The Report Committee for Kyle Chi Sung
Certifies that this is the approved version of the following Report:

Synchronization of Audio Tracks with Acoustic Markers

APPROVED BY
SUPERVISING COMMITTEE:

Vijay K Garg, Supervisor

Christine L Julien

Synchronization of Audio Tracks with Acoustic Markers

by

Kyle Chi Sung

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin

May 2021

Acknowledgements

I would like to thank Dr. Christine Julian for her guidance and valuable feedback during the course of this work. I would also like to thank Dr. Vijay Garg for his supervision throughout the report. I also want to thank Dr. Brian Evans for providing valuable suggestions on the audio signal processing aspect of this project.

Abstract

Synchronization of Audio Tracks with Acoustic Markers

Kyle Chi Sung, M.S.E.

The University of Texas at Austin, 2021

Supervisor: Christine L Julien

Synchronization of audio tracks from different sources to a common audio track is an important step in compiling virtual performances. The process is commonly performed manually because of the variables present in different recording settings. An automatic process using only audio markers present in the recorded audio tracks would have the least assumptions on the equipment setups and the audio content. This report explores two methods of automatic synchronization with audio markers - the "earphone method" in which the singer listens to the audio markers through earphones, and only the vocal audio is recorded; and the "loudspeaker method" where the audio markers are played in a loudspeaker in the same place as the singer, and the recording contains both the vocal and piano components. The earphone method produces an audio mix with good quality, but misalignments can happen from singer mistakes and timing shift; the loudspeaker method is more resilient to those problems, but the quality of the resulting mix is inferior, and the audio markers could be distracting to the singers. Both methods can simplify the recording process of virtual choirs and other online music projects.

Table of Contents

List of Figures

1. Introduction

During the COVID-19 pandemic, gathering of large crowds in an enclosed space is discouraged. The restriction interrupts the workflow for many musicians and performers. As a result, many choirs started recording virtual performances. A virtual choir is a compilation of audio tracks recorded by its singers in various locations and times; in the postprocessing stage, the beginnings of the audio tracks are either truncated or appended with silent segments, such that when they are combined, they appear to be singing together synchronized. Usually a prerecorded piano track serves as the tuning and timing basis for the singers. Everyone records themselves singing with the piano track, and a post-production editor combines and aligns everyone's recordings, producing a compilation in which everyone appears to sing in synchrony.

The post-editing process is manual and can be quite time-intensive [1]. For amateur virtual choirs, the cost of contracting editors to combine the audio tracks are often out of reach, thus making regular practices and rehearsals impossible. The recording environments of the singers are most likely far from ideal - Individuals singing in their own rooms often do not have adequate recording equipment beside their computers and smartphones. Their recordings are characterized by distortions as a result of acoustic echo, playback lags, audio compression, performer timing and tuning errors, etc.

This report explores two recording setups and synchronization strategies. The "earphone method" relies on the singers to clap on specific clicks in the piano track, which they listen to in their earphones. In the postprocessing step, the claps are extracted from the singer's audio for alignment. The "loudspeaker method" relies on nearly inaudible audio watermarks embedded in the piano track, which is played through a

loudspeaker and recorded along with the singer's voice. In the postprocessing step, the audio watermarks and the piano component are separated from the vocal component, and the audio watermarks are used for alignment. Both of these approaches have minimal hardware/software requirements and make very few assumptions about the recording environment. Each method has its own advantages and limitations discussed in this report.

2. Relevant Works

Virtual choirs and orchestras have been a common presentation format in the online communities. For example, the popular Eric Whitacre virtual choir, which has been producing virtual choirs with singers from across the globe since 2010. The process to compile the singers' videos together is mostly manual. [2] There are algorithms to synchronize recorded audios: for example, Librosa [3] has an introduction to music synchronization uses Dynamic Time Wrapping [4]. The MusicNet project has a dataset of classical music recordings and their time labels denoting the music notes, which can also be used in alignment [5]. However, these methods have more assumptions about the format of the recorded music, which limits the versatility and ease of use for amateur musicians. There are apps made specifically for virtual choirs[1], but they also have limitations on the recording format, equipment, and the app platform.

[1] https://www.mixcord.co/pages/acapella

3. Methods

3.1 Earphone Method

In the earphone method, every singer listens to the piano track with earphones. The voice of the singer is recorded with a microphone. The earphone and the microphone do not need to be on the same device; therefore, we cannot assume that the vocal audio is recorded in sync with the piano. Only the singer's voice is recorded. This is the method used by the popular Eric Whitacre virtual choir[2]. The singers follow a video in which the conductor counts down and the singers all clap at zero, and the recorded clap is used as the audio marker for alignment. However, in actual practice the claps can be quite noisy. Many singers, especially young children, do not clap at precisely when the countdown reaches zero, and the misalignment is often very noticeable. In addition, the recordings often contain a lot of noises at the beginning, such as coughing or the scratching noises from microphone adjustments. These noises can be misidentified as the clap, causing the entire track to be severely misaligned. The performances of the Eric Whitacre virtual choir are often quite slow in pace and edited with reverbs, presumably to mask the problems above.

This method reduces the drawbacks by including 8 metronome ticks at the beginning of the piano track. The singers are instructed to listen to the first 4 ticks and clap along on the last 4 ticks. A script identifies the claps by:

1. taking the beginning of the audio track amplitude envelope and calculate its derivative. Since claps have a sharp increase of volume, the delta is a better identifier of claps than just the volume.

[2] Whitacre, Eric. Conductor Video - Eric Whitacre's Virtual Choir 6: Sing Gently. YouTube, 2 May 2020, https://www.youtube.com/watch?v=eIcufPW2wwc.

2. identifying local peaks, which are local maxima of the waveform in a 0.1 second window.
3. from these local peaks the top 12 peaks are selected as clap candidates;
4. from these 12 candidates we pick 4 and evaluate the standard deviation of the clap intervals. From all the combinations (12 choose 4), the set of candidates with the least standard deviation is identified as the claps.
5. the candidate set of claps, though not in perfect regular intervals because of human imperfections, are aligned to the last 4 metronome ticks by minimizing the mean square error between the time of the claps and the time of the metronome ticks.

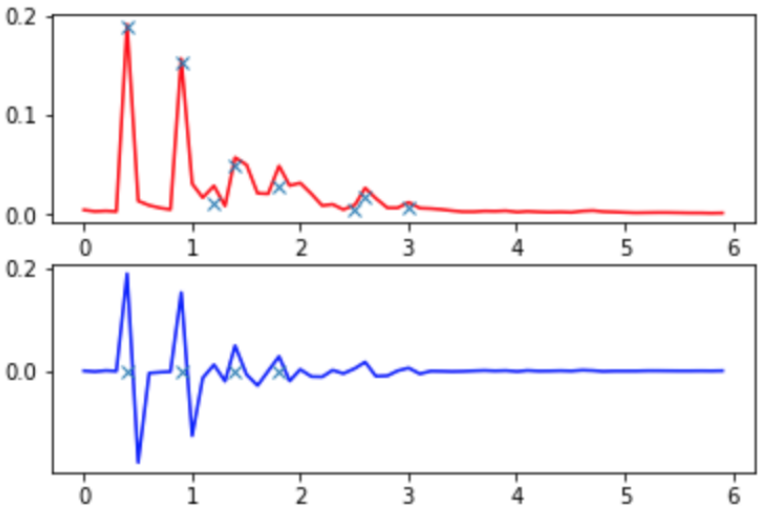

Figure 1: process of identifying claps in a noisy input.

In figure 1, the top graph shows the input signal amplitude envelope. The 12 crosses are the local maxima of the signal derivative. The bottom graph shows the derivative of the input signal. From the 12 peaks in the top graph, the 4 crosses with the most even intervals are selected as the actual claps.

3.1.1 Earphone Method Evaluation

The method works very well at identifying claps because other sporadic noises created unintentionally by the singer do not have regular intervals. Vocal audios usually do not have sharp increments in volume to be registered as claps. Even though the claps are often not loud enough to be louder than other noises, by expanding the candidates to the loudest 12 and selecting only the most regular 4, the clap loudness is much less relevant to the selection. The earphone method is implemented in python as a jupyter notebook shared on GitHub.[3] The notebook can be opened in Google Colab, an online jupyter execution environment hosted by Google, and the notebook compiles synchronized audios or videos in a given shared folder on Google Drive. With some configuration to specify the path to the shared folder and output format, the choir conductors and music teachers can ask their students to upload their recorded audio clips to a specified online shared folder, run the script, and get the compiled audio in the same shared folder. The script has been used for several short choir presentations.

3.1.2 Earphone Method Limitations

However, in actual practice, some assumptions of the method start to break down. On one occasion the script was used to compile a children's choir. The children might not follow the instruction to clap, or their claps can still be too irregular to produce accurate alignment. In another recording project, the song lasted over 7 minutes. If a singer only needed to sing at a small part of the song, he or she would still need to record from the beginning. Some singers' tracks started to get out of alignment at the end, even though the beginnings were aligned. Possible causes might include playback delays in streaming the piano track, clocking differences, or that the different audio players are not designed

[3] https://github.com/k7sung/clap2choir/blob/master/sync.ipynb

to playback in precisely the same speed across different locations and platforms. The accuracy of the earphone method depends entirely on the accurate timing of the claps at the beginning, so any distortions in other parts of the audio tracks are unaccounted.

3.2 LOUDSPEAKER METHOD

Instead of following the piano track in the earphone, the singer can also play the piano track on a loudspeaker. In this method, both the voice of the singer and the piano are recorded, and the piano component is removed later. This simplified a lot of the difficulties mentioned in the earphone method. The recording instructions for the singers become trivial - just sing along with the piano music in the background. The original audio in the recorded track can be used to align the singers' recording, as well as tracking the time drift throughout the recording. However, there are also challenges to this method. The algorithm needs to remove the background piano audio so only the singer's audio is present. Extracting timing from the piano track is also more difficult in comparison to the aligning the distinctive clap patterns in the earphone method.

3.2.1 Extracting the Vocal Component

To remove the piano audio from the singer's recording, I have tried to approach it as a problem of acoustic echo cancellation, a process often required in conference calls [5] [6]. In this scenario, the original piano track is considered as the source, which is recorded by the singer's mic again as (the echo), along with the singer's voice. I attempted using the least mean square adaptive filter to remove the echo. LMS filter requires the source and the echo audio to be pre-aligned first (see examples in [7]), which was done manually in my test. The result was not ideal - the piano component, although

weakened, is still audible, and the quality of the vocal component also deteriorated. This might be because the piano audio was nonlinearly distorted in the singer's recording, which makes LMS adaptive filter insufficient (see [5]); playing and re-recording the audio in a regular laptop speaker/microphone setup can introduce a significant number of distortions. I ended up using Spleeter[4], an audio source separation library, to remove the piano source. It has a pre-trained model that separates an audio track into vocal and some instrument components without requiring a reference piano track. In practice, it also has a tendency to misclassify and remove too much vocal component. However, in comparison to the outputs of LMS adaptive filter, Spleeter still produces clearer vocal-only tracks. Overall its outputs are sufficient for the purpose of testing the alignment of the loudspeaker method.

3.2.2 Alignment using the Piano Component

For alignment, I have tried the naive approach by simply cross-correlating the frequency spectrum of the singer's recording to the piano track. The piano component in the singer's recording should cross-correlate well with the original piano track and produce a peak when they are aligned. However, this usually ends up with the loudest parts of the music correlated together. Music pieces also often contain rhythmic, repetitive structures, which produces strong false peaks in the cross-correlation.

Since these limitations are intrinsic musical characteristics of the piano track, my next approach was to embed special audio markers in the piano track to identify the progress of the music. The audio markers should have the following characteristics:

1. the audio markers should encode timestamps to uniquely identify the playback time;

[4] https://github.com/deezer/spleeter

2. the audio markers should be nearly inaudible to humans but easily recognizable by the system (a "watermark"). However, the audio marker cannot be entirely inaudible (composes of purely ultrasonic frequencies, for example), otherwise the markers can be removed by lossy compression formats such as mp3.
3. The audio markers embedded in the piano track would also need to survive the recording process - played back without special equipment, mixed with singing voice and room echoes, and re-recorded without special setup.

I have tried some existing audio modem projects such as the Quiet Project[5]. But even when the carrier frequency is set to as low as 7kHz (well within the hearing range and shouldn't be discarded by mp3), the test data encoded in the audio did not survive the re-recording process.

3.2.3 Design of Watermark Timestamps

My final approach is based on the paper "Audio Watermarking Over the Air with Modulated Self-Correlation" [8]. The paper described a way to embed and detect resilient audio watermarks. My system used the same techniques described in the paper to insert multiple unique watermarks in the piano track at regular intervals. They act as timestamps indicating seconds passed since the beginning. To extract the alignment timing, the system has to detect the presence of a timestamp as well as correctly identify which particular timestamp it has detected. The timestamps are carried over 16 channels in the range from 10k to 13khz, which is on the high end of human hearing range, so the watermarks are less audible to humans, and also less likely to have interference with the music components. The upper range of 13khz is based on experimenting with the

[5] https://quiet.github.io/

frequency response of the build-in speaker and microphone of a 2015 MacBook Pro laptop. Frequencies above 13kHz tend to be less reliable. Every timestamp only contains 8 bits of information, but these 8 bits are duplicated on the lower 8 and upper 8 channels to alleviate the problem of lossy channels. Every watermark is repeated 5 times. The existence of a timestamp is detected by autocorrelation to detect these repeats. By using autocorrelation for detection, the frequency response of the audio channels becomes irrelevant (section 3.2 in [8]). In addition, some of the encoded bits are flipped according to a predetermined pattern randomly generated (the "modulation" in [8]). The bits will be flipped back later in the decoding stage in the receiver end, so that any non-watermark audio introduced into the channels in the recording stage can be weakened by the flipped bits (figure 2).

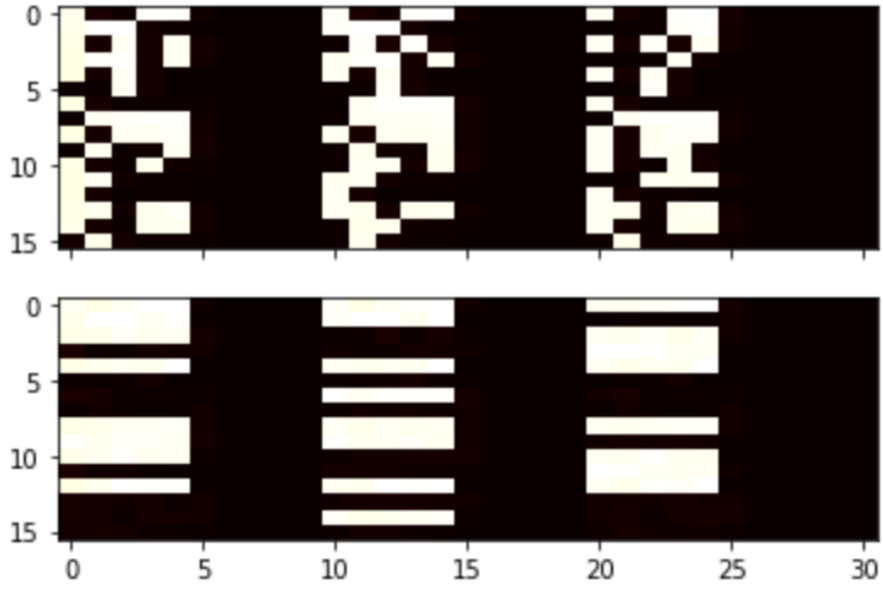

Figure 2: modulation of timestamps

In figure 2, the top graph shows the frequency spectrum of the watermark transmitted over 16 channels over time in an ideal environment. The bottom graph shows

that the signals, after some bits are flipped in a predetermined pattern, produce repeated patterns that would autocorrelated well, and can be decoded to timestamp symbols.

In the decoding stage, when the presence of a watermark is detected, the watermark is compared with all the timestamp patterns previously embedded in the piano track. Ideally the timestamp with the best pattern match is selected, but sometimes an incorrect timestamp could be selected. These incorrect timestamps are filtered out by comparing with other timestamps in the recording, and only timestamps increasing at a predictable pattern are accepted (figure 3). The top graph shows the 16 channels extracted from the frequency spectrum in the audio segment; The middle graph shows the likely locations of watermarks over time; The bottom graph shows the similarity of the watermarks to the 70 candidate timestamp patterns. Among the likely patterns (brown and red dots), only the candidates that progress regularly in a line are chosen (red).

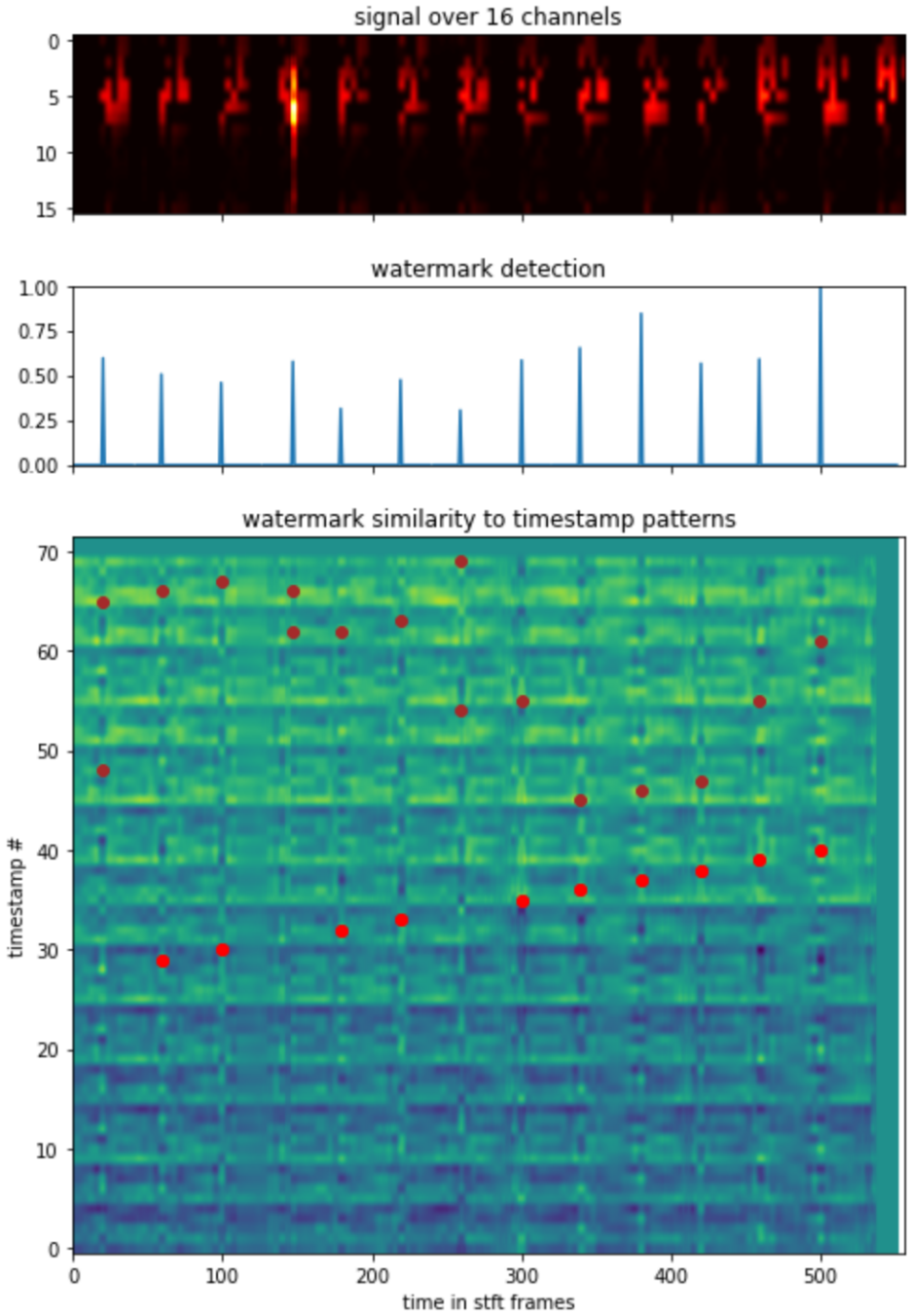

Figure 3: process of identifying a timestamp in a singer's track.

3.2.4 Loudspeaker Method Test Results

To test the loudspeaker method, 20 recordings were collected from 7 volunteers to record a short hymn lasting 97 seconds. The volunteers sing with the prepared piano track along random segments of their choice in various recording setups without special instructions. The project system is able to recover the correct timestamps from 16 of the 20 recordings and to reconstruct an *a capella* performance of the entire hymn. The audio files are included in the appendix. Of the four recordings that failed, two of them contain watermark signals that are too weak to be identified; two were recorded at a 22kHz sampling rate (I have previously assumed that all recordings are sampled at 44kHz), which puts some of the 16 channels above the Nyquist rate and are thus unrecoverable. If necessary, the failures could all be addressed by asking the singers to re-record with a louder speaker volume or better recording quality. By synchronizing small segments in various positions to the correct moments in the song, the test demonstrates the ability of the loudspeaker method to accommodate timing drift, which is not possible with the earphone method.

3.2.5 Loudspeaker Method Limitations

Even though the watermark volume was set at 1% of the volume of the piano track, some volunteers did notice the high pitch watermark noise. One volunteer said the noise is very noticeable and distracting. The testing result shows that a careful balancing between the comfort of the singers and the detectability of the watermarks is required. If the watermark volume is made louder, or if the watermark frequencies are made lower to

accommodate lower recording sampling rates, then the watermarks will also be more audible and distracting to the singers.

The loudspeaker method relies on Spleeter library to separate the piano components from the vocal components. A big limitation in this method is that Spleeter only supports the extraction of bass, piano, vocals and drums. However, it is possible to train Spleeter to extract other instruments with more datasets[6]. In addition, the library tends to be conservative in identifying the vocal components, resulting in some segments of vocal tracks gone missing (see the appendix for audio examples). The missing segments appear in various places throughout the recording, so that when the vocal tracks are combined, the mix might cover up these missing segments. But individual singers might still notice their voices missing in some parts. They might also notice degraded audio quality of their voices from the audio source separation process.

[6] https://github.com/deezer/spleeter/wiki/2.-Getting-started#train-model

4. Conclusion

Synchronization of audio tracks is an important step in compiling virtual performances. This report explores the earphone method and the loudspeaker method, both of which are suitable for casual performers. The earphone method was used in many actual choir projects. The method requires very little preparation - only a metronome is needed at the beginning for the recorder of the piano track. The method uses only the beginning of the audio for alignment, thus making it susceptible to timing drift. The loudspeaker method requires more preparation and postprocessing, but it does not require a headphone, and it allows the singer to start from anywhere. The relaxed requirements make the recording format more flexible, and the embedded timestamps in the piano track make the alignment more resilient to timing drift. This report only present findings with vocal and piano tracks. However, since Spleeter potentially supports audio separation of other instruments, both methods can be applied to other types of performances besides choir pieces. Overall the project simplified the process of recording virtual performances, bringing musicians from various places together by making a virtual performance fast and automatic, thus making frequent practices, rehearsals and presentations possible for everyone.

Appendix

The files related to the loudspeaker method synchronization test is listed below, hosted on Google Drive:

- The 20 collected audio files:
  https://drive.google.com/drive/folders/1PBBBxuUb_9k90Ibt_N1O4c-1hvpJ10ww?usp=sharing
  - Two of which were sampled at 22050Hz, therefore the timestamps could not be extracted: V_1.m4a, V_2.m4a
  - Two of which contain timestamp signals too weak to be identified: E_1.m4a, S_1.m4a

- The piano track used for synchronization, with watermarks embedded:
  https://drive.google.com/file/d/1QEjgbU8hfMl4AEDxkkKxTRlxeVXpjulh/view?usp=sharing
- The aligned audio tracks, with the piano and vocal components separated:
  https://drive.google.com/drive/folders/1R2tMSdNhk5YK5POqlJmFibWgGuE305CO?usp=sharing
  An example of an incomplete separation is K_2_piano.mp3, in which the vocal component is still clearly audible. Only a small segment of the vocal component is extracted to K_2_vocal.mp3.
- The final mix:
  https://drive.google.com/file/d/1QH6BrerwP8e9WExtkAVHZgkZj3gnkKvs/view?usp=sharing
- The python code and the jupyter notebook for synchronization:
  https://drive.google.com/drive/folders/1QLOCMw0QzqrXuebrGUD8ZDCkg0xlnbol?usp=sharing

References

[1] K. Wardrobe, "Dear Music Teachers - Please Stop Asking How To Create A Virtual Choir," 23 March 2020. [Online]. Available: https://midnightmusic.com.au/2020/03/dear-music-teachers-please-stop-asking-how-to-create-a-virtual-choir-video/.

[2] P. Citron, "Ludwig Van Toronto," 17 July 2020. [Online]. Available: https://www.ludwig-van.com/toronto/2020/07/17/interview-composer-conductor-eric-whitacre-talks-about-virtual-choir-6-sing-gently/.

[3] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg and O. Nieto, "Librosa: Audio and music signal analysis in python," in *14th python in science conference*, 2015.

[4] "Music Synchronization with Dynamic Time Warping," librosa.org, [Online]. Available: https://librosa.org/doc/latest/auto_examples/plot_music_sync.html. [Accessed January 2021].

[5] B. Kosanovic, "EE Times," 11 April 2002. [Online]. Available: https://www.eetimes.com/echo-cancellation-part-1-the-basics-and-acoustic-echo-cancellation/.

[6] "Example: Noise Cancellation using the LMS Algorithm," 2006. [Online]. Available: https://www.advsolned.com/example_ale_nc.html.

[7] J. Wramberg and M. Tausen, "Adaptfilt," [Online]. Available: https://github.com/Wramberg/adaptfilt.

[8] Y.-Y. Tai and M. F. Mansour, "Audio Watermarking over the Air with Modulated Self-correlation," in *Acoustic, Speech and Signal Processing (ICASSP)*, 2019.