

139_spanish-voice-trainer

Project Summary:

Custom Spanish Voice TrainerA high-performance, completely private desktop application built on macOS to accelerate Spanish language training through automated flashcard creation and intelligent pronunciation analysis.The application utilizes a local-first architecture to ensure complete data privacy, storing all configurations, historical student analytics, and multimedia binary files strictly on the user's local drive.

The Core Technical Stack

- Environment & Package Management: uv (Rust-based Python package manager) for ultra-fast, isolated virtual environments and dependency locking.
- Database & Persistence: SQLite to map phrase metadata, local media file paths, and chronological user practice scores without external database infrastructure.
- Audio & Signal Processing: sounddevice for hands-free, voice-activated microphone capture; librosa and fastdtw (Dynamic Time Warping) to extract phoneme features (MFCCs) and score user pronunciation accuracy against reference files.
- Asset Generation: edge-tts to stream high-quality, neural text-to-speech Spanish audio clips, and Pillow (PIL) to auto-render widescreen flashcard JPEGs matching the phrases.
- User Interface: Developed in two phases—starting as a clean, text-based Command Line Interface (CLI) before migrating to a dual-mode desktop GUI built with PyQt6.

How the System Works

The application operates across two distinct, integrated operational frameworks managed via a unified QStackedWidget interface:

1. Content Creator Mode The user inputs a Spanish phrase and its English translation. The system automatically triggers the asset generator to output a custom high-quality flashcard image and a native-sounding neural audio file. The localized text strings and file paths are instantly committed to the SQLite database.
2. Student Training ModeThe system pulls cards from the database, displays the visual flashcard image, and plays the target Spanish pronunciation. The student speaks into their MacBook microphone. The system detects when the student begins and stops talking via a voice-activation threshold, records the sample, and runs a Dynamic Time Warping alignment algorithm to provide an objective pronunciation match score (e.g., 87% accuracy).
3.  Long-Term Capability: Custom Video Compilations Because all image assets share standard HD video dimensions (1280 X 720) and all audio samples are tracked deterministically in the database, the core engine can be commanded to interface with ffmpeg-python. It can seamlessly compile entire batches of database assets into standalone, continuous .mp4 video lessons complete with timed visual pauses and silent audio gaps, providing an additional passive learning medium for language immersion.