Data Mining and Machine Learning

Applicati r ASR:

Feature re f speech

Peter Jančovič

UNIVERSITY OF BIRMINGHAM

Objectives

- Front-end analysis for ASR – feature representation of speech
  - To understand motivation and stages for ‘typical’ parameteris ed for ASR
  - Mel Freque (MFCCs)

What is “Front-End Analysis”

- First stage in any speech recognition system
- Goal is to convert the raw acoustic speech waveform into a form which is suitable (or even optimal) for cognition
- In general p ms, front-end analysis is feature extrac
- Where do we start?

The Human Auditory System

taken from J N Holmes, “Speech Synthesis and Recognition”, Van Nostrand Reinhold (1988)

The Basilar Membrane

Australian National University – http://online.anu.edu.au/ITA/ACAT/drw/PPofM/hearing/hearing3.html

Frequency response of the basilar membrane

School for advanced studies, Triste, Italy –

http::/poirot.sissa.it/multidisc/cochlea/utils/basilar.htm

Lessons from Psycho-Acoustics

- Human speech perception begins with frequency analysis on the basilar membrane
- Frequency is not perceived on a linear scale – hence use of non-linear perc **mel** scale, **bark** scale,…

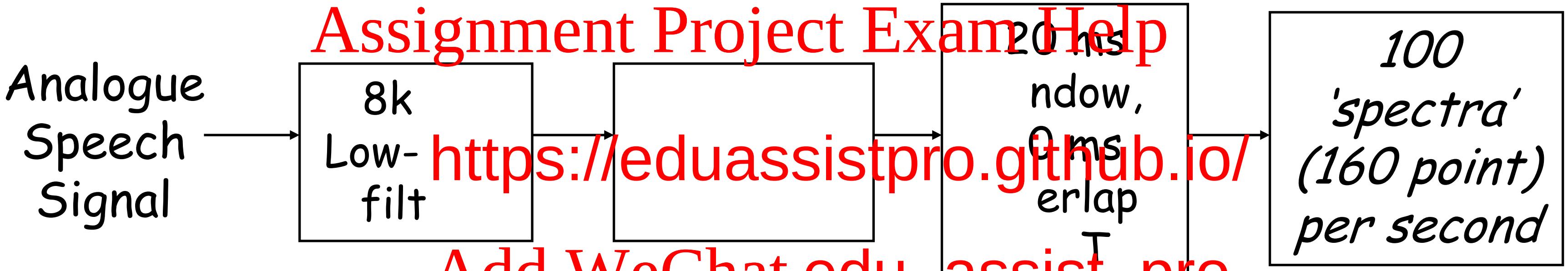
- Individual point on the basilar an be modelled as band-pass filter – a **critical ba** licit bandwidth of such an ‘auditory filter’
- Loudness perceived on logarithmic scale
- Phase of limited significance for speech recognition

Front-end analysis for ASR

- Speech waveform typically low-pass filtered at 4kHz to 8kHz
- Sampled 8,000 to 16,000 samples per second
- Frequency ana
  - 20 ms anal
  - 10 ms overlap between wi
  - Hamming window
  - Discrete Fourier Transform

UNIVERSITY OF BIRMINGHAM

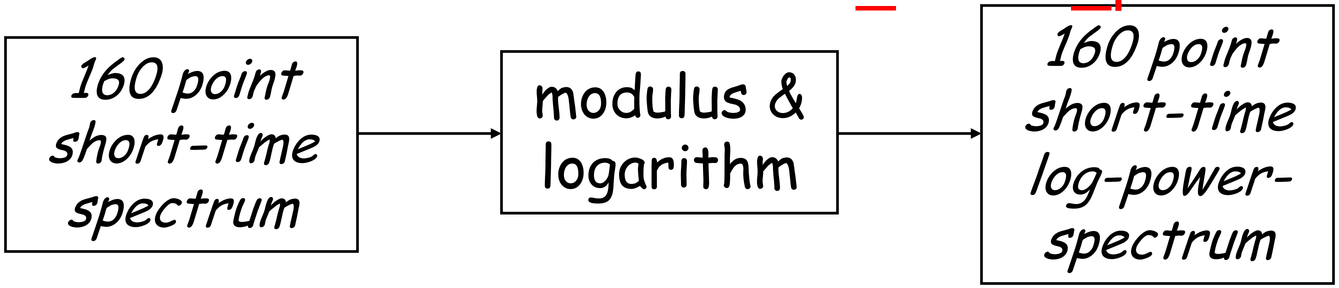
Frequency analysis for ASR



Example: 8kHz bandwidth system

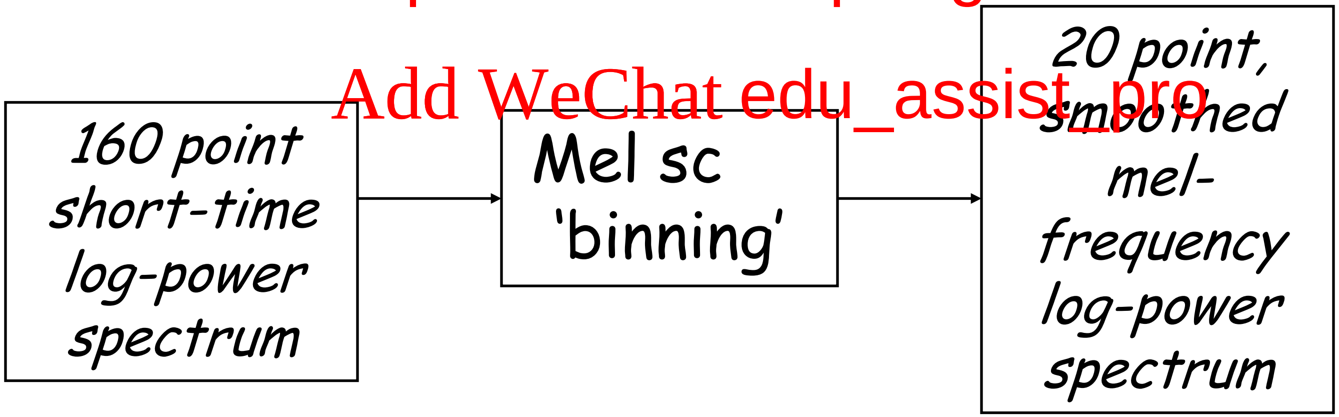
Log Power Spectrum

- Phase ignored by taking the **modulus** of the complex spectrum
- Logarithm applied
  - For consi oustic results
  - To compr


160 point short-time spectrum
modulus & logarithm
160 point short-time log-power-spectrum

Mel-scale & smoothing

- The **mel spectrum** can be computed by **averaging** the short-time Fourier spectrum over ‘bins’ whose width depends on frequency…
- …or by using band-pass filters with appropriate, frequency-dependent, ba



160 point short-time log-power spectrum

Mel sc ‘binning’

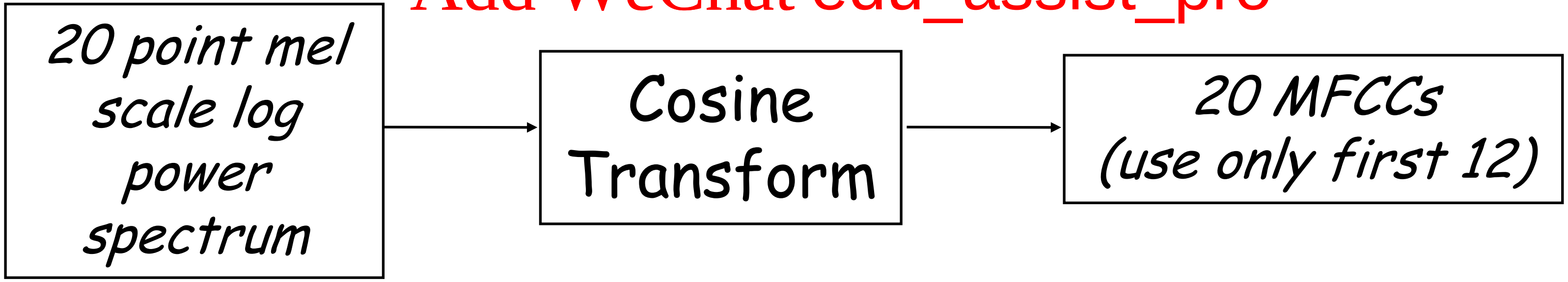
20 point, smoothed mel-frequency log-power spectrum

Mel Scale Filterbank

From Steve Young, "The HTK Book", Cambridge University Engineering Department

Cepstrum

- Cosine transform applied to remove correlation between components of mel-scale log power spectrum
  - Mel Cepstrum: MFCC = Mel Frequency Cepstral Coefficients
  - Mathemat

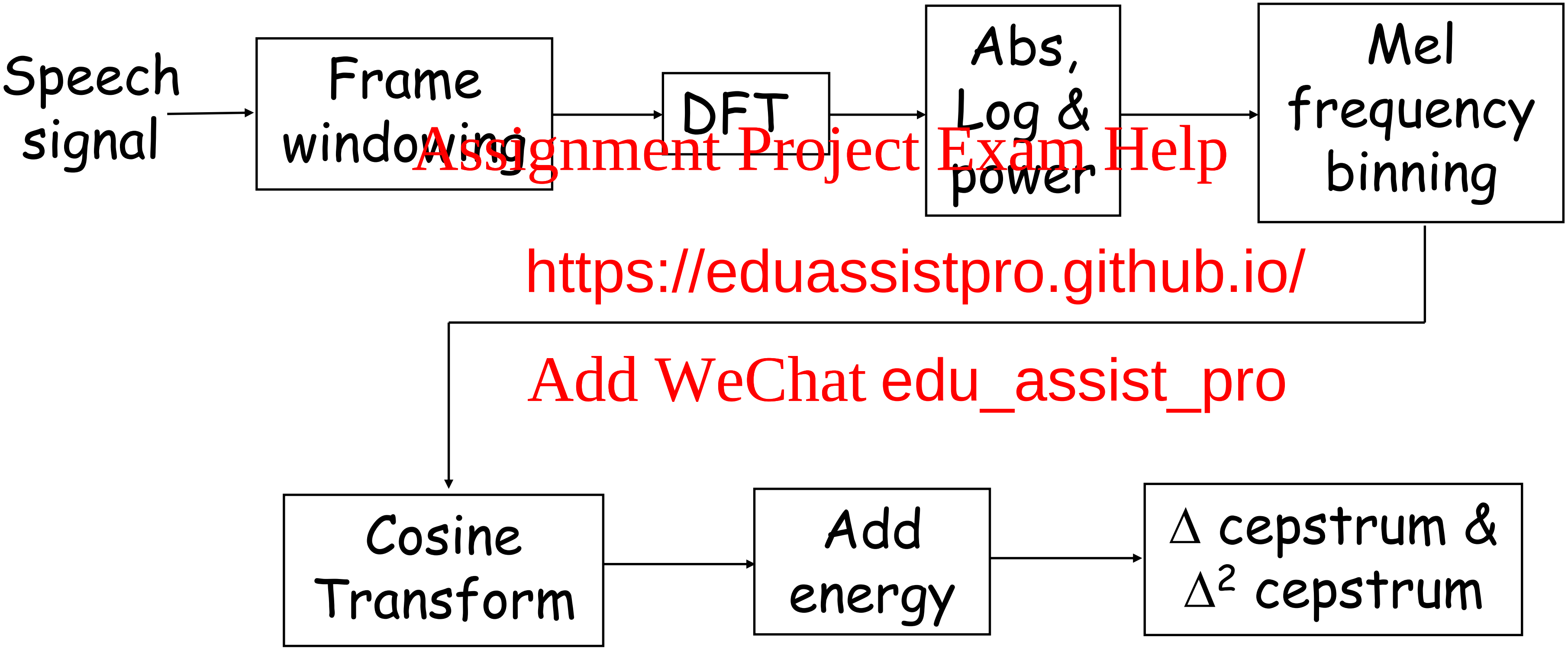

20 point mel scale log power spectrum → Cosine Transform → 20 MFCCs (use only first 12)

Energy & Delta Coefficients

- Add energy as 13th parameter
- Compute estimate of time-derivative of each parameter – pstrum)
- Compute est tion of parameter – delta^2 cepstru pstrum)
- Cepstum + Δ Cepstrum + Δ^2 Cepstrum = ‘standard’ 39 dimensional representation (e.g. in HTK)

UNIVERSITY OF
BIRMINGHAM

Front-end analysis – summary



Speech signal

Frame windowing

DFT

Abs, Log & power

Mel frequency binning

Cosine Transform

Add energy

Δ cepstrum & Δ^2 cepstrum

Summary

- Introduction to front-end speech processing for ASR
  - Motivations from human hearing
  - Description of ‘typical’ front-end representation
    - Short-ti
    - Mel scal
    - Cosine transform
    - Δ and Δ^2 parameters

UNIVERSITY OF
BIRMINGHAM