Literature Review of Signal Processing Techniques for Audio-to-Sheet Music Transcription

Team #7, Tune Goons
Emily Perica
Ian Algenio
Jackson Lippert
Mark Kogan

April 4, 2025

Contents

1 Introduction

This report outlines the mathematical techniques considered, implemented, and utilized in the final version of our WAV analysis project. The project aimed to robustly determine note durations and generate MusicXML, with an ultimate goal of creating PDF sheet music. Our primary approach to pitch detection was autocorrelation, while note onsets were determined using both peak detection in energy and spectral energy differences derived from the Short-Time Fourier Transform (STFT).

2 Techniques Utilized in the Final Project

These mathematical methods were actively employed in the final implementation:

2.1 Mono Conversion via Stereo Averaging

Since WAV files may contain stereo channels, we converted them to mono by averaging the left and right channels, ensuring a unified signal for processing.

2.1.1 How Stereo Averaging Works

Stereo signals contain two independent channels, often representing left and right audio. To process them as a single signal, we compute their average:

$$x_{mono}(n) = \frac{x_{left}(n) + x_{right}(n)}{2} \tag{1}$$

This preserves the overall amplitude structure while ensuring consistency in pitch detection and onset analysis.

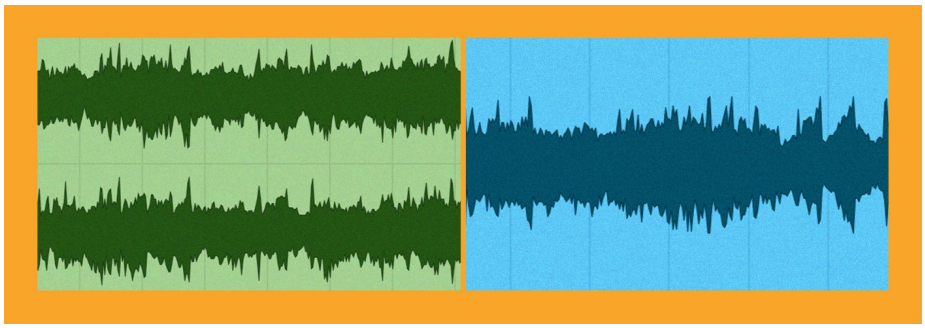

Figure 1: Conversion of Stereo Audio to Mono-Audio by averaging channels [1].

2.2 Autocorrelation for Pitch Detection

Autocorrelation was used to determine the fundamental frequency of the audio signal by analyzing periodic patterns. This method proved effective for identifying stable pitches but had limitations when dealing with fast note transitions.

2.2.1 How Autocorrelation Works

Autocorrelation is a time-domain signal processing technique used to detect periodicity in a signal [2]. It involves multiplying a signal by a time-shifted version of itself and summing the result over all possible shifts [2, 3]. The resulting function exhibits peaks at integer multiples of the signal's fundamental period, allowing the estimation of the dominant frequency. This method is particularly useful in pitch detection, as musical notes exhibit periodic waveforms.

Mathematically, the autocorrelation function $R(\tau)$ of a discrete signal $x[n]$ is defined as:

$$R(\tau) = \sum_{n=0}^{N-1} x[n]x[n+\tau] \tag{2}$$

where τ represents the lag and N is the signal length.

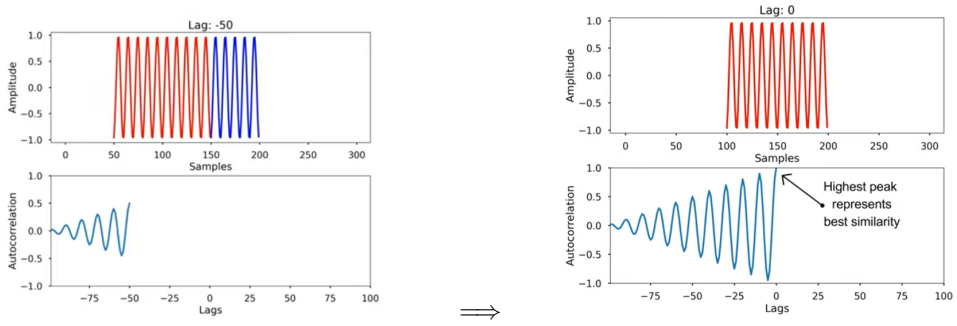

Figure 2: Signal peak occurs when offset signal is most similar to original signal [4].

2.3 Peak Detection in Energy for Note Onsets

By tracking changes in energy levels over time, we detected note onsets. Peaks in energy served as markers for the start of new notes, a fundamental aspect of transcription.

2.3.1 How Peak Detection Works

Peak detection involves identifying significant local maxima in a signal [5]. This is commonly done by analyzing the derivative of an energy envelope and applying a threshold-based criterion to filter out minor fluctuations. The energy envelope can be computed using the root mean square (RMS) or the sum of squared amplitudes within a short sliding window.

Mathematically, energy $E(n)$ of a signal in a window of length M is given by:

$$E(n) = \sum_{m=0}^{M-1} x^2[n-m] \tag{3}$$

Peak detection then involves identifying points where $E(n)$ surpasses an adaptive threshold.

2.4 STFT for Spectral Analysis and Onset Detection

We employed a Short-Time Fourier Transform (STFT) to generate spectrograms and analyze spectral energy changes. Differences in spectral energy were used as an alternative onset detection method, complementing the energy peak detection approach.

2.4.1 How STFT Works

The STFT is a method of analyzing the frequency content of a signal over time by applying the Fourier Transform to short, overlapping segments of the signal [6]. It provides a time-frequency representation, allowing the observation of spectral variations that correspond to note onsets [7].

Mathematically, the STFT of a signal $x(t)$ with a window function $w(t)$ is defined as:

$$X(t,f) = \int_{-\infty}^{\infty} x(\tau) w(\tau - t) e^{-j2\pi f \tau} d\tau \tag{4}$$

where $w(t)$ is typically a Hanning or Hamming window to reduce spectral leakage.

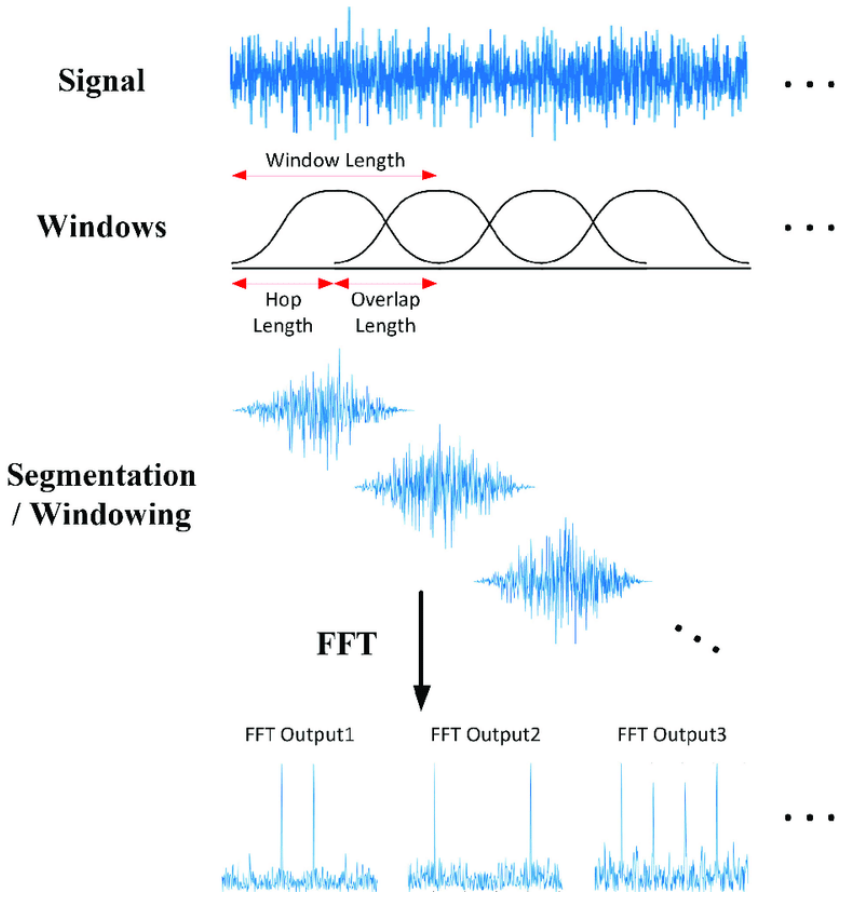

Figure 3: Breakdown of STFT Process used to construct spectrogram [8].

2.5 Hanning Window

Both Hanning [9] and Hamming [7] windows were tested as potential smoothing functions for spectral analysis. They were used in different configurations of STFT, though ultimately, the Hanning functino was chosen based on performance.

2.5.1 How Windowing Functions Work

Window functions taper the edges of a signal segment before applying Fourier analysis, reducing spectral leakage. The Hanning window is defined as:

$$w(n) = 0.5 - 0.5\cos\left(\frac{2\pi n}{N-1}\right) \tag{5}$$

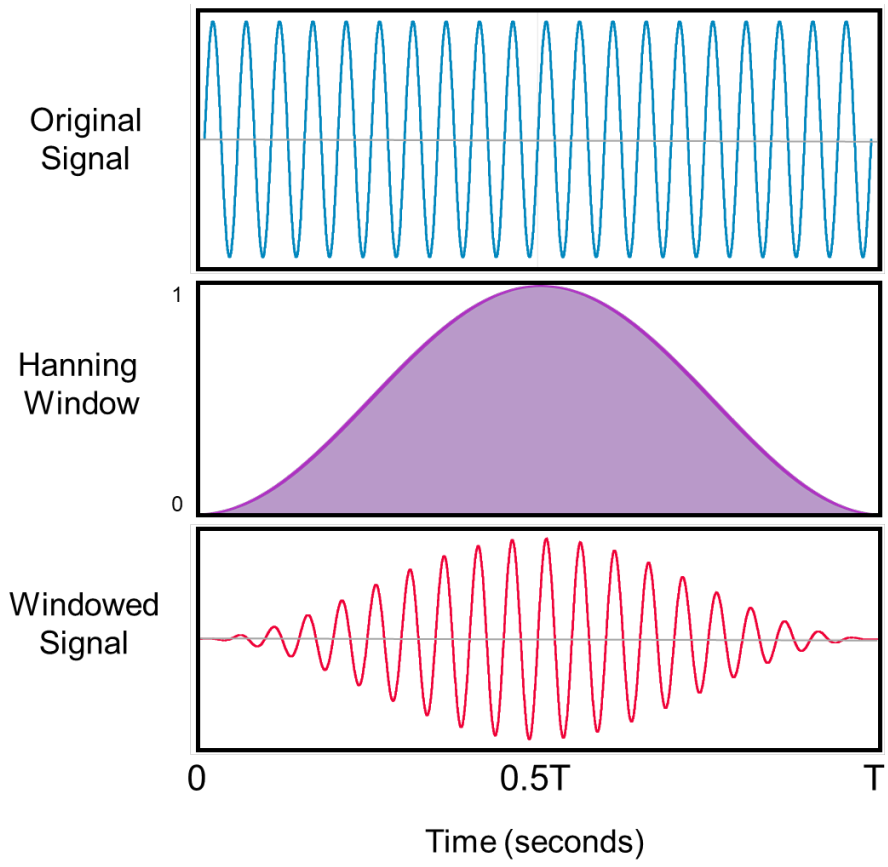

Figure 4: Hanning function used to smooth out signal onsets [10].

2.6 Krumhansl-Schmuckler Key-Finding Algorithm

This algorithm was used for determining the key signature of a musical piece based on pitch class profiles.

2.6.1 How the Key-Finding Algorithm Works

The Krumhansl-Schmuckler algorithm matches pitch distributions in an audio signal to predefined key profiles, computing a correlation score for each possible key [11]. The key with the highest correlation is selected.

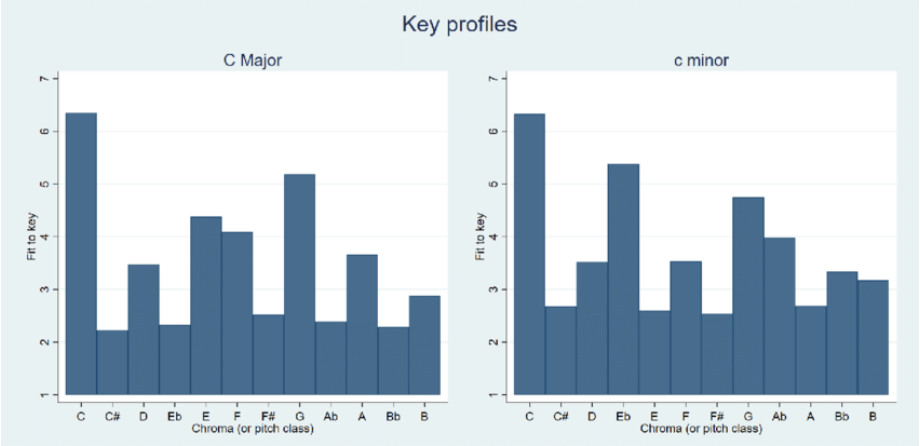

Figure 5: Expected Relative Note Frequency profiles of C and C-Minor key, used in Krumhansl-Schmukler algorithm to create correlation scores [12].

3 Techniques Considered but Not Implemented

Several mathematical techniques were considered but ultimately not implemented in any phase of development.

3.1 Cross-Correlation

Cross-correlation was evaluated as a potential enhancement for pitch detection and time-domain alignment.

3.1.1 Mathematical Definition

Given two signals $x(t)$ and $y(t)$, their cross-correlation is defined as:

$$(R_{xy} * x)(\tau) = \int_{-\infty}^{\infty} x(t)y(t+\tau)dt. \tag{6}$$

This operation measures the similarity of $x(t)$ and a time-shifted version of $y(t)$. In the discrete case, for signals sampled at intervals n, it is given by:

$$R_{xy}[k] = \sum_{n} x[n]y[n+k]. \tag{7}$$

Cross-correlation is extensively used in time-delay estimation and pattern recognition. The Fast Fourier Transform (FFT) can accelerate its computation through:

$$R_{xy} = \mathcal{F}^{-1}\{X^*(f)Y(f)\}, \tag{8}$$

where $X(f)$ and $Y(f)$ are the Fourier transforms of $x(t)$ and $y(t)$, and $X^*(f)$ is the complex conjugate of $X(f)$.

Could have been used for improved time alignment of detected note events, reducing timing inaccuracies in sheet music generation.

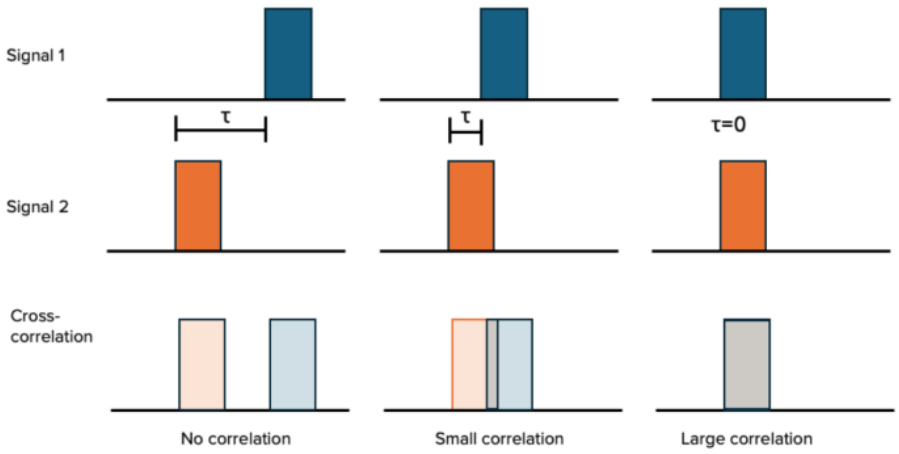

Figure 6: Diagram highlighting Cross-Correlation functionality. Similar to Auto-correlation but used for different signals, rather than offsets of same signal [13].

3.2 Wavelet Denoising

Wavelet denoising is a signal processing technique that uses wavelet transforms to separate a signal into different frequency components and apply thresholding to reduce noise. It is particularly effective for denoising non-stationary signals, such as audio, where the noise is typically localized in time or frequency.

3.2.1 How Wavelet Denoising Works

Wavelet denoising involves three main steps:

1. **Wavelet Transform**: The signal is decomposed into a series of wavelet coefficients using a discrete wavelet transform (DWT) or continuous wavelet transform (CWT). The DWT splits the signal into **approximation** coefficients (low frequencies) and **detail** coefficients (high frequencies) at multiple scales.

Mathematically, the DWT of a signal $x(t)$ is represented as:

$$x(t) = \sum_{j} \sum_{k} c_{jk} \psi_{jk}(t)$$

where $\psi_{jk}(t)$ are the wavelet functions at scale j and position k, and c_{jk} are the wavelet coefficients.

2. **Thresholding**: The wavelet coefficients corresponding to the high-frequency noise are identified and thresholded. The thresholding technique reduces or removes coefficients that correspond to noise while preserving those that correspond to signal features. There are two types of thresholding:
- **Hard Thresholding**: Coefficients smaller than a set threshold are set to zero.
- **Soft Thresholding**: Coefficients smaller than a threshold are shrunk towards zero, reducing the magnitude of the larger coefficients while eliminating smaller ones.

Mathematically, soft thresholding is given by:

$$\hat{c}_{jk} = \text{sign}(c_{jk}) \cdot \max(0, |c_{jk}| - \lambda)$$

where λ is the threshold value.

3. **Inverse Wavelet Transform**: After thresholding, the signal is reconstructed by applying the inverse wavelet transform, which combines the modified coefficients back into the time domain to produce the denoised signal.

3.2.2 Advantages of Wavelet Denoising

- Wavelet denoising is highly effective at removing localized, high-frequency noise while preserving important signal features like sharp transients, which is important for tasks like music transcription. - The method can be adapted to

different levels of noise by adjusting the thresholding parameters and using different wavelet functions for specific types of signals.

Could have been used for denoising input audio, effectively reducing high-frequency noise that could interfere with accurate note onset detection and improving the clarity of secondary notes for transcription.

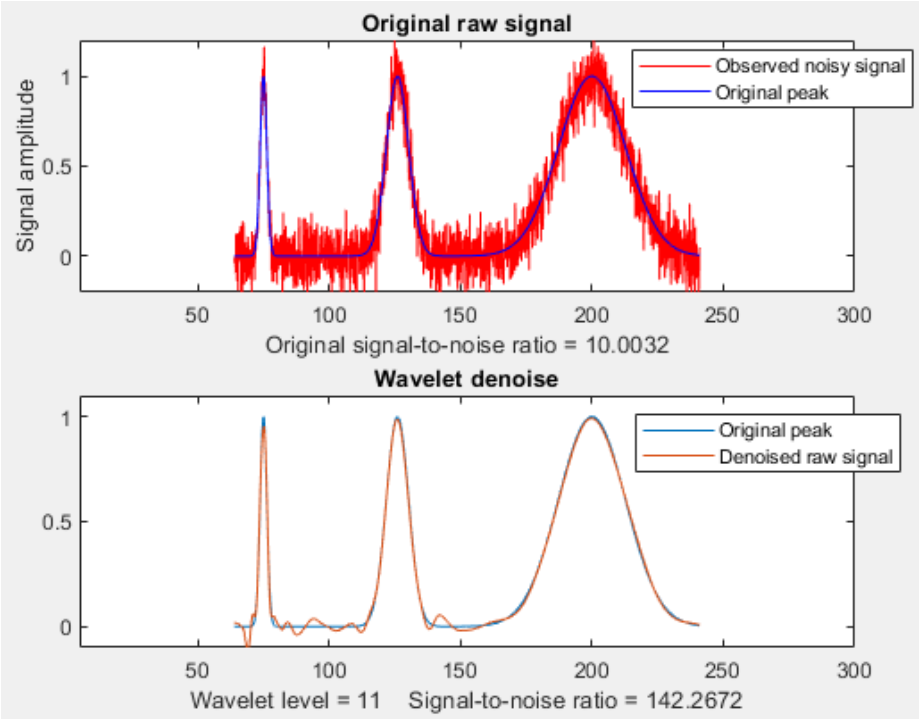

Figure 7: Signal Denoising via soft wavelet thresholding [14].

3.3 YIN Algorithm for Pitch Detection

The YIN algorithm [15], a robust pitch detection technique, was considered.

3.3.1 Mathematical Basis

YIN refines autocorrelation by computing a difference function:

$$d(\tau) = \sum_{t=1}^{T} (x_t - x_{t+\tau})^2. \tag{9}$$

A cumulative mean normalization step is then applied:

$$\tilde{d}(\tau) = d(\tau) / \left(\frac{1}{\tau} \sum_{j=1}^{\tau} d(j) \right). \tag{10}$$

Pitch corresponds to the minimum of $\tilde{d}(\tau)$. This improves over standard autocorrelation by reducing octave errors and improving resolution at low frequencies.

Could have been used to refine pitch estimation, reducing frequency detection errors and improving note transcription accuracy.

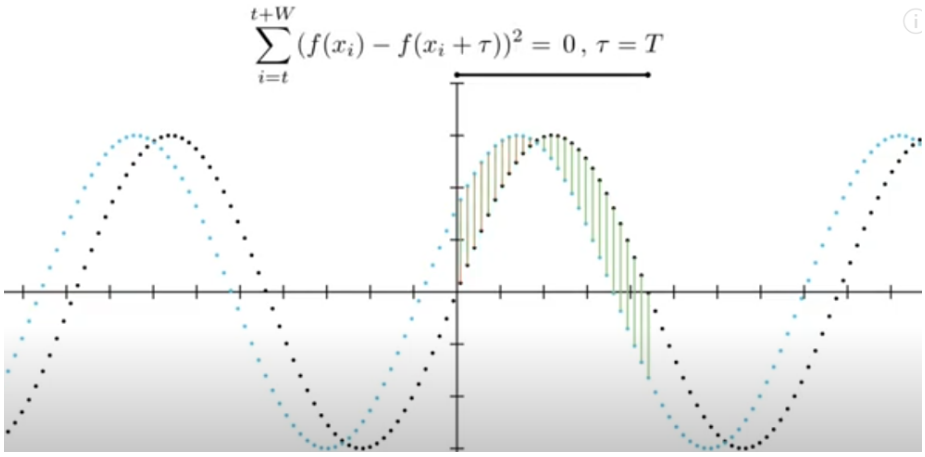

Figure 8: Difference in signal offsets calculated to determine cumulative mean difference, allowing better handling of noisy inputs or harmonics [16].

4 Conclusion

This project employed a combination of signal processing techniques to achieve accurate pitch detection and onset analysis. While several methods were considered, the final implementation balanced computational efficiency with accuracy. Future improvements could explore more advanced onset detection mechanisms and integrate additional filtering techniques to refine transcription accuracy.

References

[1] iZotope, "Mono vs. stereo in audio mixing," https://www.izotope.com/en/learn/mono-vs-stereo.html, n.d.

[2] L. Rabiner and R. W. Schafer, "On the use of autocorrelation analysis for pitch detection," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 1, pp. 24–33, 1977.

[3] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[4] S. Technion, "A visualization of the autocorrelation function," https://www.youtube.com/watch?v=uf679Qo-bB4, n.d.

[5] R. Zhou and J. D. Reiss, "Music onset detection," in *Machine Audition: Principles, Algorithms and Systems*, W. Wang, Ed. IGI Global, 2010, pp. 297–316.

[6] E. Benetos, A. Holzapfel, and Y. Stylianou, "Pitched instrument onset detection based on auditory spectra," in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*. Crete, Greece: ISMIR, 2009, institute of Computer Science, FORTH, Greece, and Multimedia Informatics Lab, Computer Science Department, University of Crete, Greece.

[7] S. Dixon, "Onset detection," in *Proc. of the 9th Int. Conference on Digital Audio Effects (DAFx'06)*, 2006.

[8] F. 1, "Short-time fourier transform (stft) overview," https://www.researchgate.net/figure/Short-time-Fourier-transform-STFT-overview_fig1_346243843, n.d.

[9] ScienceDirect, "Hanning window," https://www.sciencedirect.com/topics/engineering/hanning-window#:~:text=The%20Hanning%20window&text=The%20general%20shape%20obviously%20gradually,scallop%20loss%20is%20also%20lower., accessed: 2025-04-03.

[10] S. D. I. S. Community, "Overlap: What, why and how to use it," https://community.sw.siemens.com/s/article/Overlap-What-Why-and-How-to-use-it, n.d.

[11] D. Temperley, "What's key for key? the krumhansl-schmuckler key-finding algorithm reconsidered," http://davidtemperley.com/wp-content/uploads/2015/11/temperley-mp99.pdf, 1999, accessed: 2025-04-03.

[12] E. M. Review, "Tonal and "anti-tonal" cognitive structure in viennese twelve-tone rows," https://ojs.library.osu.edu/index.php/EMR/article/view/7655/5748, n.d.

[13] Jball, "What is cross-correlation, and how does it advance spectrum analysis?" https://liquidinstruments.com/blog/cross-correlation-and-spectrum-analysis/, 2025, 2025, February 4.

[14] I. to Signal Processing, "Wavelets and wavelet denoising," https://terpconnect.umd.edu/~toh/spectrum/wavelets.html, n.d.

[15] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 112, no. 4, pp. C500–C505, 2002.

[16] V. for Science, "Detecting pitch automatically - the intuition behind the yin pitch detection algorithm," https://www.youtube.com/watch?app=desktop&v=W585xR3bjLM, n.d.-a.