


In the name of God  
the Compassionate, the Merciful



# Automatic Speech Recognition (ASR) and Text- to-Speech

Zahra Rahaie, PhD

# Examples

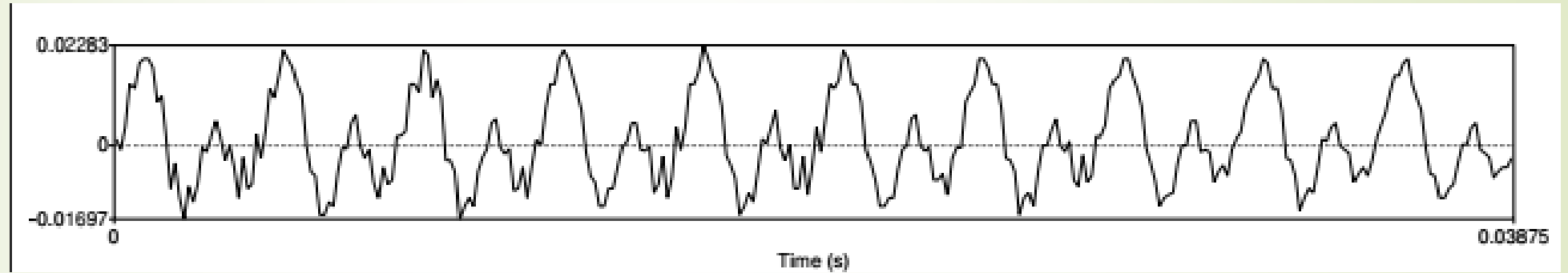
- Read Speech
- Conversational Speech



# Datasets

- ▶ LibriSpeech: read-speech 16 kHz dataset with over 1000 hours of audio books from the LibriVox project, with transcripts aligned at the sentence level
- ▶ Switchboard: telephone : 2430 conversations averaging 6 minutes each, totaling 240 hours of 8 kHz speech and about 3 million words
- ▶ CALLHOME 120 unscripted 30-minute
- ▶ Santa Barbara Corpus of Spoken American English
- ▶ CORAAL is a collection of over 150 sociolinguistic interviews with African American speakers
- ▶ CHiME **Challenge** is a series of difficult shared tasks with corpora that deal with robustness in ASR.
- ▶ HKUST Mandarin Telephone Speech corpus has 1206 ten-minute telephone conversations between speak
- ▶ AISHELL-1 corpus contains 170 hours of Mandarin read speech

# Feature Extraction for ASR: Log Mel Spectrum



- A waveform of an instance of the vowel [iy] (the last vowel in the word “baby”).
- The y-axis shows the level of air pressure above and below normal atmospheric pressure.
- The x-axis shows time. Notice that the wave repeats regularly

# Digitizing Sound Wave..

- = convert the **analog** representations (first air pressure and then analog electric signals in a microphone) sampling into a **digital** signal
- This **analog-to-digital conversion** has two steps: **sampling** and **quantization**.
- = To **sample** a signal, measure its **amplitude** at a particular time; the sampling **rate** is the number of samples taken **per second**.
- The maximum frequency wave that can be measured is one whose frequency is half the sample rate (since every cycle needs two samples) = **Nyquist frequency**

# Example: Sampling

- Most information in human speech is in frequencies below **10,000 Hz**; thus, a 20,000 Hz sampling rate would be necessary for complete accuracy.
- But telephone speech is filtered by the switching network, and only frequencies less than **4,000 Hz** are transmitted by telephones. Thus, an 8,000 Hz sampling rate is sufficient for telephone-bandwidth speech like the Switchboard corpus
- while **16,000 Hz** sampling is often used for microphone speech



# Quantization

- ▶ All values that are closer together than the minimum granularity (the **quantum size**) are represented identically. We refer to each sample at time index  $n$  in the digitized, quantized waveform as  $x[n]$
- ▶ Amplitude measurements are stored as integers, either 8 bit (values from -128–127) or 16 bit (values from -32768–32767).
- ▶ sample rate and sample size:
  - ▶ telephone speech is often sampled at 8 kHz and stored as 8-bit samples
  - ▶ microphone data is often sampled at 16 kHz and stored as 16-bit samples

# Num of Channels

- For stereo data or for two-party conversations, we can store both channels in the same file or we can store them in separate files.

The next parameter is individual sample storage—**linearly or compressed**

- Compression format: mu-law: human hearing is more sensitive at small intensities than large ones; the **log** represents small values with more faithfulness at the expense of more error on large values. ( $\mu=255$ )

$$F(x) = \frac{\text{sgn}(x) \log(1 + \mu|x|)}{\log(1 + \mu)} \quad -1 \leq x \leq 1$$

- linear** (unlogged) values are generally referred to as linear PCM values (PCM stands for pulse code modulation)



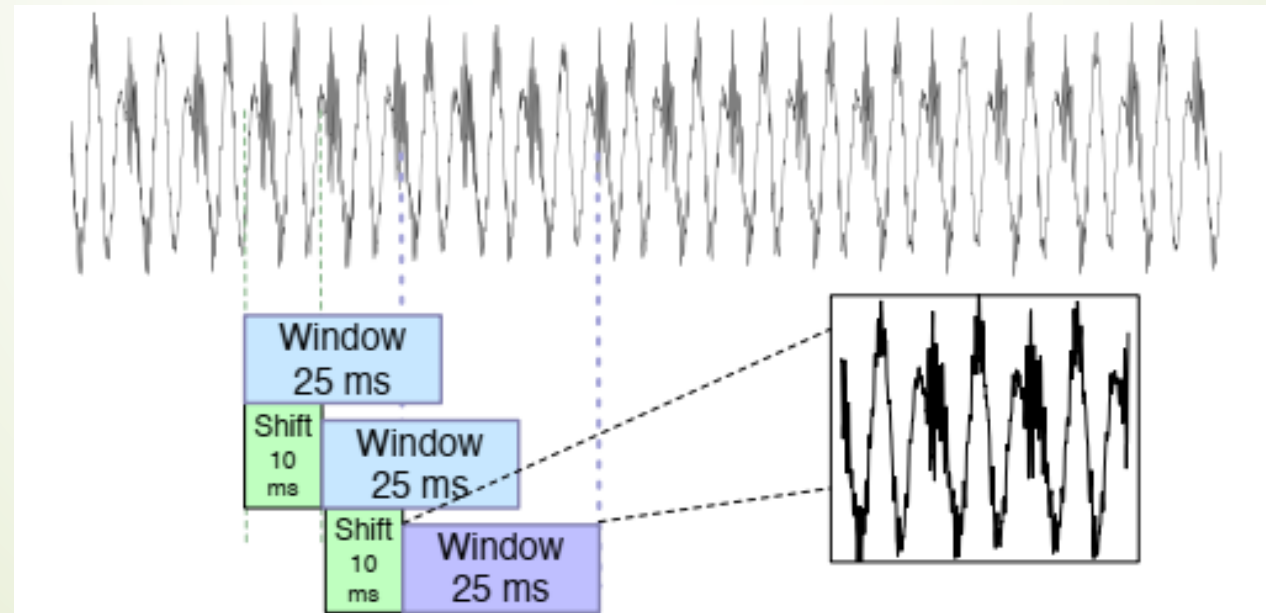
# Windowing

- ▶ need to extract spectral features from a small window of speech that characterizes part of a particular phoneme
- ▶ Inside this small window, we can roughly think of the signal as **stationary** (= statistical properties are constant over time within this region)
- ▶ We extract this roughly stationary portion of speech by using a window which is non-zero inside a region and zero elsewhere, running this window across the speech signal and multiplying it by the input waveform to produce a windowed waveform

$$y[n] = w[n]s[n]$$

# Continue..

- The speech extracted from each window is called a **frame**. The windowing is characterized by **three** parameters: the window **size** or frame size of the window (its width in milliseconds), the frame **stride**, (also called shift or offset) between successive windows, and the **shape** of the window



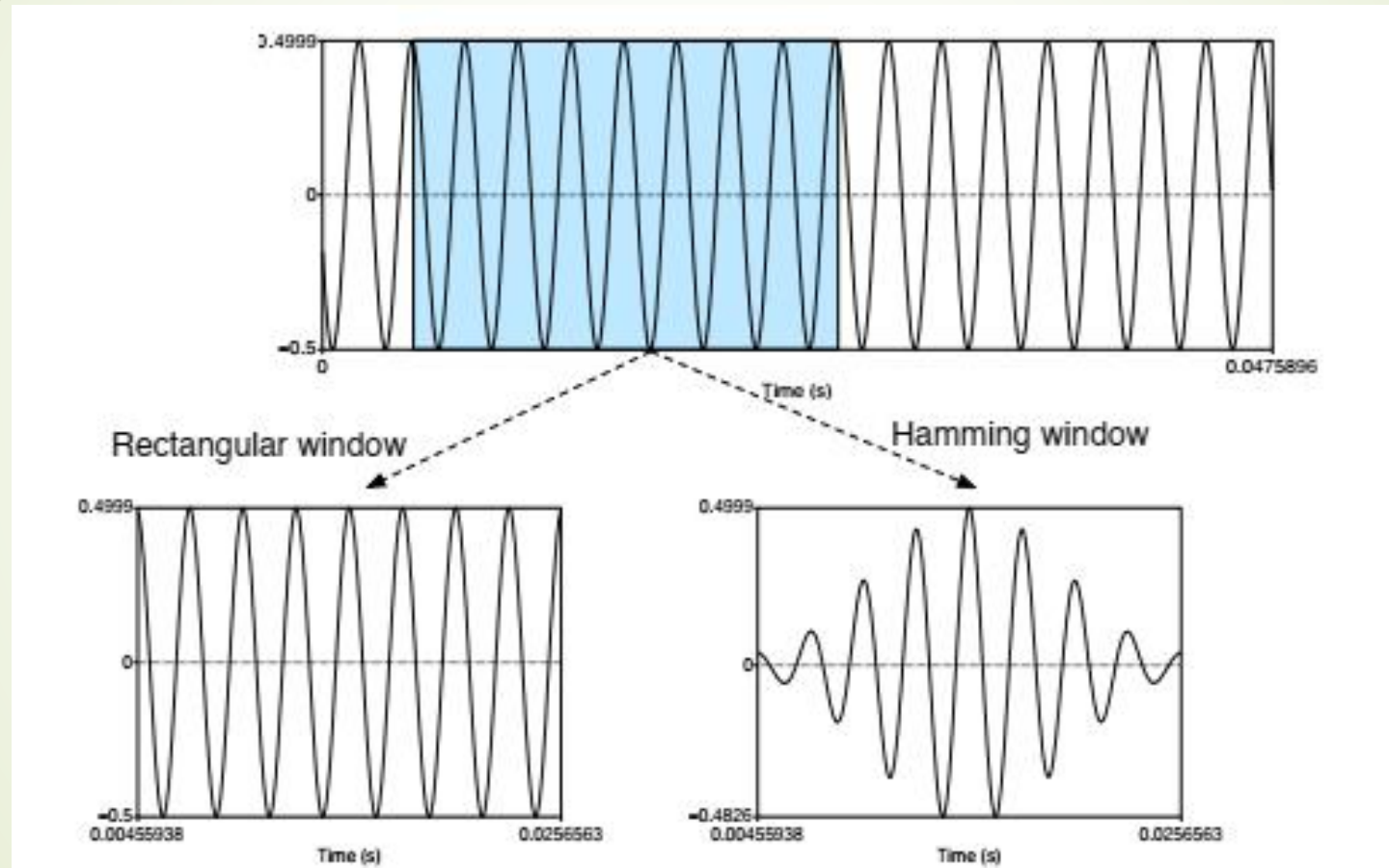
# Types

- **Rectangular:** abruptly cuts off the signal at its boundaries, which creates problems when we do Fourier analysis.
- **Hamming:** which shrinks the values of the signal toward zero at the window boundaries, avoiding discontinuities.

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

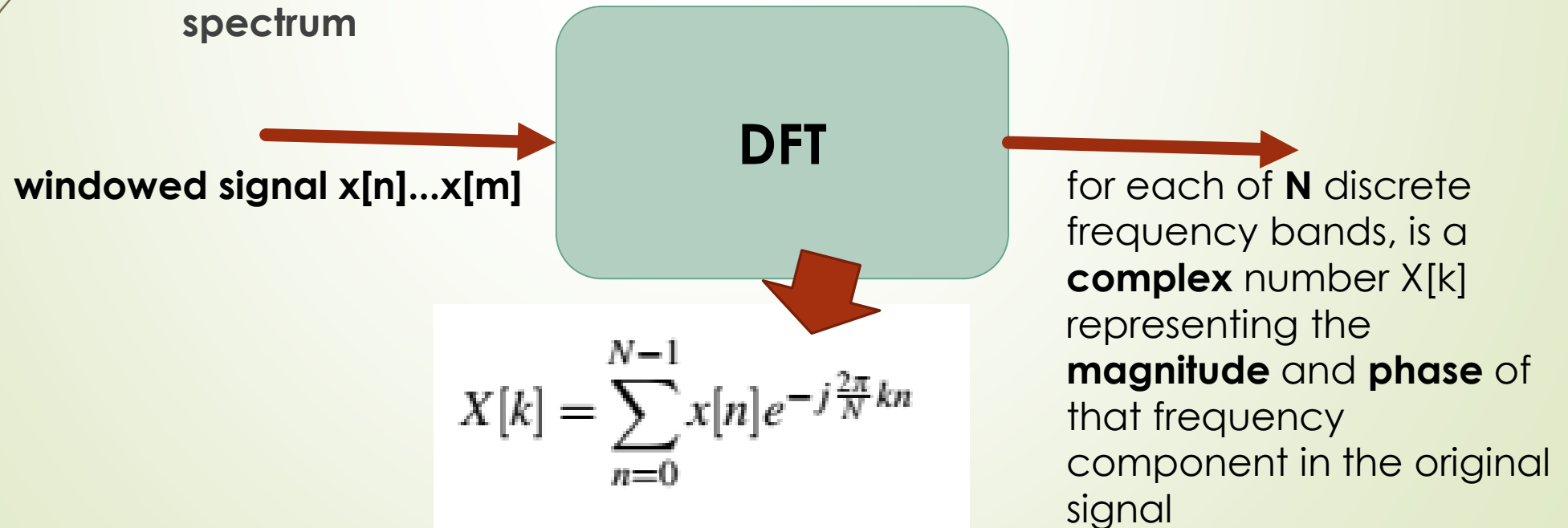
$$\textit{Hamming} \quad w[n] = \begin{cases} 0.54 - 0.46 \cos\left(\frac{2\pi n}{L}\right) & 0 \leq n \leq L-1 \\ 0 & \text{otherwise} \end{cases}$$

# Example

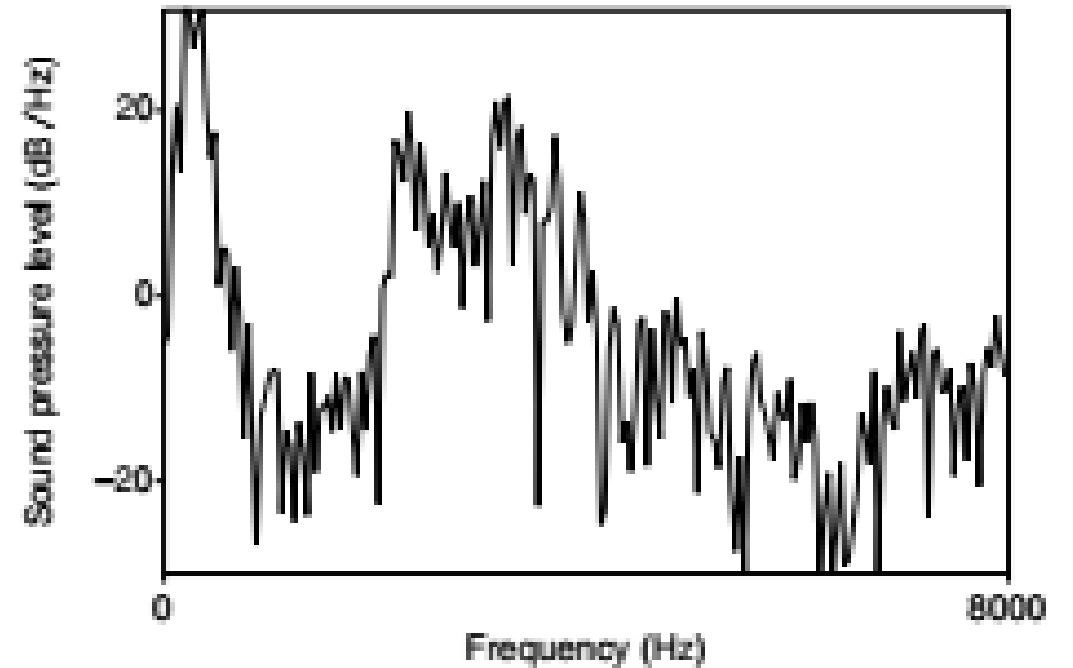
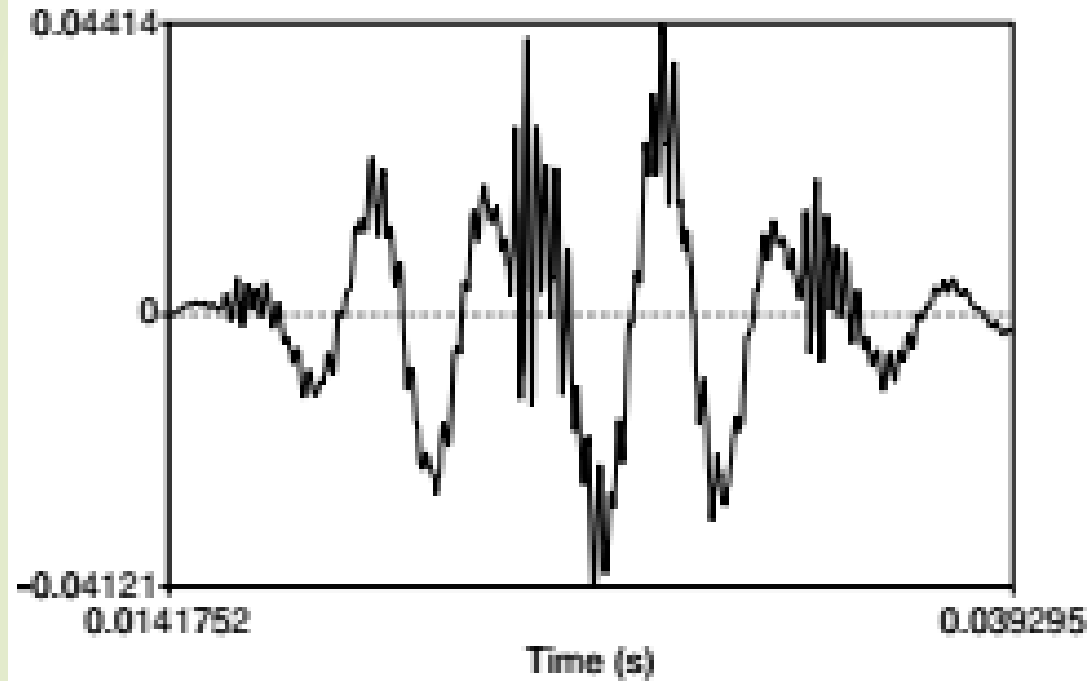


# DFT (Discrete Fourier Transform)

- The next step is to extract **spectral** information for our windowed signal
- We need to know how much **energy** the signal contains at different frequency bands
- If we plot the magnitude against the frequency, we can visualize the **spectrum**



# Spectrum, Example



- A commonly used algorithm for computing the DFT is the **fast Fourier transform** or FFT. This implementation of the DFT is very efficient but only works for values of **N that are powers of 2**



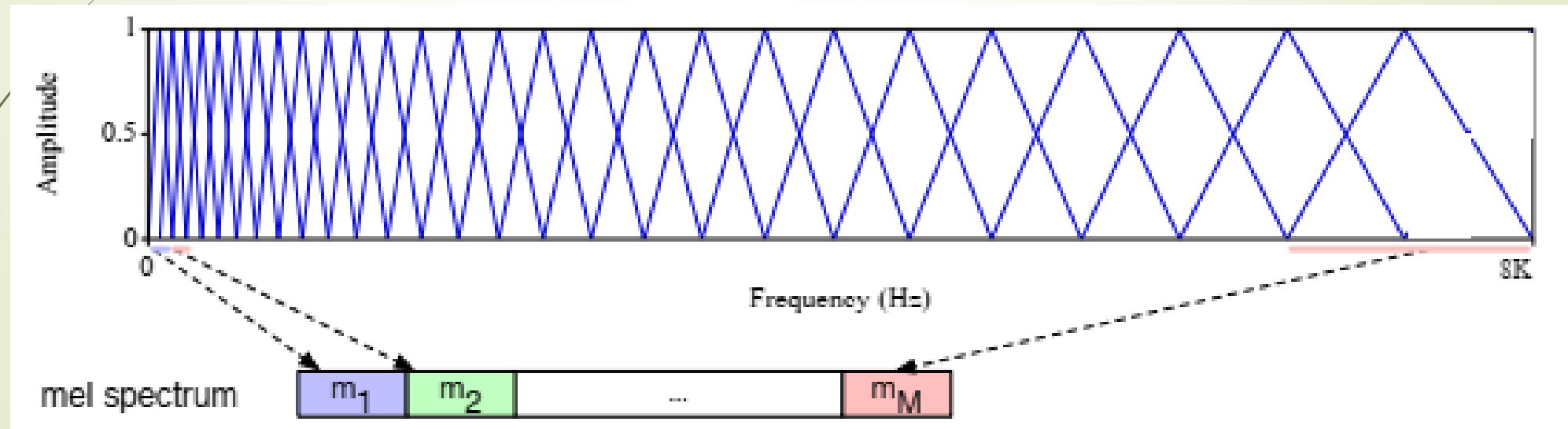
# Mel Filter Bank and Log

- ▶ The results of the FFT tell us the **energy** at each **frequency** band.
- ▶ Human hearing, however, is **not equally sensitive** at **all frequency** bands;
- ▶ This bias toward **low frequencies** helps human recognition, since information in low frequencies (like formants) is crucial for distinguishing **vowels or nasals**, while information in high frequencies (like stop bursts or fricative noise) is less crucial for successful recognition.
- ▶ The **mel** frequency  $m$  can be computed from the raw acoustic frequency by a log transformation:

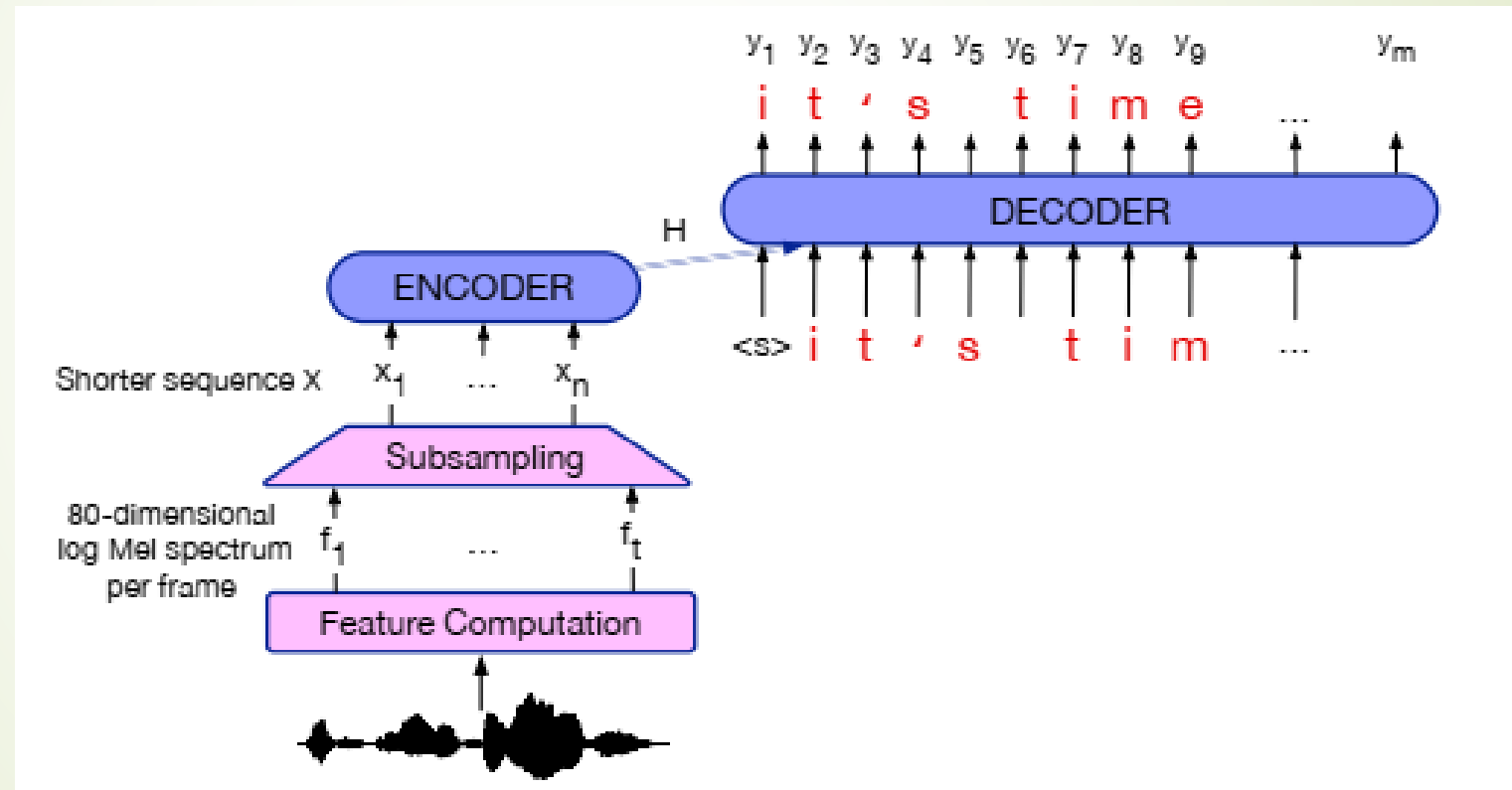
$$mel(f) = 1127 \ln\left(1 + \frac{f}{700}\right)$$

# Mel

- creating a bank of filters that collect energy from each frequency band, spread logarithmically so that we have **very fine resolution at low frequencies, and less resolution at high frequencies.**



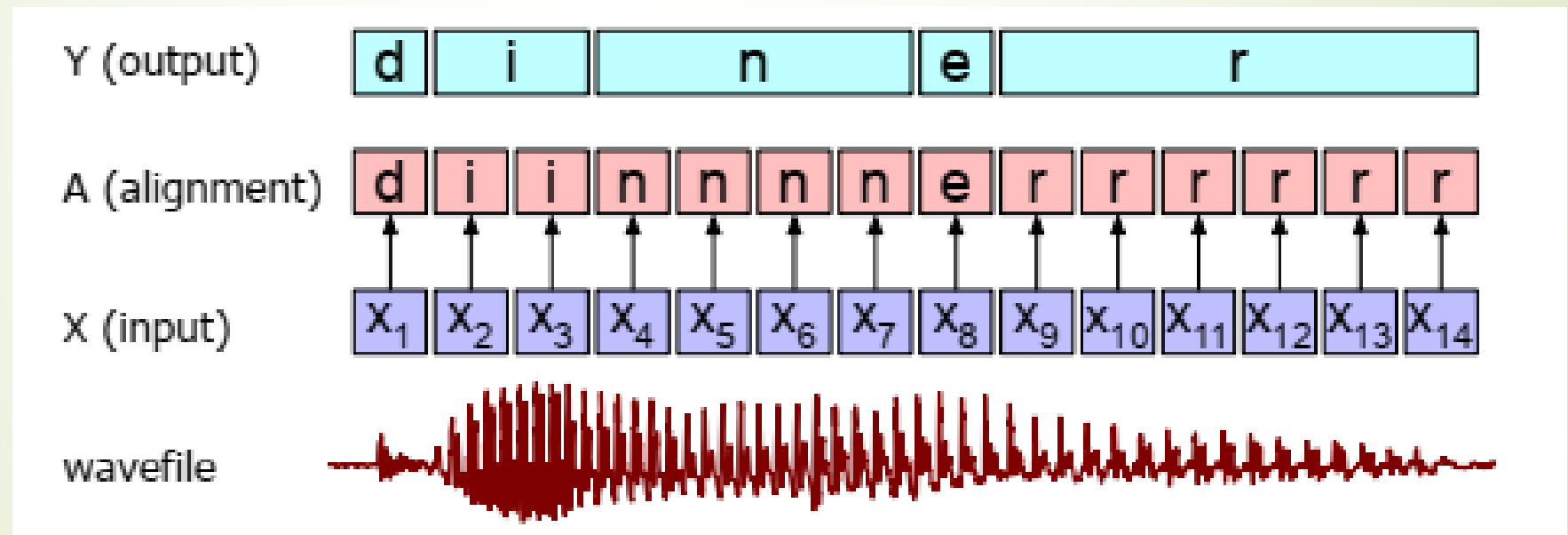
# Schematic architecture for an encoder-decoder speech recognizer



# Connectionist Temporal Classification

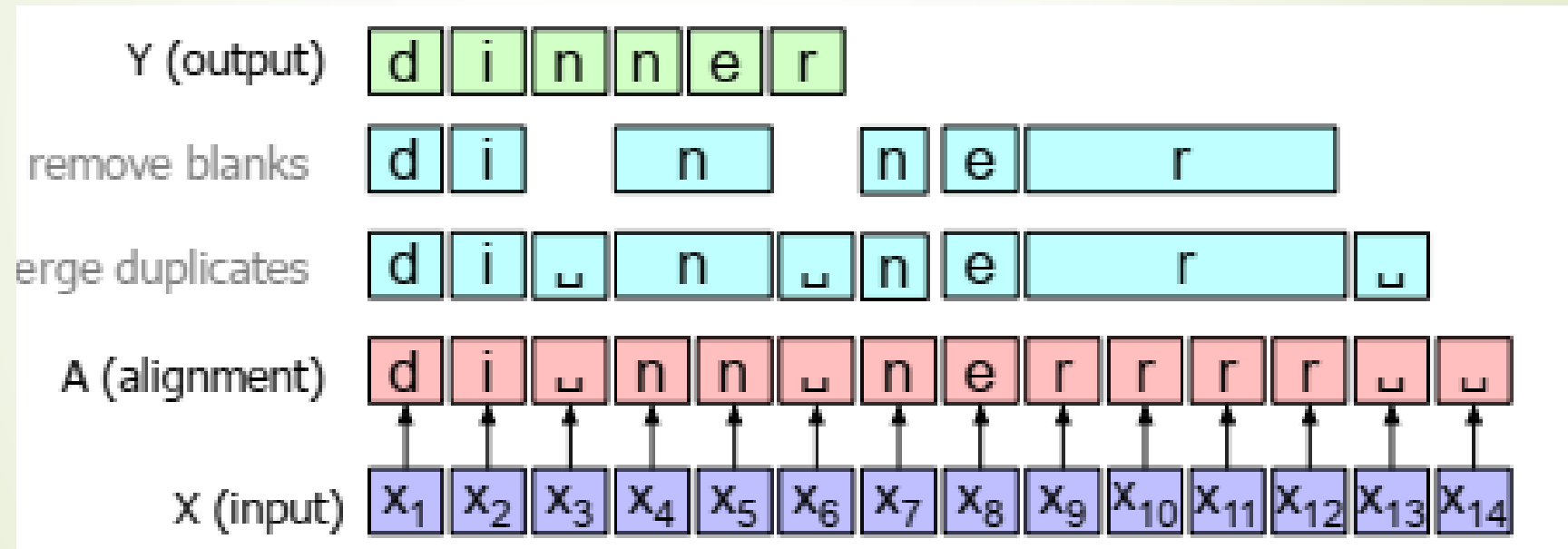
از نظر زمانی متصل

- BUT, Collapsing doesn't handle double letters. dine or dinner?

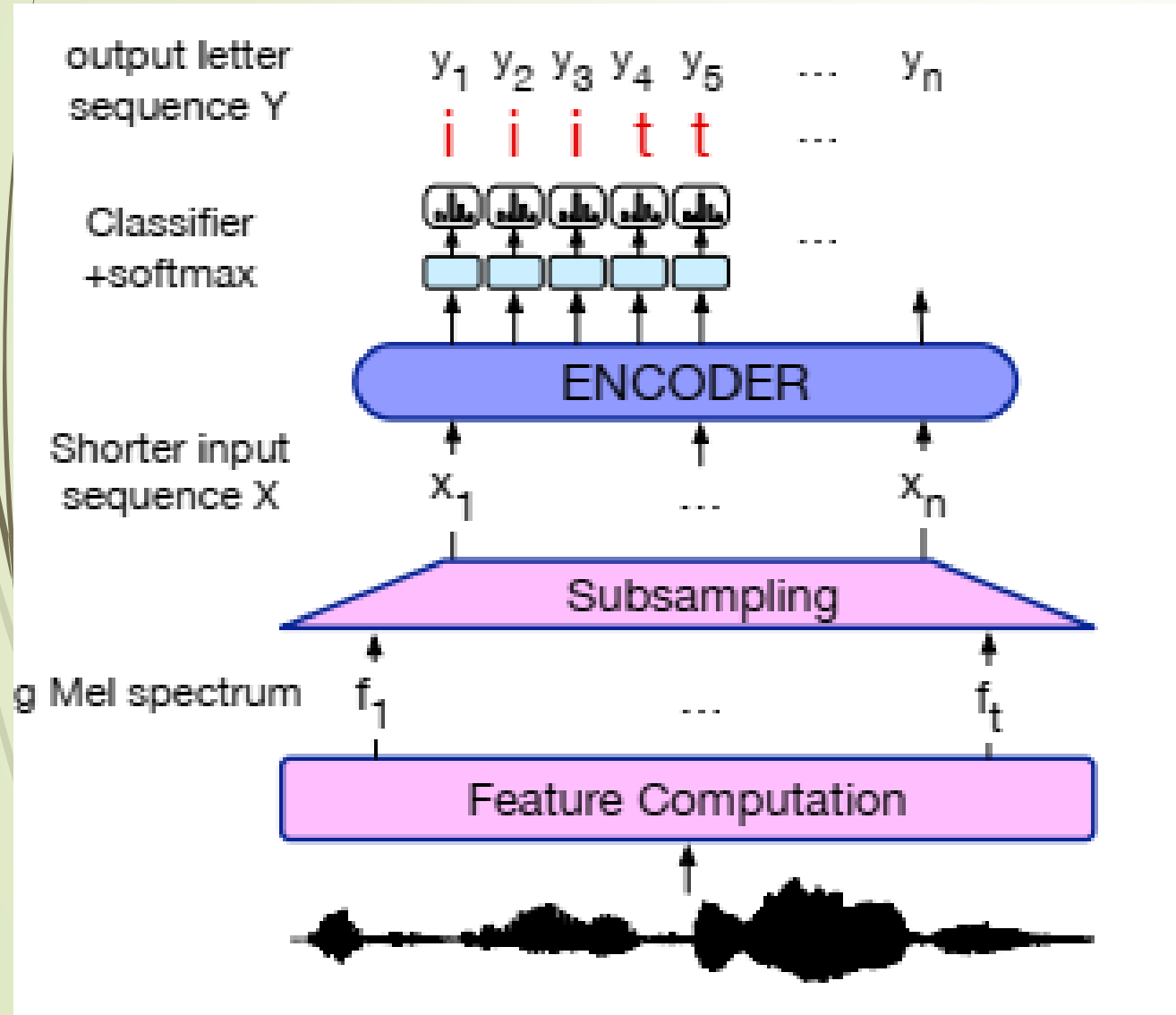


naive

The CTC collapsing function  $B$ , repeated (consecutive) characters in an alignment  $A$  are removed to form the output  $Y$ .



# CTC inference model



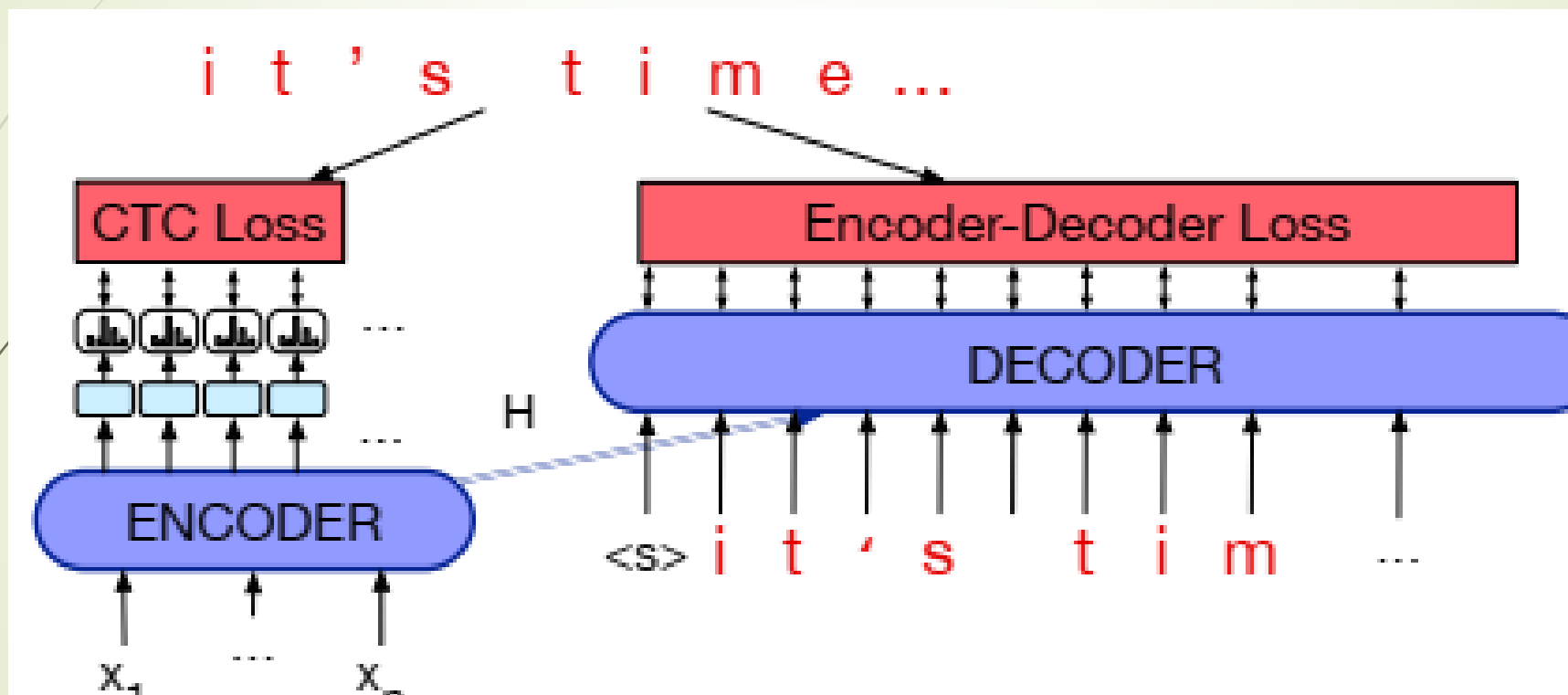
Using an encoder-only model, with decoding done by simple softmaxes over the hidden state  $h_t$  at each output step.

## Training:

use negative log-likelihood loss with a special CTC loss function



# Combining..



# ASR Evaluation: Word Error Rate

➤ minimum edit distance

$$\text{Word Error Rate} = 100 \times \frac{\text{Insertions} + \text{Substitutions} + \text{Deletions}}{\text{Total Words in Correct Transcript}}$$

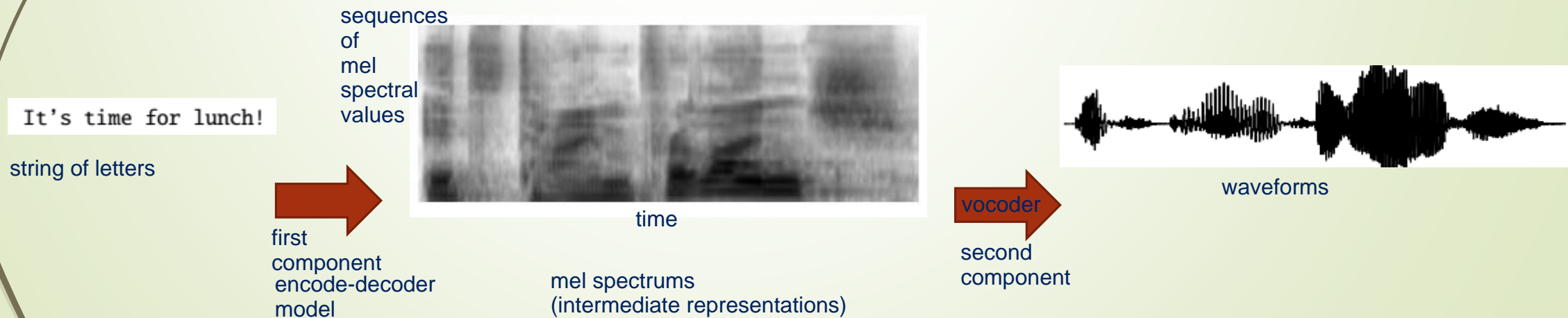
REF:	i	***	**	UM	the	PHONE	IS		i	LEFT	THE	portable	****	PHONE	UPSTAIRS	last	night
HYP:	i	GOT	IT	TO	the	*****	FULLEST	i	LOVE	TO	portable	FORM	OF		STORES	last	night
Eval:	I	I	S		D	S		S	S		I	S		S			

This utterance has six substitutions, three insertions, and one deletion:

$$\text{Word Error Rate} = 100 \frac{6 + 3 + 1}{13} = 76.9\%$$

# Text-to-speech (TTS) systems

- We generally break up the TTS task into **two** components.
- The **first** component is an encoder-decoder model for spectrogram prediction: it maps from strings of letters to mel spectrographs: sequences of mel spectral values over time.
- The **second** component maps from mel spectrograms to waveforms. Generating waveforms from intermediate representations like spectrograms is called **vocoding** and this second component is called a **vocoder**.



# TTS Preprocessing: Text normalization

- ▶ TTS systems require text normalization preprocessing for handling non-standard words: numbers, monetary amounts, dates, and other concepts that are verbalized differently than they are spelled.

**seventeen fifty:** (in "*The European economy in 1750*")

**one seven five zero:** (in "*The password is 1750*")

**seventeen hundred and fifty:** (in "*1750 dollars*")

**one thousand, seven hundred, and fifty:** (in "*1750 dollars*")

# Semiotic class

کلاس نشانه شناسی

وابسته به علائم مرض

کلامی کردن

- Often the verbalization of a non-standard word depends on its **meaning** = **semiotic** class

semiotic class	examples	verbalization
abbreviations	gov't, N.Y., mph	government
acronyms read as letters	GPU, D.C., PC, UN, IBM	G P U
cardinal numbers	12, 45, 1/2, 0.6	twelve
ordinal numbers	May 7, 3rd, Bill Gates III	seventh
numbers read as digits	Room 101	one oh one
times	3.20, 11:45	eleven forty five
dates	28/02 (or in US, 2/28)	February twenty eighth
years	1999, 80s, 1900s, 2045	nineteen ninety nine
money	\$3.45, €250, \$200K	three dollars forty five
money in tr/m/billions	\$3.45 billion	three point four five billion dollars
percentage	75% 3.4%	seventy five percent

# Normalization

- In languages with grammatical **gender**, normalization may depend on morphological properties
- 1 mangue -> une mangue
- 1 ananas -> un ananas
- Normalization can be done by
  - rule
    - 1-Tokenization: regular expressions
    - 2-verbalization
  - encoder-decoder model



# TTS Evaluation

- **Mean opinion score (MOS)**, a rating of how good the synthesized utterances are, usually on a scale from 1–5.
- **AB test**: compare two systems
  - In AB tests, we play the same sentence synthesized by two different systems (an A and a B system). The human listeners choose which of the two utterances they like better. We do this for say 50 sentences (presented in random order) and compare the number of sentences preferred for each system

# Other Speech Tasks

- **Wake-word detection:** wake up a voice-enabled assistant like Alexa, Siri, or the Google Assistant
- Thus wake word detectors need to be **fast, small footprint software that can fit into embedded devices**. Wake word detectors usually use the same frontend feature extraction we saw for ASR, often followed by a whole-word classifier
- **Speaker Diarization:** who spoke when. This can be useful for transcribing meetings, classroom speech, or medical interactions.
- **Speaker Recognition:** such as for security when accessing personal information over the telephone, and **speaker identification**, where we make a one of N decision trying to match a speaker's voice against a database of many speakers. These tasks are related to **language identification**, in which we are given a wavefile and must identify which language is being spoken; this is useful for example for automatically directing callers to human operators that speak appropriate languages.

# Conclusion

- Speech Recognition
- Text to Speech

# References

- Chapter 16 of NLP book

