|  |
| --- |
|  |
| Sonetrix |
| Where Sound, Sight, and Words Converge to Create Immersive Experiences. |

|  |
| --- |
| Nourhene Hanana  13/04/2025 |

# Application purpose :

Unlock the potential of multisensory integration with a project that transforms audio input into dynamic, interactive experiences. With this tool, users can choose from three distinct tasks based on their audio recording:

1. **Task 1**: Transcribe the audio into the language in which it was originally spoken.
2. **Task 2**: Translate the audio into a selected target language, such as French or Italian. ( many other languages can be used but we can explore the limitations when detailing the models architecture)
3. **Task 3**: Generate a visual representation, producing an image that describes the content of the recording along with its translation.

This personal project explores the seamless fusion of audio, text, and image, creating a multi-modal solution that showcases the possibilities of combining these different forms of media.

# High Level Architecture:

The screen displays a form where the user can submit an audio recording in **MP3 format**, and choose a task to perform. Currently, the task is selected from a **dropdown menu**, but in future versions, it will be handled through a **prompt-based interface**.

Once a task is selected, the application receives both the audio recording and the user’s task request, and executes the corresponding operation by interacting with the appropriate models:

1. **Transcription**:  
   The **OpenAI Whisper** model will be used to transcribe the audio into text, preserving the original language of the recording.
2. **Translation**:  
   The user must select a **target language** from a dropdown list of 52 languages (this will later be inferred automatically from the user prompt).
   * First, Whisper will transcribe the audio and **detect the original language**.
   * Then, the **Facebook MBART model**, fine-tuned for enhanced performance on **English-French** and **English-Arabic** translations, will translate the transcribed text into the target language.
   * *Example*: If the input audio is in English and the user selects **French** as the target language, the output will be a **French translation** of the original English recording.
3. **Image Generation**:  
   An intermediate step uses Whisper to transcribe the audio into text. This transcribed text is then passed to **diffusion models** (such as those from **Dreamlike.art**) to generate a **fantasy-style image** that visually represents the content and meaning of the input audio.

Une image contenant texte, diagramme, Police, capture d’écran

Le contenu généré par l’IA peut être incorrect.

Figure Sonetrix High Level Architecture

# Whisper Open AI Model:

Paper: **Robust Speech Recognition via Large-Scale Weak Supervision Link (https://arxiv.org/abs/2212.04356)**

Une image contenant texte, capture d’écran, diagramme, Police

Le contenu généré par l’IA peut être incorrect.

* 1. What is Whisper ?

Whisper is an automatic speech recognition system developed by OpenAI. It’s a multilingual, multitask model that can:

* Transcribe speech to to text
* Detect the language of the spoken audio
* Translate speech into English from other languages.

It was released on September 21, 2022 with open source code and models on GitHub. It’s one of the most robust open-source ASR sysyetms, trained on 680K hours of multilingual and multitask supervised data collected from the web. The latter made Whisper trained with weak supervision.

* 1. Why Whisper is trained with Weak Supervision? ( Refers to Introduction in the Article)

ASR models used to work on unlabeled data scaling up to 1,000,000 hours of training ([Zhan et al.2021](https://www.researchgate.net/publication/349804353_Zhang_et_al_2021_Ear_Hearing)), but model failed to perform decoding when transforming the tokens back into human readbale language.

Working with labeling data such a very high quality data like librispeech specialized for ASR, they have closed the 5140 hours of training which still so far from 1000000 hours of unlabeled data training. To close the gab they obtained for weak supervision where they added different source of audio to text recognition labeled data that was in the sake of having a tradeoff between quality and quantity of the data.

OpenAI trained Whisper on **680,000 hours of audio data scraped from the web**, such as:

* YouTube videos
* Podcasts
* Publicly available datasets  
  And many of these came with **auto-generated or user-generated transcripts**

So the tsupervision is considered “weak” because, the transcriptions may not be 100% accurate, some translations may be noisy or approximate and some metadata ca be wrong. Despite this, Whisper worked on filering pipeline where ti turn raw audio plus the noisy text into useful and significant training pairs, they obtained for different robust stages of data pipelines to ensure data quality: (Refers to the Approach 2.1 Data Processing). Listing some of them:

Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.

* Validate the language:

Audio samples VoxLingual107 ([Valk et Alumae,2021: A dataset for Spoken Language Recognition](https://arxiv.org/abs/2011.12998)) were passed through a language detection process using Whisper prototype version as a language detection for audio and then compared to CLD2 which is Google compact language detector2 as language detection tool that identifies the language of a written text. So the process in detail would be:

Une image contenant texte, diagramme, capture d’écran, ligne

Le contenu généré par l’IA peut être incorrect.

* Avoid transcriptase ( machine generated transcripts)

Open Ai has developed heuristics to detect and remove machine generated transcripts from the training dataset as they output only a limited subset of written language which removes aspects that are difficult to predict from

* Error Analysis and De-duplication:

After training the initial model, Open AI conducted an additional filtering by analyzing the error rates across training data. High-error and large data sources were manually reviewed to identify **poorly aligned** or **misaligned transcripts**. They also applied **de-duplication** to remove **duplicate or near-duplicate** entries from the dataset. This additional step helped refine the training data for better model performance.