Module 3: Audio Signal Processing

In this module, we delve into techniques for pitch estimation and manipulation, as well as methods for vocal processing, which are crucial in many areas of audio engineering, including music production, post-production, and sound design.

- **Semitones:** A semitone, also known as a half step or half tone, is the smallest standard interval used in Western music and music theory. It is the fundamental building block of the chromatic scale, which is made up of 12 semitones.

- **Chromatic Scale**: This scale includes all notes in an octave, each a semitone apart. For example, moving from C to C# (or Db) is a movement of one semitone.

- **In Audio Processing**: When pitch shifting or scaling in audio processing, adjusting by "semitones" involves increasing or decreasing the pitch by these half-step increments. Raising the pitch by one semitone means increasing the frequency to the next note in the chromatic scale, and lowering by one semitone means decreasing the frequency to the previous note.

- **Pitch**: Pitch refers to the perceived frequency of a sound; it is how high or low a sound seems to a listener. In technical terms, pitch relates to the frequency of the sound waves:
    - High-pitched sounds have high frequencies (more cycles per second).
    - Low-pitched sounds have low frequencies (fewer cycles per second).

- **Frequency:** Frequency is the number of vibrations (or cycles) per second of a sound wave. It is measured in Hertz (Hz). Frequency directly correlates with the pitch perceived by the human ear; higher frequencies sound higher in pitch, while lower frequencies sound lower.

- **Octave:** An octave is a series of eight notes occupying the interval between (and including) two musical pitches, one having twice the frequency of vibration of the other. In Western music, an octave is divided into 12 semitones. Doubling the frequency of any note will take you up by one octave (e.g., from A4 at 440 Hz to A5 at 880 Hz).

- **Time Stretching:** Time stretching refers to the process of changing the speed or duration of an audio signal without affecting its pitch. This is used in various audio applications, such as making music tracks fit a specific time segment or slowing down a piece for practice purposes.

- **Resampling:** Resampling is changing the sample rate of an audio file. This can involve increasing the number of samples per second (upsampling) or reducing it (downsampling). In pitch shifting, resampling is used after time stretching to bring the modified signal back to its original sample rate, thereby changing the pitch but maintaining the original duration.

➢ Pitch Estimation and Manipulation

What is Pitch?

Pitch refers to the human perception of the frequency of a sound. It is how we interpret the highness or lowness of a tone, based primarily on the frequency of the sound waves. Higher frequencies correspond to higher pitches and vice versa. While frequency is a physical and measurable attribute of sound waves, pitch is a perceptual property and can be influenced by factors such as sound intensity and the surrounding acoustic environment.

Why is Pitch Estimation Important?

Pitch estimation is critical in many fields and applications, including:

1. **Music Technology**: In music production and engineering, accurate pitch analysis helps in tuning instruments, modifying vocal tracks, and harmonizing music.
2. **Speech Processing**: Understanding pitch is crucial for speech recognition systems, speaker identification, and linguistic analysis, as it can convey intonation, stress, and emotion in speech.
3. **Telecommunications**: Pitch estimation can enhance the clarity and intelligibility of voice communications in devices like mobile phones and teleconferencing systems.
4. **Healthcare**: In medical fields, pitch analysis of vocal sounds can be used in the diagnosis and monitoring of conditions that affect the voice or breathing, such as asthma or Parkinson's disease.

How These 4 Methods Estimate Pitches:

1. **Autocorrelation Method**:
    - **Principle**: This method calculates the correlation of a signal with delayed versions of itself to identify periodicity.
    - **Process**: It computes the autocorrelation function and identifies peaks that indicate the fundamental frequency. The first significant peak after the zero-lag peak usually corresponds to the pitch period.
    - **Usefulness**: Effective for signals with clear, periodic components, such as musical tones.
    - **Functionality**:
        - Autocorrelation measures the similarity of a signal with its delayed versions to identify periodicity.
            - **Process**: It involves calculating the autocorrelation function

$$R(\tau) = \sum_{n} s(n)s(n+\tau)$$

        - Peaks in the autocorrelation function indicate the presence of a periodic component, where the first significant peak beyond zero lag corresponds to the fundamental period (inverse of the fundamental frequency).
        - This method is very effective for signals with clear and consistent periodicity, such as musical notes.

2. **FFT (Fast Fourier Transform) Method**:
    - **Principle**: FFT transforms a time-domain signal into its frequency components, revealing the spectrum of frequencies present.
    - **Process**: By analyzing the magnitude spectrum obtained from FFT, the method identifies peaks which correspond to the fundamental frequency and its harmonics.
    - **Usefulness**: Useful for analyzing complex sounds containing multiple frequencies, but it can be susceptible to errors from harmonics or noise.

    - **Functionality**:

        - FFT converts a time-domain signal into its frequency components, unveiling the spectrum.
        - **Process**: Compute the FFT of the signal to obtain a complex spectrum, then calculate the magnitude spectrum from it.

$$S(f)=FFT(s(n))$$

        - Identify the peaks in the magnitude spectrum to infer the fundamental frequency. This often involves looking for the highest peak that corresponds to the fundamental frequency or using harmonic analysis to deduce it from the relationship between harmonics.
        - FFT is beneficial for analyzing complex sounds containing multiple frequencies but may require windowing techniques to improve resolution and reduce spectral leakage.

3. **LPC (Linear Predictive Coding) Method**:
    - **Principle**: LPC models the vocal tract as a series of filters and predicts future samples of a signal based on a linear combination of past samples.
    - **Process**: It calculates LPC coefficients to minimize prediction error, and the residual signal (error) is analyzed, often using autocorrelation, to estimate the pitch.
    - **Usefulness**: Particularly effective for speech signals, as it efficiently models the formants and harmonics of human speech.

- **Functionality**: In LPC, the signal s(n)s(n)s(n) is estimated by a linear function:

$$\widehat{S(n)} = \sum_{i=1}^{p} a_i s(n-i) + Gz(n)$$

where a_i are the predictor coefficients, p is the order of the predictor, G is the gain, and z(n) is the input to the prediction filter.

- The goal is to minimize the prediction error, which is the difference between the actual signal and its predicted value.
- LPC is particularly effective in encoding speech for transmission, reducing data requirements by efficiently modeling the formants of human speech.

4. **YIN Algorithm**:
   - **Principle**: Tailored for robust pitch detection, especially in musical contexts.
   - **Process**: YIN computes a difference function that measures the difference between the signal and its delayed versions. It then refines this into a cumulative mean normalized difference function (CMND) to detect the fundamental pitch reliably.
   - **Usefulness**: Known for its accuracy and resistance to noise, making it ideal for music where other methods may fail due to complexity or distortion.
   - **Functionality**:
     - YIN is tailored for robust pitch detection in musical pitch analysis.
     - **Process**: First, it computes a difference function that measures the squared difference between the signal and its delayed versions for various lags.
     - Then, it computes a Cumulative Mean Normalized Difference Function (CMND) from the difference function.

$$d'(t) = \frac{d(t)}{\frac{1}{t}\sum_{j=1}^{t} d(j)}$$

- The pitch is estimated by identifying the first local minimum in the CMND curve, which indicates the best lag corresponding to the fundamental period.
- YIN is effective in situations where other algorithms fail due to its robustness against noise and the presence of harmonics.

➢ Pitch Scaling and Shifting: Techniques and Tools

Pitch scaling and shifting are essential techniques in audio processing, particularly in music production, post-production, and vocal training. These methods allow for adjusting the pitch of audio samples while preserving other qualities such as tempo or timbre.

Pitch Scaling

- **Definition**: Pitch scaling involves changing the pitch of an audio sample without affecting its playback speed or duration.
- **Application**: This is especially useful in music education, where musicians might want to practice a piece in a different key without changing the tempo. It's also used in music production to match the key of different samples or tracks.
- **Example**: If a singer recorded a song in the key of C major and it needs to be adjusted to B major, pitch scaling can shift the entire performance down by two semitones without altering the song's duration.

Pitch Shifting

- **Definition**: Pitch shifting transposes the pitch of an audio file either higher or lower, while maintaining the same speed.
- **Application**: This is useful for key changes in songs or creating harmonic layers. Producers often use pitch shifting to fit instrumental loops into a new key or to create vocal harmonies from a single vocal line.
- **Example**: In a studio setting, if a backing vocal needs to harmonize with the lead but was originally sung at the same pitch, pitch shifting can be used to raise the backing vocal by a third or a fifth, creating a harmony without recording new vocals.

➢ Vocal Processing

1. Vocal Removal Techniques: Phase Cancellation, Spectral Editing

Vocal removal is used to isolate or suppress vocals from music tracks, commonly for creating karaoke tracks or remixes.

Methods:

- **Phase Cancellation**
  - **Concept**: This technique exploits the stereo nature of most music tracks, where vocals are typically mixed equally into both the left and right channels. By subtracting one channel from the other, any sound that is identical in both channels can be effectively canceled out.
  - **How It Works**:
    - In a stereo track, if the left channel is L and the right channel is RRR, and if vocals are identical on both tracks, the subtraction L−R will cancel out the vocals.
    - **Equation**:

      $$\text{Output} = L - R$$

    - This method works best when the vocals are perfectly centered and other instruments are panned to the sides. It's less effective if the vocal reverb or other effects spread differently across the stereo field.
- **Spectral Editing**
  - **Concept**: This method involves identifying and manipulating specific frequency bands where the vocals predominantly reside.
  - **How It Works**:
    - Using a spectral editor or an equalizer, specific frequency ranges can be attenuated or boosted. Since human vocals typically occupy a certain range of frequencies (usually between 300 Hz to 3400 Hz), these can be specifically targeted.
    - **Process**:
      1. Analyze the spectrum of the audio track to identify vocal frequencies.
      2. Apply a notch filter or an equalization cut around these frequencies to reduce or remove vocal presence.
    - This method allows for more precise control compared to phase cancellation and can be used even on mono tracks or tracks where phase cancellation is ineffective.

2. Temporal Separation: Differentiating Between Similar Sounds

Temporal separation involves distinguishing and isolating sounds based on when they occur or their duration and rhythmic characteristics within an audio track.

Methods:

- **Gating**
  - **Concept**: Gating is a dynamic audio processing technique used to control the volume of an audio signal by setting a threshold level. Sounds that are below this threshold are significantly attenuated or muted.
  - **How It Works**:
    - A gate allows signals that exceed a preset amplitude threshold to pass through unaffected, while signals that fall below this threshold have their amplitude reduced.
    - **Equation**:

$$Output = \begin{cases} Input & if \quad Input \geq Threshold \\ 0 & if \quad Input \leq Threshold \end{cases}$$

    - This technique is particularly useful for removing background noise or isolating sounds that have distinct amplitude differences from unwanted noise or other sounds.

- **Advanced Deep Learning Models**
  - **Concept**: Utilizing machine learning models to analyze and classify segments of audio based on their temporal features.
  - **How It Works**:
    - Models like convolutional neural networks (CNNs) or recurrent neural networks (RNNs) can be trained on a dataset of isolated sounds to learn temporal and spectral features.
    - These models can then predict or segregate similar sounds from a complex audio mix based on learned features, effectively isolating desired sounds from a track.

- These techniques require substantial computational resources and a well-annotated training dataset but offer high precision and adaptability across different audio types.