Real-Time Digital Signal Processing Final Report
Music Vocals Remover 2K

Michael Urbanski (mu120) Adrian Kwan (akpc426)

Introduction and problem statement

This report documents the attempts to implement a singing voice removal algorithm on the Texas Instruments TMS320F28179D Digital Signal Processing (DSP) microcontroller. The design specification called for an implementation accepting series of 8kHz single-channel audio samples as input, and outputting series of samples of the same frequency, but with the human voice removed. The method of choice was required to operate on the DSP board in real time and with less than 2 seconds of delay. As documented in the report, a successful algorithm was not developed in the given time frame, and this report focuses on documenting and analysing the attempted solutions.

Proposed solutions

The method of choice, blind source separation using harmonic-percussive (HP) median filtering, was described in the initial report, but the theory and the accompanying commentary is expanded on in this report, as the efforts to implement the solution provided more insight into the technique. This technique is based on using the magnitude spectrogram of the input signal in order to separate harmonic sounds from percussive ones. In order to obtain the spectrograms, a Short Time Fourier Transform (STFT) is performed on the input signal. This transform implements a Fast Fourier Transform (FFT) with window moving through time, providing information not only on what frequencies make up the signal, but also how the frequency content changes through time. This can help identify the transitions between singing and instrumental parts of the song, and filter these differently depending on time. On the spectrogram, the harmonic sources appear as horizontal lines, while the percussive sources appear as vertical lines (Figure 1).

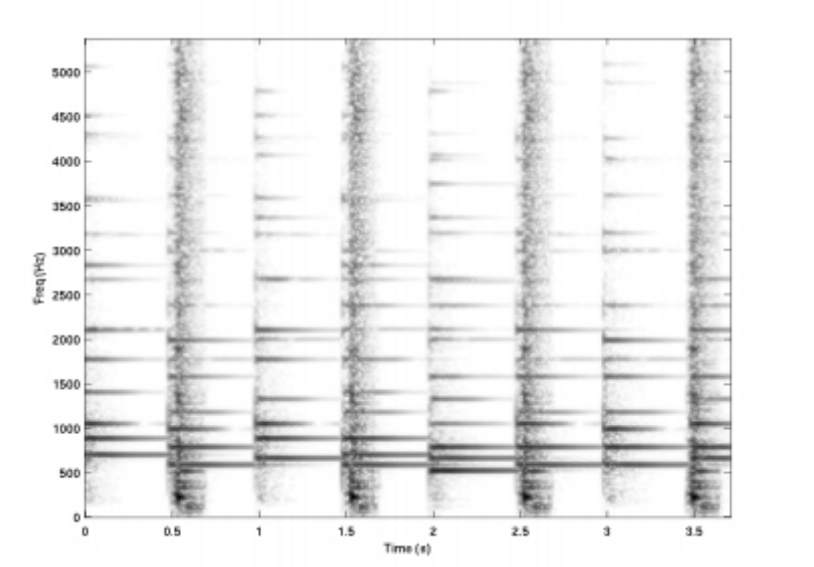

Figure 1: Spectrogram of a drum and piano showing vertical (percussive) ridges and horizontal (harmonic, pitched) lines. Taken from [1].

These can then be attenuated by using the median filtering along either dimension of the matrix of data behind the spectrogram. The main premise behind the voice removal algorithm, is that the human voice appears as a harmonic sound if an low frequency resolution STFT is used, and as percussive sound if a high frequency

resolution STFT is used. The technique was originally proposed by D. FitzGerald and M. Gainza in [1], where the authors were working on a 44.1kHz recording. The proposed frequency resolutions were equivalent to FFT sizes of 512 and 16384 bins respectively. In this project, the recording is sampled at 8kHz, so the FFT sizes were scaled down to 3072 and 90 bins. Similarly to the technique in [1], the algorithm uses first a high frequency resolution STFT and a median filter in order to separate the harmonic sources from percussive and vocal ones. This harmonic mixture is then preserved, while the percussive and vocal mixture is passed onto a low frequency resolution STFT with a median filter, which separates the percussive sources from the vocal ones. Each filtering operation also includes a time-frequency mask based on Wiener filtering, to get rid of the artefacts imposed by the filter. Both harmonic and percussive mixtures can then be inverted back into time domain and added together, yielding the audio track with the vocal part removed. The diagram of the system presented in [1] is shown in Figure 2, to visually demonstrate the idea before presenting the Simulink models.

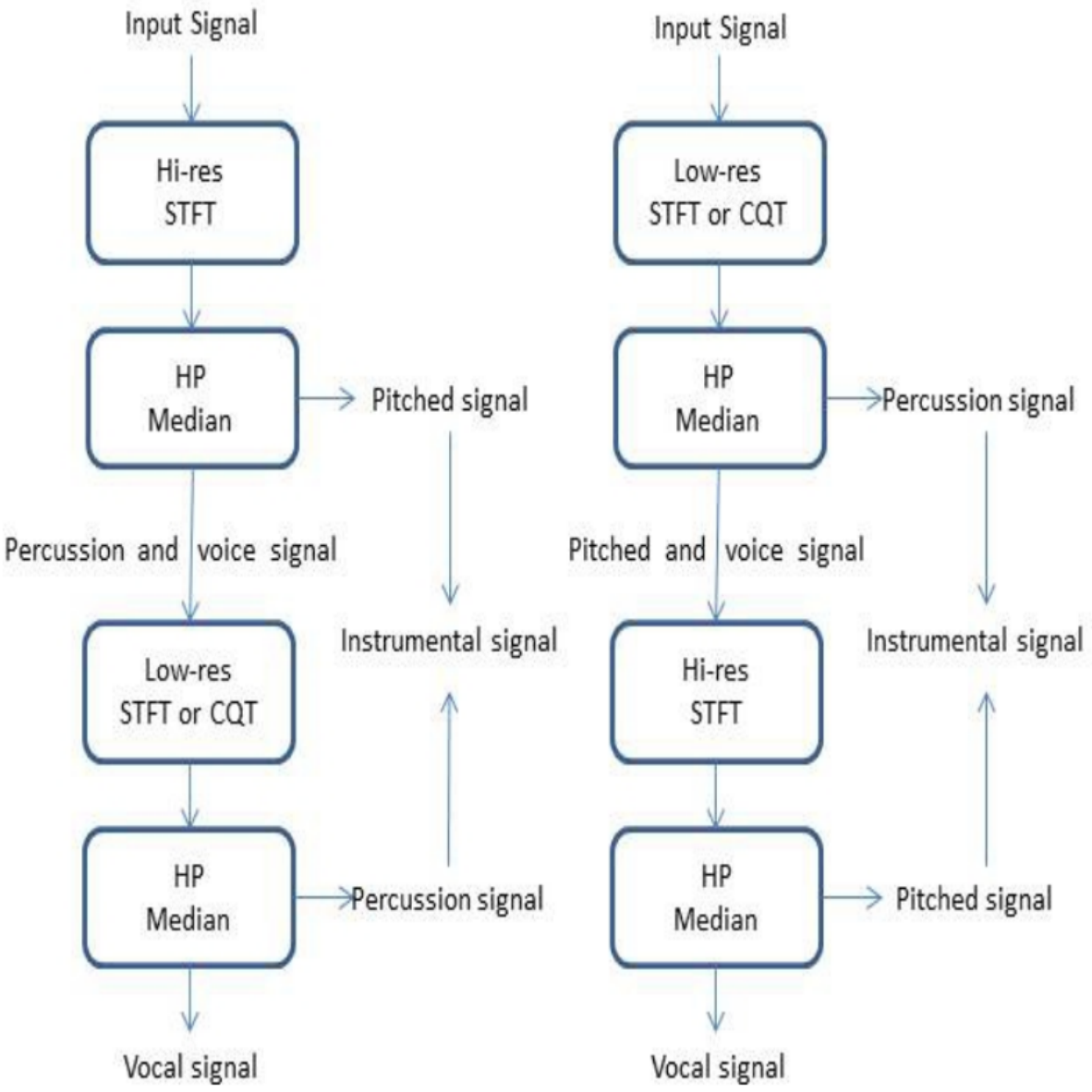

Figure 2: Algorithm flowchart as defined in [1]. CQT stands for constant Q transform.

Alternative designs were also explored as contingency options, noting that the main approach, though deemed suitable and promising, is quite complicated and leaves room for error in the implementation. From the sample tracks, it was noted that they mostly belonged to the pop/ folk genre, with a clear vocal accompanied by acoustic instrument like guitar. Hence, it was proposed that simple filters, eg. high-pass/ band-pass filters could give a somewhat promising result in case the main design cannot be fully implemented on the DSP, by having high attenuation in the frequency range for the vocals. To begin with the design, the frequency range of vocals and other common instruments identified in freemusicarchive.org were looked into [2]. Different from the normal speaking voice, vocals typically are within the 200 to 1000 Hz fundamental frequency range for female

voices and 80 to 800 Hz for that of male voice. While the bass instruments, for instance, would have farther apart frequency ranges and are thus easier to separate using a simple filter , instruments like guitar with energy concentrated in similar frequency range to vocals, ie. 80 Hz to 1000 Hz [2], posed challenges in terms of the filter design. In particular, it was expected that the simple filter could not fully separate vocal and instrument tracks due to the presence of resonances, including overtones (higher harmonics), as well as varying level of energy in different harmonics across the frequency spectrum for different samples, instead of mainly in the fundamental frequencies.

For example, from the frequency chart below [3], similar to the values in [2], shows that guitar has a fundamental frequency range (10 - 1000Hz) that overlaps to a large extent with the male voice, while its harmonics extend less and other instruments each differs in how high its frequency (including overtones) range go. In general, woodwinds, which is commonly used in jazz and pop music, go up to the thousands in the frequency range and can also play at lower ranges, while brass instruments typically are in lower ranges and are similar relative to vocals. For strings and keyboards, less of the track would be left out due to their wide and varying range. Nevertheless, the harmonics again asked for more complex designs for a clean vocal removal algorithm.

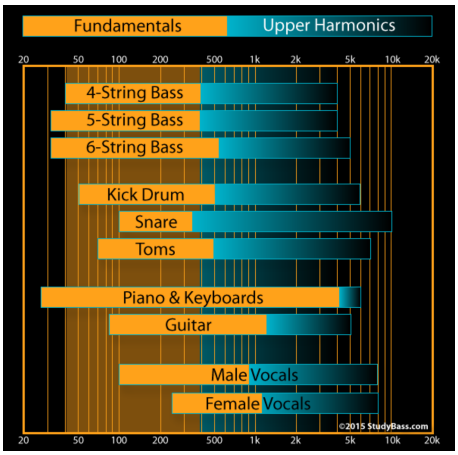

Figure 3: fundamental frequency and harmonics range of vocals and instruments [3]

A simple standalone STFT implementation with reference to mainly the STFT and ISTFT block model provided in the course has also been tested. In principle,the FFT coefficient matrix, ie. coefficient vectors for each FFT frame, was scaled down in hopes of removing the vocal and keeping it small enough to fit in the DSP, which eventually could not run properly on the DSP due to the inability to allocate sufficient memory and adjust the buffer, alongside other challenges to be discussed in the next section.

HP Median Filtering - PC Implementation

The implementation setup used for validation on a PC is shown in Figure 4. The left half of the diagram covers the separation of harmonic sources from the percussive and vocal ones. The signal from the audio file is buffered at a modest 512 samples per channel and passed through a high frequency resolution STFT, using a Hamming window. The window and STFT parameters are summed up alongside the low frequency resolution STFT parameters in Table 1. Following the STFT, the signal enters the median filter and mask subsystem, shown in Figure 5. Larger versions of these figures are available in the appendix.

Before filtering, an absolute value of the signal is taken to convert it to the real domain. The upper branch in Figure 5 is filtering across frequency bins, attenuating the vertical lines (percussive sources), while the bottom branch attenuates the horizontal lines (harmonic sources). The original spectrogram uses the frequency bins as rows and time slices as columns, therefore in order to filter along the second dimension (columns, or time slices) the signal is transposed on the lower branch of the diagram. The filtering yields the harmonic enhanced time slices H_h and percussive enhanced frequency bins P_i, where h and i are indexes of time slices and frequency bins respectively. The time-frequency masking is then done as described in equations 1 and 2, followed by recovering the separated complex spectrograms (equations 3 and 4).

$$H_h = M(S_h, l_h) \tag{1}$$

$$P_i = M(S_i, l_p) \tag{2}$$

where M stands for the *median* operation, l_h and l_p stand for the length of median frames for harmonic and percussive filters respectively.

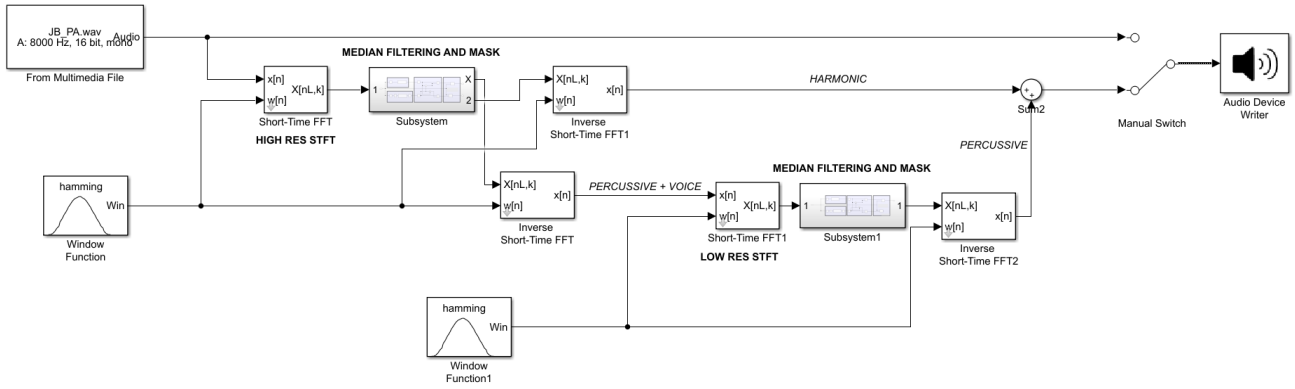

Figure 4: Simulink model of the system used for PC validation.

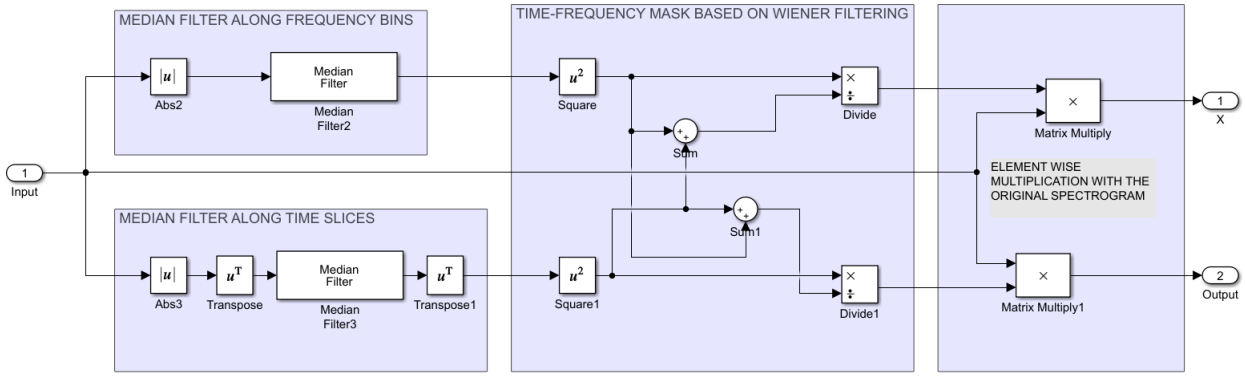

Figure 5: Median filtering and mask subsystem for the high frequency resolution harmonic filtering.

$$M_{Hh,i} = \frac{H_{h,i}^2}{H_{h,i}^2 + P_{h,i}^2} \tag{3}$$

$$M_{Ph,i} = \frac{P_{h,i}^2}{H_{h,i}^2 + P_{h,i}^2} \tag{4}$$

where M_H and M_P are the harmonic and percussive masks respectively, h and i are once again used to index the time slices and frequency bins, this time as a double index.

Both outputs are then inverted back into the time domain using the inverse STFT. The harmonic signal is passed on, while the percussive signal, containing also the vocals, is passed through a low frequency resolution STFT (90 bins), shown in the right half of Figure 4. The subsystem used for the low frequency resolution median filtering is shown in Figure 6. The only difference from subsystem in Figure 5 is that only the percussive sources are passed on as output, while the vocals are discarded. The full set of parameters used for the windowing and transforms is shown in Table 1.

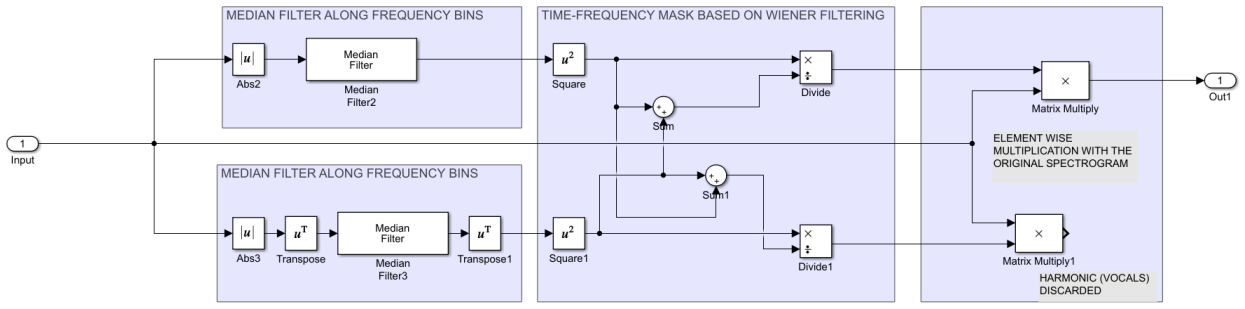

Figure 6: Median filtering and mask subsystem for the low frequency resolution percussive filtering.

Parameter	High frequency resolution STFT	Low frequency resolution STFT
STFT length	3072	90
Analysis window length	3072	90
Overlap	1536	45
Samples per channel	512	

Table 1: STFT and window parameters used in the model.

HP Median Filtering - Results

The single-channel sound mixture can be analysed using the spectrograms to reveal information about the quality of the filtering and vertical/horizontal line attenuation. As mentioned in the previous sections, moving median filtering along the first dimension of the spectrogram (frequency bins) attenuates the vertical slices, while filtering along the second dimension (time slices) attenuates the horizontal ones. The spectrograms of the inputs at each stage can therefore serve as a visual guide representation of the quality of separation. The spectrogram using the first 30 seconds of the original mixture of *Passive-Aggressive* by John Bovey is shown in Figure 7.

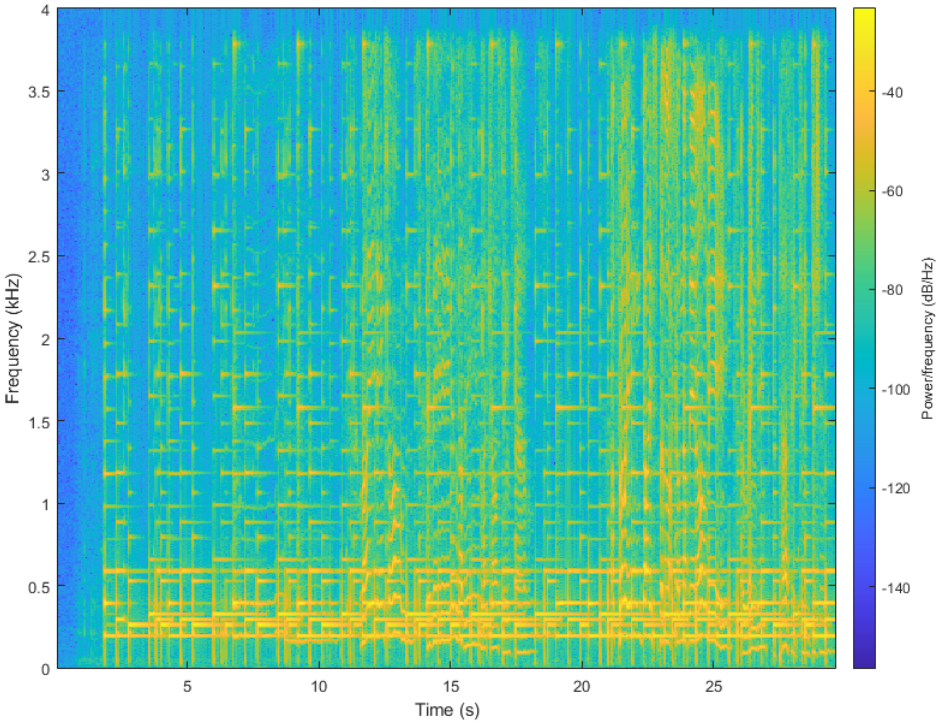

Figure 7: Spectrogram of the first 30 seconds of *Passive-Aggressive* by John Bovey (using 1024 bin STFT with a moving window of 1024 samples, 512 sample overlap.

The harmonic instrumental parts are present mostly at low frequencies (0-1.5kHz), visible as the horizontal bars on the plot. It is important to note at this point that attention must be paid to which axis is used for time in the spectrogram. In this case time is the horizontal axis and frequency is the vertical axis. This is not always the default setup in popular software. The percussive sounds are impulses stretching across most of the frequency spectrum in the plot, visible as the vertical lines. This spectrogram was created using a 1024 bin STFT, which is in between the thresholds of 90 and 3072 bins specified as low and high resolutions for a recording with an 8kHz sampling rate. For comparison, the spectrograms using the STFT and window parameters from Table 1 are shown in Figure 8A (low frequency resolution) and 8B (high frequency resolution). The main striking difference is the quality of the spectrogram, with the low frequency STFT plot (8B) being visibly more granular and pixelated. At such a low frequency resolution, the horizontal lines are wider, and more sounds are classified as harmonic. In Figure 8A the horizontal lines are much finer, and so can filter out the truly harmonic instrumental sounds from the pseudo-harmonic ones (vocals).

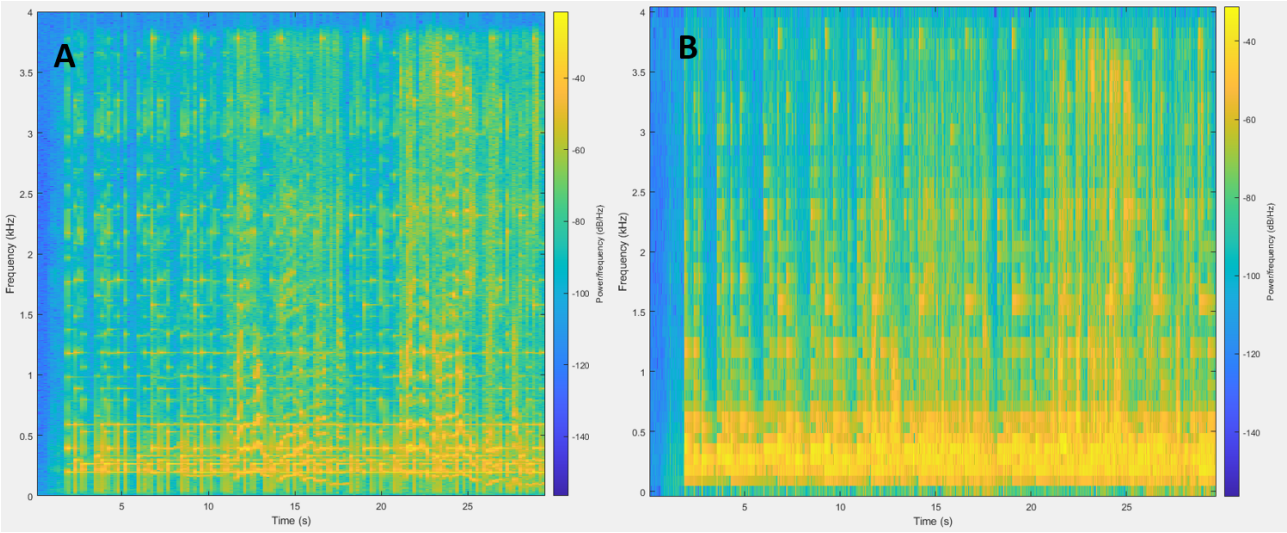

Figure 8: Spectrogram of the first 30 seconds of *Passive-Aggressive* by John Bovey, using A) 3072 bin STFT with a moving window of 3072 samples, 1536 sample overlap, B) 90 bin STFT with a moving window of 90 samples, 45 sample overlap.

Both of the spectrograms in Figure 8 were taken before the filtering was applied, and are shown only for comparison. In practice, the high frequency resolution STFT was used on the original mixture, and the low frequency resolution STFT on the remaining percussive and vocal mixture. Figure 9 is showing the results of filtering and masking for the harmonic (9A) and percussive, vocal (9B) separation after the first, high resolution filtering. The same STFT parameters were used to plot both spectrograms in Figure 9, as both mixtures were separated using the same, high frequency resolution STFT. The separation seems to be relatievely effective after the first filter. There are significant differences in the subfigures A and B, the horizontal lines are far more prominent in Figure 9A, while the vertical lines are mostly filtered out. Figure 9A seems to have more prominent vertical lines, and a much more prominent contribution from the vocals. This suggests that the separation worked to a certain degree, and listening to both signals (after performing an inverse STFT) corroborates that. Both recordings still contain human voice, the harmonic portion contains a faint and muffled vocal part, while the percussive part understandably features a prominent vocal part. This falls short of the performance target after the first filtering technique, but could still provide partial vocal suppression if the following percussive-vocal separation is successful. As explained further on, this was not the case, and the following separation operation was less effective.

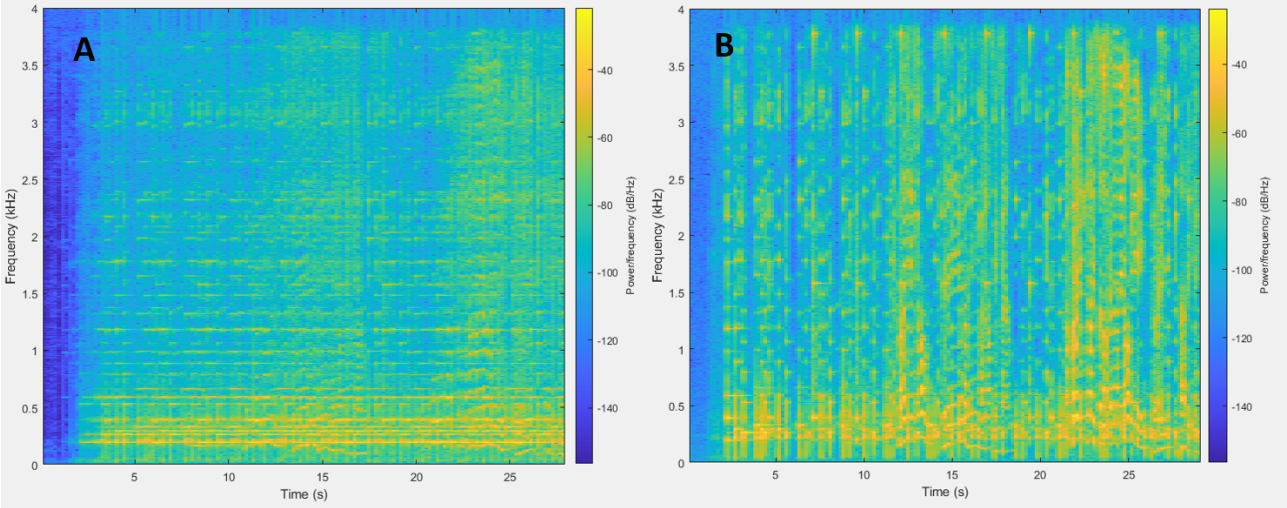

Figure 9: Spectrogram of the first 30 seconds of *Passive-Aggressive* by John Bovey, using a 3072 bin STFT with a moving window of 3072 samples, 1536 sample overlap. A) Harmonic components, B) Percussive and vocal components.

The percussive and vocal mixture was then passed through a low frequency resolution STFT and another set

of median filters and mask operations. This yielded the spectrograms shown in Figure 10. Figure 10A shows a strong attenuation of the vertical lines, and Figure 10B shows more prominent horizontal lines, suggesting that the median filtering itself was implemented correctly. However, listening to both audio signals which produced these spectrograms (after inverting both recordings into time domain) proved that the filtering did not achieve the vocal separation. There is a significant difference in how both recordings sound, but both still contain the vocal part. Moreover, the high frequency resolution harmonic separation, though it also did not manage to fully separate harmonics from the rest of the mixture, came much closer to suppressing the vocal part than the low frequency resolution percussive separation. Combining the harmonic and percussive parts yielded a mixture that still contains the singing voice, this time there is a muffled contribution from the harmonic part and a more clear contribution from the percussive part. The spectrogram of the combined harmonic-percussive mixture, which is the final output of the algorithm is shown in Figure 11.

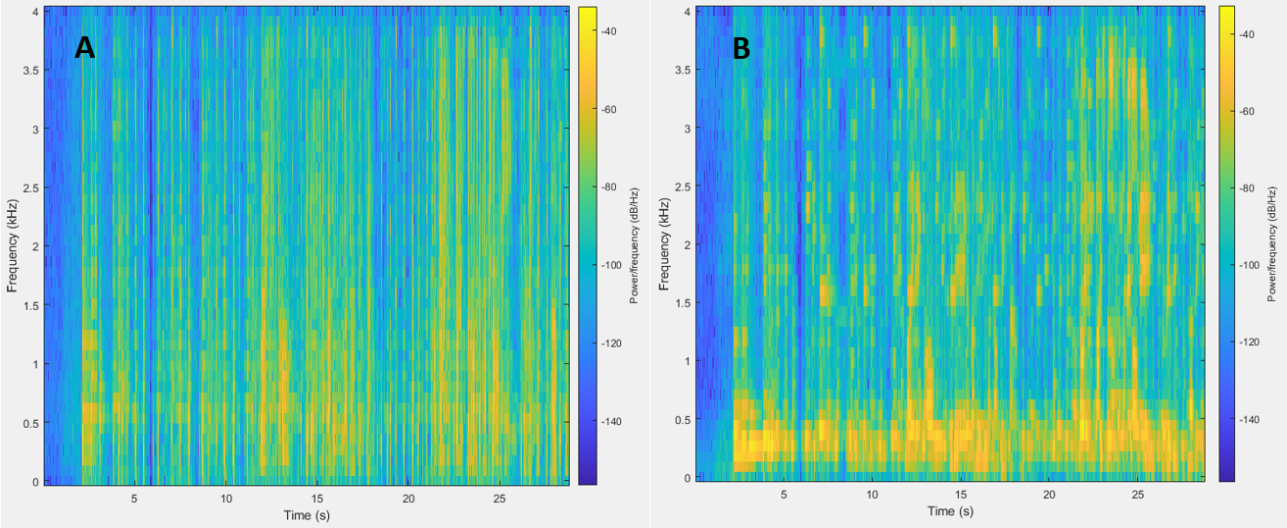

Figure 10: Spectrogram of the first 30 seconds of *Passive-Aggressive* by John Bovey, using a 90 bin STFT with a moving window of 90 samples, 45 sample overlap. A) Attenuated percussive component, B) Vocal component.

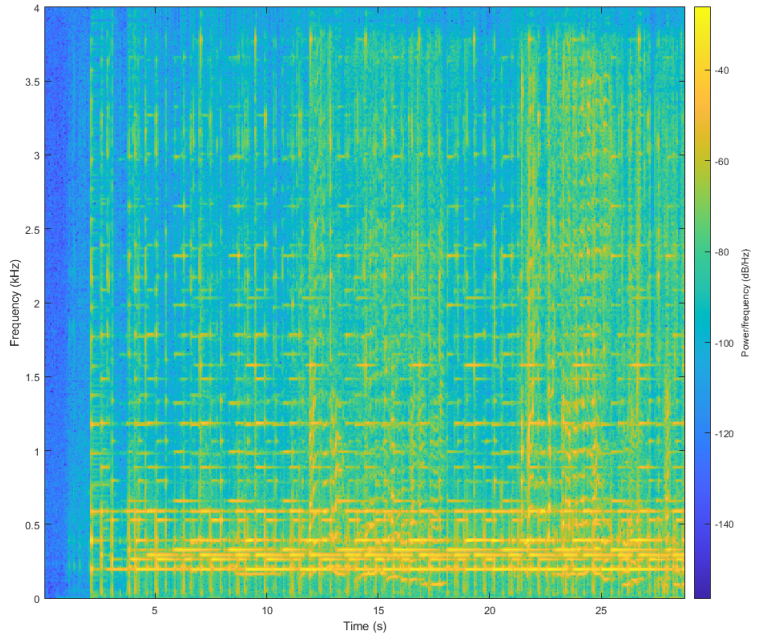

Figure 11: Spectrogram of the final output of the vocals removal algorithm, using the first 30 seconds of *Passive-Aggressive* by John Bovey.

The conclusion was made that though the filtering operations work as expected and the theory behind the STFT is well understood, the technique is not performing the source separation as well as expected based on source material. The input frequency and the setup differs from those used in [1], but the parameters were scaled to reflect those changes. In order to better understand the problems associated with this implementation, the parameters were varied to determine their impact on the quality of separation. Different sizes of the window of analysis and lengths of the median filter frame were used in order to produce the spectrograms in Figure 12. All of the spectrograms shown in Figure 12 come from the harmonic signal output by the first filter only, as it is easier to analyse than the combined mixture. The parameters used are shown in Table 2.

Parameter	Case A	Case B	Case C	Case D
Window length	512	1536	1536	1536
Filter frame length	17	17	10	30

Table 2: Parameters used when producing the spectrograms in Figure 12

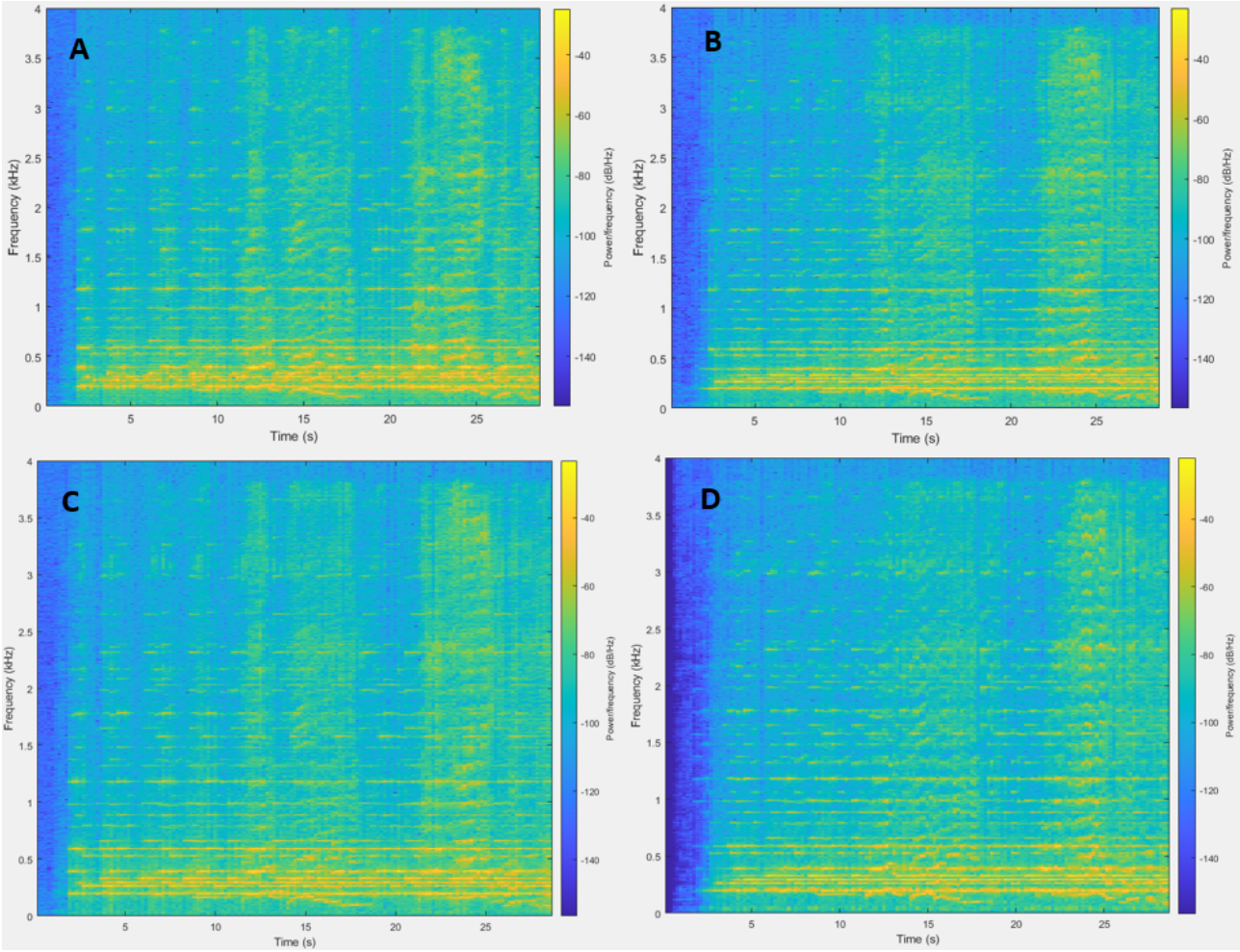

Figure 12: Spectrogram of the final output of the vocals removal algorithm, using the first 30 seconds of *Passive-Aggressive* by John Bovey.

Figure 12B has a higher spectral resolution than 12A, while the opposite can be said about their temporal resolutions. This makes sense as increasing the analysis window length means more frequencies fit into a single window, increasing the spectral resolution. At the same time, using a larger window means that fewer of those windows fit in the time-frame, leading to a smaller temporal resolution. Upon listening to recordings A and B, recording A seemed to have more impurities, and a lower quality separation, but the difference was not overwhelming. This suggests a larger window of analysis is advantageous for this task. Figure 12D seems to show a more aggressive filtering approach, which widened the horizontal lines, but also introduced some blurring into the spectrogram. As mentioned in [1], the filter frame length should affect the quality of separation, with the frame length between 15 and 30 samples providing optimal results. Despite the large difference between filter length samples in 12C and 12D, and the fact that the length in 12C is beyond the stated optimal range,

both harmonic recordings showed few differences when listened to. The filter frame length seems to have affected the separation quality in the spectrograms, but had little audible effect on the audio sample.

Other parameters, such as the order of the time-frequency mask and STFT size in samples were also studied. Changing the mask order had little effect on the audible quality of separation. Increasing the STFT size did have some impact on the quality of separation, but ti was not significant enough to reach a satisfiable end result before violating the constraints of the Nyquist frequency. The size of the STFT must be less than twice the size of the audio track sampling frequency for the model to work. Increasing the size of the STFT also introduced a large delay when testing the model on a PC, which could translate into a delay when the model is deployed on a DSP.

The investigation was largely inconclusive, as no solution to the problem could be found. If the recording was sampled at a higher rate, the STFT size could be increased further, leading to possibly a semi-acceptable quality of separation after the first filter. This might not affect the effectiveness of the second filtering operation (splitting the track into percussive and vocal parts), which means that even if the track was resampled there is no evidence that the technique would work. After the preliminary research and scaling the parameters to suit the 8kHz input, the technique was promising and the computational requirements seemed realistic considering the hardware available. As mentioned above, in order to obtain even a semi-satisfiable result, the spectral resolution of the STFTs would have to be increased, leading to larger matrices and buffer sizes than expected. This would cause significant delays when deploying the algorithm on the DSP. This was not identifiable at the beginning of the project and only became apparent once the algorithm was constructed and tested offline. The source material for the technique [1] mentions additional methods improving the quality of separation, such as non-negative matrix cofactorisation, but these would only increase the computational requirements, while the base technique was meant to give satisfiable results. In order to attempt a DSP implementation of the voice removal algorithm, an alternative and much simpler solution was explored, which is outlined in the next section.

Alternative solution for the DSP

Implementation

The filter design was implemented as a modification of the audio player model provided in the course in figure 13.

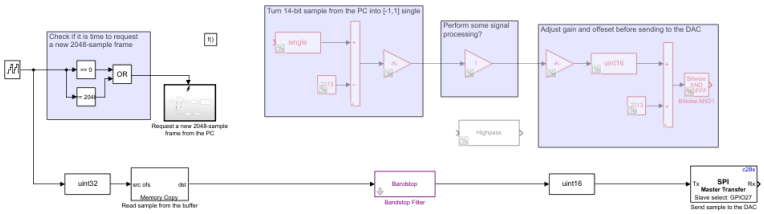

Figure 13: Simulink model of BPF implementation on DSP

The built-in Parks-McClellan optimal FIR filter order estimation was used for the initial design, while the "bandstop" block was used directly to facilitate easier testing and changing of parameters. The design for this solution is a direct-form FIR Band-stop filter in a frame based manner that aims to filter the fundamental frequency range and main harmonics for vocals while maintaining a good sound quality.

With the input sampled at 8000 Hz, the first passband ends at 50 Hz and the stopband begins at 250 Hz, in order to attenuate the low frequency vocals component without completely removing the fundamental frequencies of instruments in similar ranges. The stopband ends at 3600 Hz and the second passband begins at 4000 Hz, while the passband ripple is allowed to be 5 dB and the stopband attenuation is at 30 dB (limited to -30 dB). This design was the backbone to the model submitted for the listening test, with a simpler design that can be readily implemented and validated to show notable difference.

Results and Discussion

Implementing the simpler filter design, the sample tracks showed mostly instruments with fundamental frequency ranges higher than that of vocals and a lack of low-frequency instruments like toms, snares and bass drums. Therefore, a high-pass filter that attenuates frequencies below 800 Hz, with a transition band of 200 Hz, was first tested on the DSP board. It gave thin sound and poor sound quality with a lot of noise introduced. Raising the stopband frequency, eg. to 2000 Hz, gave similar results, with muffled sound and a greatly reduced clarity. Too much of the fundamental frequencies or the main harmonics with concentrated energy of the key instruments

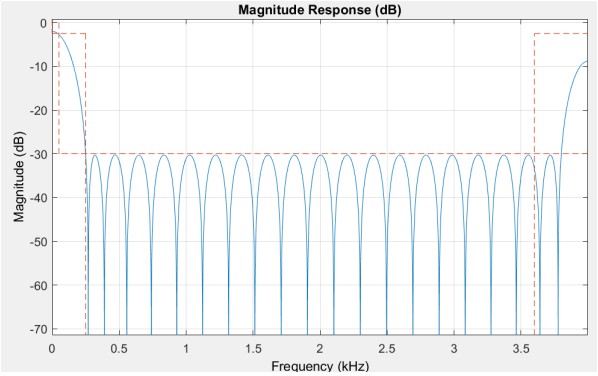

Figure 14: Magnitude response of Implemented BPF filter

might have been removed along with the vocals, whereas the higher harmonics of vocals extend deep into the frequency range. While it is not clearly shown in the tests due to the more complex soundtracks, the significant phase shifts introduced in theory might have been detrimental to the quality of the music, apart from having removed the rumbles and the most important fundamental frequencies, particularly for bassline and instruments like the kick drum.

A filter was then adjusted to do band-stop filtering with the parameters shown in the last section (Alternative solution - Implementation). It was found from testing that the vocals from the sample tracks potentially largely exist outside of the fundamental frequencies and are mixed with other harmonics and fundamental frequencies in the spectrum from guitars, keys and other instruments, where filtering in the fundamental range showed less difference.

For the parameters of the filter, the stopband attenuation was set to be 30 with passband ripples set to be 5 dB. A larger ripple has been allowed for the sake of achieving a steeper roll-off that is desirable, ie. a small transition band, where a small stopband attenuation serves the same purpose. An overly high stopband attenuation significantly muffles the sound, while the narrow transition band shows harsher removal characteristics.

A higher passband ripple gave higher intensity output that was unstable with more noise, where driving it low worsened the clarity at a great extent. A steep roll-off and thus an overly small transition band significantly required a higher order filter that lengthened the calculation time and made the sound quality deteriorate, thus the first passband ends at 50 Hz and the stopband begins at 250 Hz as discussed and preserves some fundamental frequencies. Setting the stopband to end at lower frequencies, eg. 2000 Hz, only filters out some of the instrument tracks slightly with less difference on the vocals, potentially indicating that the frequency range of the vocals (including the higher harmonics) may be wider that the other major instruments in the sample tracks used. As a result, a compromise was made to set the stopband at 3600 Hz and have the passband begin at 4000 Hz, which significantly reduces the magnitude and thus the sound intensity of all signals in the track with notable difference heard from the DSP. To a limited degree, it filtered out a small portion of the human voice, making it thinner and lower in volume compared to other instruments relative to without any filtering at the cost of a lower volume (sound intensity) output. The main melody is occasionally filtered out with missing segments, with the vocals heard in the background.

As shown in figure 15, 'Passive Aggressive' by John Bovey, one of the sample tracks (the red signal on top without any filtering) were used to shed some light on the effect of the bandstop filter (the blue signal at the bottom). Evidently, due to the large stopband from 250 Hz to 3600 Hz, a large portion of signal has been filtered out, while the lower frequency range below 300 Hz (including the transition band) has been largely unchanged. Hence, this meets the expected "band stop" behavior that preserves the sound made by low frequency instruments. This also explains the overall much lower sound intensity and volume of the processed output, which was inevitable with a simple filter as the aim was to remove the vocals to a notable degree that was found to span a wide range of frequencies. The performance was not optimal, as the main instruments' sounds were seen to be removed to a large degree as well due to the overlap and producing a notable difference in the vocals heard was the priority. From actual testing, only the main melody and some of the vocals could be still heard after filtering. Nevertheless, this filter implementation was able to meet the design specification of a less than 2 second delay without long bufferings, hence enabling processing in real time.

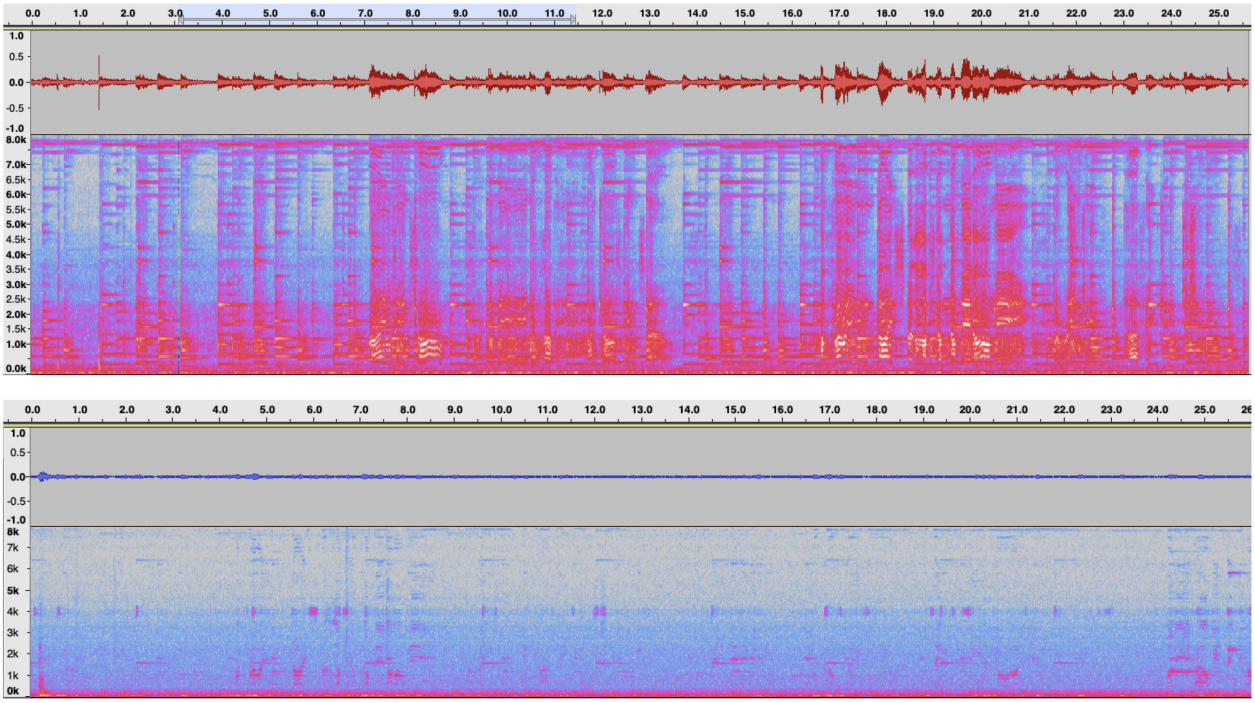

Figure 15: Original Soundtrack (top in red), bandstop filter implementation on music filtering (bottom in blue)

The results vary across samples, for instance slightly better filtering was achieved for the male vocals, eg. "Passive Agressive by John Bovey", " Capulet by The Rope River Blues Band" potentially due to more distinct frequency ranges and the overall lower fundamental frequency and overtone range of male voices. Also, singers, usually for opera singing and concert performance, often are shown to give a spectral peak at 2-4 kHz, ie. "singers' formant" [4], that is more sensitive and best perceptible to human ears. While this may be slightly different for pop songs, usually with flat harmonics, this difference in singing style with different "richness" in high or low overtones adds to the challenge in designing a simple filter for different song styles without a great compromise in sound clarity.

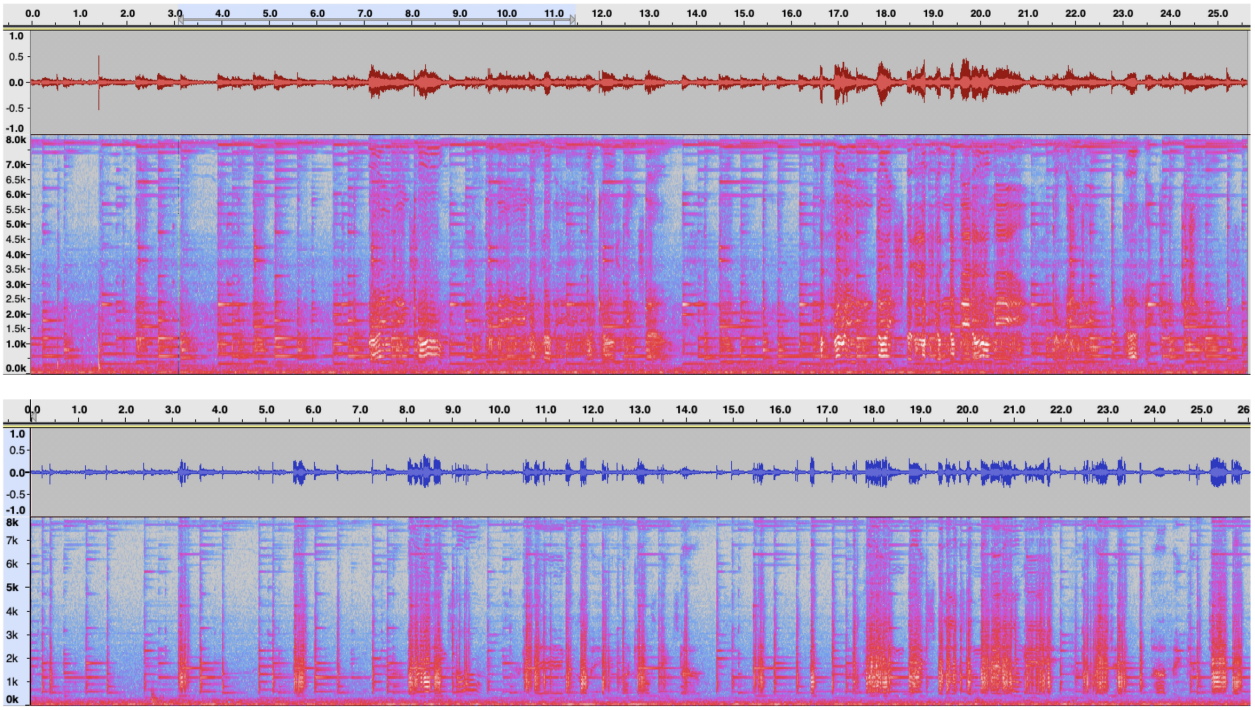

Figure 16: Original Soundtrack (top in red), Small stopband bandstop filter implementation on music filtering (bottom in blue)

Moreover, if removing the vocals was not the solely prioritised objective, a smaller stopband, ie. from 250 Hz to 2000 Hz (passband starting at 2300 Hz) could be employed to give more details on the instrument sounds. The filter with a smaller band was thus tested on the same track 'Passive Aggressive'. As shown in figure 16, more fundamental frequencies and also higher harmonics were preserved, where the degree of filtering was shown to differ across time instances potentially due to the processing. Due to the transition band being smaller, more signal and noise were passed through, which also kept more of the sounds made by instruments in the similar frequency range as the vocals. The vocals, however, could be perceived more clearly with much less difference as a result. All other parameters were kept the same.

Butterworth filter implementation has been attempted using the same filter design approach to hopefully get maximally flat response with no ripple expected to output a similar sound quality, as well as a Chebyshev type II filter that sets the stopband attenuation and removes passband ripples. Nevertheless, in actual implementation, to give a good enough response from the implementations derived for these filters, the DSP could not process the filter coefficients in time smoothly, likely due to the fact that they are IIR filters. Approximating them by truncating the infinite impulse response with windowing functions, thus as FIR Filters, only gave similar results and therefore the previous implementation was used.

In conclusion, with the available DSP board and limited calculation power, the alternative solution could only provide limited results in terms of removing the human voice and inevitably attenuates the heavily overlapping instrument tracks. In theory, the proposed solution should be feasible after adjusting the parameters from the preliminary report based on the feedback and able to give promising results, which eventually could not be successfully implemented due to the calculation power and other factors discussed despite repeated attempts and trials.

References

[1] D. FitzGerald and M. Gainza, "Single channel vocal separation using median filtering and factorisation techniques," 2010.

[2] S. G. Blythe, P. Beuret, and P. Blythe, "Attention, balance and coordination," *The ABC of learning success. JohnWiley & Sons Ltd, Chichester. Hallahan, Daniel P. Lloyd, John and Kauffman, James. M Weiss, Margaret. P. Martinez, Elizabeth. A.(2005). Learning Disorders (Basics, Characteristics, and Effective Teaching). Translated by Alizadeh, Hamid*, 2009.

[3] "Base frequency range," 2020, [Online; accessed 19-March-2021]. [Online]. Available: https://www.studybass.com/gear/bass-tone-and-eq/bass-frequency-range/

[4] J. Wolfe, M. Garnier, and J. Smith, "Voice acoustics: an introduction," *Retrieved June*, vol. 3, p. 2011, 2010.

Appendix

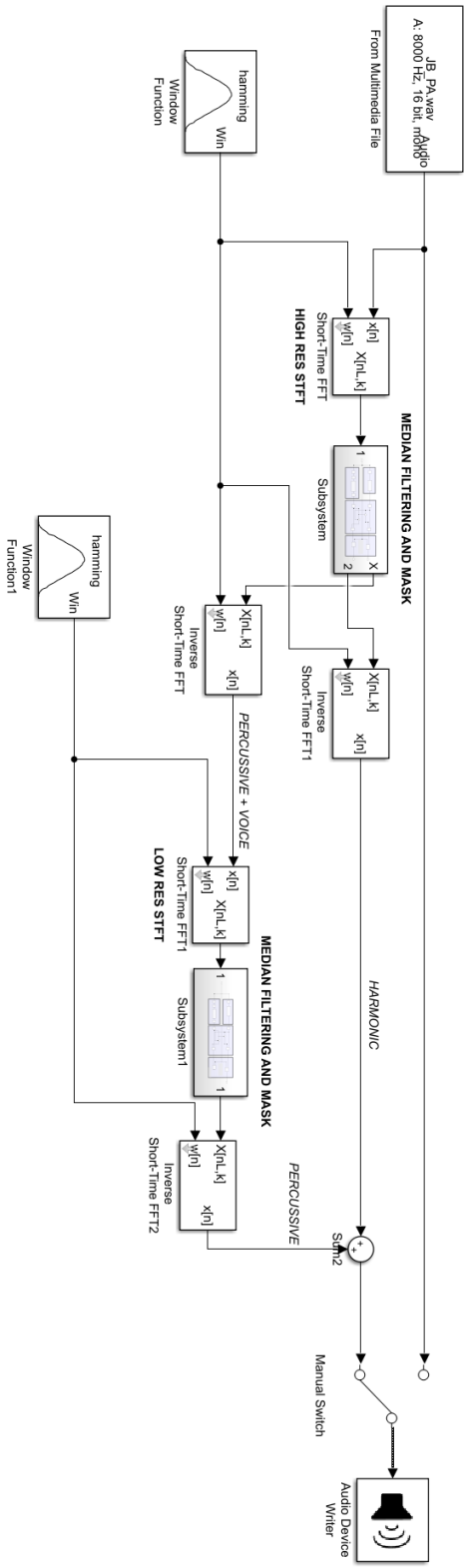

Figure 17: Enlarged version of Figure 4

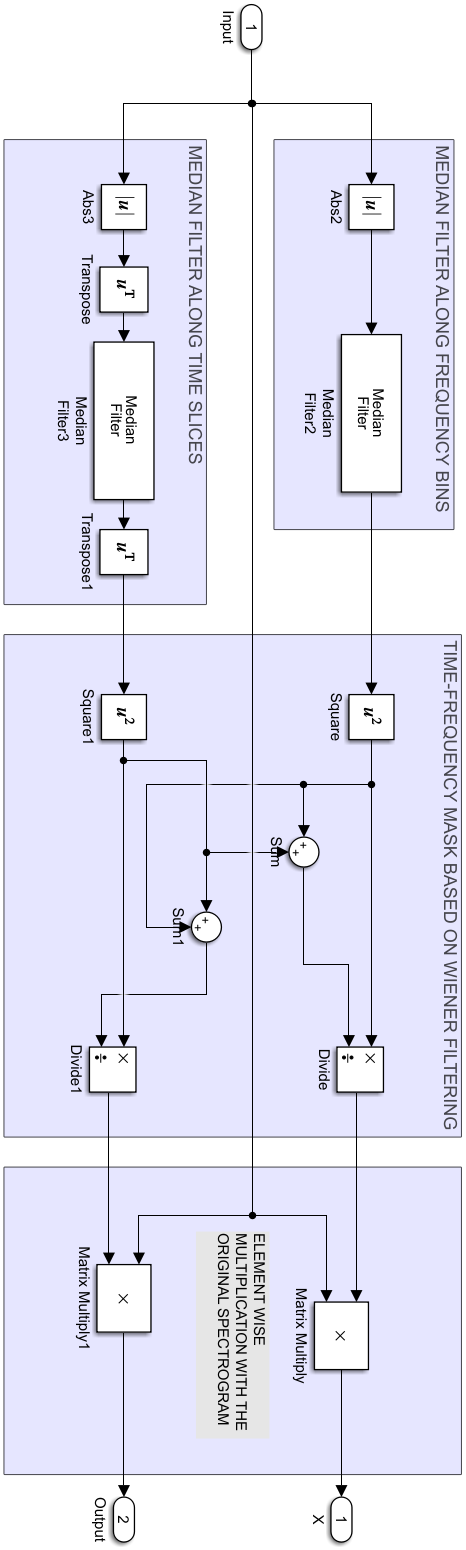

Figure 18: Enlarged version of Figure 5