SEGP Report
AI-powered web-based radio advertisement generation

March 2025

1 Introduction

1.1 Project Motivation and Objectives

In a world where audio advertising is incredibly powerful yet time-consuming to produce, small business owners and marketers may struggle to create engaging, on-brand audio ads when put under time constraints. Traditional ad creation usually involves many steps and multiple parties - there is a need to communicate with copywriters, voice actors, audio engineers and so on. As a result, this process can be slow and expensive, creating an entry barrier for small businesses to make as big of an impression on the market as already established businesses. To give a concrete example, producing a radio ad traditionally costs anywhere between £200 and £5000 for a 30-second ad.[1] Depending on revisions, scheduling, and recording studio bookings, the total time commitment required often takes several days to weeks.[2] In contrast, our AI-assisted platform can deliver an ad script plus generated audio in a few minutes - all within a single web interface. Even accounting for a platform subscription, the total costs are significantly lower than traditional ones, improving accessibility to this field.

Our goal was to streamline this workflow by developing an AI-assisted web-based tool that would help complete the following steps:

- Generate customised ad scripts based on a simple user prompt that includes product information, target audience, and marketing goals.
- Refine scripts with AI assistance, using a sophisticated validation system for controlled modifications.
- Allow for further personalisation through custom background music/sound effects, the ability to give feedback to our tool on specific parts of the script, and speaker consistency.
- Add audio directions alongside script content describing how the script should be read or sound like.
- Produce expressive, high-quality audio from these scripts.
- Track version history of both scripts and audio outputs for comparison and iteration.
- Provide an end-to-end workflow, including all the necessary steps in ad production, fast enough for same-day delivery at a fraction of the cost.

By fine-tuning an existing text-to-speech (TTS) model to incorporate emotional expressiveness into generated speech, our aim is to deliver expressive and natural sounding audio ads at a fraction of the traditional cost and time. We want to allow users to maintain creative control though targeted AI refinements, all while ensuring quality with our multi-layer validation system.

1.2 Contributions

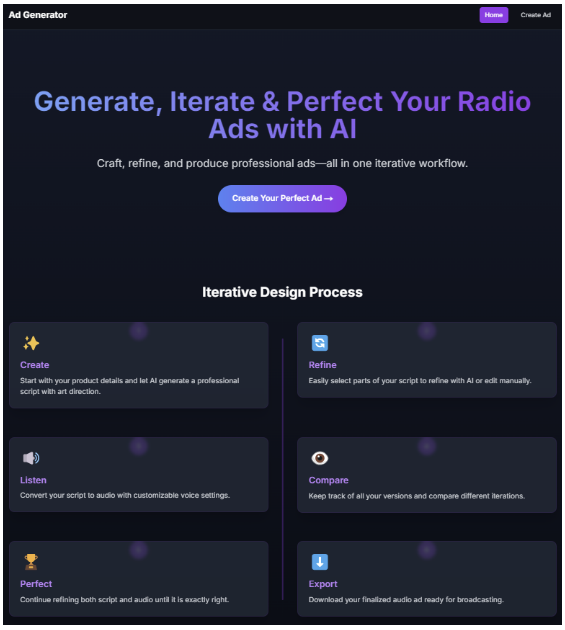
To allow our user to easily and freely specify how a script should be read, we chose to accept natural language text prompts as the audio description instead of forcing them to select features from a predefined list. To aid this, we chose to fine-tune the open-source TTS model Parler-TTS.

- **A Fine-Tuned TTS Model with Emotional Expressiveness** Parler-TTS has been pre-trained to extract fearures such as speaker gender, pitch and pace from free text. We fine-tuned this to accept emotion as an additional input feature, resulting in more varied and engaging speech output.

[1] `https://www.wondercraft.ai/blog/uk-radio-ad-costs-breakdown`

[2] `https://www.audioads.co.uk/faqs`

- **Prompt generation from datasets** As a part of the original Parler-TTS framework, we also used a library called DataSpeech, which enabled us to transform multi-column dataset rows into natural language prompts that Parler-TTS can interpret effectively.
- **Ensuring Speaker Consistency via SpeechBrain** Parler-TTS uses a random voice per generation, which results in successive generations of the same script having different voices. Using a voice classifier from SpeechBrain, we ensured that ads featuring more than one speaker maintain consistent speaker identities throughout. Additionally, this feature also allowed for the iterative timeline we implemented - this way, if a user wanted an additional line in the script, but to keep the voice the same for example, they would get consistent output.
- **User-Facing Web Application** Marketers can generate scripts, fine-tune them and produce final audio directly in the browser, with options like background music, trimming and script rephrasing.



Generate, Iterate & Perfect Your Radio Ads with AI

Craft, refine, and produce professional ads—all in one iterative workflow.

Iterative Design Process

Create

Start with your product details and let AI generate a professional script with art direction.

Refine

Easily select parts of your script to refine with AI or edit manually.

Listen

Convert your script to audio with customizable voice settings.

Compare

Keep track of all your versions and compare different iterations.

Perfect

Continue refining both script and audio until it is exactly right.

Export

Download your finalized audio ad ready for broadcasting.

Figure 1: Website homepage (partial)

2 Deliverables

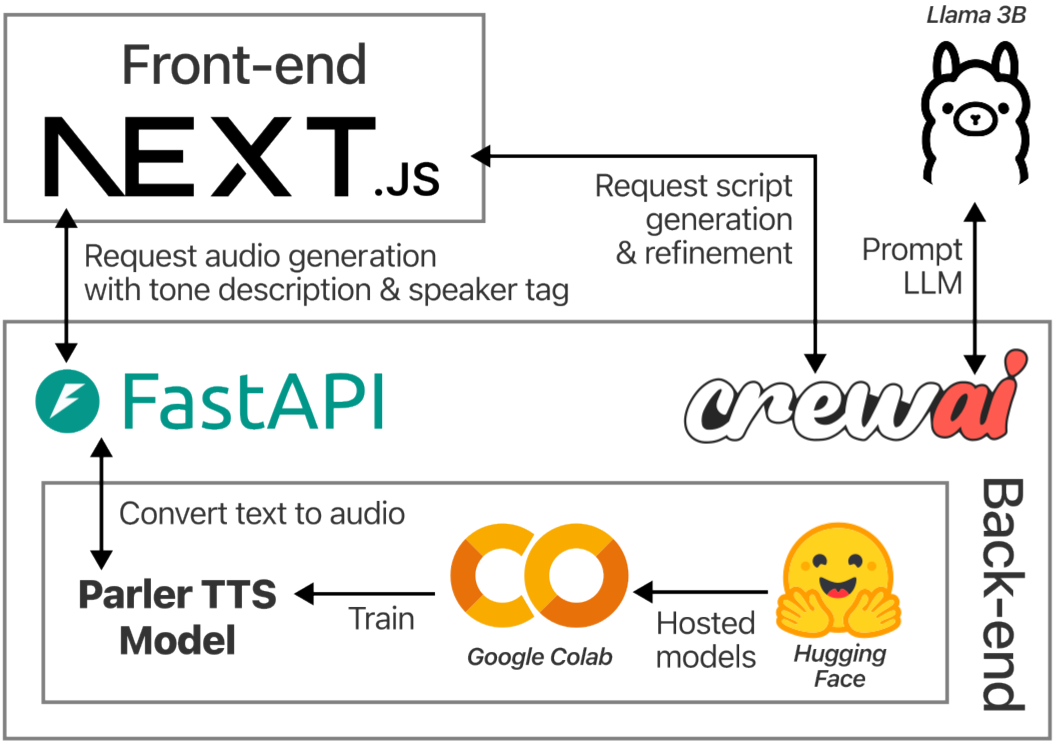
Our deployed webapp can be found here: `https://ad-craft-app.vercel.app/`

We initially worked in this repository: `https://github.com/SpeedyOrc-C/IC-SEGP-2025`, but our final web app was built in a different repository: `https://github.com/abhivir-42/marketing-app-ad-gen`
Our web-based demo walks users through the main features of the app, that were all desgined with our overarching goals in mind:

- **Input form:** Enter product details (name, audience, background information).
- **AI script generation:** Our back-end uses an LLM agent to curate initial scripts.
- **Refinement and Editing:** The user can highlight lines for rephrasing or regenerate the script after adjusting requirements.
- **TTS output:** The fine-tuned Parler-TTS synthesises speech, integrating emotion and flow, and ensuring consistent speaker identity.

2.1 System Architecture

The application follows a modern architecture with clear separation of concerns:



3 Evaluation

3.1 Quality and User Needs

We performed informal user tests with startup owners and marketers who found the interface intuitive and appreciated the ability to quickly generate multiple versions of a short radio ad. In further conversations with marketing agents, they recommended that we display all core features directly on the homepage for a smoother, more centralised workflow. They also suggested introducing the version control feature (for generated audio and scripts) and a built-in editing interface to add background music, sound effects and trimming. Based on this feedback, we refactored the UI to tailor to the suggestions.

Furthermore, we ran a small listening test where 20 volunteers rated clarity, emotion and appeal on a 1-5 scale. We made sure to select participants based on the target audience of each ad. The fine-tuned Parler-TTS often scored 4 in perceived naturalness, but due to generation consistency issues, it also often scored around 2. Our model performed well on ads tagged with the ”happy” emotion, and worse on ads tagged with the ”angry” or ”frustrated” emotion - this is likely a result to limited data in our datasets for these emotions.

The quality of our fine-tuning dataset is a limitation of this project for several reasons. These include:

- **Availability of emotional speech datasets:** Parler-TTS was originally trained on 45,000 hours of audiobook data, which focuses on clarity of speech and not on representing different emotions. Likewise, most audio datasets we found that are openly available were as such. This meant that we didn't have a good amount of data representing each emotion type needed for high-quality fine-tuning.
- **Insufficient speaker data:** The datasets we used often did not have a large enough amount of data per individual speaker, which meant that consistency in voice was negatively affected. Additionally, the data that we did have was often noisy and low quality, resulting in some generated audio being unintelligible.
- **Scalability for Longer Scripts:** Generating audio for scripts longer than 30 seconds increases runtime substantially. Additionally, the datasets that provided emotional speech used very short sentences which made it difficult for the model to generate audio for a longer input. Ideally, the fine-tuning dataset should have a mix of varying speech lengths.
- **Accents and Dialects:** In addition to emotion, because our training data did not include substantial accent/dialect samples, the system primarily supports standard U.S. English and a few other mainstream variants.

3.2 Legal, Social and Ethical Considerations

One of the most important and frequent concerns with AI-generated audio is the possibility of replacing human voice actors and production specialists. As TTS systems become increasingly lifelike and cost-effective, the demand for these traditional services may decline, leading to potential job losses or shifts in job nature.

However, it is important to note that human creativity remains valuable for tasks such as brand strategy, nuanced acting, and comedic or dramatic storytelling, which AI models struggle to replicate convincingly. Additionally, our goal with the platform is to encourage as much creative personalisation as possible, aiding the user in providing their own creative feedback and ideas. Furthermore, new opportunities might emerge around AI supervision, such as ensuring quality and ethical standards and working together with creative professionals instead of replacing them.

Moreover, there are concerns about data privacy and security. When users upload information about their company and the nature of their required ad, the information might contain sensitive or private details. Thus, there is a duty to protect this information. To ensure this, it is important to implement effective encryption that safeguards the data in transit, have data retention policies that specify how long user content is kept (with an option of permanent deletion on request), and finally have transparency around data usage. Consent must be obtained for any recorded voices used in fine-tuning and training, and proper anonymisation procedures for shared or user-contributed data that might contain personal identifiers must be put in place.

In general, it is important to disclose that the content has been AI-generated to avoid the risk of misleading customers and listeners, and maintain strict terms of service to ensure proper usage of the tool. It is also a good idea to involve human moderators for high-stakes ads (e.g., medical ads) to validate the authenticity and correctness of the content.

3.3 Technical Challenges and Solutions

1. Script Integrity During Refinement
Challenge: Ensuring AI refinements only modify user-selected sentences without affecting other content.
Solution: Implemented the multi-layer validation system that

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

2. AI Agent Orchestration
Challenge: Coordinating multiple specialised AI agents for cohesive output.
Solution: Utilised CrewAI framework to:

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

3. 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

4. Frontend-Backend Consistency
Challenge: Maintaining consistent state between frontend and backend.
Solution: Implemented:

- Clear API contracts with Pydantic models
- Redundant validation on both ends
- Detailed error handling and reporting
- Stateless architecture with complete request context

3.4 Model Evaluation

The loss function used for the model evaluation of Parler-TTS was cross-entropy loss. Since cross-entropy tracks how well the predicted tokens match the ground truth, a decreasing loss curve during training implies that the model is learning to generate increasingly accurate tokens that reconstruct the target audio. A low cross-entropy is necessary but not always sufficient for high perceived audio quality. While it strongly correlates with correct token prediction, subjective factors such as emotional expressiveness, prosody or timbre can benefit from additional evaluations. Monitoring cross-entropy loss over time also helps identify if the model might be overfitting. For instance, if the training loss is dropping but validation loss does not improve, it suggests that the model is memorising specific training examples rather than generalising. We ran an experiment by changing some hyperparameters to test if this was the case, and came to the conclusion that the model was in fact not overfitting.

To go beyond simple cross entropy, we also implemented Multi-Scale STFT loss. This approach measures how closely the spectral components of the synthesised audio match those of the target audio across multiple time-frequency scales, ensuring that both short term details (e.g., transients) and longer, sustained frequencies (e.g., formants) are well reconstructed. Unlike raw token-level objectives, Multi-Scale STFT loss directly addresses the perceptual quality of speech, making it particularly effective at improving naturalness.

We chose this loss function because it balances implementation practicality and strong perceptual gains. It is also well-established in vocoder and TTS pipelines—many state-of-the-art models employ a similar technique to achieve high-fidelity, human-like outputs. By focusing on spectral alignment rather than strict sample-level accuracy, Multi-Scale STFT loss preserves the timbral and prosodic nuances of speech while remaining robust to minor misalignments in time.

3.5 Future Directions

User testing of the finished webapp and the results from our model evaluation indicate that, while the system provides quick turnaround and good naturalness and expressiveness, there is room for improvement in emotional nuance and clarity.

We have identified the following as the steps for improvement:

- **Expanded dataset:** Acquire a larger labeled dataset representing diverse emotional states, accents and sufficient data for each speaker to improve prosody control.
- **Automated personalisation:** If a user provides a list of store locations across the country, it might be beneficial to automatically generate several versions of the same radio ad. Each version would be played in a different town/location, with the name of the location included in the ad.
- **Integrated analytics analysis:** Give the user the option to provide their analytics data on their ads, and analyse the different engagement patterns to figure out which versions of their ad perform better. With this approach, the system can curate ads with the maximisation of predicted performance in mind.