Speech contains many characteristics that are specific to each individual, many of which are independent of the linguistic message for an utterance. some of these characteristics from the perspective of speech recognition, for which they generally are a source of degradation. For instance, each utterance from an individual is produced by the same vocal tract, tends to have a typical pitch range (particularly for each gender), and has a characteristic articulator movement that is associated with speaker, dialect, or gender. All of these factors have a strong effect on the speech that is highly correlated with the particular individual who is speaking. For this reason, listeners are often
able to recognize the speaker identity fairly quickly, even over the telephone. Artificial systems recognizing speakers rather than speech have been the subject of much research over the past 30 years, and multiple commercial systems are currently in use.

Speaker recognition has many potential applications, including the authentication (e.g., telephone and banking applications), access control, parole monitoring, fraud detection, and intelligence.

Speaker recognition generally requires the calculation of a score reflecting the similarity between two speech segments: a test segment and a training (or enrollment) segment. The basic task is that of detection: to tell whether the segments were spoken by the same of by different speakers. This task can be directly used for verification or authentication purposes. The detection score can also be used as the basis for speaker identification, but here we will concentrate on the more general detection task. The main challenge in speaker recognition is to distinguish the variability due to the difference in speakers from the variability due to other factors. These confounding variabilities can be intrinsic, such as the physical, medical or emotional state of the speaker,
the content, the language spoken, the effort at which speech is produced; or extrinsic, such as the recording conditions including acoustics, transducers, recording equipment, transmission channel and noise. For better performance, there can be multiple training segments, preferably recorded in different sessions.

GENERAL DESIGN OF A SPEAKER RECOGNITION SYSTEM

Since the basic task of speaker detection is a two-way classification problem, the core approach is to try to estimate the likelihoods $P(X|S)$, that the speech X is produced by speaker S, and $P(X|\neg S)$, that the speech is produced by someone else. The speech data X here is usually represented by a sequence of features extracted from the speech signal, and the likelihood functions for S and $\neg S$ are formed by some mathematical model M. A basic similarity measure then is formed by the likelihood ratio

$$s = \log \frac{P(X|M,S)}{P(X|M,\neg S)}. \tag{41.1}$$

Note that it is also possible to directly model the likelihood ratio in a single model. The model parameters for $M(S)$ and $M(\neg S)$ are estimated using the training speech segment and from typically a large set of "background" speech from many speakers that are known to be different from S.

The score s can be used for decisions, i.e., whether or not the speech X was uttered by the speaker S, directly by thresholding the score. However, in practice the likelihood functions tend to be dependent on the particular sample of the test data X, and similarly the model parameters are sensitive to the particular sample of training training data. Therefore, this influence is lowered by normalizing the scores s by computing scores over a cohort of non-target models and test segments, respectively.

Multiple systems can be fused by computing a weighted sum of the individual normalized system scores, leading to better performance. As a final step, a threshold must be chosen that will minimize the expected cost of decision errors. This is a process known as *calibration* and is governed by the relative costs of false positives and false negatives, and the prior probability of the target speaker.

At various points in this design, a collection of "background" speech is required, for instance for modeling the feature space, the non-target speakers $\neg S$, for score normalization and for fusion and calibration. A proper choice of data for these steps is an essential part of the design of a speaker recognition system.

Speaker recognition systems utilize various spectral features extracted from speech signals to distinguish between different speakers. These features are crucial for accurately identifying and verifying the identity of a speaker. Here are some commonly used spectral features and their roles:

1. **Short-Time Fourier Transform (STFT)**:

- The STFT is often used as the initial step in extracting spectral features from speech signals.
- It represents the frequency content of the speech signal over short time intervals by decomposing the signal into its constituent frequency components.

2. **Mel-Frequency Cepstral Coefficients (MFCCs)**:

- MFCCs are widely used spectral features in speaker recognition.
- They are derived from the magnitude spectrum of the STFT using a series of processing steps, including the Mel frequency scale and discrete cosine transform (DCT).
- MFCCs capture the characteristics of the human auditory system and are effective in representing the spectral envelope of speech signals.
- They are robust to variations in speech due to factors like speaker identity, channel conditions, and background noise.

3. **Linear Predictive Coding (LPC) Coefficients**:

- LPC coefficients are derived from the prediction of future samples in a speech signal using a linear prediction model.
- These coefficients represent the spectral envelope of speech and are particularly useful for capturing the formant structure of speech sounds.
- LPC coefficients are sensitive to speaker-specific characteristics and can be used for speaker recognition tasks.

4. **Perceptual Linear Prediction (PLP) Coefficients**:

- PLP coefficients are derived by modeling the human auditory system's perception of speech.
- They are based on a combination of linear prediction and perceptual weighting techniques.
- PLP coefficients provide a more compact representation of speech compared to MFCCs and are often used in speaker recognition systems.

5. **Relative Spectral Phase (RSP)**:

- RSP represents the phase difference between different frequency bands of a speech signal.
- It captures fine-grained temporal information and is robust to variations caused by factors like channel distortion and background noise.
- RSP has been shown to improve the performance of speaker recognition systems, especially in challenging conditions.

6. **Spectral Subband Centroids (SSCs)**:

- SSCs are computed by dividing the speech spectrum into subbands and calculating the centroids of each subband.
- They capture information about the distribution of spectral energy across different frequency regions.
- SSCs are useful for discriminating between speakers with different vocal characteristics.

Overall, these spectral features play a crucial role in speaker recognition systems by capturing different aspects of speech signals that are relevant for distinguishing between speakers. By extracting and analyzing these features, speaker recognition systems can accurately identify and verify speaker identities in various applications, including security systems, forensic analysis, and human-computer interaction.

Certainly! Let's delve into the mathematical formulations behind some of the spectral features commonly used in speaker recognition systems:

1. **Short-Time Fourier Transform (STFT)**:

The STFT of a signal $x(t)$ is computed by applying the Fourier Transform to short, overlapping windows of the signal. Mathematically, for a window function $w(t)$, the STFT $X(\omega,\tau)$ is calculated as:

$$X(\omega,\tau)=\int_{-\infty}^{\infty} x(t)\cdot w(t-\tau)\cdot e^{-j\omega t}dt$$

Where:

- ω represents frequency.
- τ represents time.
- $x(t)$ is the signal.
- $w(t)$ is the window function.

2. **Mel-Frequency Cepstral Coefficients (MFCCs)**:

MFCCs are computed using the following steps: a. Pre-emphasis: $x_{\text{preemph}}(t)=x(t)-\alpha\cdot x(t-1)$ where α is a pre-emphasis coefficient typically set to 0.97. b. Frame blocking: The pre-emphasized signal is divided into frames of fixed duration. c. Windowing: Each frame is windowed using a window function. d. FFT: Compute the magnitude spectrum by taking the FFT of each windowed frame. e. Mel Filterbank: Apply a Mel filterbank to the magnitude spectrum. f. Logarithm: Take the logarithm of the filterbank energies. g. Discrete Cosine Transform (DCT): Apply DCT to decorrelate the filterbank energies.

Mathematically, the MFCC computation involves a series of operations as described above.

3. **Linear Predictive Coding (LPC) Coefficients**:

LPC modeling involves modeling a speech signal $x(t)$ as the output of a linear time-invariant (LTI) system. Mathematically, this can be represented as:

$$x(t)=\sum_{i=1}^{p} a_i x(t-i)+e(t)$$

where:

- p is the order of the prediction.
- a_i are the LPC coefficients.
- $e(t)$ is the prediction error.

LPC coefficients are typically estimated using methods such as the autocorrelation method or the covariance method.

4. **Perceptual Linear Prediction (PLP) Coefficients**:

PLP coefficients are computed similarly to MFCCs but incorporate perceptual weighting. The PLP spectrum is obtained by applying a filterbank to the power spectrum of the pre-emphasized signal. The PLP cepstral coefficients are obtained by applying a discrete cosine transform (DCT) to the log PLP spectrum.

These mathematical formulations provide a foundation for understanding how spectral features are extracted from speech signals in speaker recognition systems. By manipulating and analyzing these features, systems can effectively distinguish between different speakers.

Wiener filter:

Wiener filtering is a widely used method for estimating a clean signal from its noisy version. It's particularly effective when the statistical properties of the signal and the noise are known or can be estimated. The Wiener filter minimizes the mean square error between the estimated clean signal and the true clean signal.

Given a noisy signal $y(t)$, the Wiener filter estimates the clean signal $x(t)$ as follows:

$$x\hat{}(t)=W(y(t))=H(\omega)Y(\omega)$$

Where:

- $x\hat{}(t)$ is the estimated clean signal.
- $W(y(t))$ is the Wiener filter.
- $H(\omega)$ is the frequency response of the Wiener filter.
- $Y(\omega)$ is the Fourier Transform of the noisy signal $y(t)$.

The frequency response of the Wiener filter $H(\omega)$ is given by:

$$H(\omega)=S_{yy}(\omega)S_{xx}(\omega)$$

Where:

- $S_{xx}(\omega)$ is the power spectral density (PSD) of the clean signal.
- $S_{yy}(\omega)$ is the PSD of the noisy signal.
- $S_{xx}(\omega)$ and $S_{yy}(\omega)$ can be estimated using methods such as periodogram estimation or parametric methods like autoregressive (AR) modeling.

The Wiener filter minimizes the mean square error (MSE) between the estimated clean signal and the true clean signal, given by:

$$\text{MSE}=E[|X(\omega)-\hat{X}(\omega)|^2]$$

Where:

- $X(\omega)$ is the Fourier Transform of the true clean signal $x(t)$.
- $\hat{X}(\omega)$ is the Fourier Transform of the estimated clean signal $\hat{x}(t)$.
- $E[\cdot]$ denotes the expected value operator.

To implement Wiener filtering in practice, one typically estimates the PSDs $S_{xx}(\omega)$ and $S_{yy}(\omega)$ from the noisy signal $y(t)$. Then, the frequency response $H(\omega)$ is computed, and the estimated clean signal $\hat{x}(t)$ is obtained by multiplying $H(\omega)$ with the Fourier transform of the noisy signal $Y(\omega)$. Finally, the inverse Fourier transform of $\hat{x}(t)$ yields the estimated clean signal in the time domain.

It's important to note that Wiener filtering assumes stationary signals and stationary noise. If the statistical properties of the signal and noise vary over time, adaptive Wiener filtering methods may be employed.