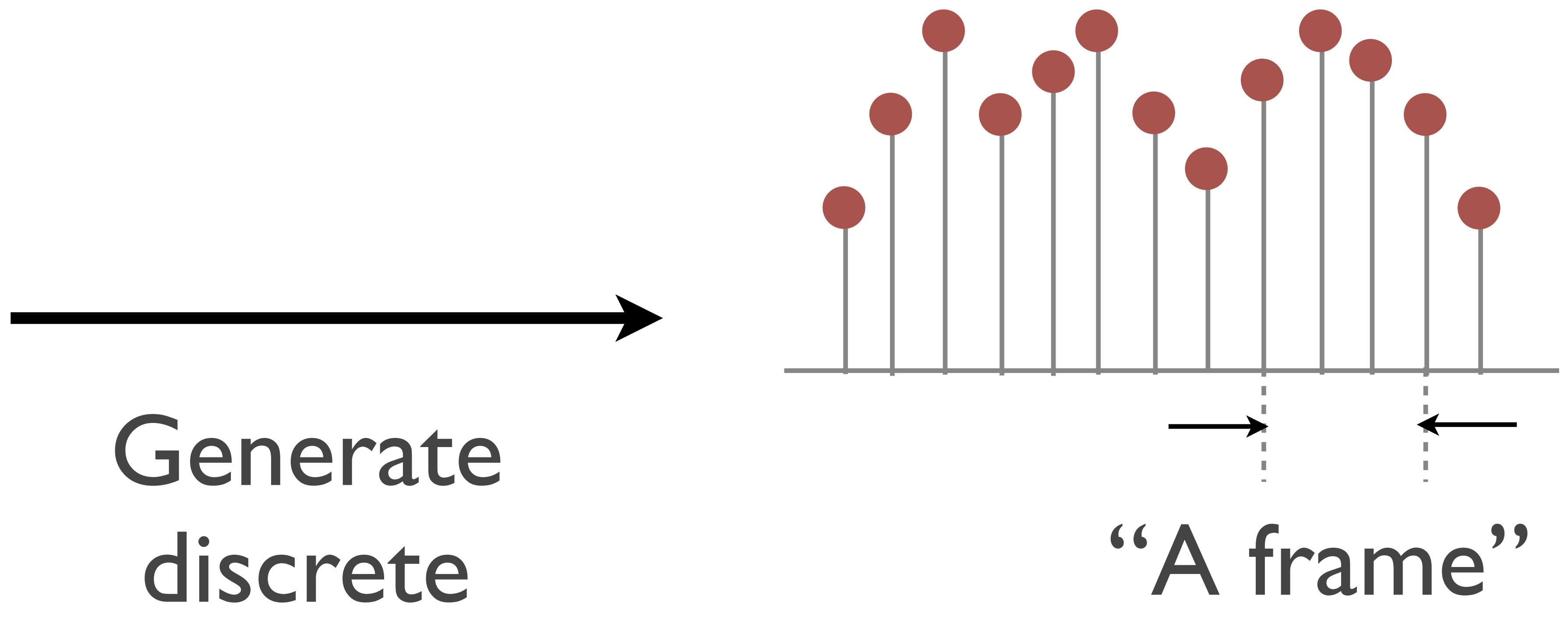
Acoustic Feature Analysis for ASR

Lecture 13

CS 753

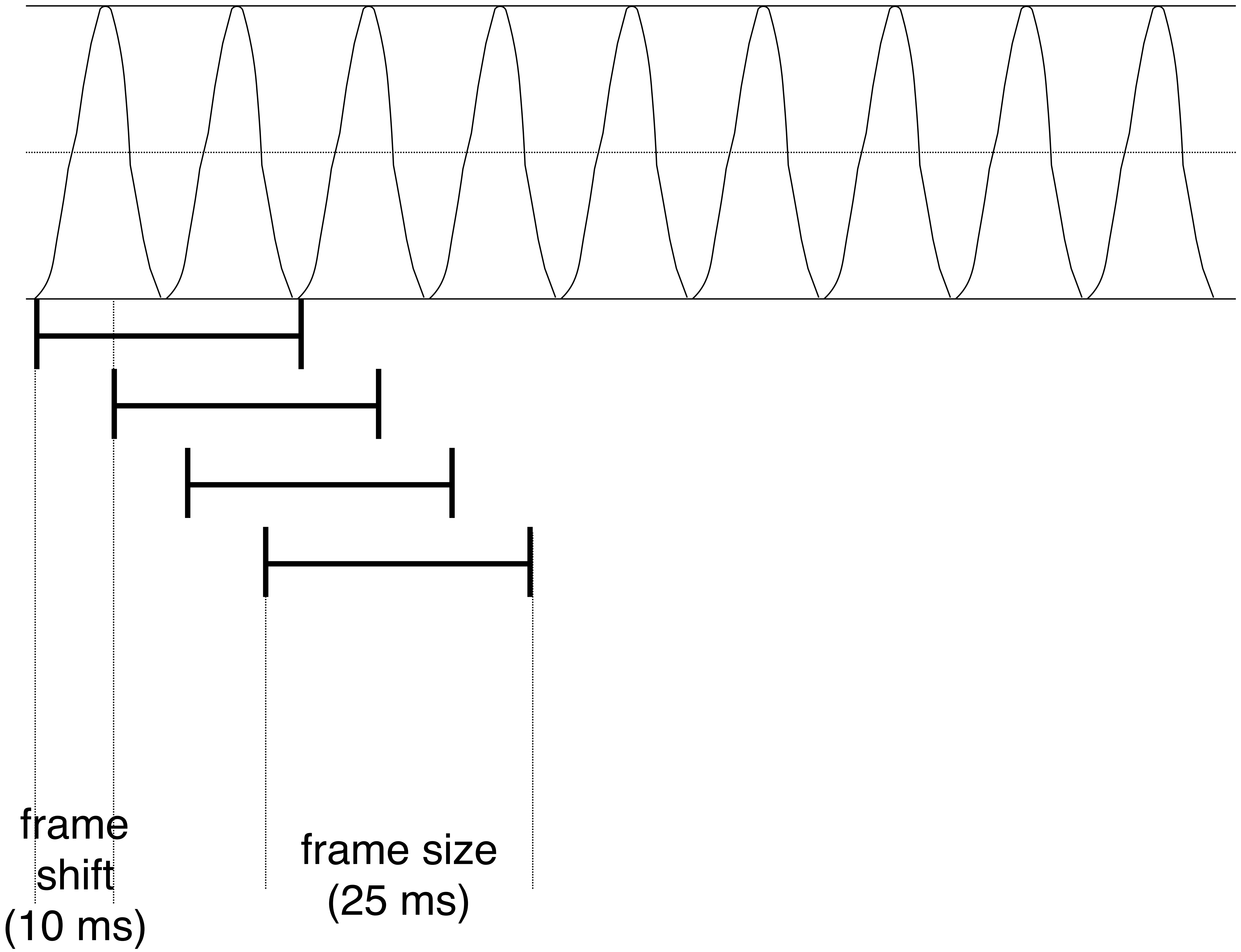
Instructor: Preethi Jyothi

Speech Signal Analysis

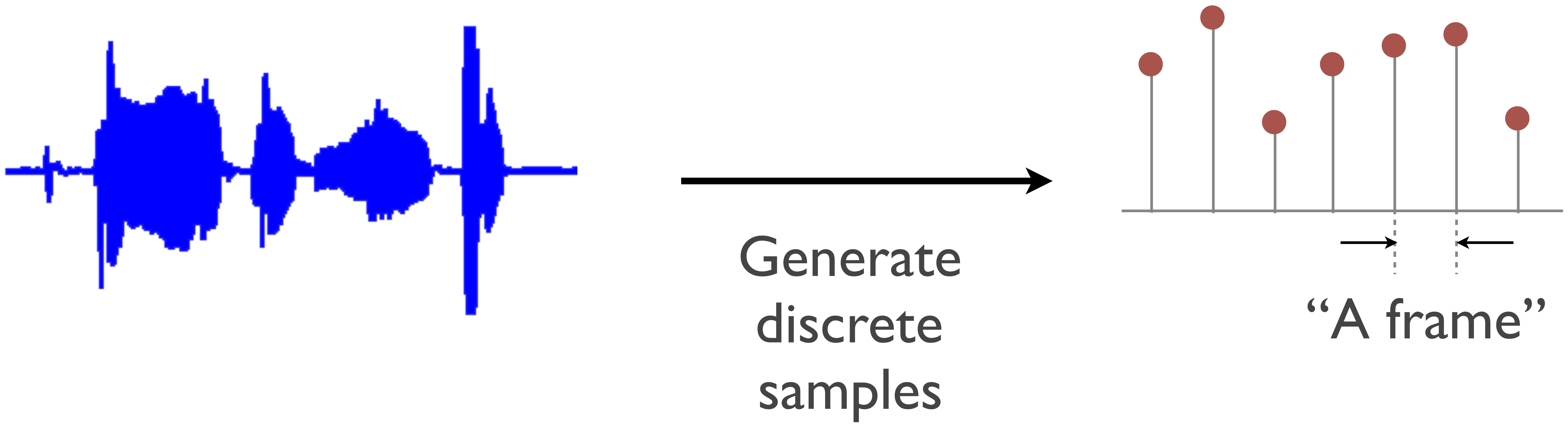


- Need to focus on short segments of speech (speech frames) that more or less correspond to a subphone and are stationary
- Each speech frame is typically 20-50 ms long
- Use overlapping frames with frame shift of around 10 ms

Frame-wise processing



Speech Signal Analysis



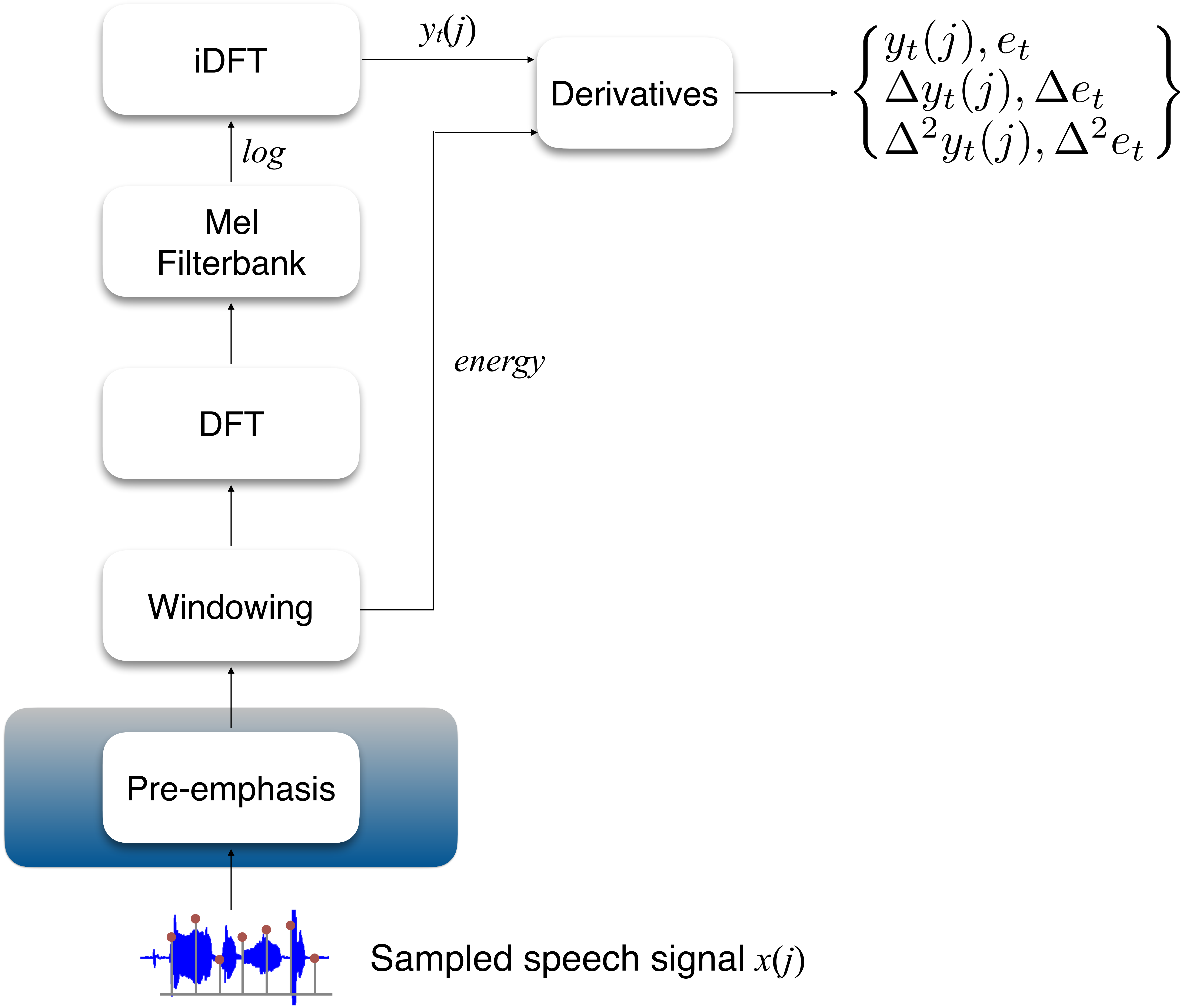
- Need to focus on short segments of speech (speech frames) that more or less correspond to a phoneme and are stationary
- Each speech frame is typically 20-50 ms long
- Use overlapping frames with frame shift of around 10 ms
- Generate acoustic features corresponding to each speech frame

Acoustic feature extraction for ASR

Desirable feature characteristics:

- Capture essential information about underlying phones
- Compress information into compact form
- Factor out information that's not relevant to recognition e.g. speaker-specific information such as vocal-tract length, channel characteristics, etc.
- Would be desirable to find features that can be well-modelled by known distributions (Gaussian models, for example)
- Feature widely used in ASR: Mel-frequency Cepstral Coefficients (MFCCs)

MFCC Extraction



Pre-emphasis

- Pre-emphasis increases the amount of energy in the high frequencies compared with lower frequencies
- Why? Because of spectral tilt
  - In voiced speech, signal has more energy at low frequencies
  - Attributed to the glottal source
- Boosting high frequency energy improves phone detection accuracy

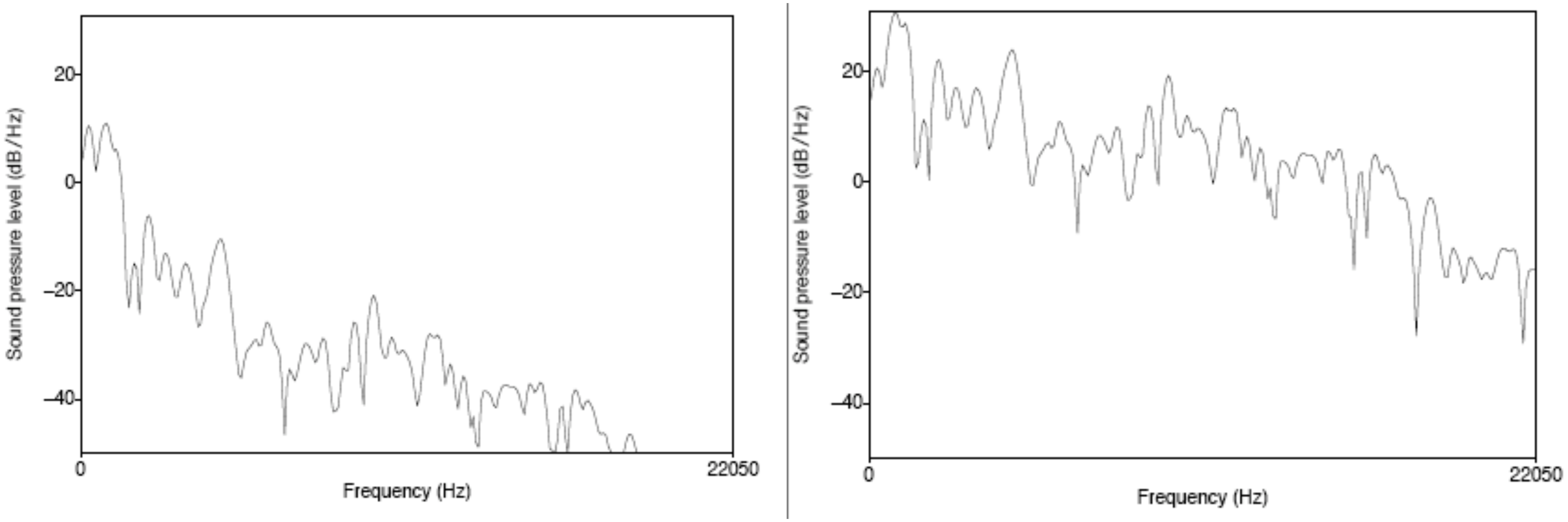
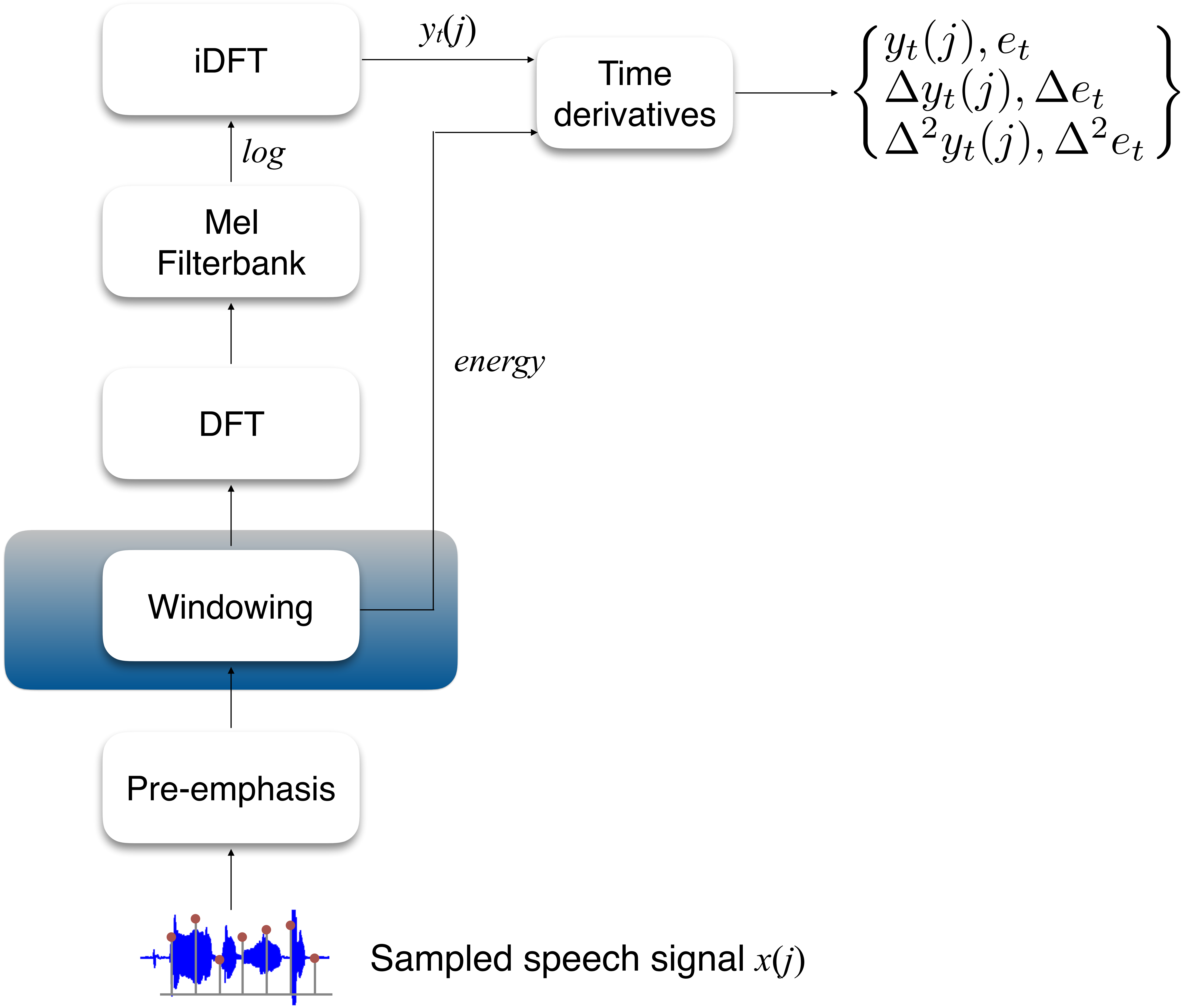


Image credit: Jurafsky & Martin, Figure 9.9

MFCC Extraction



Windowing

- Speech signal is modelled as a sequence of frames (assumption: stationary across each frame)
- Windowing: multiply the value of the signal at time n, $s[n]$ by the value of the window at time n, $w[n]$: $y[n] = w[n]s[n]$

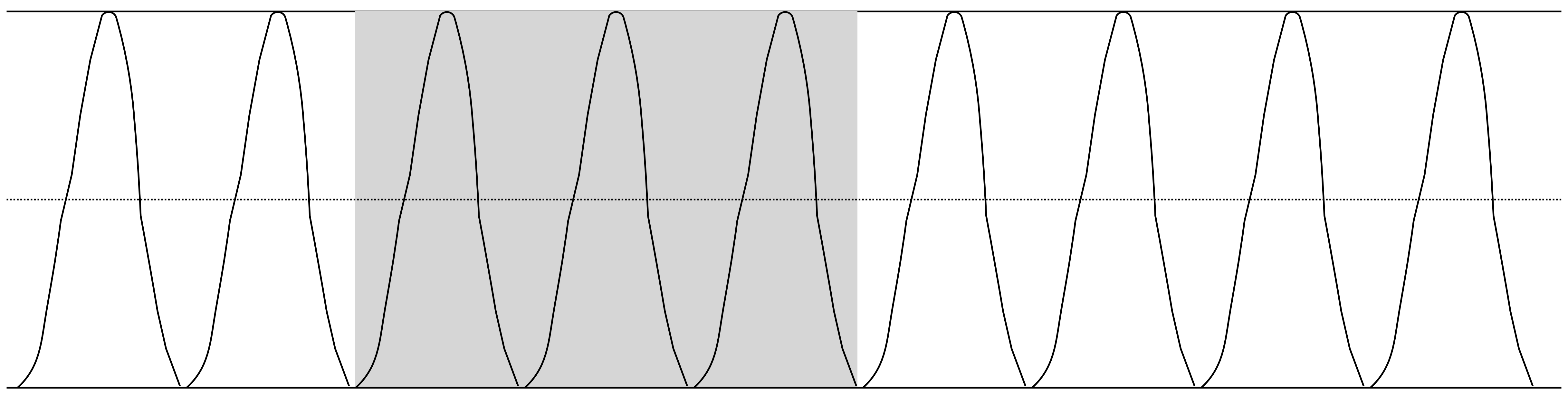
Rectangular:

$$w[n] = \begin{cases} 1 & 0 \leq n \leq L-1 \\ 0 & \text{otherwise} \end{cases}$$

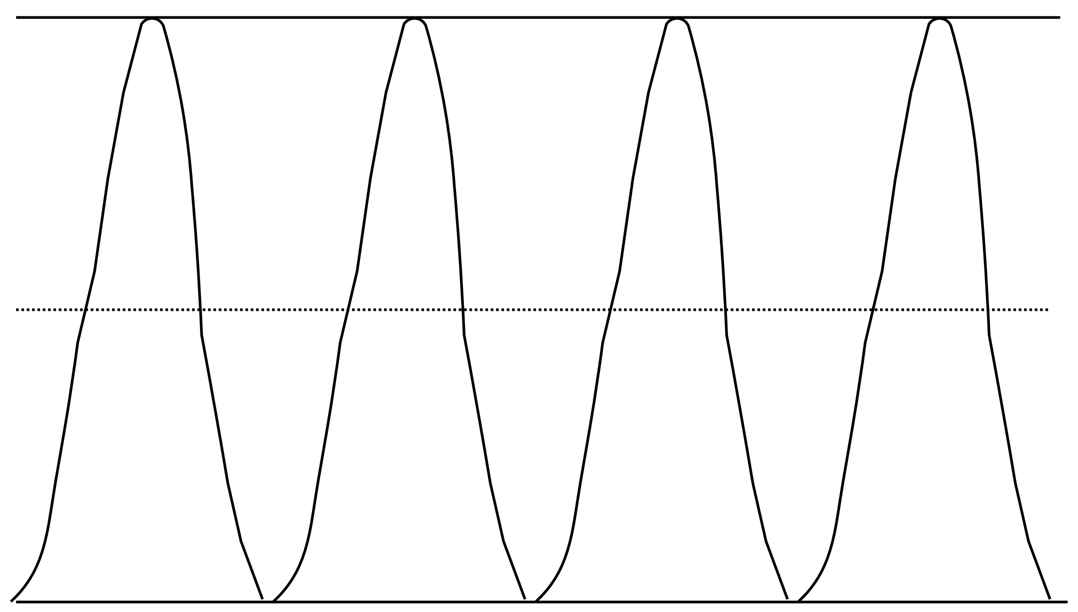
Hamming:

$$w[n] = \begin{cases} 0.54 - 0.46\cos\frac{2\pi n}{L} & 0 \leq n \leq L-1 \\ 0 & \text{otherwise} \end{cases}$$

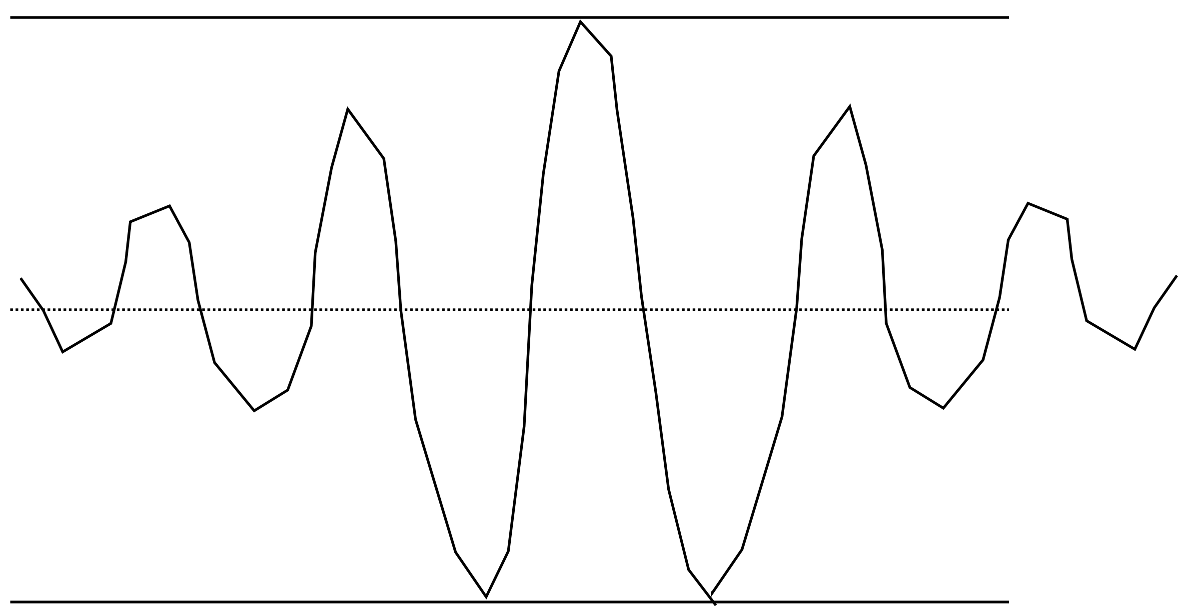
Windowing: Illustration



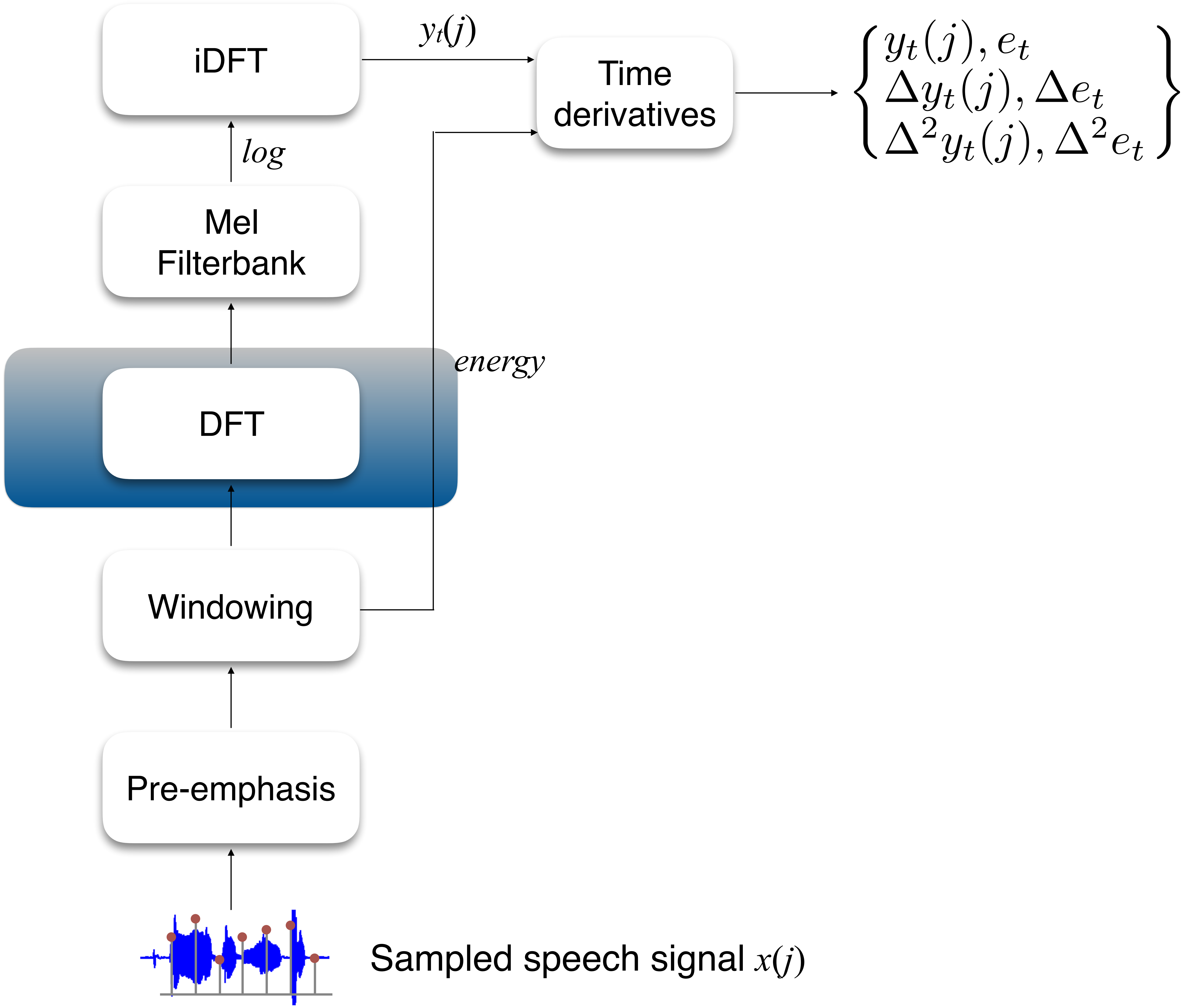
Rectangular window



Hamming window



MFCC Extraction



Discrete Fourier Transform (DFT)

Extract spectral information from the windowed signal:
Compute the DFT of the sampled signal

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j \frac{2\pi}{N} kn}$$

Input: windowed signal $x[1],\ldots,x[n]$
Output: complex number $X[k]$ giving magnitude/phase for the kth frequency component

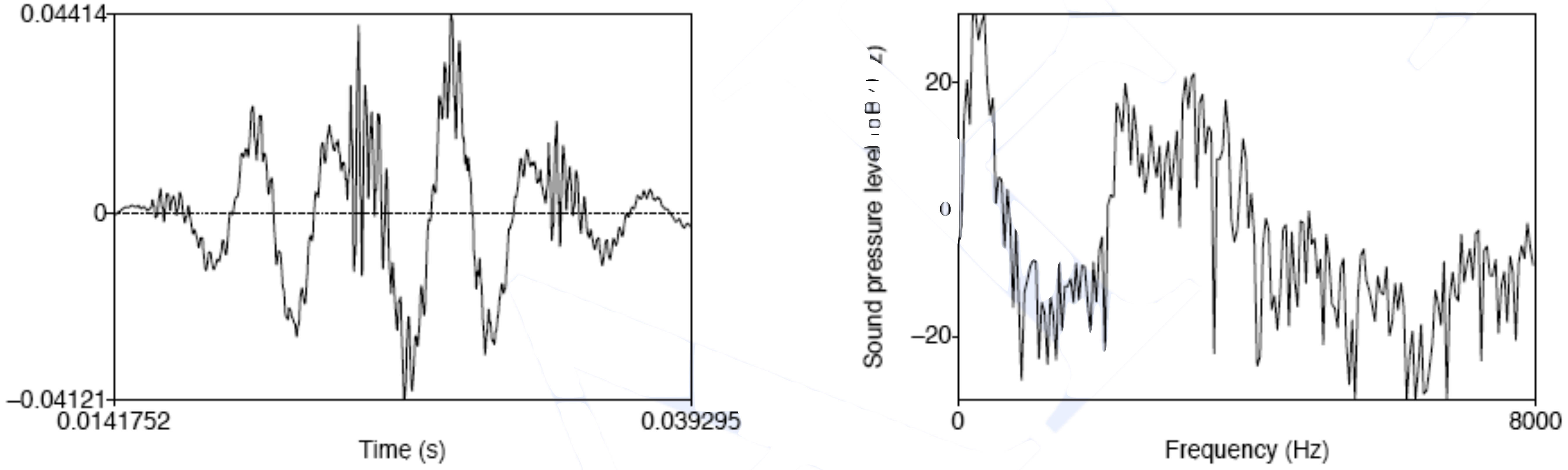
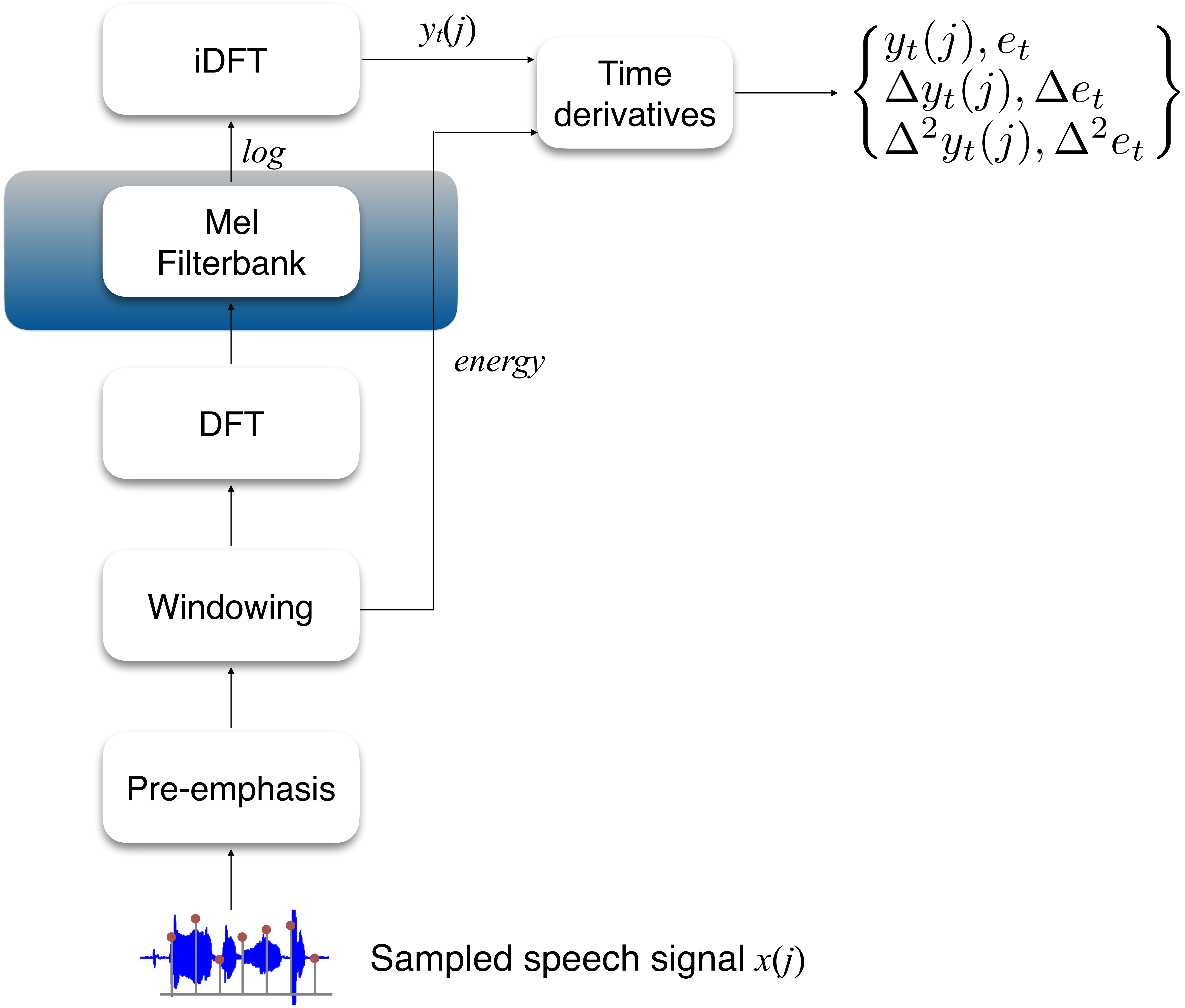


Image credit: Jurafsky & Martin, Figure 9.12

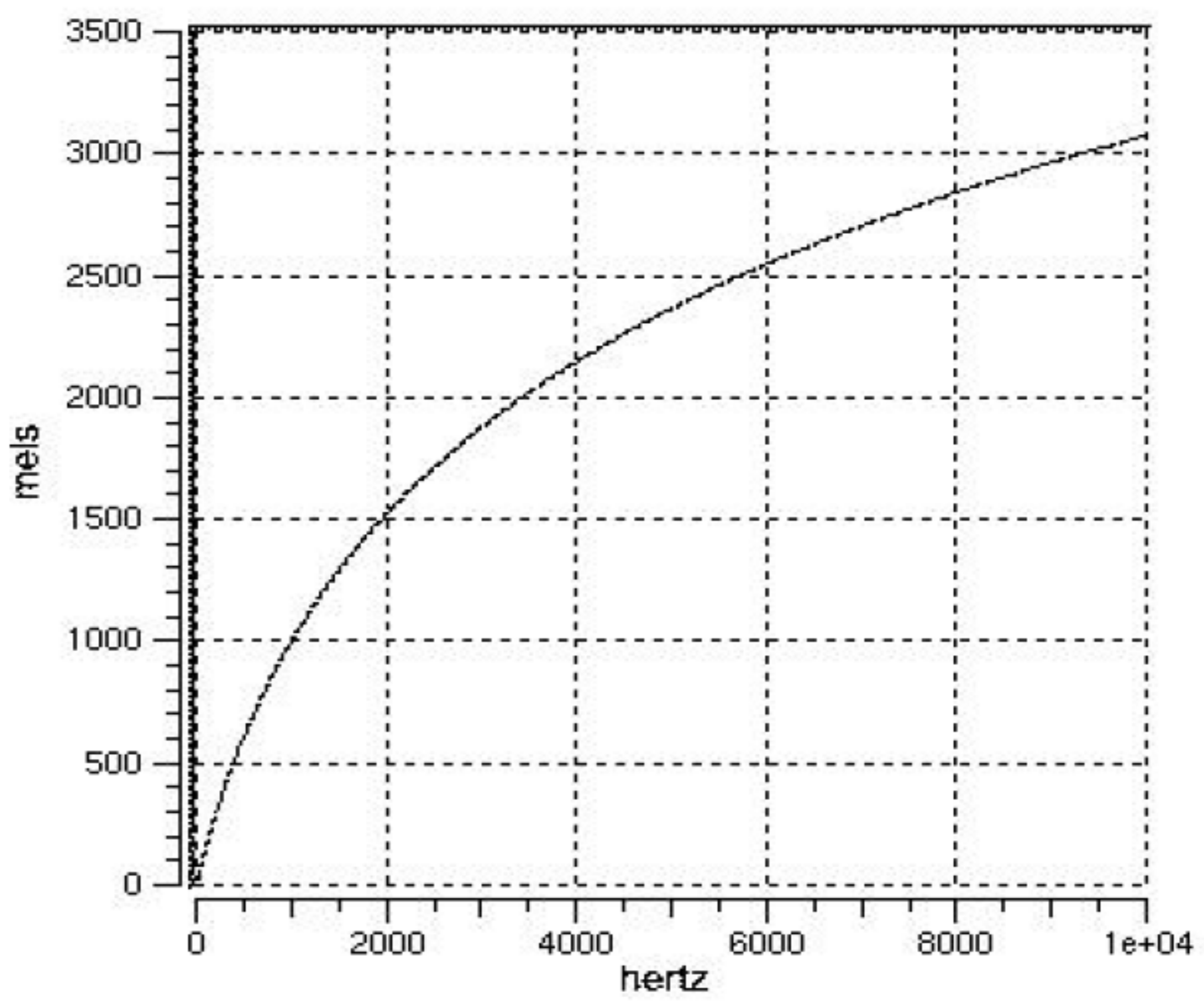
MFCC Extraction



Mel Filter Bank

- DFT gives energy at each frequency band
- However, human hearing is not sensitive at all frequencies: less sensitive at higher frequencies
- Warp the DFT output to the mel scale: mel is a unit of pitch such that sounds which are perceptually equidistant in pitch are separated by the same number of mels

Mels vs Hertz



Mel filterbank

- Mel frequency can be computed from the raw frequency f as:

$$\mathrm{mel}(f) = 1127 \ln(1 + \frac{f}{700})$$

- 10 filters spaced linearly below 1kHz and remaining filters spread logarithmically above 1kHz

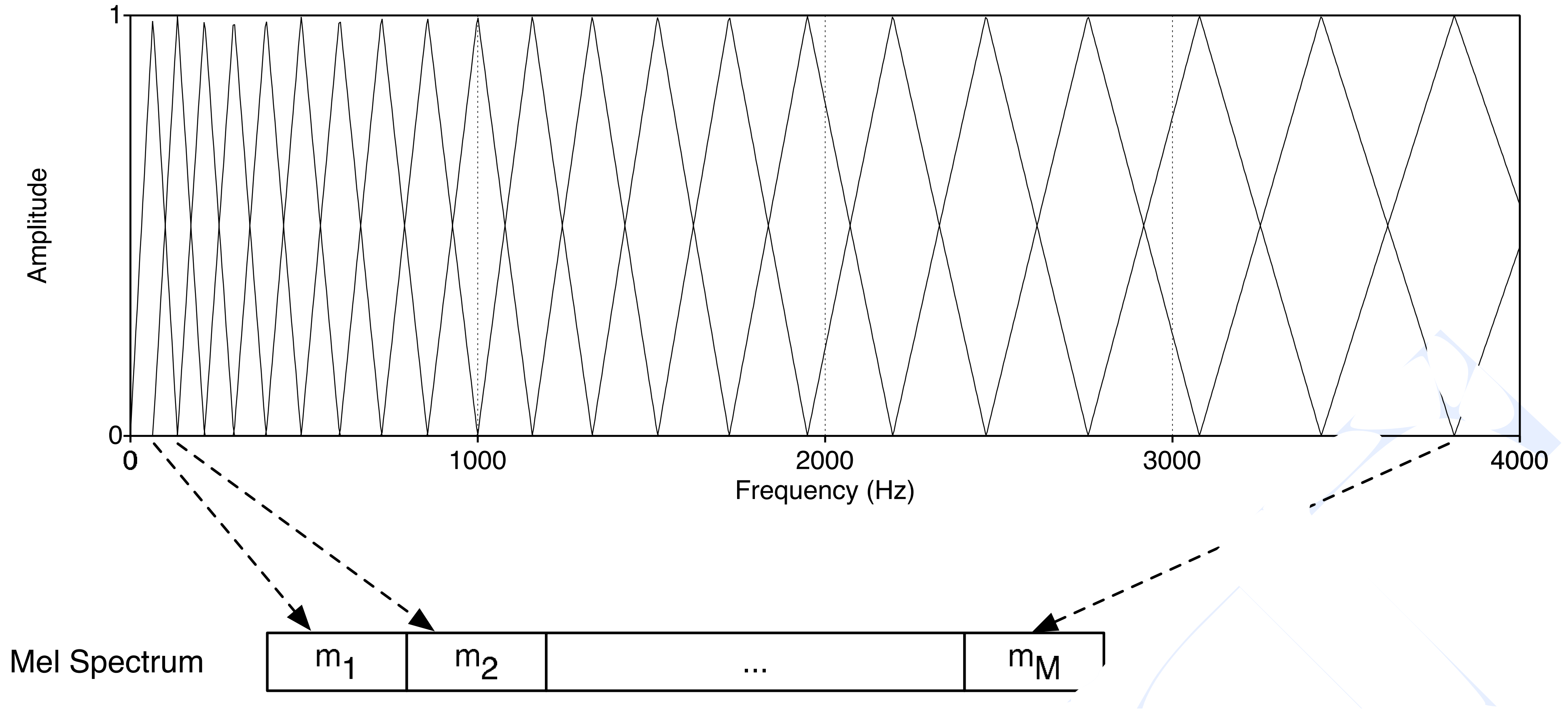


Image credit: Jurafsky & Martin, Figure 9.13

Mel filterbank inspired by speech perception

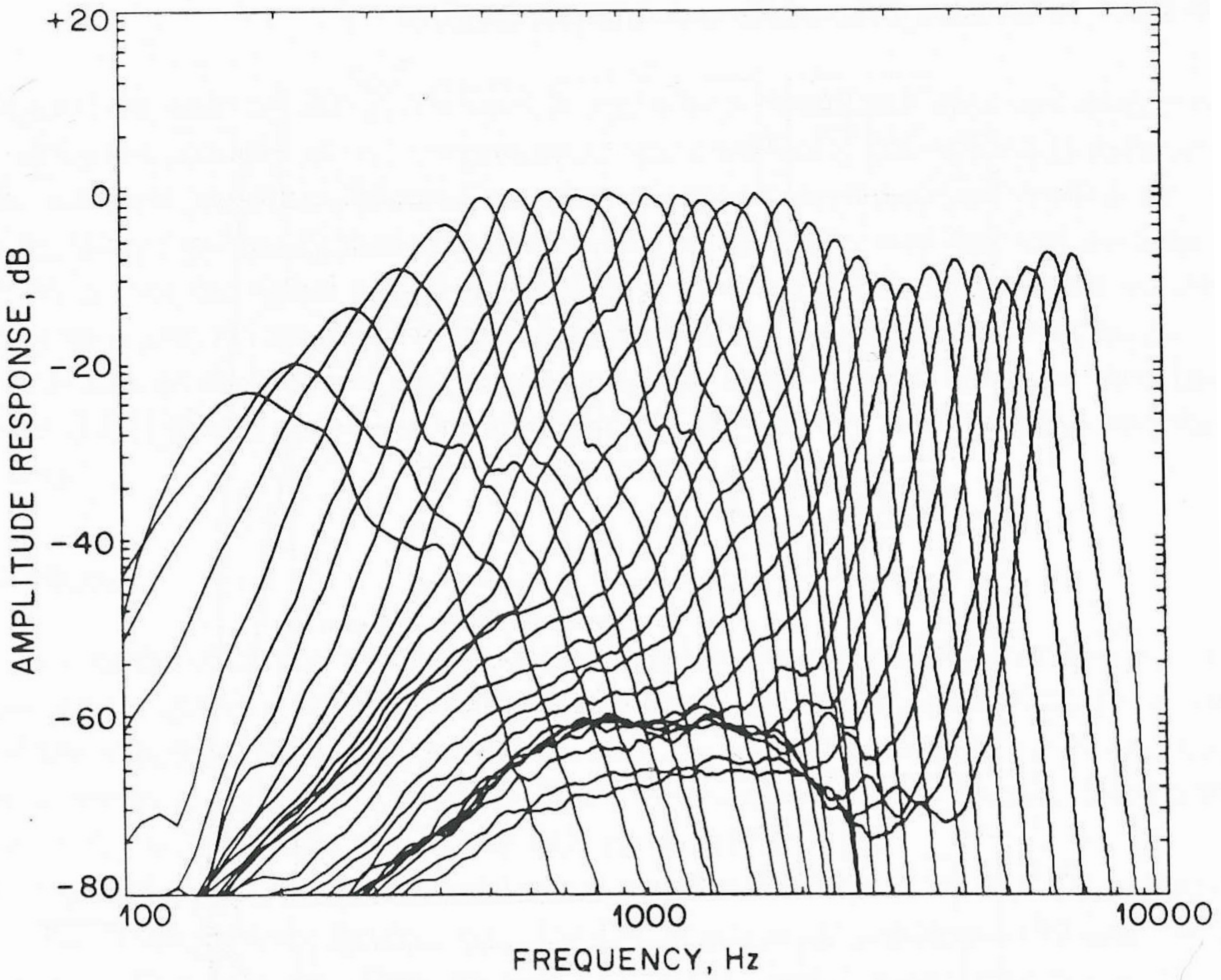


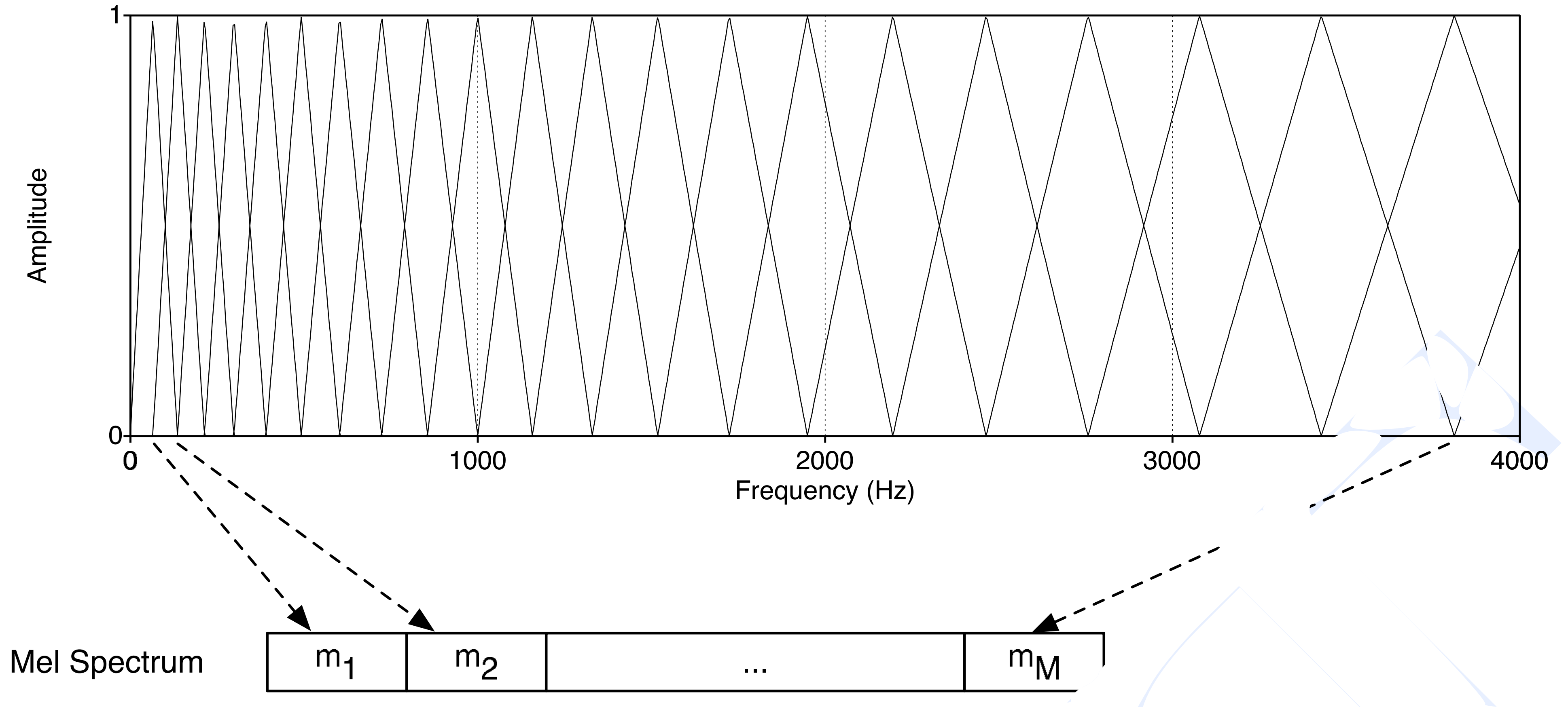
Figure 3.50 Frequency response curves of a cat's basilar membrane (after Ghitza [13]).

Mel filterbank

- Mel frequency can be computed from the raw frequency f as:

$$\mathrm{mel}(f) = 1127 \ln(1 + \frac{f}{700})$$

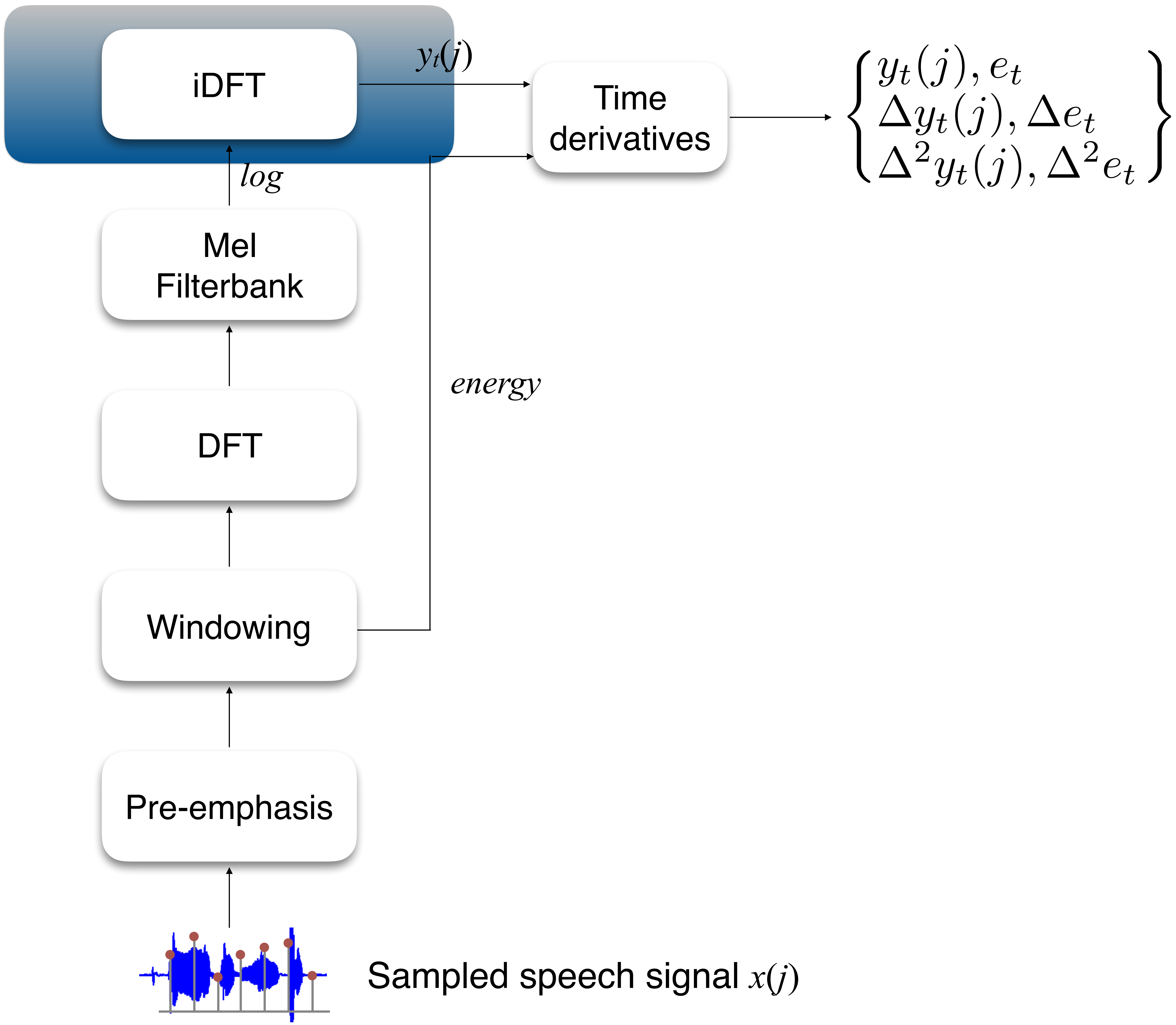
- 10 filters spaced linearly below 1kHz and remaining filters spread logarithmically above 1kHz



- Take log of each mel spectrum value 1) human sensitivity to signal energy is logarithmic 2) log makes features robust to input variations

Image credit: Jurafsky & Martin, Figure 9.13

MFCC Extraction

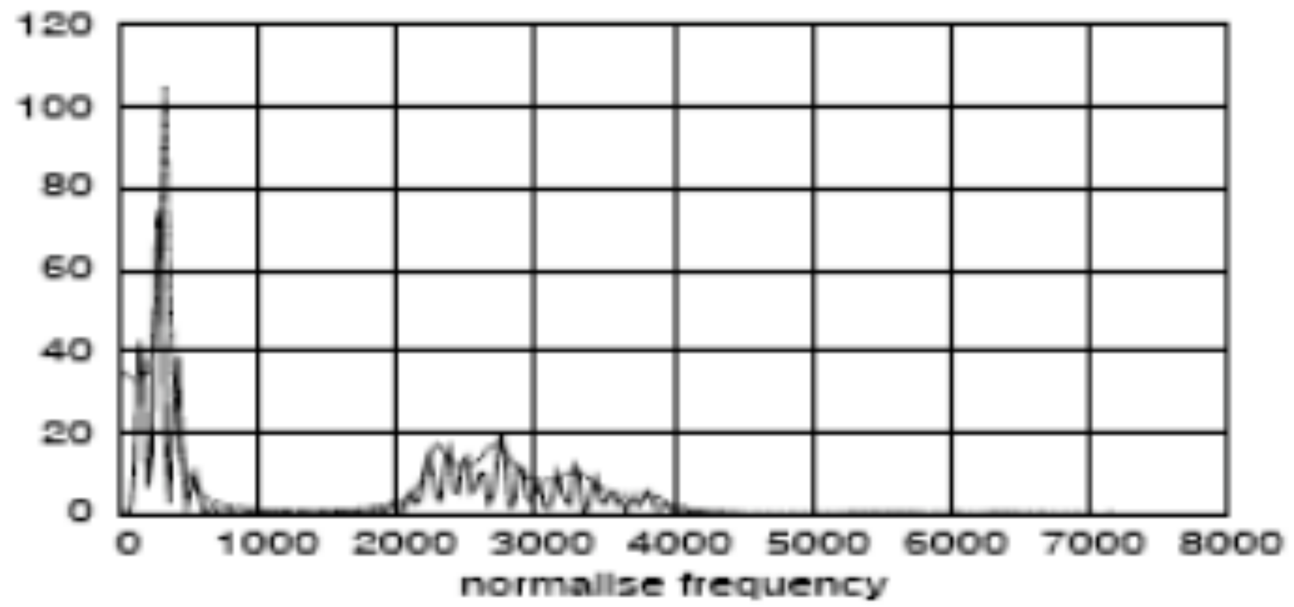


Cepstrum: Inverse DFT

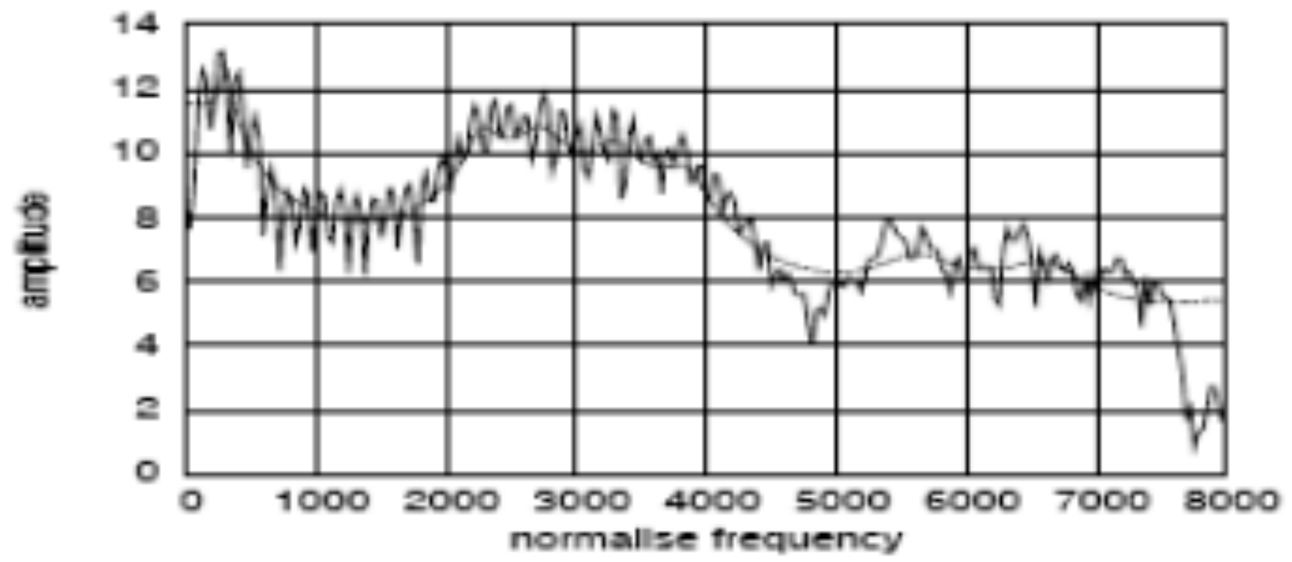
- Recall speech signals are created when a glottal source of a particular fundamental frequency passes through the vocal tract
- Most useful information for phone detection is the vocal tract filter (and not the glottal source)
- How do we deconvolve the source and filter to retrieve information about the vocal tract filter? Cepstrum

Cepstrum

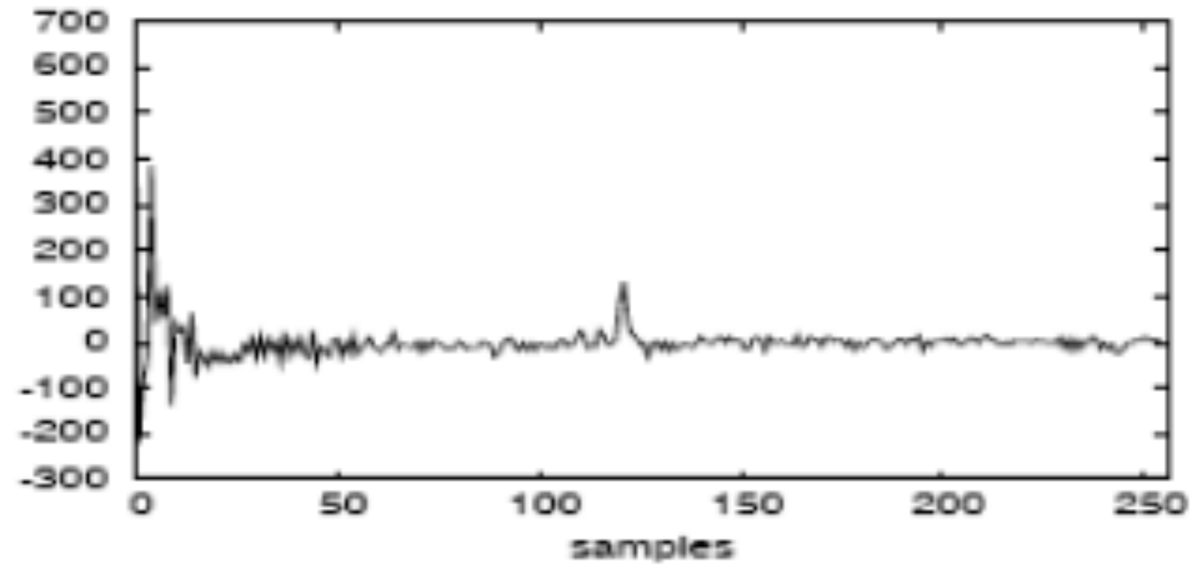
- Cepstrum: spectrum of the log of the spectrum



magnitude spectrum



log magnitude spectrum



cepstrum

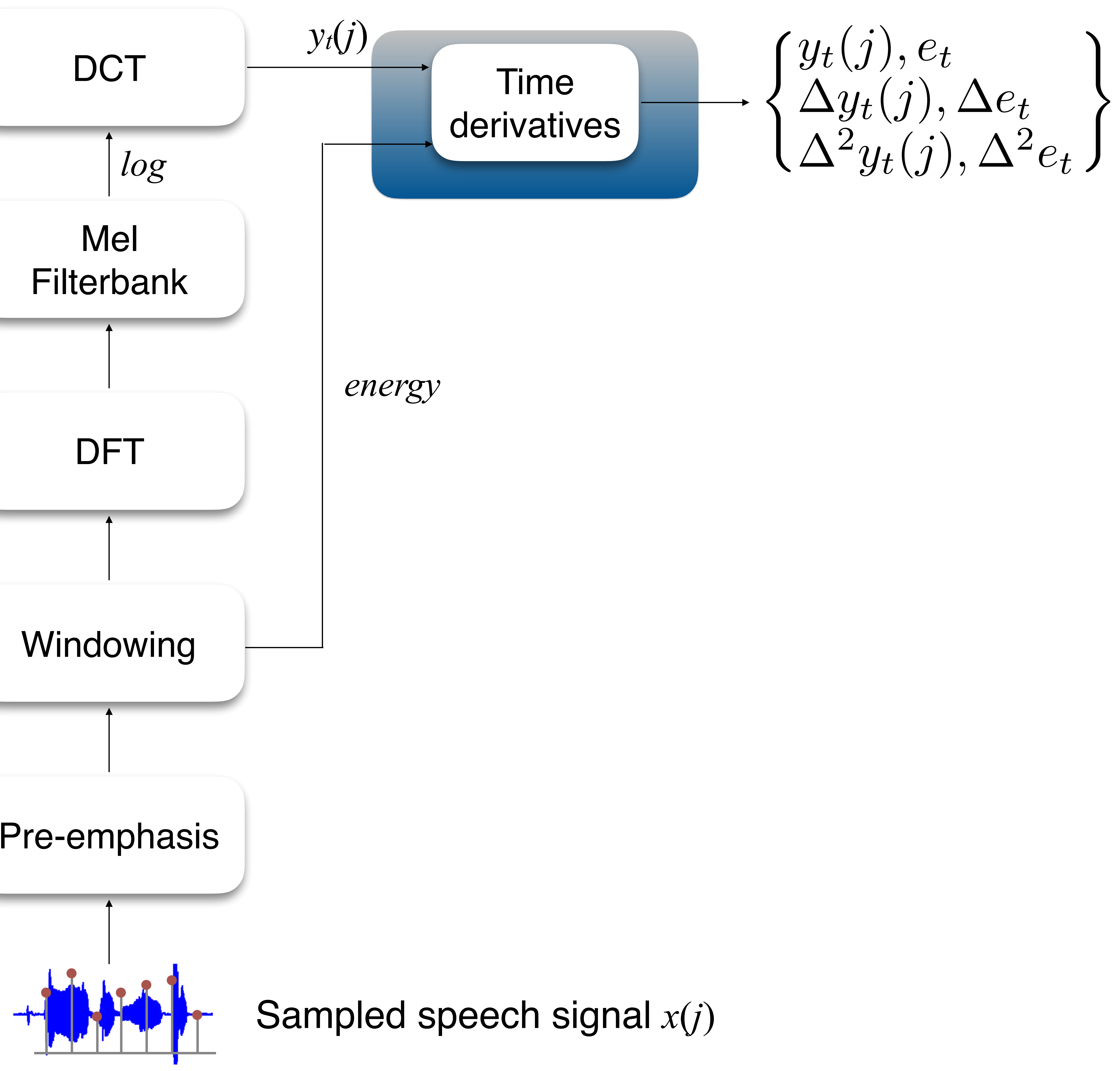
Image credit: Jurafsky & Martin, Figure 9.14

Cepstrum

- For MFCC extraction, we use the first 12 cepstral values
- Variance of the different cepstral coefficients tend to be uncorrelated
  - Useful property when modelling using GMMs in the acoustic model — diagonal covariance matrices will suffice
- Cepstrum is formally defined as the inverse DFT of the log magnitude of the DFT of a signal

$$c[n] = \sum_{n=0}^{N-1} \log \left(\left| \sum_{n=0}^{N-1} x[n] e^{-j \frac{2\pi}{N} kn} \right| \right) e^{j \frac{2\pi}{N} kn}$$

MFCC Extraction



Deltas and double-deltas

- From the cepstrum, use 12 cepstral coefficients for each frame
  - 13th feature represents energy from the frame — computed as sum of the power of the samples in the frame
- Also add features related to change in cepstral features over time to capture speech dynamics:

$$\Delta x_t = x_{t+\tau} - x_{t-\tau} \quad \text{(if } x_t \text{ is feature vector at time t)}$$

- Typical value for τ is 1 or 2.
- Add 13 delta features (Δx_t) and 13 double-delta features ($\Delta^2 x_t$)

Recap: MFCCs

- Motivated by human speech perception and speech production
- For each speech frame
  - Compute frequency spectrum and apply Mel binning
  - Compute cepstrum using inverse DFT on the log of the mel-warped spectrum
  - 39-dimensional MFCC feature vector: First 12 cepstral coefficients + energy + 13 delta + 13 double-delta coefficients

Other features

- Perceptual Linear Prediction (PLP) features
- Mel filter-bank features (used with DNNs)
- Neural network-based "bottleneck features" (covered in lecture 8)
    - Train deep NN using conventional acoustic features
    - Introduce a narrow hidden layer (e.g. 40 hidden units) referred to as the bottleneck layer, forcing the neural network to encode relevant information in this layer
    - Use hidden unit activations in the bottleneck layer as features

Features used for speaker recognition

- E.g. from a recent speaker identification (VoxCeleb) task.
- Input features, F: Spectrograms generated in a sliding window fashion using a Hamming window of width 25ms and step 10ms
- F used as input to a CNN architecture
- Mean and variance normalisation performed on every frequency bin of the spectrum (crucial for performance!)

Accuracy	Top-1 (%)	Top-5 (%)
I-vectors + SVM	49.0	56.6
I-vectors + PLDA + SVM	60.8	75.6
CNN-fc-3s no var. norm.	63.5	80.3
CNN-fc-3s	72.4	87.4

Nagrani et al.,“VoxCeleb: a large-scale speaker identification dataset”, Interspeech 2017

About pronunciations

- There exist a number of different alphabets to transcribe phonetic sounds
- E.g. ARPAbet (used in CMUdict)
- International Phonetic Alphabet (IPA) for all languages

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2018)

CONSONANTS (PULMONIC)

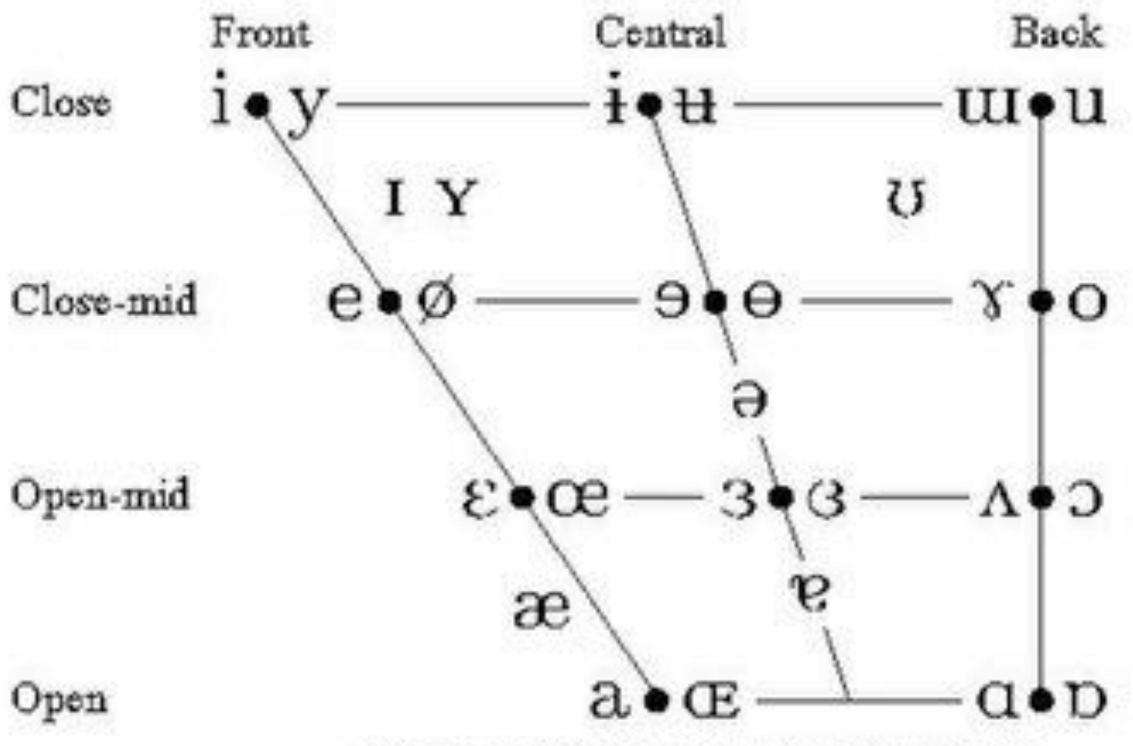
	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
ʘ Bilabial	ɓ Bilabial	ʼ Examples:
ǀ Dental	ɗ Dental/alveolar	pʼ Bilabial
ǃ (Post)alveolar	ʄ Palatal	tʼ Dental/alveolar
ǂ Palatoalveolar	ɠ Velar	kʼ Velar
ǁ Alveolar lateral	ʛ Uvular	sʼ Alveolar fricative

VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

OTHER SYMBOLS

ʍ Voiceless labial-velar fricative

w Voiced labial-velar approximant

ɥ Voiced labial-palatal approximant

ʜ Voiceless epiglottal fricative

ʢ Voiced epiglottal fricative

ɕ ʑ Alveolo-palatal fricatives

ɺ Voiced alveolar lateral flap

ɧ Simultaneous ʃ and x

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary: t͡s k͡p

SUPRASEGMENTALS

ˈ Primary stress

ˌ Secondary stress

ˌfoʊnəˈtɪʃən

Pronunciation Dictionary/Lexicon

- Pronunciation model/dictionary/lexicon: Lists one or more pronunciations for a word
- Typically derived from language experts: Sequence of phones written down for each word
- Dictionary construction involves:
    1. Selecting what words to include in the dictionary
    2. Pronunciation of each word (also, check for multiple pronunciations)

Graphemes vs. Phonemes

- Instead of a pronunciation dictionary, one could represent a pronunciation as a sequence of graphemes (or letters). That is, model at the grapheme level.

- Useful technique for low-resourced/under-resourced languages

- Main advantages:

  1. Avoid the need for phone-based pronunciations

  2. Avoid the need for a phone alphabet

  3. Works pretty well for languages with a systematic relationship between graphemes (letters) and phonemes (sounds)

Grapheme-based ASR

Language	ID	System	WER (%)		
			Vit	CN	CNC
Kurmanji Kurdish	205	Phonetic	67.6	65.8	64.1
		Graphemic	67.0	65.3	
Tok Pisin	207	Phonetic	41.8	40.6	39.4
		Graphemic	42.1	41.1	
Cebuano	301	Phonetic	55.5	54.0	52.6
		Graphemic	55.5	54.2	
Kazakh	302	Phonetic	54.9	53.5	51.5
		Graphemic	54.0	52.7	
Telugu	303	Phonetic	70.6	69.1	67.5
		Graphemic	70.9	69.5	
Lithuanian	304	Phonetic	51.5	50.2	48.3
		Graphemic	50.9	49.5	

Image from: Gales et al., Unicode-based graphemic systems for limited resource languages, ICASSP 15

Grapheme to phoneme (G2P) conversion

- Produce a pronunciation (phoneme sequence) given a written word (grapheme sequence)
- Learn G2P mappings from a pronunciation dictionary
- Useful for:
  - ASR systems in languages with no pre-built lexicons
  - Speech synthesis systems
  - Deriving pronunciations for out-of-vocabulary (OOV) words

G2P Conversion

- One popular paradigm: Joint sequence models [BN12]
  - Grapheme and phoneme sequences are first aligned using EM-based algorithm
  - Results in a sequence of graphones (joint G-P tokens)
  - Ngram models trained on these graphone sequences
- WFST-based implementation of such a joint graphone model [Phonetisaurus]

[BN12]:Bisani & Ney , "Joint sequence models for grapheme-to-phoneme conversion",Specom 2012

[Phonetisaurus] J. Novak, Phonetisaurus Toolkit