*	speaker_id string · *classes*
I would li… 4.5%	sad 15.8%	moderate 87%	Female 28.9%	young 60.9%	Crema1078 0.7%
The surface is slick.	neutral	moderate	Female	adult	Crema1028
I'm on my way to the meeting.	angry	moderate	Male	old	Crema1087
It's eleven o'clock.	happy	low	Male	adult	Crema1039
It's eleven o'clock.	fear	low	Female	young	Crema1018
Don't forget a jacket.	neutral	moderate	Female	young	Crema1079
I wonder what this is about	disgust	moderate	Female	young	Crema1063

Over 1000 downloads on huggingface already!

Data collection



noise string · *classes* slightly c… 9.4%	reverberation string · *classes* very close… 20.8%	speech_monotony string · *classes* slightly e… 32%	sdr_noise string · *classes* slightly n… 20.2%	pesq_speech_quality string · *classes* wonderful … 8.1%	text_description string · *lengths* 315↔365 7.3%
slightly clean	very close-sounding	slightly expressive and animated	slightly noisy	wonderful speech quality	This speech is uttered by an older female speaker with a slightly expressive and animated tone. Her voice carries a moderate intensity and she speaks at a moderate speed. The recording is slightly noisy but the sound is very clear and close-sounding, suggesting a good quality microphone or recording environment. The emotion conveyed is one of anger.

Data processing

-For each of the open-source datasets, extracted feature values from file names, different table formats etc.

-Two main datasets exist: one with the audio transcript and all the feature values, and one with the actual audio.

-The rows of each dataset have a 1:1 correspondence with each other.

Data processing

filepath audio · *duration (s)* 1.25 – 7.14	transcript string · *classes* I think I'… 4.5%	emotion string · *classes* angry 15.8%	intensity string · *classes* moderate 87%	gender string · *classes* Male 32.3%	age string · young
0:00 / 0:02	I think I've seen this before.	angry	moderate	Male	young
0:00 / 0:03	I think I've seen this before.	angry	moderate	Female	adult
0:00 / 0:02	I think I've seen this before.	neutral	moderate	Female	young
0:00 / 0:02	I think I've seen this before.	neutral	moderate	Female	adult
0:00 / 0:02	I'm on my way to the meeting.	disgust	moderate	Female	young

Challenges and solutions

Script integrity during refinement

Challenge: Ensuring AI refinements only modify user-selected sentences without affecting other content.

Solution: Implemented a multilayer validation system that

- Tracks original and modified content
- Reverts unauthorised changes
- Provides transparent feedback
- Maintains script structure

AI agent orchestration

Challenge: Coordinating multiple specialised AI agents for cohesive output.

Solution: Used CrewAI framework to:

- Define specialised agent roles
- Structure sequential and parallel tasks
- Manage information flow between agents
- Consolidate results into coherent output

Voice direction integration

Challenge: Integrating art direction with script text in a maintainable format.

Solution: Developed a paired data structure that:

- Maintains 1:1 relationship between script lines and direction
- Preserves formatting during refinement
- Supports rendering in the user interface
- Enables text-to-speech interpretation

Voice consistency over successive generations

Challenge: Parler-TTS generates a random voice per generation, leading to inconsistency over successive generations for the same speaker.

Solution: Trained a voice classifier that:

- Labels a random voice generation with the speaker ID of the closest sounding speaker in our dataset.
- Prepends this speaker ID into the Parler-TTS prompt to ensure consistency over later generations for the same speaker.

Product evaluation

Informal user testing with startup owners and marketers. We used this to:

- Refactor the UI to display all core features directly on the homepage
- Introduce the version control feature
- Introduce a built-in editing interface to add extra effects and trimming

We also conducted listening test with volunteers:

- To rate clarity, emotional expressiveness and appeal from 1 to 5
- Often scored 4 in perceived naturalness but also scored around 2 due to consistency issues.
- Performed best on ads with “happy” labels, worse for “angry” or “frustrated” due to a lack of data.

Performance limitations

The performance of our system is limited mainly by the low availability of open-source datasets with expressive speech.

Limitations include:

- Insufficient speaker data.
- Increased background noise.
- Poor scalability for longer inputs.
- Incorrect transcripts in the dataset.
- Poor performance on less common words.

With access to larger, more diverse, high-quality datasets, its clarity and consistency will be much better.

Limitations with training

Parler TTS uses a text-based tokenizer.

See = Sea [siː]
One = 1 [ʊʌn]
First = 1st
Word = word
Content ≠ Content
Read [ɹiːd] ≠ Read [ɹɛd]

Possible solutions (extensions)

Use a dictionary to normalize words to phonemes
Ask an LLM to predict pronunciation for heteronyms
Allow users to type in phonemes

Extensions

- Using an expanded fine-tuning dataset
- Training the model on phonemes rather than words
- Automated personalisation
- Advanced audio features
- Integrated analytics analysis
- Using the larger Parler-TTS model and fine-tuning using LoRA for feasibility.
- Video ads!

What makes AdCraft great

- All-in-one package
- Easy to get started
- Quick and easy A/B testing
- Highly flexible
- Highly customisable
- Fast
- Affordable!