Automated Spatial Audio Tuning with AI

2020240081 Juhee Choi
2022240038 Jaeha Lee
2022240053 Muncheol Chae
2022240084 Sewon Lim

Introduction

Issues

The South Korean film <The Thieves> is the movie faced substantial criticism from numerous audiences due to ***the dissonance caused by after-record sound.*** *Film sound can be broadly categorized into three methods: prescoring, synchronous recording, and after-record, each distinguished by the recording timing of audio and video.*

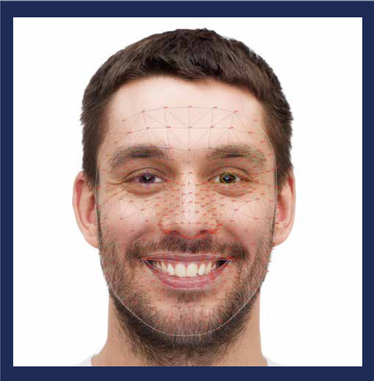
In South Korea's film industry, after-record (recording after filming) is predominantly favored, it is also true that after-record presents ***critical limitations.*** *The single-directional microphones used in synchronous recording can accurately represent the acoustic space, whereas the microphones employed in after-record cannot do, which means sound distance and left-right direction are* ***inadequately conveyed.*** *This dissonance interrupts the audience's immersion in the movie, creating a disparity between the visual and auditory aspects of the film.*

Objectives

Through this project, we propose an AI model that can address ***issues such as interruptions in film immersion caused by after-record.*** *The most significant implication of our project is to solve the problem of after-record's lacking spatial presence. People can feel how audio harms the immersion of the video by directly comparing a typical after-record version with a modified one through our program.*

Methods

Mediapipe



*We planned to develop the program by using the AI framework provided by Google, called '***MediaPipe****', which is a Framework for building machine learning pipelines for processing time-series data like video, audio, etc.*

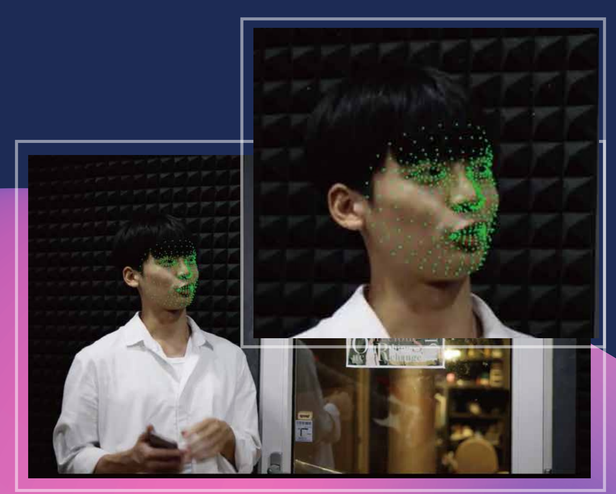
For our project, we used the Face Mesh method from MediaPipe to obtain ***the location values (x, y) of landmarks, especially those on the lips.*** *With this information, we planned to automatically perform audio tuning. For example, if a lip landmark is located on the right side, we can increase the volume on the right side with x coordinates. It can allow us to* ***represent the left-right direction of the recorded audio and create a sense of spatial presence.***

Results

Mechanism of the program

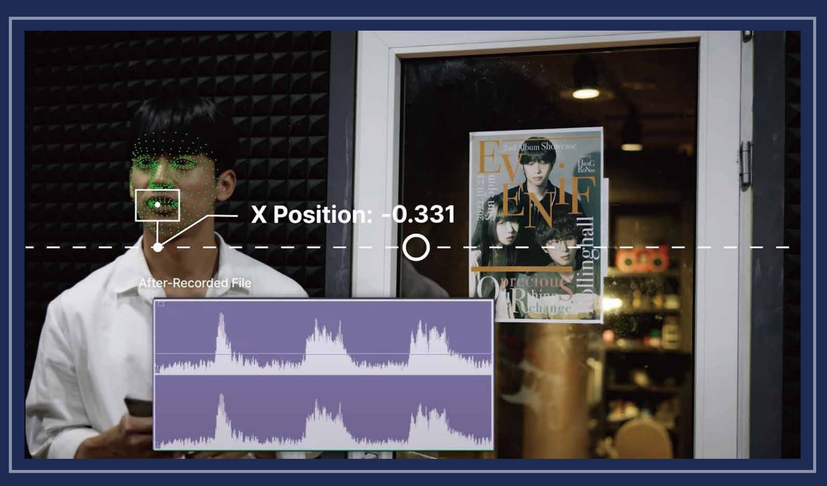
This technology operates based on the standard 16:9 resolution commonly used in the video industry. The center of a 16:9 frame is defined as the origin, denoted as point O. The screen is characterized by ***two axes: x and y, representing horizontal, and vertical dimensions, respectively.***

Relative to the origin, as an object's position shifts to the right, it increases the stereo sound's R-value while simultaneously decreasing the L-value, creating a natural sound. In cases like human speech, audio levels should be maintained between -6 dB and -18 dB. Therefore, at the ends of the video, which are ±100 positions, correspond to R/L values of -6 dB/-18 dB. In the practical implementation process, we also opted for a ***face mesh solution****, which allows us to specify the location of the mouth in the video based on the mesh structure.*



Conclusions

Practical Implications



By resolving the dissonance between the sound and visuals in a film, it is possible to enhance ***audience immersion.*** *Furthermore, by streamlining the labor-intensive aspects of film post-production and mitigating the complexity of the process, this approach enables film stakeholders and directors to concentrate on elevating the overall quality of the film. Ultimately, the development of this program is poised to have a positive impact on the* ***revitalization of the Korean film industry****.*

Further Recommendations

This model offers a foundation for developing ***plugin programs*** *that can be used in video editing environments such as Adobe Premiere Pro, rather than editing audio files in a Python environment and attaching them to the video. As a precedent, Adobe's After Effects program is equipped with* ***AI-powered tracking features.*** *If such video editing programs incorporate these functionalities as plugins, it would make them even more user-friendly and accessible, allowing users to leverage advanced capabilities seamlessly within the editing software.*