BTQT3 - Xử lý thông tin giọng nói

1. Member list

MSSV	Họ và Tên
21522173	Từ Lễ Huy
21522062	Lìu Thế Hiền
21521652	Đặng Chí Tường
22521183	Nguyễn Đặng Minh Quan
21521041	Lê Trọng Tuấn Kiệt

2. Content

2.1. Fast Fourier Transform (FFT)

2.1.1. Definition and Purpose

Fast Fourier Transform (FFT) is a mathematical method used to convert a time-domain signal into its corresponding frequency-domain representation. Speech signals consist of varying frequencies, and FFT is particularly useful for decomposing these signals into their harmonic components. The primary purpose of FFT is to analyze the harmonic structure of speech signals, which helps in understanding their frequency content.

2.1.2. Process

- Signal Acquisition: The speech signal is first recorded and digitized using sampling techniques. Sampling ensures the continuous speech signal is converted into discrete data, which can be processed by FFT.

- Windowing: The signal is divided into short time frames (usually 20-40ms) to assume stationarity, as speech signals are non-stationary over longer periods.

- Applying FFT: The FFT algorithm is applied to each frame to compute the frequency spectrum, which represents the amplitudes of different frequencies present in the signal.

- Spectral Analysis: The resulting frequency spectrum is analyzed to identify harmonic components, formants, and other important features of the signal.

2.1.3. Applications

- Spectrogram Generation: FFT is used to create spectrograms, which visualize the frequency content of speech over time. Spectrograms are essential in speech research and analysis to observe frequency patterns.

- Pitch and Tone Analysis: By analyzing the harmonic structure, FFT helps in determining the fundamental frequency (pitch) and tone of a speaker.

- Feature Extraction for Speech Recognition: FFT is often used as the first step in feature extraction pipelines for systems like Automatic Speech Recognition (ASR)

2.2. Cepstral Analysis

2.2.1. Definition and Purpose

Cepstral analysis is a technique used to extract information about the spectral envelope of speech signals. It is particularly effective at separating the vocal tract characteristics (filter) from the excitation source (glottal pulses). By doing so, it provides essential features for speech recognition, speaker identification, and voice coding.

2.2.2. Process

- Fourier Transform: The speech signal is transformed into the frequency domain using the FFT.

- Logarithmic Power Spectrum: The power spectrum (magnitude of the signal) is computed, and its logarithm is taken to compress the dynamic range of the signal.

- Inverse Fourier Transform (IFT): The log-spectrum is converted back to the time domain (called the quefrency domain) using the IFT. The result is the cepstrum, which separates the source and filter components of the signal.

- Liftering: The low quefrency components correspond to the source (pitch), while the high quefrency components represent the filter (vocal tract characteristics). Liftering isolates the desired components

2.2.3. Applications

- Mel-Frequency Cepstral Coefficients (MFCCs): The cepstrum is used to compute MFCCs, which are widely used in speech recognition systems due to their ability to represent the vocal tract's spectral envelope.

- Speaker Verification: Cepstral coefficients capture speaker-specific features, making them ideal for verifying a person's identity.

- Speech Enhancement: Cepstral analysis is used to suppress noise and improve the quality of speech signals in noisy environments.

2.3. Linear Predictive Coding (LPC)

2.3.1. Definition and Purpose

Linear Predictive Coding (LPC) is a speech analysis technique that models the vocal tract as an all-pole filter to estimate its parameters. It predicts the current speech sample as a linear combination of past speech samples. LPC is widely used for analyzing formants (spectral peaks), compressing speech signals, and synthesizing speech.

2.3.2. Speech Compression Using LPC

LPC compresses speech signals bLPC compresses speech signals by reducing the amount of data required to represent the signal. Instead of storing the entire speech waveform, LPC stores only the filter coefficients (representing the vocal tract) and the excitation signal (residual). This is achieved through the following steps:

- Frame Segmentation: The speech signal is divided into short overlapping frames (e.g., 20ms).

- Parameter Estimation: LPC coefficients are calculated for each frame, representing the vocal tract filter.

- Residual Signal Extraction: The excitation signal (residual) is obtained by subtracting the predicted signal from the original signal.

- Coding: The LPC coefficients and residual are encoded and transmitted or stored. At the receiver or decoder, the speech signal is reconstructed using the stored data

2.3.3. Two Versions of LPC

2.3.3.1. Autoregressive Method

The autoregressive method is a straightforward approach to estimating LPC coefficients. It assumes that the speech signal is stationary over short time frames and models the current speech sample as a linear combination of past samples. The key principle behind this method is minimizing the prediction error, which is the difference between the actual speech signal and the predicted signal, using the least-squares approach. This involves solving a set of linear equations known as the Yule-Walker equations.

Advantages

One of the main advantages of the autoregressive method is its simplicity of implementation. This makes it computationally efficient and ideal for many real-time speech processing applications. Additionally, it works well for short, stationary segments of speech signals, where the statistical properties of the signal remain stable.

Disadvantages

However, the autoregressive method has some significant drawbacks. It is highly sensitive to noise, meaning even small errors in the input signal can result in inaccurate LPC coefficient estimation. This sensitivity can lead to instability in the model, especially when applied to non-stationary signals or signals with background noise.

2.3.3.2. Covariance Method (Burg’s Method)

The covariance method, also referred to as Burg’s method, is an enhanced version of LPC that improves the accuracy of coefficient estimation by minimizing prediction errors over the entire segment of the signal. Unlike the autoregressive method, it does not strictly assume stationarity of the signal within short frames. Instead, it minimizes both forward and backward prediction errors simultaneously, ensuring a more robust analysis. This is achieved through a recursive computation process, where LPC coefficients are estimated iteratively for greater numerical stability.

Advantages

The covariance method is more robust and accurate compared to the autoregressive method. It is particularly well-suited for processing non-stationary speech signals, where the properties of the signal change over time. Additionally, its ability to minimize forward and backward errors simultaneously ensures greater precision in estimating LPC coefficients. This makes it a preferred choice in applications requiring high-quality speech analysis.

Disadvantages

Despite its advantages, the covariance method has some limitations. It is computationally more complex than the autoregressive method, which can make it less suitable for real-time applications or systems with limited computational resources. Moreover, the added complexity can make the implementation more challenging, requiring careful handling to ensure numerical stability and efficiency.

3. References

1. *The generalized correlation method for estimation of time delay | IEEE Journals & Magazine | IEEE Xplore*
2. Fast Fourier transform - Wikipedia
3. Mel-frequency cepstrum - Wikipedia
4. Linear predictive coding - Wikipedia
5. Speech recognition - Wikipedia