Speech signal acquisition

(microphone/telephone, sampling frequency, quantization)

e(n): Vocal fold vibration (voice source) change leads to change in type of speech sound voiced/unvoiced and fundamental frequency (F0) in voiced sounds.

v(n): Vocal tract shape changes lead to change in the resonance patterns, i.e. Formants, consequently speech sound.

Short-term analysis

(Window/frame size, Window/frame shift)

Focus on changes in both F0 and Formants: 20-40 ms window size

*s(n) = e(n) * v(n)*

Focus on only Formants: window size can be short, e.g., 5 ms

Focus on only F0: window size can be long, e.g., 40-60 ms

Typical window shift: 10 ms (can change w.r.t speech processing problem)

Reason: Non-stationary signal, i.e., statistical characterisitics of the speech signal s(n) changes overtime due to changes in vocal fold vibration and change in vocal tract shape.

Time domain analysis

Mean (typically 0)
Energy (variance)
Zero crossing rate
Autocorrelation: periodic (voiced), estimate F0
"Energy" vs Time

Fourier transform

Frequency domain analysis

Power spectrum
Energy vs Frequency
S(f) = E(f) x V(f)

Spectrogram
Energy vs Time vs Frequency
Wideband: short window size
Narrowband: long window size

20-40 ms window

Peaks in spectral envelop of S(f) characterize Formants, i.e. V(f)

Finer variations in S(f) characterize E(f): F0 and its harmonics (voiced), "noisy" (unvoiced).

Linear prediction (LP) analysis

LP coefficients characterize spectral envelop, i.e., v(n)

LP residual characterizes voice source, i.e., e(n)

Cepstral recursion

Cepstral analysis

Lower cepstral coefficients characterize spectral envelop, i.e., v(n)

Higher cepstral coefficients characterize voice source, i.e., e(n)

Cepstral analysis after applying Mel-frequency based auditory filterbanks on short-term spectrum S(f) yields mel-frequency cepstral coefficients (MFCC).
Cepstral analysis after applying auditory filterbank and loudness related knowledge on S(f) yields perceptual LP (PLP) cepstral coefficients *(needs LP route for cepstral coefficients because of cubic compression operation)*
MFCC and PLP CC characterize spectral envelop, i.e., v(n) information.

Approximate temporal derivatives capture dynamic information. *This can be seen as a filtering operation in time.*

Short-term speech signal analysis yields sequence of feature vectors for speech processing