Software Requirement Specification Document for Audio Upscaling With text To speech Sound Synthesis

Mahmoud Mamdouh, Maged Magdy ,Omar Adel, Ahmed Mohamed
Supervised by: Dr. Diaa Salama , Eng. Nada Nofal

February 15, 2024

Table 1: Document version history

Version	Date	Reason for Change
1.0	2-december-2023	SRS First version's specifications are defined.
1.1	15-december-2023	New system modules added.
1.2	27-december-2023	Update Class Diagram.
1.3	12-January-2024	Finished Remaining Sections.

GitHub: `https://github.com/MahmoudMamdouh3/Audioupscale_svs_GraduationProject`

Contents

Abstract

Nowadays old audio, especially music is annoying to listen to in the advanced era we live in. Many deep learning methods and models have been proposed but none of them have completely solved the problem of delivering clear and high-fidelity audio to enhance the old one. We propose a different approach of applying a combination of audio upscaling, singing voice synthesis (SVS), Text-to-speech (TTS), sound editing, voice cloning, and Ai generated voices to enhance the audio while generating the missing data to produce high-fidelity audio. This document will briefly discuss the initial requirements and objectives of the system, similar systems, and the project's deliverables.

1 Introduction

1.1 Purpose of this document

The purpose of this document is to present and highlight the distinct requirements needed to develop our project's desired web application. This document illustrates the main elements required in the system's development process, including the functional and non-functional requirements. The intended audience for this document includes all stakeholders, from students to professors, as well as any developer who works on this project. We also provide future project users with a complete description of each processing stage, inputs, and outputs.

1.2 Scope of this document

This document demonstrates the flow of the system and its purpose. It also focuses on the functional and non-functional requirements, a detailed explanation of the data design, and the initial design of the UI.

1.3 Business Context

Our project proposal is to develop an audio enhancement technology that has various market applications. We have identified three potential market segments that can benefit from our solution: the music industry, the movie industry, and the gaming industry. The music industry is the most obvious market segment that can benefit from our audio enhancement technology. Our solution can improve the sound quality of music recordings, which can attract audiophiles and enthusiasts who are willing to pay a premium for high-quality sound. Additionally, our product can help music producers and sound engineers better understand the nuances of audio mixing and mastering, improving their creative output. The movie industry is another market segment that can benefit from our audio enhancement technology. Our solution can improve the sound quality of movie soundtracks, resulting in a more immersive cinema experience for viewers. Additionally, your product can help movie producers and sound designers create more realistic and engaging soundscapes that enhance the overall cinematic experience. The gaming industry is a rapidly growing market segment that can benefit from our audio enhancement technology. Our solution can improve the sound quality of gaming audio, making the gaming experience more immersive and engaging for players. The virtual sound added to the digital games can make the players feel more connected with the surroundings and vibes, making the gaming experience more realistic.

2 Similar Systems

2.1 Academic

As noted in Audio Super Resolution in the Spectral Domain [1], the main concern is that the majority of deep learning-based techniques for producing audio in the speech and music domains only operate with low-resolution target audio. Moreover, audio signals are extremely long, especially when they are of high quality, which makes modeling them computationally costly. The paper noted that they will present a technique for producing high-frequency content in the spectral domain by combining a set of reconstruction, adversarial, and feature losses that function on the spectral and time representations of the signal with a convolutional U-Net model that only operates in the frequency domain. Furthermore, they introduced a way to up-sample a signal in the spectral domain that avoids concatenation between existing and generated frequency bands. Additionally, they presented a method for spectral domain up-sampling of a signal that circumvents concatenation between produced and existing frequency bands. Their suggested approach has computational limitations. Their proposed method had limitations in terms of computational complexity

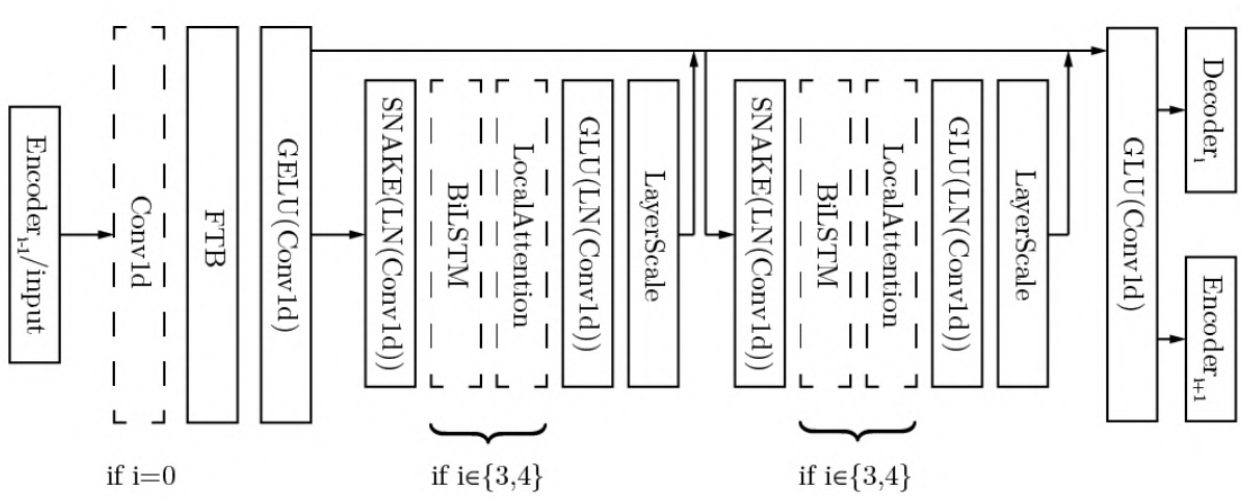

Figure 1: Encoder layer. **Aero**

and memory requirements, which made it challenging to implement in real-time and streaming processing applications. Additionally, the proposed method is not suitable for all types of audio signals.
Results: The results suggest that the proposed method is superior to the evaluated baselines considering both objective and subjective metrics.
Dataset: The dataset used in this paper is the "VCTK" dataset which contains around 44 hours of speech from 110 speakers, sampled at 48 kHz.

Target Sound Extraction with Variable Cross-modality Clues: In this paper **SoundExt**, they noted that automated target sound extraction (TSE) often uses a model prepared on a fixed structure of targeted sound clues, such as a sound class label, which limits how users can interact with the model to specify the target sounds they want. On top of that, a single clue may be insufficient to represent a specific targeted sound. To overcome this, the authors of the paper suggested a new approach to automated target sound extraction that leverages numerous clues across different modalities to improve performance and robustness. The authors designed a unified TSE system that can extract the target sound by flexibly combining multiple

clues from different modalities available at test time. The paper also seeks to address the challenges of processing clues of different modalities, dealing with the alignment between the clues and the input audio features, and handling various numbers of input clues.

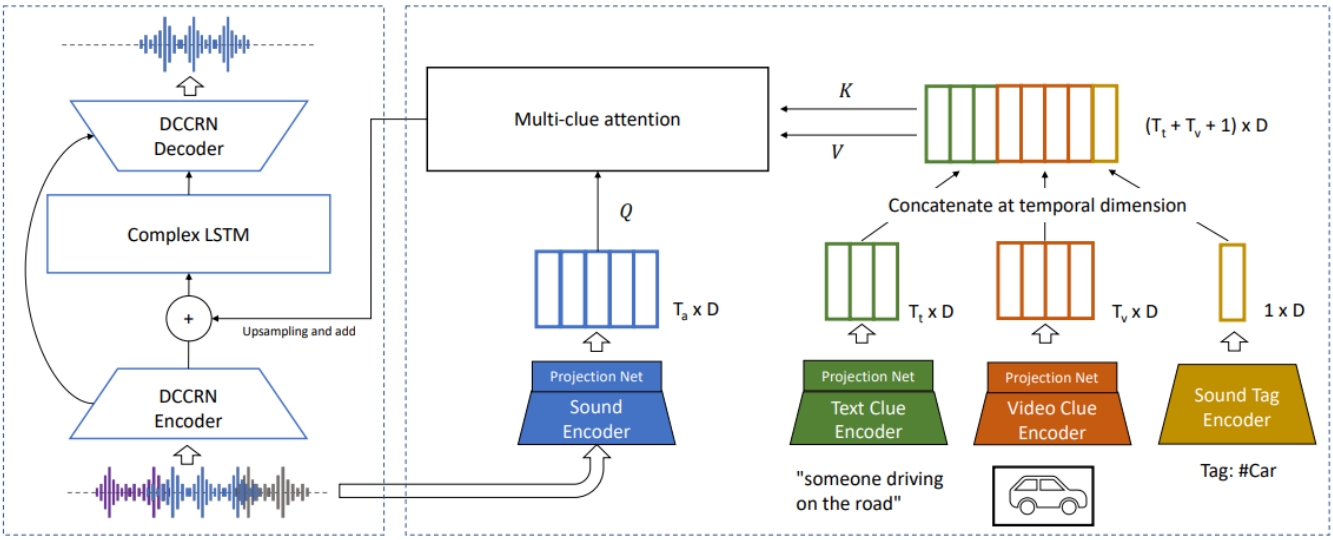

Figure 2: Multi-clue target sound extraction. **SoundExt**

They had a great disadvantage in that the proposed approach relied on the availability and accuracy of multiple clues from different modalities, which may not always be feasible or reliable in real-world scenarios. Additionally, the proposed approach required significant computational resources and training data to achieve optimal performance.

Results: They concluded that their proposed system had the highest SNRi score for both seen and unseen test sets when it operated with all of the clues. Even with just two added clues, the multi-clue model surpassed all the single-clue suggested models. However, the proposed model performed almost the same as with the single-cue baselines for the seen test set when only one clue was supplied, although it significantly performed better for the unseen test set. The paper also shows that the proposed system is robust to compromised clues, where one or two of the input clues were artificially compromised.

Dataset: They first created a dataset for their model which was based on "AudioSet", which is a large-scale audio dataset drawn from YouTube videos, and it has 527 sound classes labeled by humans. Most of the data in AudioSet are 10-second video clips with soundtrack.

ADVERSARIAL AUDIO SYNTHESIS: The authors in this paper **GAN**, noticed that huge databases of sound effects are explored by musicians and artists to find certain audio recordings suitable for specific scenarios they need. This procedure is painstaking and may result in a negative outcome if the ideal sound effect that they need does not exist in the library. So, they introduce WaveGan, which is a new approach to unsupervised synthesis of raw-waveform audio, and to evaluate its performance in synthesizing audio from various domains, including animal sounds, vocals, and different instruments like piano and drums. The authors also compare their model to other methods of audio synthesis and provide a formula for modifying other synthesizing methods of images to operate on waveforms.

The main drawback was that it can produce audio that is not always perceptually convincing, and a huge amount of data to train is required to accomplish good results. Additionally, the inception scores for WaveGAN are weaker than those of other methods.

Results: The results of this paper show that WaveGAN is capable of synthesizing audio from various domains, including speech, drums, bird vocalizations, and piano. The authors compare WaveGAN to other methods of audio synthesis and find that it outperforms some methods and is comparable to others. They also provide a formula for modifying other image generation methods to operate on waveforms.
Dataset: The dataset employed in this paper is the "Speech Commands Dataset," containing recordings of individual words by multiple speakers under uncontrolled recording conditions. The study specifically focused on a subset within this dataset, which encompasses spoken digits from "zero" through "nine." This subset is referred to as the "Speech Commands Zero Through Nine."

Singing Voice Enhancement in Monaural Music Signals Based on Two-stage Harmonic/Percussive Sound Separation on Multiple Resolution Spectrograms: In this research paper **6646221**, the authors introduce an innovative approach to enhance singing voices in monaural music audio signals, a particularly challenging task. The focus of the study is on capturing the fluctuations in a singing voice, and the authors address this challenge by detecting these fluctuations through the analysis of two spectrograms with distinct resolutions. One spectrogram emphasizes temporal details at the expense of frequency resolution, while the other prioritizes frequency resolution at the cost of temporal details. The authors leverage the differing shapes of fluctuating components in these two spectrograms. This concept leads to the proposal of a novel singing voice enhancement technique referred to as two-stage harmonic/percussive sound separation (HPSS) as shown in 3.

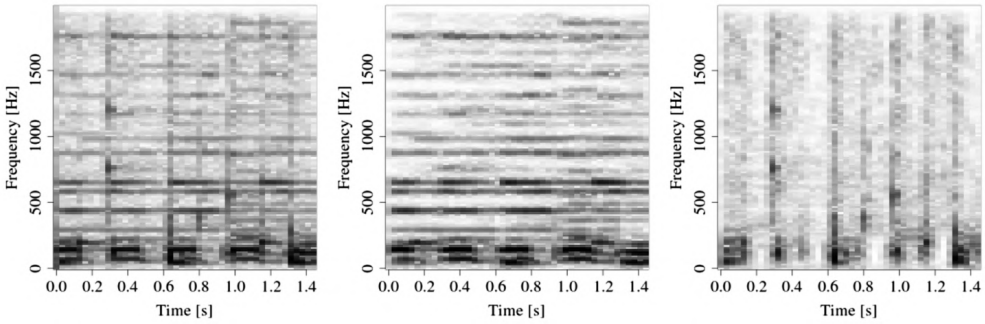

Figure 3: two-stage harmonic/percussive sound separation (HPSS)

Results: The results of the experiments indicate a notable improvement of approximately 4 dB in the Signal-to-Distortion Ratio (SDR), a widely-used criterion for this task. This improvement surpasses that achieved by existing methods. Furthermore, the authors assessed the method's performance as a preprocessing step for melody estimation in music. The experimental outcomes demonstrate a significant enhancement in the performance of a basic pitch estimation technique when utilizing the proposed singing voice enhancement method. These findings provide strong evidence supporting the effectiveness of the proposed approach.

Dataset: The research leveraged two primary datasets: the MIR-1K dataset and the LabROSA dataset. The MIR-1K dataset, specifically tailored for singing voice separation, comprises 1000 song snippets featuring music accompaniment and singing voice recorded as left and right channels, respectively. This dataset includes manually annotated pitch contours in semitone, indices and types for unvoiced frames, lyrics, and vocal/non-vocal segments. Additionally, it integrates speech recordings of the lyrics by the same individual who performed the songs. Each snippet's duration varies from 4 to 13 seconds, contributing to a total dataset length of 133 minutes. These snippets are extracted from 110 karaoke songs, selected freely from a pool of 5000 Chinese pop songs. The performances involve researchers from the MIR lab, consisting of 8

females and 11 males, most of whom lack professional music training. The MIR-1K dataset was utilized for a comprehensive assessment of singing voice enhancement on a large scale. Simultaneously, the LabROSA dataset was employed to evaluate the method as a preprocessing step for audio melody extraction. From the LabROSA dataset, nine out of thirteen pieces were chosen, considering the condition that the melody is sung, while the remaining four pieces were excluded due to their melodies being performed by instruments. All the data used in the study were in monaural format, with a sample rate set at 16 kHz.

Fusion methods for speech enhancement and audio source separation: In this paper **7451223**, the authors introduce a versatile fusion framework that has demonstrated significant potential in classification tasks. This framework leverages the diversity of available separation techniques to enhance the quality of separation. The authors achieve new source estimates by combining individual estimates from various separation techniques and assigning weights through a set of fusion coefficients. Three alternative fusion methods are explored, involving standard non-linear optimization, Bayesian model averaging, and deep neural networks. The authors apply these methods to both speech enhancement and singing voice extraction.

Results: The experiments conducted for both speech enhancement and singing voice extraction reveal that all the suggested methods surpass conventional model selection. Specifically, employing deep neural networks for estimating time-varying coefficients results in significant quality enhancements, with improvements of up to 3 dB in signal-to-distortion ratio (SDR) when compared to model selection.

Dataset: For the speech enhancement experiments, the authors utilized the second CHiME challenge corpus, comprising speech utterances from 34 speakers mixed with real domestic environment noise. The data was divided into four distinct datasets: a clean training set with 500 utterances in clean conditions for each speaker, a training set with 600 utterances mixed with background noise at six different Signal-to-Noise Ratios (SNRs), a validation set with 300 utterances similarly mixed, and a test set with 300 utterances mixed at six different SNRs.

In the singing voice experiment, aiming to separate the main voice signal from musical accompaniment, the authors employed a Music dataset gathered from the ccMixter community music remixing website. This dataset included 49 full-length stereo tracks spanning various musical genres. The tracks were randomly divided into five groups for cross-validation, and each of the 49 tracks was segmented into non-overlapping chunks lasting 20 to 30 seconds, resulting in a total of 308 excerpts.

Time-Frequency Filter Bank: A Simple Approach for Audio and Music Separation: In this study**8063868**, the authors explore the challenge of isolating the human voice from a mixture of vocals and sounds from various musical instruments. The human voice may manifest as singing in a song or as part of a news broadcast accompanied by background music. The paper introduces a generalized approach based on Short Time Fourier Transform (STFT), enhanced by a filter bank, to extract vocals from the musical backdrop. The primary objective is to devise a filter bank that minimizes background aliasing errors, employing approximated scaling factors for optimal reconstruction conditions. The experiments utilize stereo signals in the time-frequency domain. The input stereo signals undergo frame-based processing through the proposed STFT-based technique as shown in 4. The STFT-based output is then subjected to the filter bank to reduce background aliasing errors. The reconstruction involves applying an inverse STFT followed by the OverLap-Add method, resulting in the final output containing only the vocals.

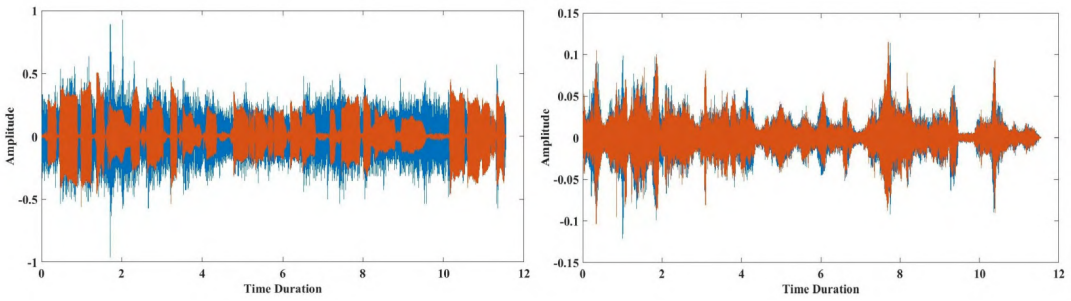

Figure 4: Input/output stereo signal

Results: The experiments show that the proposed approach performs better than the other state-of-the-art approaches, in terms of Signal-to-Interference Ratio (SIR) and Signal-to-Distortion Ratio (SDR), respectively, However The method might exhibit diminished effectiveness in scenarios where musical instruments are predominantly positioned near the center, where the vocals are located. Consequently, the background music may overpower the singing voice. In such instances, interference between the music and vocals may occur, and complete elimination of the background music might not be achievable.

Dataset: The study utilized the TIMIT dataset, encompassing recordings from 630 speakers across eight dialects of American English. Each speaker read ten phonetically rich sentences, and the dataset is supplemented with transcriptions at both word and phone levels for the spoken content.

Additionally, the MIR-1K dataset played a crucial role, comprising 1000 song clips featuring music accompaniment and singing voices recorded as left and right channels. This dataset includes manual annotations of pitch contours in semitones, indices, and types for unvoiced frames, as well as lyrics and vocal/non-vocal segments. Moreover, it incorporates speech recordings of the lyrics by the same individual who performed the songs. The duration of each clip ranges from 4 to 13 seconds, contributing to a total dataset length of 133 minutes. These clips are derived from 110 karaoke songs, selected freely from a pool of 5000 Chinese pop songs. The performances involved researchers from the MIR lab, comprising 8 females and 11 males, most of whom lack professional music training. This dataset was employed for experimentation and consisted of artificially generated combinations of songs.

Audio super-resolution using neural network :

problem statement: the problem this paper is trying to solve is the bandwidth extension. The purpose of the project is to construct a high-quality audio from a low-quality. **Research Objective:** The main objective of this paper is to explore new lightweight modeling algorithms for audio, specially for the bandwidth extension problem and this will be done by increasing the sampling rate of audio signals using deep convolutional neural networks.

Data Sets: there are 3 data sets used in this research paper and they are as listed below;

- VCTK dataset, this contains recordings from 109 native english speakers with more than one accent.
- MUSHRA dataset, this is a collection of audio samples that has been upsampled using different techniques.
- The Piano Dataset, contains recording of piano notes and this is mainly used to evaluate the method on non-vocal data **Methodology:**
- Authors use a deep convolutional neural network for audio super-resolution.
- A pre-trained VGG network extracts features from audio signals, incorporated into the super-resolution network.

- Perceptual loss function optimizes the network for visually and audibly similar outputs to the original.
- Evaluation on single and multi-speaker tasks using VCTK dataset shows improvements in SNR and LSD over baselines.
- Human evaluation study with MUSHRA dataset indicates their method outperforms baselines in perceived audio quality.
- Evaluation on Piano dataset reveals state-of-the-art performance achieved by their method.

2.2 Business Applications

1. Media.io: Media.io has the capability to rapidly and efficiently diminish and clean audio noise from various file formats, including MP3, VMA, and others. Moreover, it can effectively handle background noise like static, traffic, weather, or fan sounds present in the audio file.

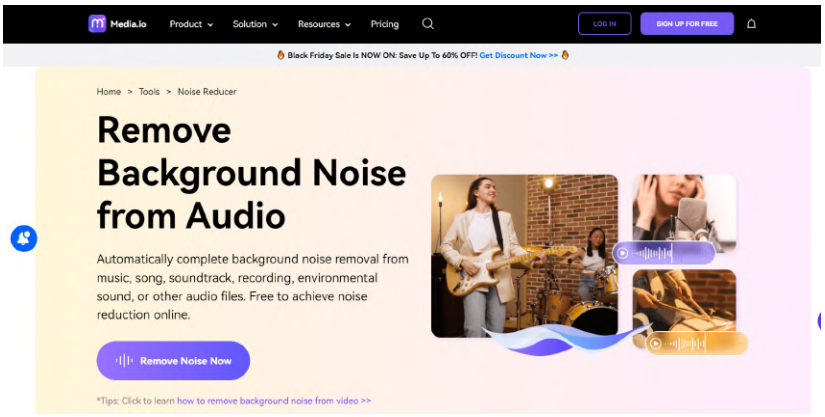

Figure 5: media

2. neural.love: Neural.love has the capability to enhance audio quality seamlessly without requiring the installation of additional software, utilizing advanced AI technology. The sophisticated neural network employed by the platform enables the augmentation of audio sample rates, reaching up to 48 kHz. This technology can refine voice clarity, eliminate background noise from recordings, and selectively remove vocals or human voices, leaving behind only music, ambient sounds, nature sounds, and more.

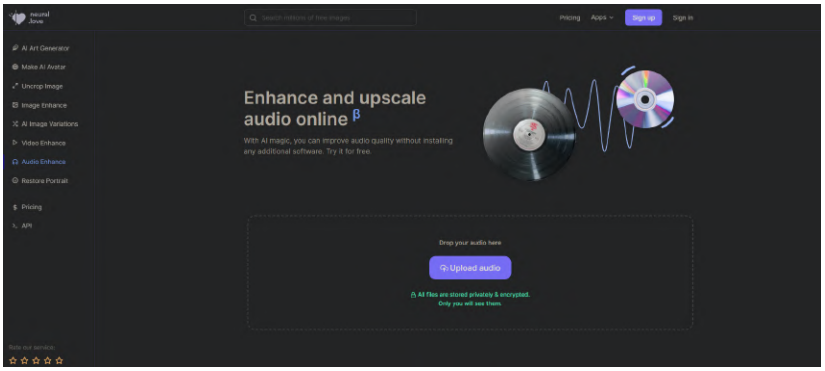

Figure 6: neural.love

3. Dolby.io:

   Dolby.io provides a highly advanced audio processing API, featuring the Enhance API to ensure your recordings have a consistent and accurate tonal quality. This API automatically takes care of tasks such as eliminating background noise and hums, maintaining consistent loudness levels, and minimizing undesirable sounds. Notably, there's no requirement to separate the audio from your video. The platform also incorporates Adaptive speech isolation. All you need to do is send your media file, and Dolby.io will return the enhanced audio within the same file type.

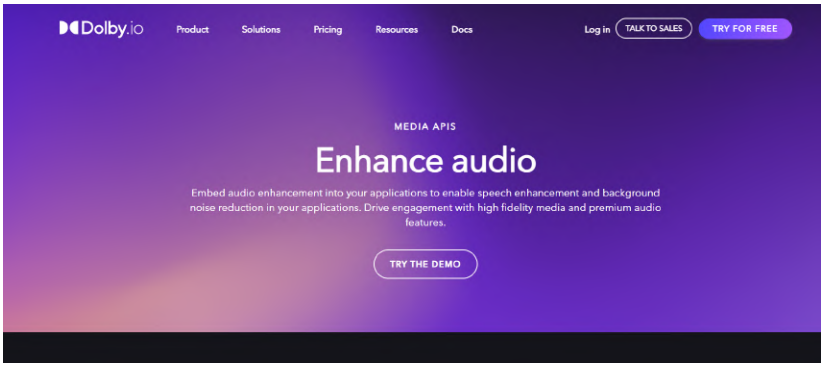

Figure 7: Dolby.io

3 System Description

3.1 Problem Statement

The main problem that a lot of people face is that some audios may have missing bits due to poor recording equipment in the past or corruption from faulty drives, which results in producing noise and unclear sound. Therefore, the project's main focus is generating those missing bits to produce a higher-quality audio.

3.2 System Overview

The user will input an audio file that needs to be enhanced. Next, the system will separate the vocals from the instruments, then reduce the noise and adjust the frequency to start operating on better input. Afterward, the slightly improved vocal will undergo various audio-improving techniques including singing voice-synthesis, text-to-speech, and upscaling models to adjust higher frequency while generating a clone of the singer's voice to greatly improve and produce high-fidelity vocal. Lastly, the system will combine the vocal with the slightly improved, earlier separated instruments (with neglected noise), to finalize the system's work and deliver pure audio back to the user to enjoy it.

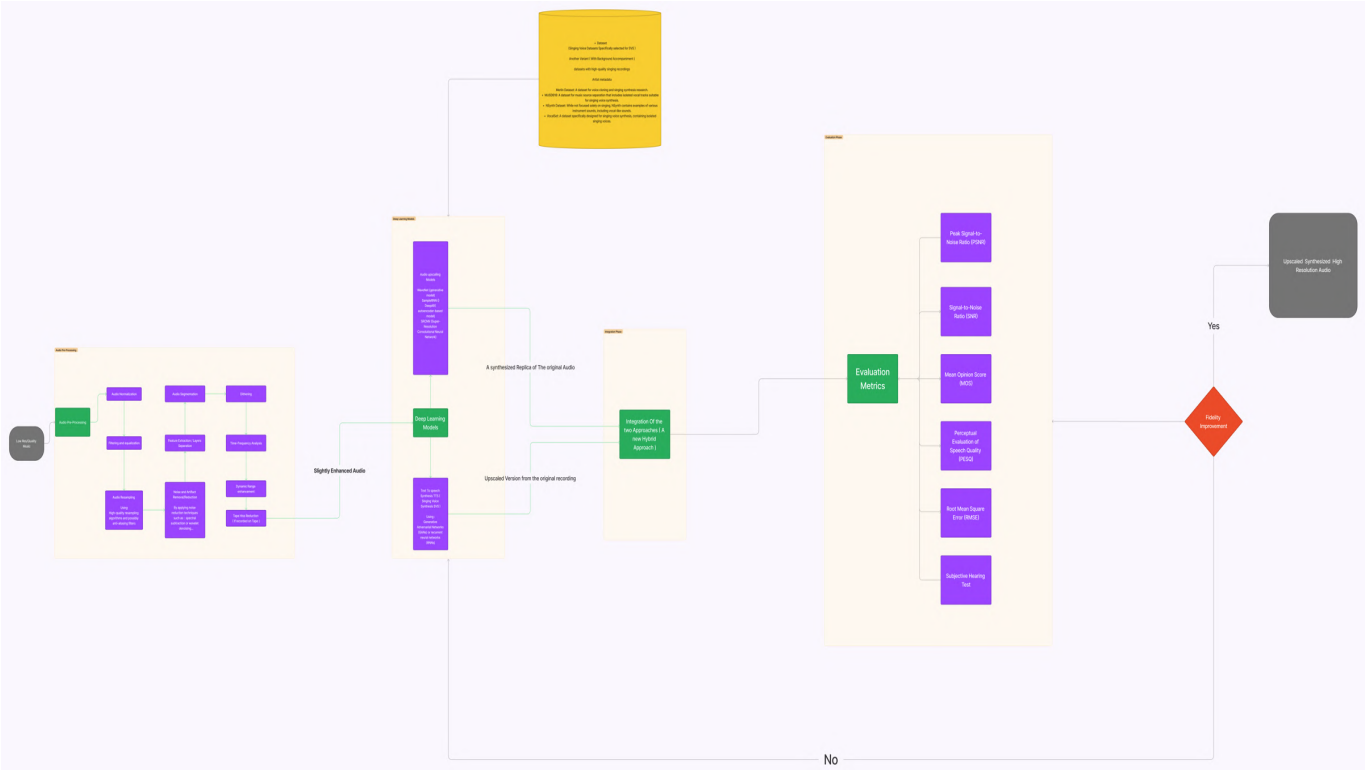

Figure 8: System Overview

https://drive.google.com/drive/folders/1hA-VW0mrusrycgER9JA73cRjm-x1RPVY?usp=sharing

3.3 System Scope

- The project aims to help music enthusiasts improve their old audio frequency
- The system will use audio enhancement techniques and algorithms to boost the quality and clarity of sound signals
- The system will be trained on Arabic songs of various quality levels, using relatively new models such as TTs and SVS
- The system will minimize the processing time and deliver the enhanced audio in no time

- The system's main scope is to upsample the low-quality audio files and generate quality-boosted audio using different models trained on large datasets of different singers

3.4 System Context

As shown in figure 9 ,The application will start by requesting the user to identify themselves, Then They will upload the audio file. On the other hand, The system will perform pre-processing techniques on the input audio file then it will employ audio upscaling using deep learning models, lastly output the upscaled and synthesized file. the user will also have the option to put text as input , after the application of pre-processing techniques, apply text to speech to that input file using deep learning models. Additionally, any adjustments made to the web app by the admin will be properly carried out.

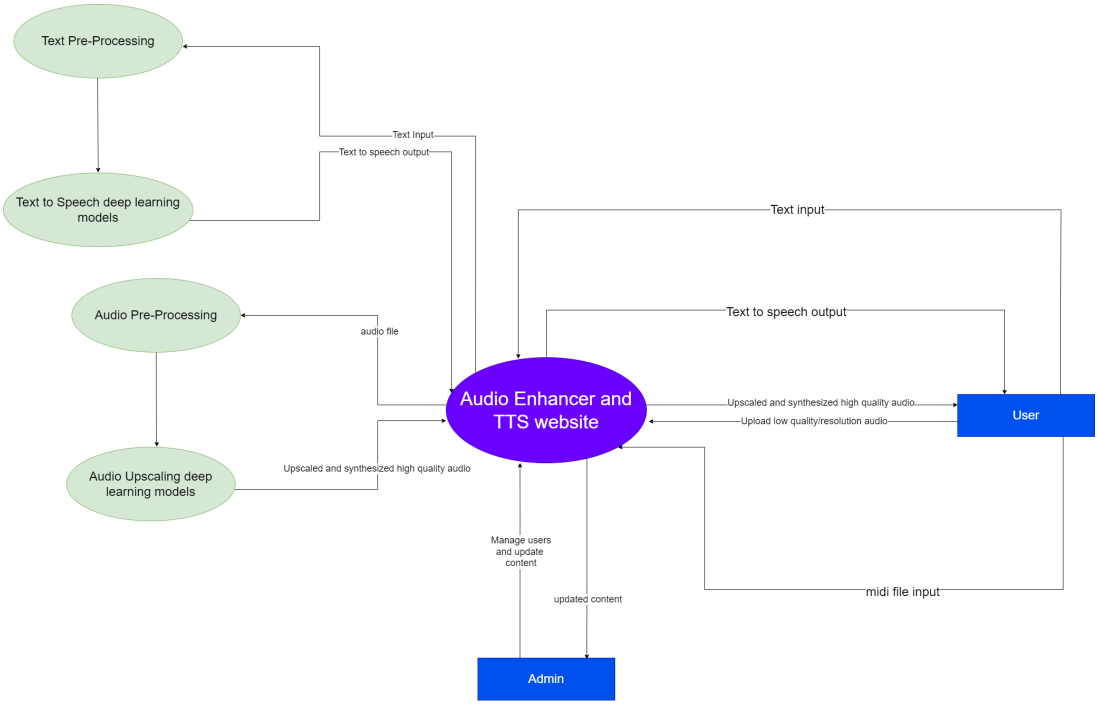

Figure 9: System context

3.5 Objectives

- Developing a high-quality replication of the vocals from the original recording or audio file is achieved through advanced speech technology methodologies, specifically employing singing voice synthesis and text-to-speech sound synthesis. This process ensures a natural and superior quality resemblance.
- Utilizing upscaling technology, commonly referred to as Super Resolution, enhances the bit-depth and sample rate of audio files and recordings. This technique contributes to an elevated audio experience.
- Enhancing the overall quality and perceived quality of user-uploaded audio files is achieved through the integration of generative AI and deep learning technologies and algorithms. This includes upscaling (Super Resolution), text-to-speech sound synthesis (TTS), and singing voice synthesis (SVS).
- Innovative configurations and parameter adjustments are introduced to facilitate the harnessing of transfer learning between the two technologies, ultimately leading to heightened accuracy in audio replication.

3.6 User Characteristics

- The user shall be familiar with common audio file formats (e.g., MP3, WAV) and an understanding of audio quality concepts.
- The user should be able to upload audio files to the website for upscaling and employing the text to speech and download the enhanced files afterward.
- The user should have a stable internet connection to upload and download audio files efficiently.
- The user should have access to a device with a web browser that is compatible with our website.

4 Functional Requirements

4.1 System Functions

The below use case diagrams demonstrates the system functional requirements. The system is composed of three user types: Customer, admin, and System.

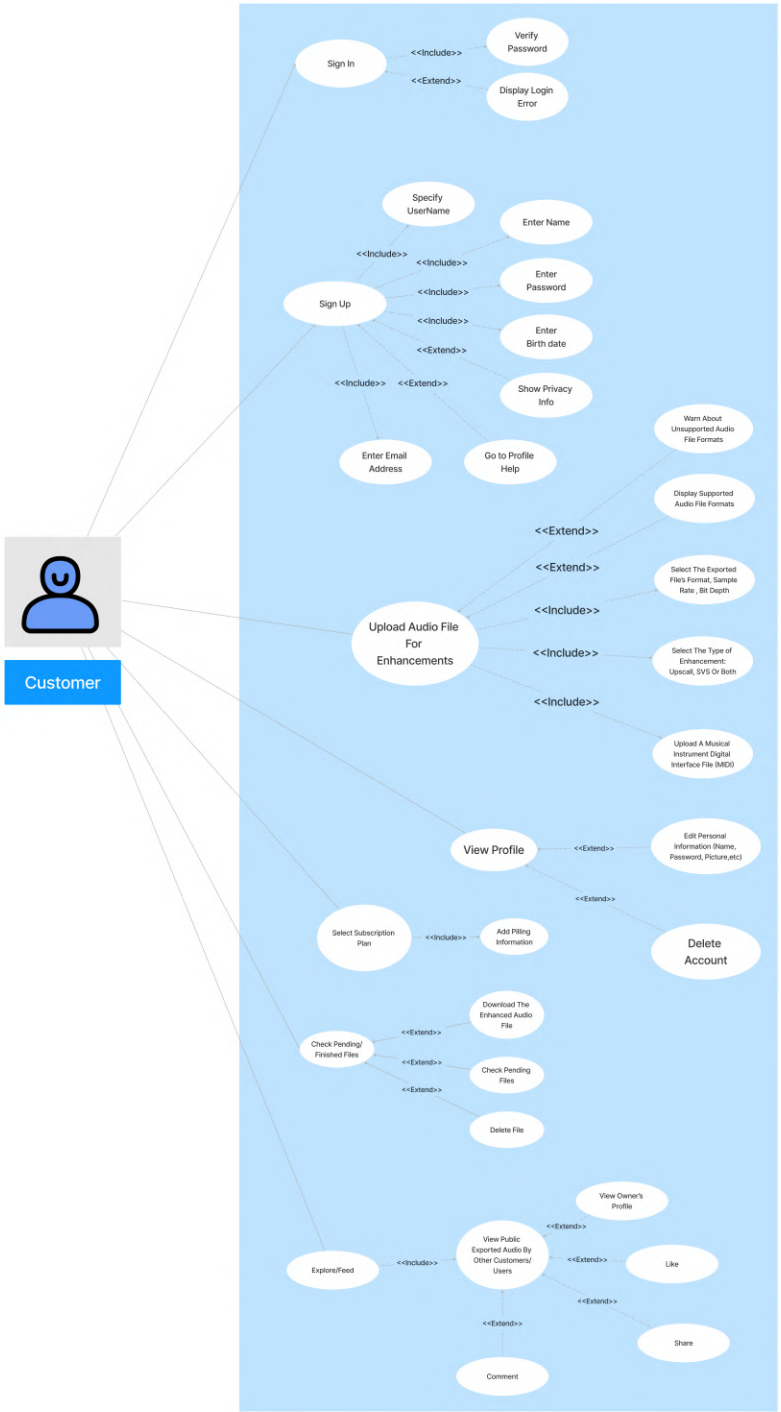

Figure 10: Use Case Diagram (User)

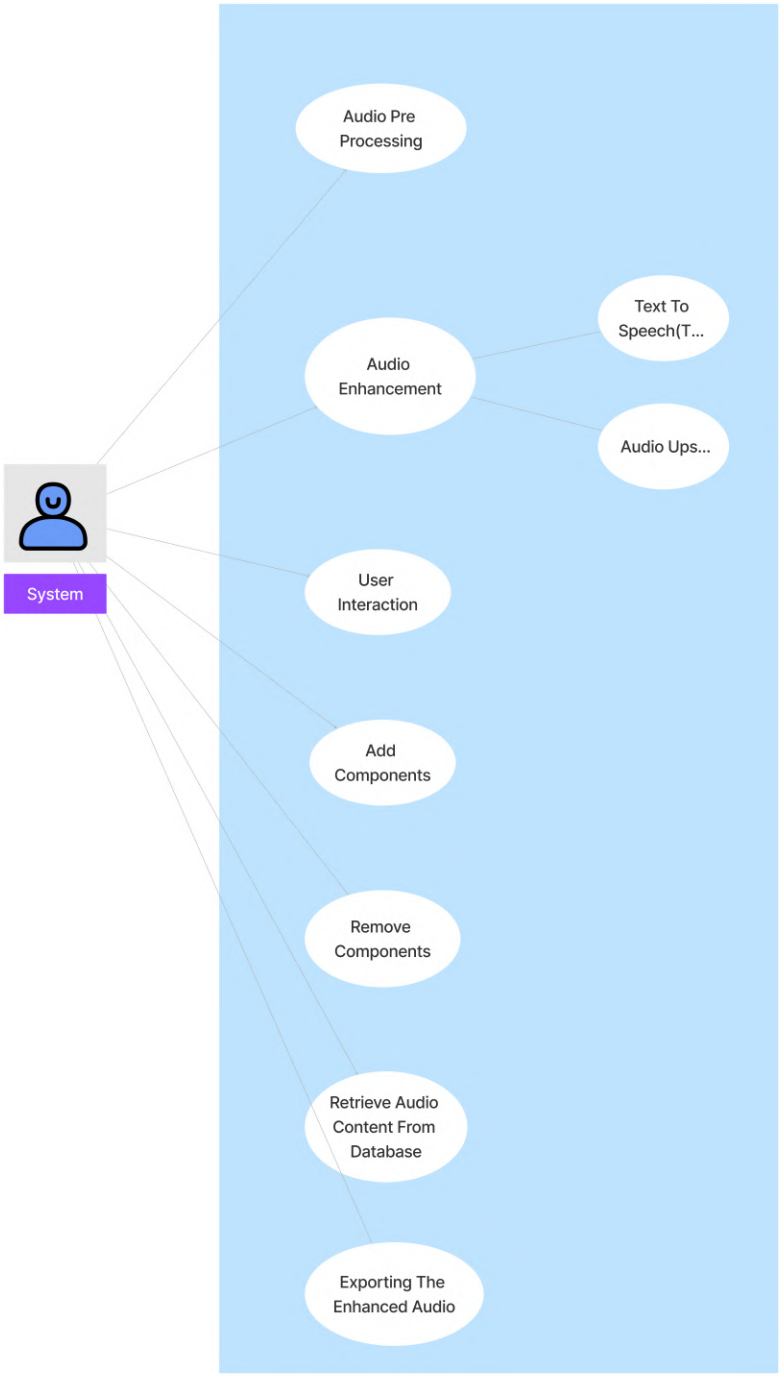

Figure 11: Use Case Diagram (System)

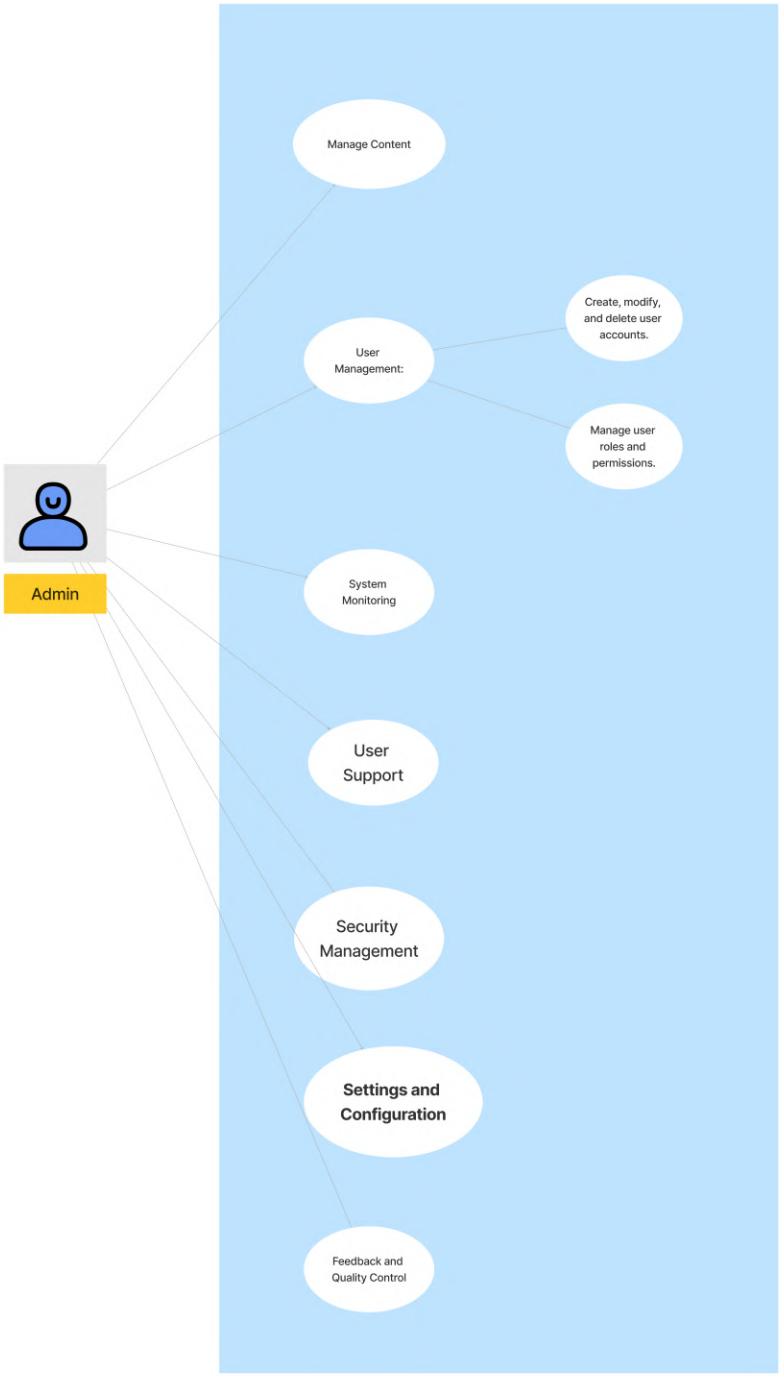

Figure 12: Use Case Diagram (Admin)

4.2 Detailed Functional Specification

Table 2: uploadAudioFile Function Description

Name	uploadAudioFile
Code	UAF-001
Priority	High
Critical	Yes
Description	Allows users to upload audio files
Input	Audio file (upload)
Output	Success/Failure
Precondition	-The customer must be logged in -The customer must specify the file type -The customer must specify the type of audio enhancements to be made -The customer must upload aswell a MIDI File
Post-condition	The User Will Receive A Message indicating the success or failure of the Upload
Dependency	None
Risk	Loss of uploaded data

Table 3: audioEnhancement Function Description

Name	audioEnhancement
Code	AE-001
Priority	High
Critical	Yes
Description	Enhances audio files using superresolution and singing voice synthesis
Input	Audio file
Output	Enhanced audio File integrated with synthesised Vocal Replica
Precondition	The system must have access to the audio file
Post-condition	Enhanced audio is generated with applied superresolution and singing voice synthesis
Dependency	None
Risk	-Loss of audio quality during enhancement process, -Generating Artifacts and Hallucinations

Table 4: audioPreProcessing Function Description

Name	audioPreProcessing
Code	APP-001
Priority	Medium
Critical	No
Description	Pre-processes audio files for further enhancements
Input	Audio file
Output	Pre-processed audio
Precondition	- The system must have access to the audio file
Post-condition	Audio file is pre-processed and ready for further enhancements
Dependency	None
Risk	Loss of audio quality during pre-processing, Incompatibility with certain audio formats

Table 5: Configuration Function Description

Name	configuration
Code	CFG-001
Priority	High
Critical	Yes
Description	Allows the administrator to configure the system and its components
Input	Configuration settings
Output	Updated system configuration
Precondition	- The administrator must have the necessary permissions
Post-condition	System and components are configured according to the administrator's settings
Dependency	None
Risk	Misconfiguration leading to system instability, Unauthorized access to configuration settings

5 Design Constraints

Dataset limitations: if the artist has little data to train on If the artist has no variation in singing styles If there is no variation in the quality of the songs, this will lead to a difficult training process

language restrictions : If the artist sings in any different language that the tts or upscaling models is not trained on like Arabic, Spanish, etc. Data loss: due to the enhancement process used in our project, one of the steps of enhancement is to separate the audio into layers, therefore, the quality of each layer will decrease especially the musical instrument. Expensive computations will also limit our project as we need powerful GPU and CPU to compute different models of TTS and SVS.

5.1 Standards Compliance

large audio files that exceed the limit of the processing ability may cause computation errors. the user should have a stable internet connection as the processing will take a few minutes to generate an output audio file.

5.2 Hardware Limitations

as the processing of the model will need a powerful device so this will limit the computation phase if there are devices with powerful CPUs and GPUs. An audio interface is also needed for large audio file processing. Monitor speakers also will be a restriction because it will be useful in subjective testing.

5.3 Other Constraints as appropriate

lack of generated audio quality varieties as the system does not support different qualities with extra charge

6 Non-functional Requirements

- Usability : Having a user-friendly view is essential to make our user comfortable enough to use the app. Users shall find it easy to use the interface; the functionalities will not be hard which will take the users no time to understand it .
- Maintainability : the system should be sustainable , which means it can evolve to meet the changing needs .
- Availability : The system will be a web application that can work on any browser. An internet connection is the only thing required for the users to reach any tools within the system.
- Portability: The web application shall be compatible with major web browsers (e.g., Chrome, Firefox) and support multiple operating systems, in order to ensure that it is accessible to a broad user base.
- Reliability : The system shall have strong error handling and recovery mechanisms to reduce the impact of software failures and unexpected issues.
- Security: The system shall enforce secure authentication mechanisms, ensuring that user credentials are stored securely. Additionally, role-based access control shall be employed to restrict access to sensitive functionalities based on user roles and responsibilities.

7 Data Design

7.1 Dataset

The VCTK dataset contains 110 English speakers with different accents. Each speaker recorded about 400 sentences. The training and testing files of the dataset are divided into clean and noisy each for speech enhancement methods. All of the files are saved in WAV format which is about 11GB in size, and text files containing information about their age, gender, accents, region and the words they were recording where included. In addition to that, another text file was found in a Github repository **silent-label**, which contained all the silent parts to be removed in each audio.

7.2 Database

All data involved with the user's information and data of the models will be saved and stored in the database. The relationships and the structure of the data are shown in figures 13 and 14 below.

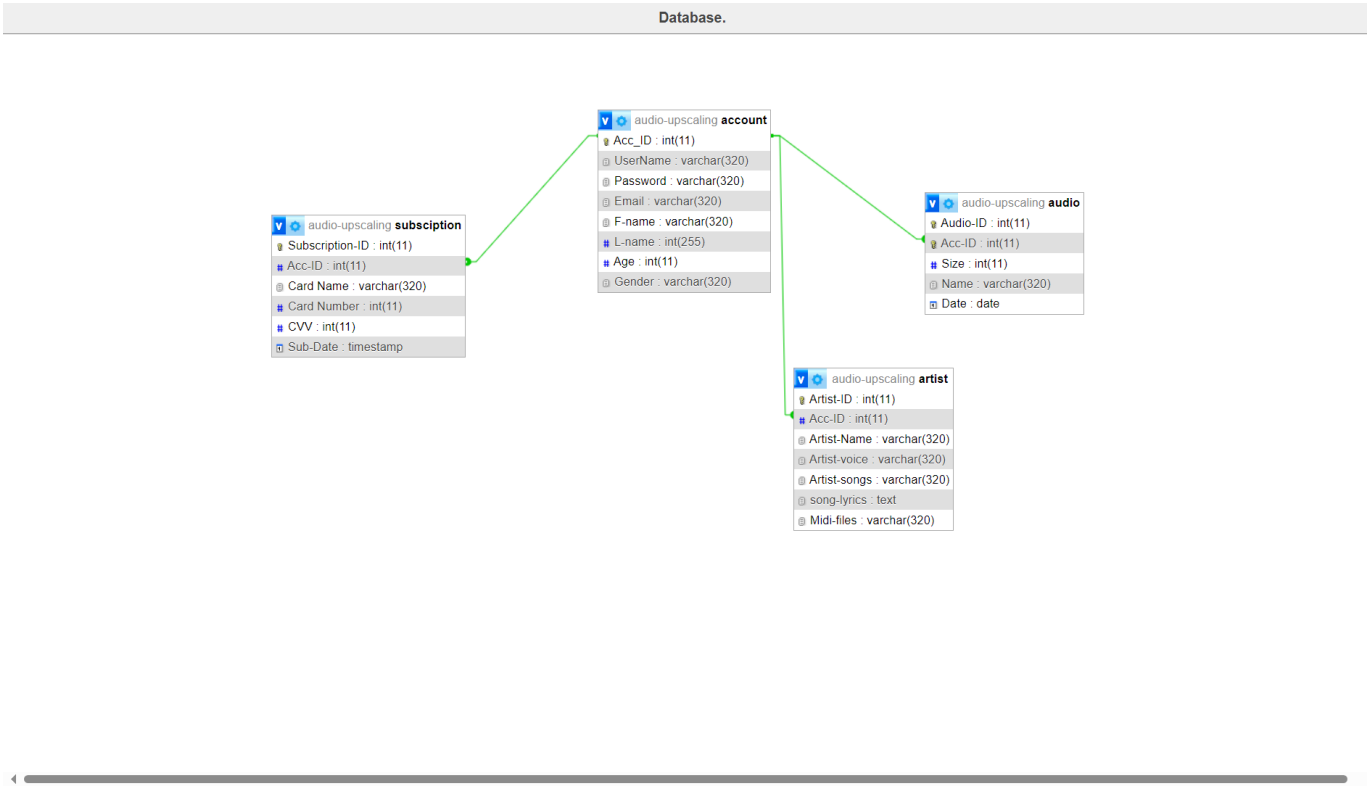

Figure 13: Database

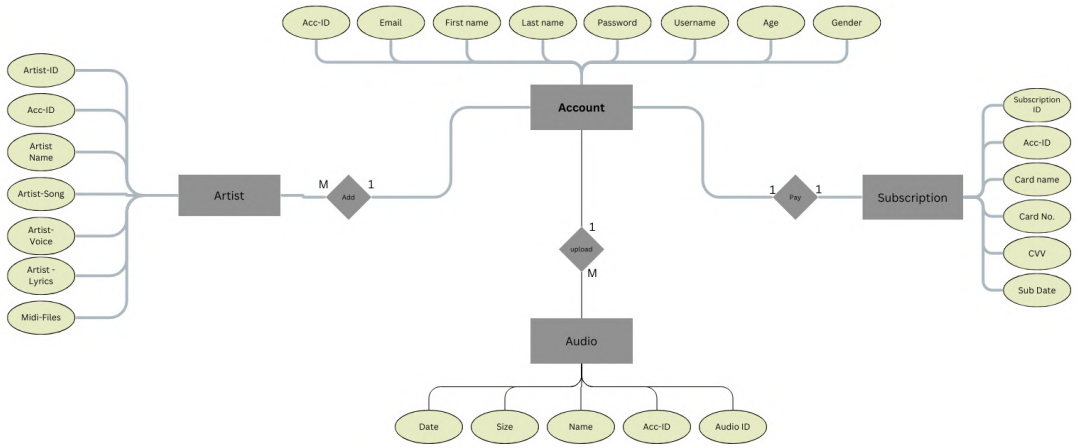

Figure 14: EER Diagram

8 Preliminary Object-Oriented Domain Analysis

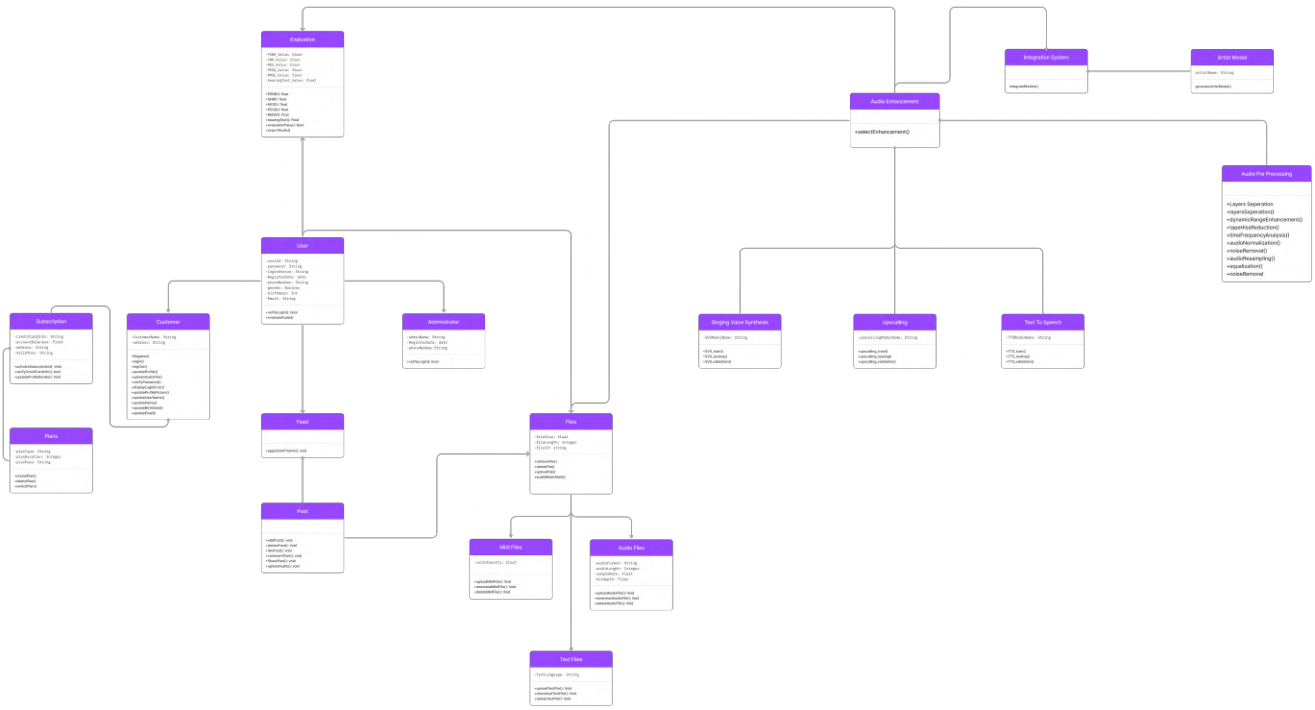

Figure 15: Initial UML Class Diagram

Link: https://drive.google.com/drive/folders/1JiFapQeW$_nSCfxI5NPZahjwd9qKy9spi?usp = sharing$

9 Operational Scenarios

- First scenario : First of all the users need to sign up, they should do some steps to create an account. They should choose a unique username as the database will not accept redundancy. Secondly, the user should enter their name, password, birthdate and email address . Users can also view their profile If they want to change any information, it's optional to view the privacy info. After the user has already signed up and they came back to sign in, they will need to verify their password otherwise they will see login error and here they will need to try again until they enter the correct password. The main purpose of the user accessing the web app is to upload audio files for enhancement, to do this the user will have to upload the file they need to enhance in a form of URL. Moreover, they need to select the file format, sample rate and bit depth they want to convert to, the user could enter a file with unsupported format, at this point we will show them a pop-up message warning them about the unsupported file format, there will be also be a drop-down menu with the supported formats to choose from them. The user will have the access to download the enhanced audio file.

- Second scenario, the admin signs in using the admin username and password. Admins have access to almost everything in the web app, they manage the content of the pages if they see that something might be more reachable and in the sight of the user. Some users may use the program to enhance audio. That may be personal or harmful to someone, at this point, the admin can delete the user's account and change the permissions for the users. The users may face some troubles in the process so they can chat with the admin

for instant support.

- Third scenario. The user uploads the audio file to be enhanced but the unseen process goes through multiple steps before generating the output audio file. First of all, the system should preprocess the audio according to the artist's voice user upload, our model is made to train on singers' voices to be able to synthesize enhanced audio. After the training process, the system will figure out whether the user chose to enhance the audio using TTS(text-to-speech) or audio upscaling. Briefly, the difference between both models is that tts generates audio from input text, on the other hand, audio upscaling takes input audio files and generates an enhanced audio file. The system is made to be reachable to the user and admins as the admins can retrieve audio content from the database and also can add components that can be helpful audio enhancement process.

10 Project Plan

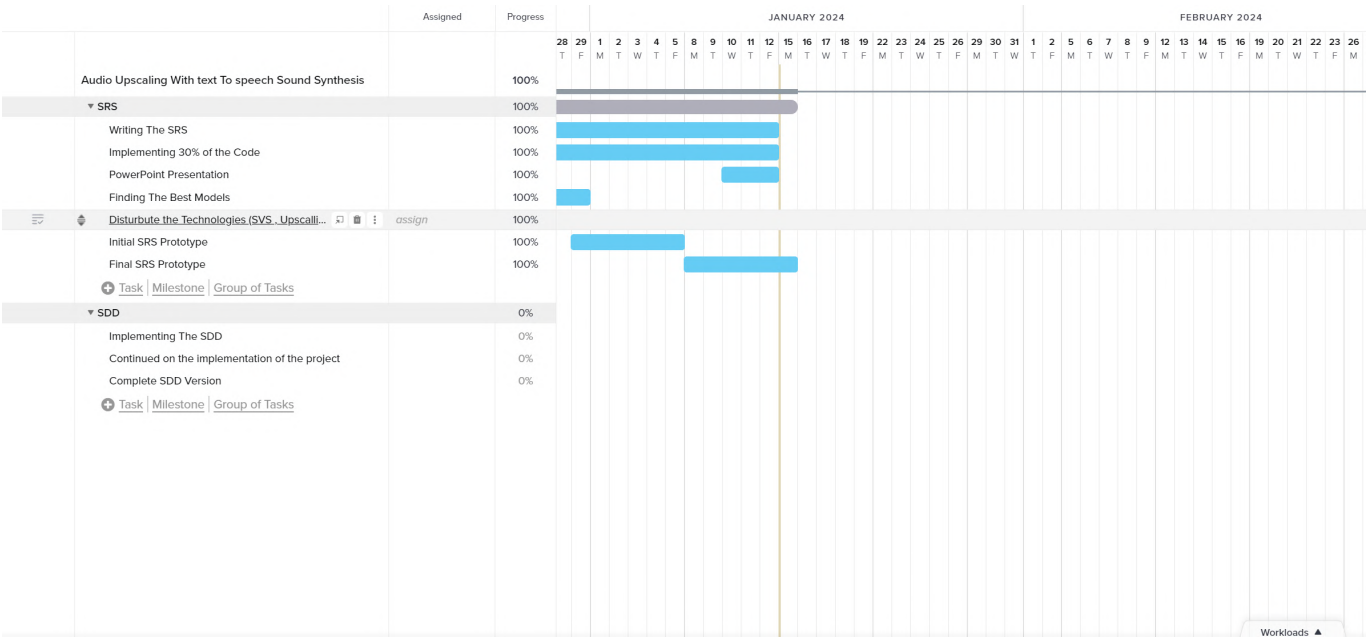

Figure 16: Project Plan

11 Appendices

11.1 Definitions, Acronyms, Abbreviations

1. **Speech Technology Methodologies (STM):**

   Definition: Speech technology methodologies refer to advanced techniques and processes employed in the manipulation and replication of vocal elements within audio files.

2. **Singing Voice Synthesis (SVS):**

   Definition: Singing Voice Synthesis is a speech synthesis technology specifically designed for replicating and generating singing vocals in audio recordings.

3. **Text-to-Speech Sound Synthesis (TTS):**

   Definition: Text-to-Speech Sound Synthesis involves converting written text into spoken words, contributing to the synthesis of natural-sounding vocal elements in audio files.

4. **Super Resolution (SR):**

   Definition: Super Resolution is an upscaling technology that enhances the bit-depth and sample rate of audio files and recordings, leading to an improved audio experience.

5. **Generative AI (GAI):**

   Definition: Generative Artificial Intelligence (AI) refers to a class of algorithms that utilize deep learning techniques to generate new and realistic content based on existing data.

6. **Deep Learning (DL):**

   Definition: Deep learning involves training artificial neural networks with multiple layers to perform complex tasks, contributing to the enhancement of audio quality through advanced algorithms.

7. **Transfer Learning (TL):**

   Definition: Transfer learning is a machine learning technique where a model trained for a specific task is re-purposed for a different but related task, fostering increased accuracy in audio replication by leveraging knowledge from related technologies.

8. **Audio Replication (AR):**

   Definition: Audio replication involves the process of recreating or reproducing audio content, focusing on achieving a natural and superior quality resemblance to the original recording.

9. **Convolutional Neural Network (CNN):**

   Definition: Convolutional Neural Network is a type of neural network particularly effective in processing grid-like data, often used in image and audio analysis.

10. **Recurrent Neural Network (RNN):**

    Definition: Recurrent Neural Network is a type of neural network designed for sequential data processing, making it suitable for tasks involving temporal dependencies, such as audio processing.

11. **Generative Adversarial Network (GAN):**

    Definition: Generative Adversarial Network is a class of machine learning models where two neural networks, a generator and a discriminator, are trained together to generate realistic data.

12. **Sample Rate:**

    Definition: Sample rate refers to the number of samples of audio taken per second, determining the accuracy and fidelity of the digital representation of the audio signal.

13. **Bit Depth:**

    Definition: Bit depth represents the number of bits in each sample of audio, influencing the dynamic range and resolution of the audio signal.

14. **Audio Normalization:**

    Definition: Audio normalization is the process of adjusting the amplitude levels of an audio signal to a standard level, ensuring consistency in playback.

15. **Audio Resampling:**

    Definition: Audio resampling involves changing the sample rate of an audio signal, influencing its playback speed without altering its pitch.

16. **Audio Segmentation:**

    Definition: Audio segmentation is the division of an audio signal into distinct segments or sections, often useful for analysis and processing.

17. **Feature Extraction / Layers Separation:**

    Definition: Feature extraction is the process of selecting relevant information from an audio signal for further analysis, and layers separation involves isolating different components within an audio file, such as vocals and instruments.

18. **Noise Removal:**

    Definition: Noise removal is the process of reducing unwanted sounds or interference from an audio signal to enhance its clarity.

19. **Dithering:**

    Definition: Dithering is the addition of low-level noise to an audio signal to minimize quantization errors during the digital-to-analog conversion process.

20. **Acapella:**

    Definition: Acapella refers to a musical performance without instrumental accompaniment, focusing solely on vocals.

21. **Instrumental:**

    Definition: Instrumental denotes a musical composition or performance without vocal elements, highlighting only the instrumental components.

22. **Time-Frequency Analysis:**

    Definition: Time-frequency analysis involves studying the varying frequency content of an audio signal over time, providing insights into its temporal characteristics.

23. **Tape Hiss Reduction:**

    Definition: Tape hiss reduction is the process of minimizing or removing unwanted background noise, commonly associated with analog tape recordings.

24. **Peak Signal-to-Noise Ratio (PSNR):**

    Definition: Peak Signal-to-Noise Ratio is a metric used to quantify the quality of a reconstructed signal concerning the original signal, often employed in audio and image processing.

25. **Signal-to-Noise Ratio (SNR):**

    Definition: Signal-to-Noise Ratio measures the ratio of signal strength to background noise level in an audio signal, indicating its clarity and fidelity.

26. **Mean Opinion Score (MOS):**

    Definition: Mean Opinion Score is a subjective metric used to assess the overall quality of an audio signal, typically obtained through human listener evaluations.

27. **Perceptual Evaluation of Speech Quality (PESQ):**

    Definition: Perceptual Evaluation of Speech Quality is an algorithmic method used to measure the perceived quality of speech signals.

28. **Root Mean Square Error (RMSE):**

    Definition: Root Mean Square Error is a statistical metric used to quantify the difference between predicted and actual values in an audio signal, providing a measure of accuracy.

29. **Subjective Hearing Test:**

    Definition: Subjective hearing tests involve human listeners evaluating the quality of an audio signal based on personal perceptions and preferences.

30. **Audio Fidelity:**

    Definition: Audio fidelity refers to the faithfulness with which an audio system reproduces the original sound, considering factors such as accuracy, clarity, and dynamic range.

31. **DAW (Digital Audio Workstation):**

    Definition: DAW, or Digital Audio Workstation, refers to software or hardware systems designed for recording, editing, and producing digital audio files, providing a comprehensive environment for music and audio production.

11.2 Supportive Documents

- Dataset **Ar-MGC: Arabic Music Genre Classification Dataset araraltawil_2021** Audio classification involves listening to and analyzing audio recordings. Referred to as sound classification, this procedure is fundamental to numerous contemporary AI technologies, such as virtual assistants, automatic speech recognition, and text-to-speech applications. In this dataset for music recode in the Arabic language, the classifier has five types of music:

  1. Rai
  2. loyal
  3. Muwashahat
  4. east
  5. poems

  all data in the JSON file are of size 799 MB.
  Using these datasets for research using the CNN model and different types of ANN:

1. Rai music using: 113 clips
2. loyal music using: 291 clips
3. Muwashahat music using: 251 clips
4. East using: 291 clips
5. poems using:320 clips

- survey

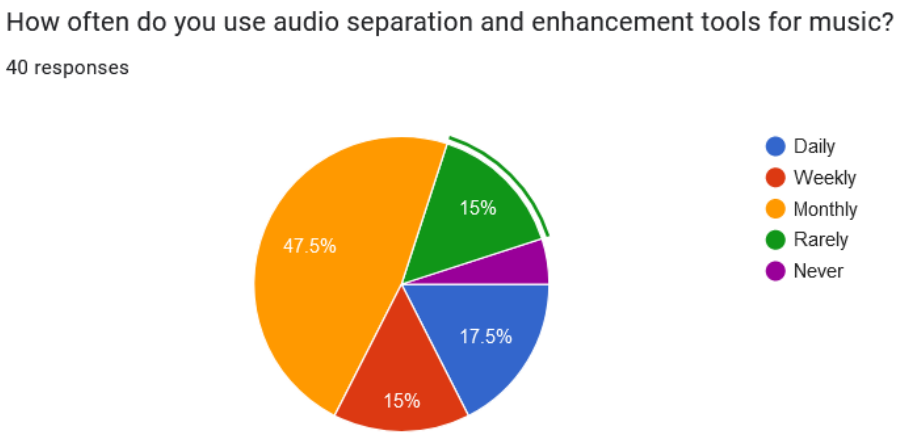

Figure 17: survey

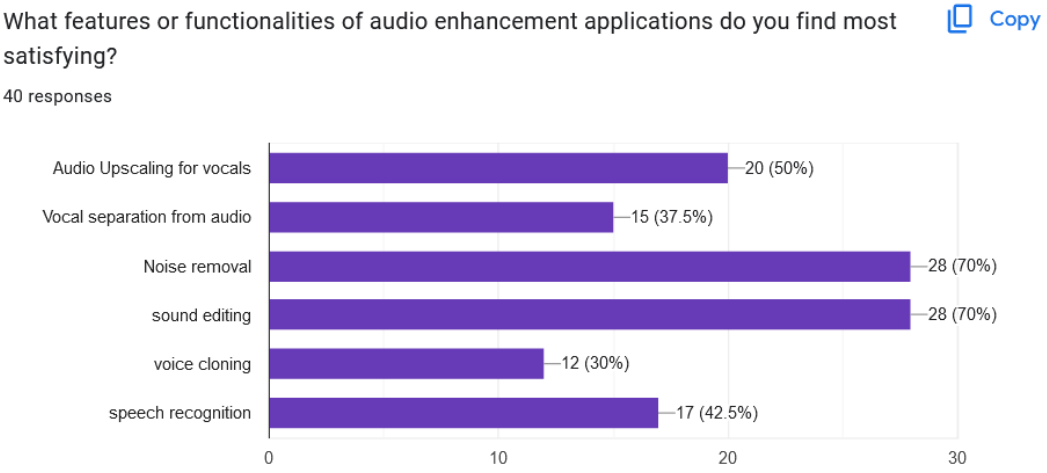

Figure 18: survey

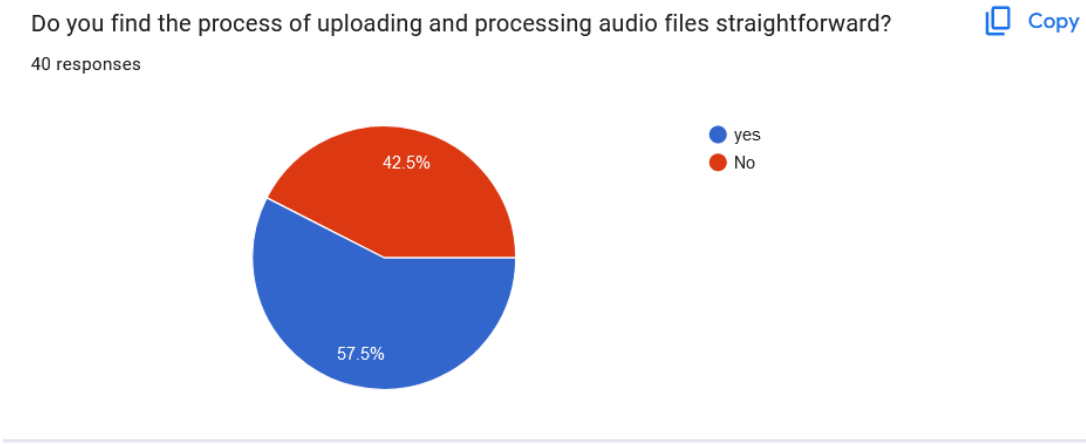

Figure 19: survey

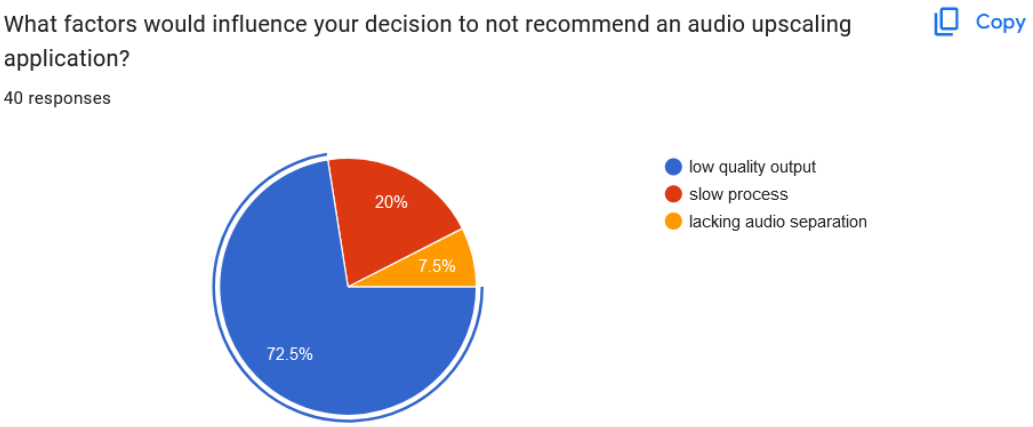

Figure 20: survey