Automatic Speech Recognition & Text-to-Speech

Kwankamol Nongpong Ph.D.

CS4430 Selected Topic in Natural Language Processing and Social Interaction
CSX4210 / ITX4210 Natural Language Processing and Social Interaction

Vincent Mary School of Science and Technology
Assumption University

Speech Processing

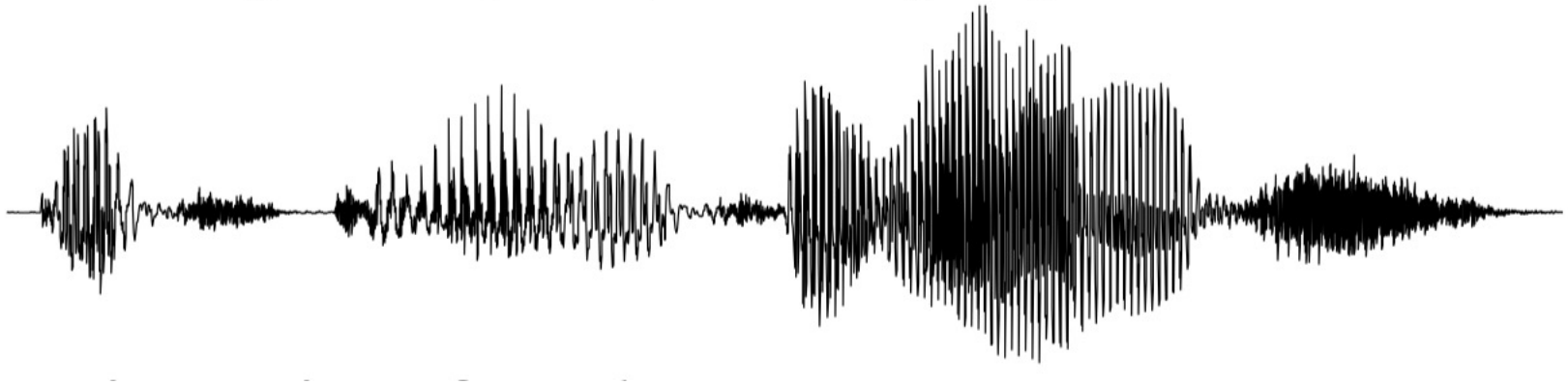
- Understand the spoken language
- Transcribe the words into writing

Radio Rex

- A toy that recognized speech from the 1920s.

The Task of ASR

- Map any waveform to an appropriate sequence of words.

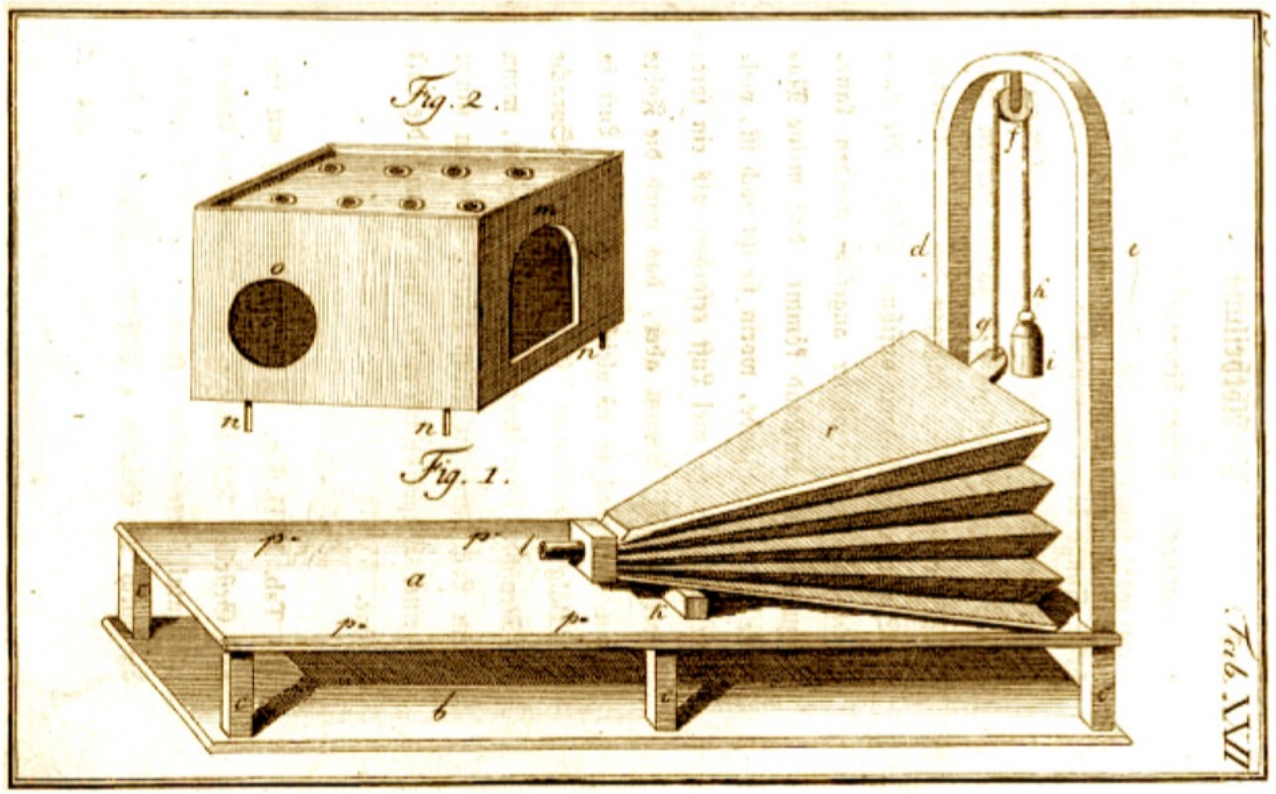


It’s time for lunch!

Applications of ASR

- Natural interface for communicating with smart home appliances, personal assistants, or cellphones.
- Call routing in telephony (call center).
- Automatically generating captions for audio or video text.
- Augmentative communication (interaction between computers and humans with some disability resulting in difficulties or inabilities in typing or audition)

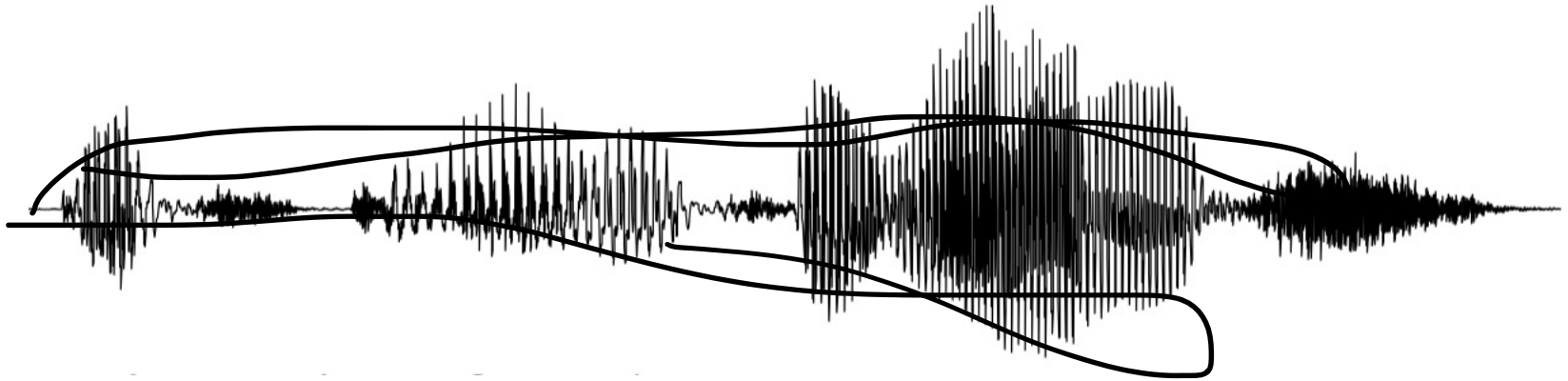
Text-to-Speech in 1790



The Task of Text-to-Speech (TTS)

- *aka* Speech Synthesis
- The reverse of ASR.

It’s time for lunch!



Automatic Speech Recognition (ASR)

Dimensions in ASR

- Dictionary size
  - 2-word vocabulary (Yes or No)
  - 11-word vocabulary (Digit Recognition)
  - 60k-word vocabulary (Transcribing videos or human conversations)
- Who the speaker is talking to
  - Human to machine vs. Human to human
  - Read speech
  - Conversational speech
- Channel and noise
  - Studio / Quiet room / Noisy street
- Accent

Speech Corpora

- LibriSpeech
  - 16 kHz
  - 1,000 hours of audio books
- Switchboard

- CALLHOME: 120 unscripted 30-minute telephone conversations between friends or family
- CORAAL: 150 Interviews with African American speakers

Word Error Rate & Character Error Rate

English Tasks	**WER%**
LibriSpeech audiobooks 960hour clean	1.4
LibriSpeech audiobooks 960hour other	2.6
Switchboard telephone conversations between strangers	5.8
CALLHOME telephone conversations between family	11.0
Sociolinguistic interviews, CORAAL (AAL)	27.0
CHiMe5 dinner parties with body-worn microphones	47.9
CHiMe5 dinner parties with distant microphones	81.3
Chinese (Mandarin) Tasks	**CER%**
AISHELL-1 Mandarin read speech corpus	6.7
HKUST Mandarin Chinese telephone conversations	23.5

Feature Extraction

- Transform waveform into a sequence of acoustic feature vectors.
- Log Mel Spectrum vectors are the most commonly used ones.

Analog-to-Digital Conversion

- Sampling
  - The sampling rate is the number of samples taken per second.
  - To accurately measure a wave, we must have at least two samples in each cycle:
    - one measuring the positive part of the wave
    - one measuring the negative part
  - More than two samples per cycle increases the amplitude accuracy, but less than two samples will cause the frequency of the wave to be completely missed.
- Quantization

Sampling

- The maximum frequency wave that can be measured is one whose frequency is half the sample rate
  - since every cycle needs two samples
- This maximum frequency for a given sampling rate is called the Nyquist frequency.
- Human speech is in frequencies below 10,000 Hz
  - 20,000 Hz sampling rate would be necessary for complete accuracy
- Telephone speech is in frequencies less than 4,000 Hz
  - 8,000 Hz sampling rate is sufficient.

Note on Sampling

- Higher sampling rates produces higher ASR accuracy
- The sampling rates for training and testing have to match.
  - If we are testing on a telephone corpus like Switchboard (8 KHz sampling), we must downsample our training corpus to 8 KHz.

Quantization

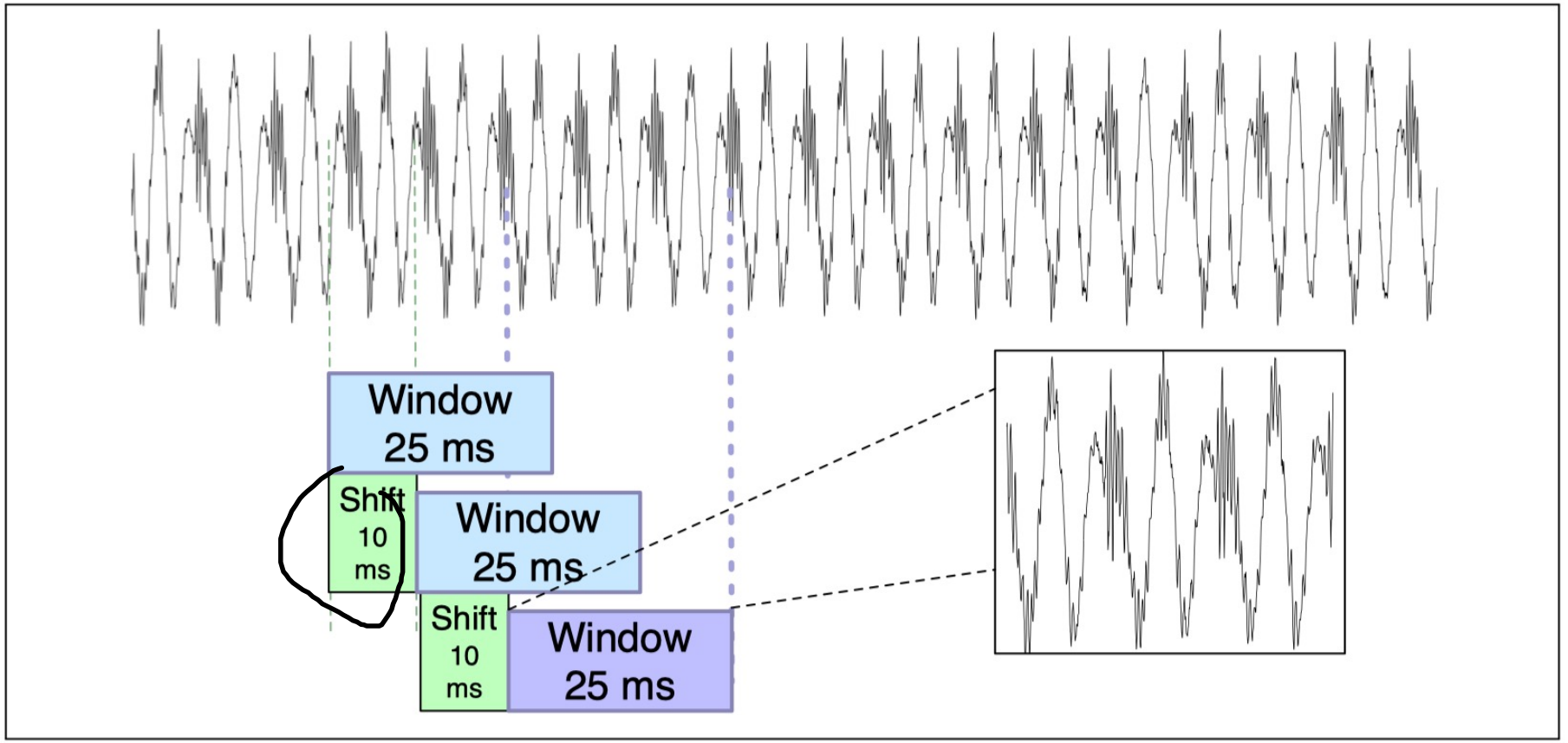
- Amplitude measurements are stored as integers,
  - 8 bit (values from -128 to 127)
  - 16 bit (values from -32768 to 32767)
- The process of representing real-valued numbers as integers is called quantization.
- All values that are closer together than the minimum granularity, i.e., the quantum size are represented identically.

Windowing (1/2)

- The spectral features must be extracted from a small window of speech.
- Its statistical properties are constant within this window/region (stationary).
  - In general, speech is a non-stationary signal.
  - Its statistical properties are not constant over time.
- 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.

Windowing (2/2)

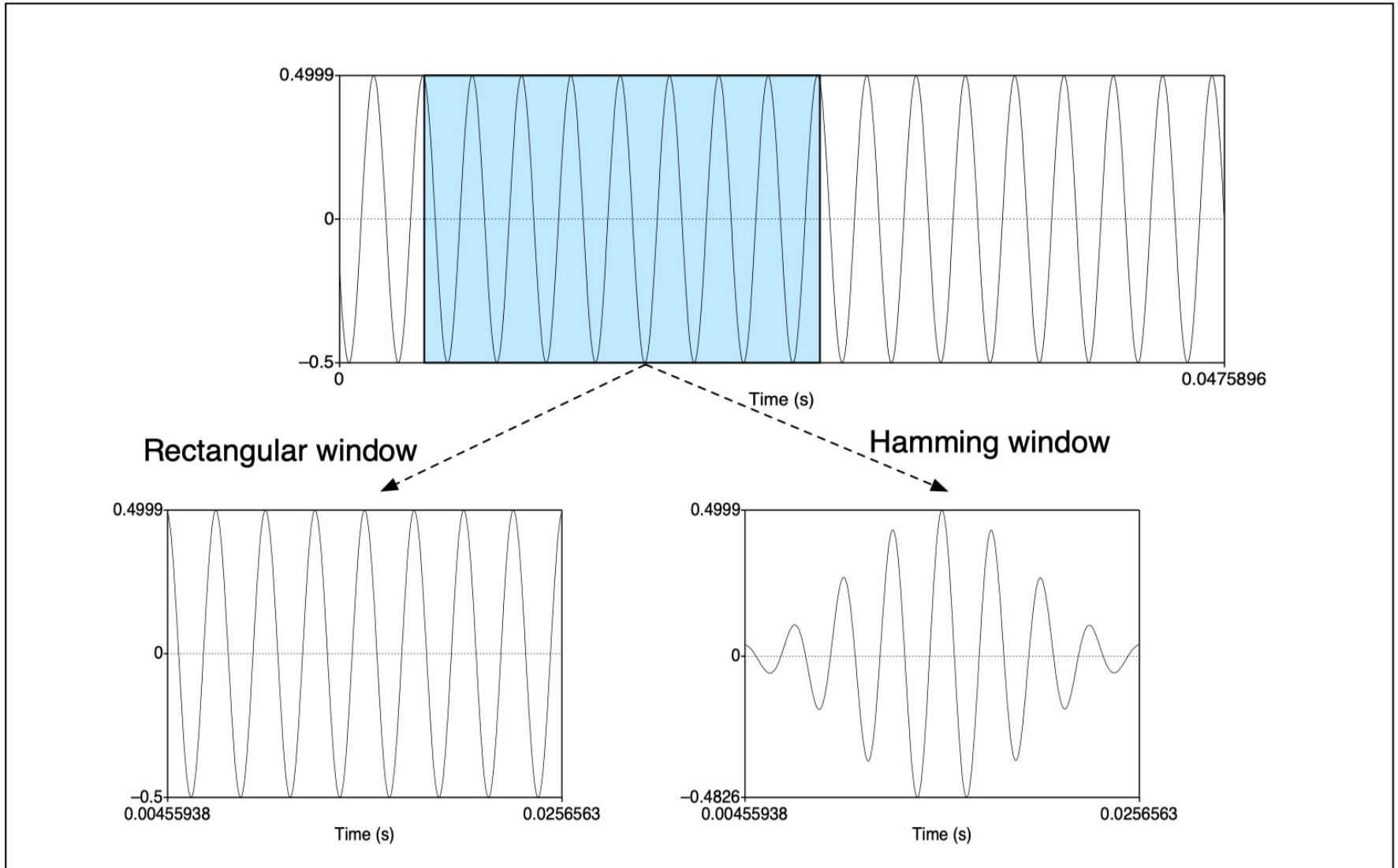
25ms window with 10ms stride



Hamming Window

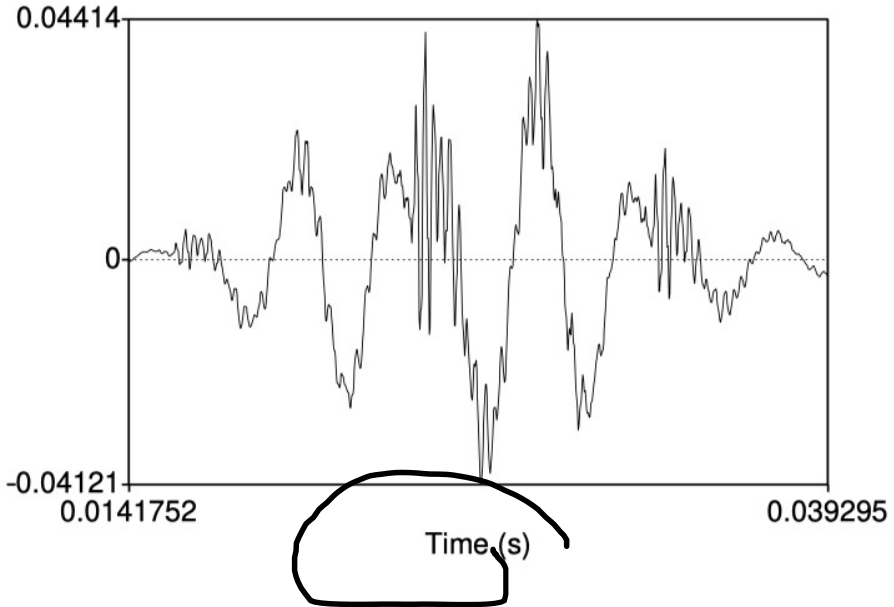
$$rectangular \quad w[n] = \begin{cases} 1 & 0 \le n \le L-1 \\ 0 & \text{otherwise} \end{cases}$$

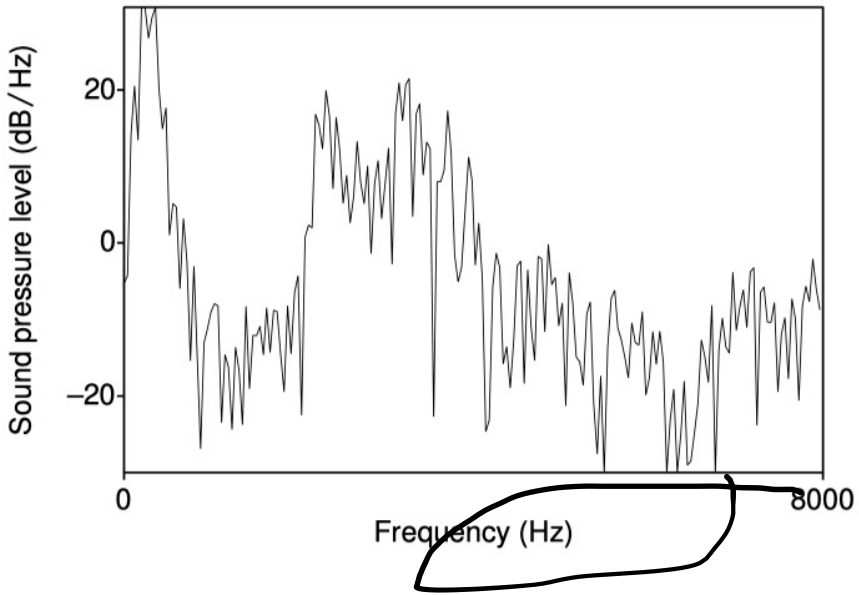
$$Hamming \quad w[n] = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{L}\right) & 0 \le n \le L-1 \\ 0 & \text{otherwise} \end{cases}$$



Discrete Fourier Transform

- Extract spectral information for our windowed signal
  - how much energy the signal contains at different frequency bands
- This can be done through Discrete Fourier Transform (DFT)
  - Fast Fourier transform or FFT





Mel Filter Bank

- The results of the FFT tell us the energy at each frequency band.
- Human hearing is not equally sensitive at all frequency bands
  - It is less sensitive at higher frequencies.
  - Modeling this human perceptual property improves speech recognition performance in the same way.
- This intuition can be implemented by collecting energies according to the mel scale, an auditory frequency scale (Chapter 25).
- A mel (Stevens et al. 1937, Stevens and Volkmann 1940) is a unit of pitch.

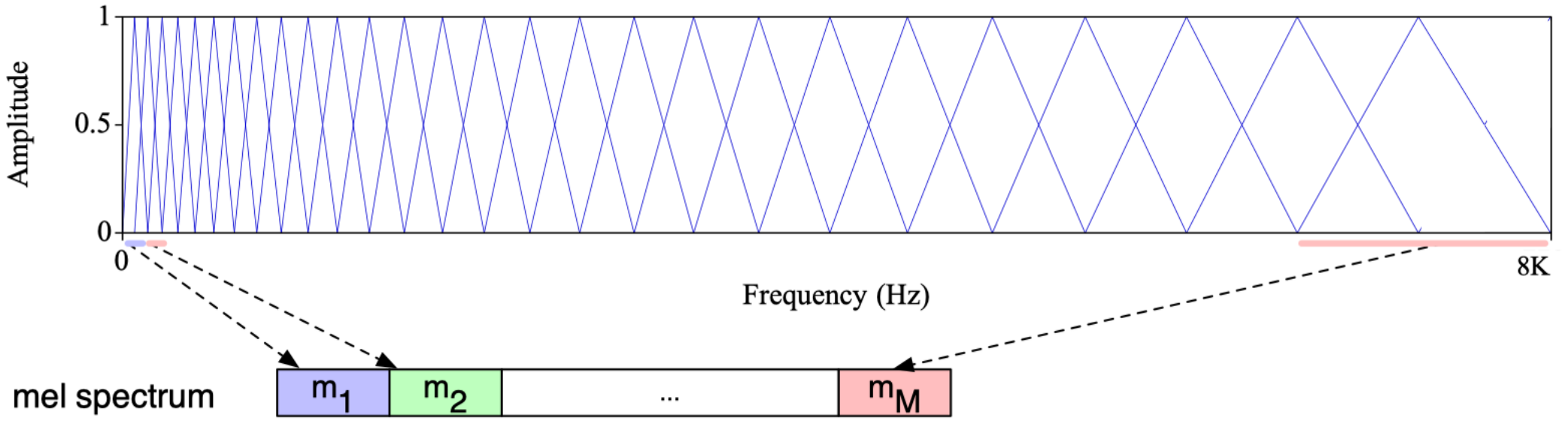
Mel Spectrum

- Pairs of sounds that are perceptually of the same distance in pitch are separated by an equal number of mels.

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

- Bank of filters that collect energy from each frequency band are created.
  - Fine resolution at low frequencies, and less resolution at high frequencies.

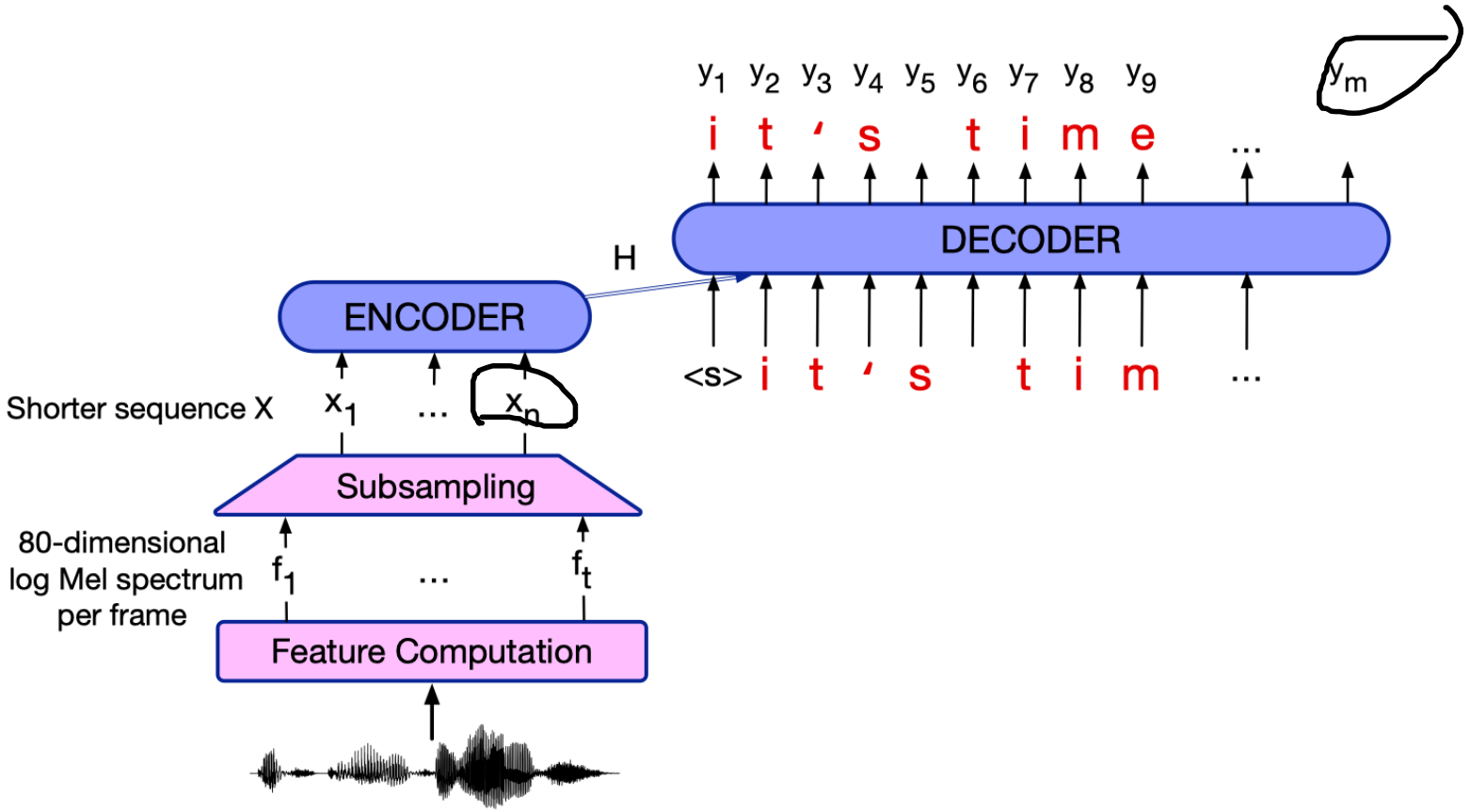
Mel Filter Bank



Speech Recognition Architecture

Basic Architecture of ASR

- Encoder-decoder implemented with either RNNs or Transformer.
  - Same architecture for Machine Translation



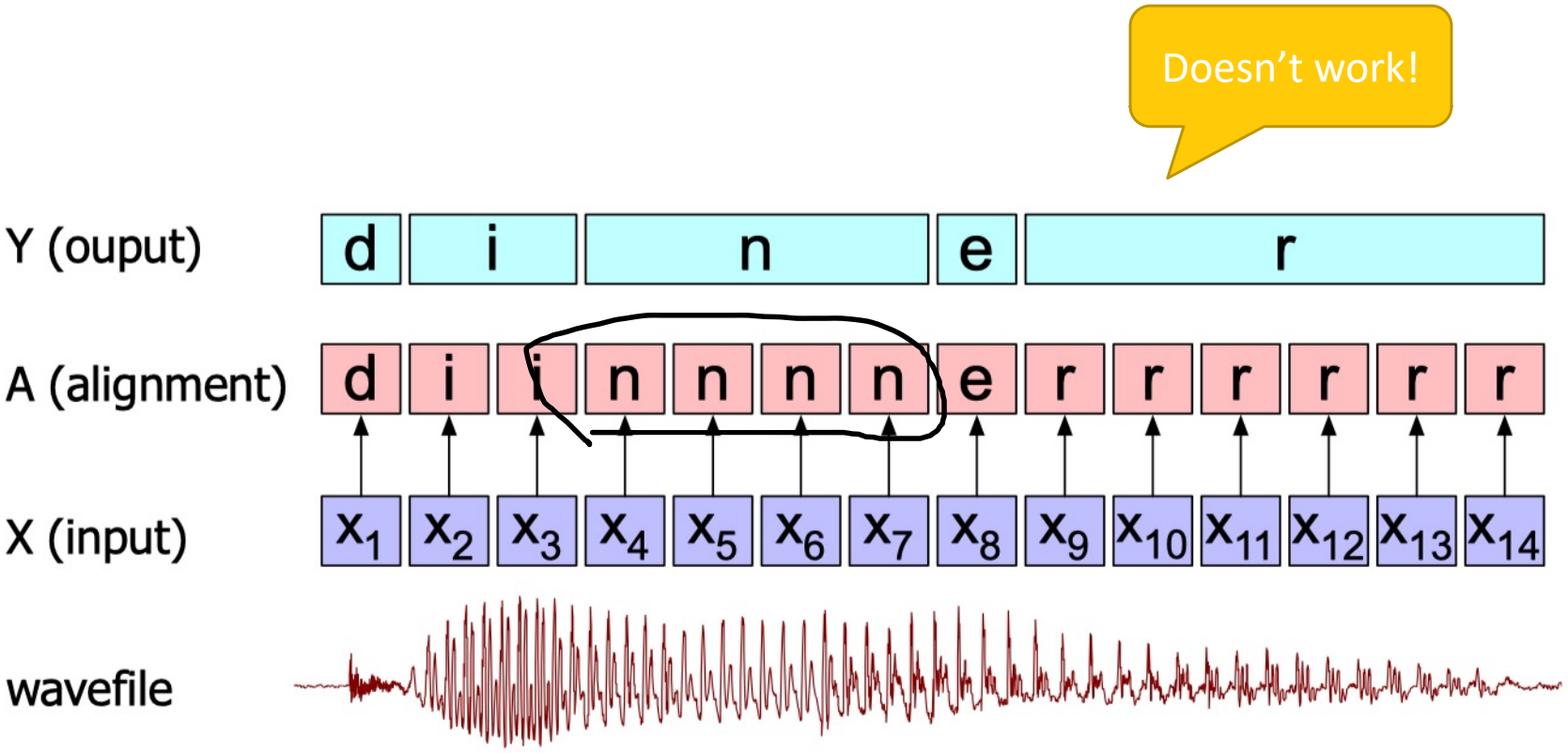
Encoder-Decoder Architecture for Speech

- The encoder-decoder architecture is particularly appropriate when input and output sequences have huge length differences.
  - as they do for speech i.e., very long acoustic feature sequences mapping to much shorter sequences of letters or words
- A single word might be 5 letters long but, supposing it lasts about 2 seconds, would take 200 acoustic frames (of 10ms each).
- There needs to be a special compression stage that shortens the acoustic feature sequence before the encoder stage.
  - ~~or CTC loss~~ function that deals very well with compression
- Encoder-decoders for speech are trained with the normal cross-entropy loss generally used for conditional language models.

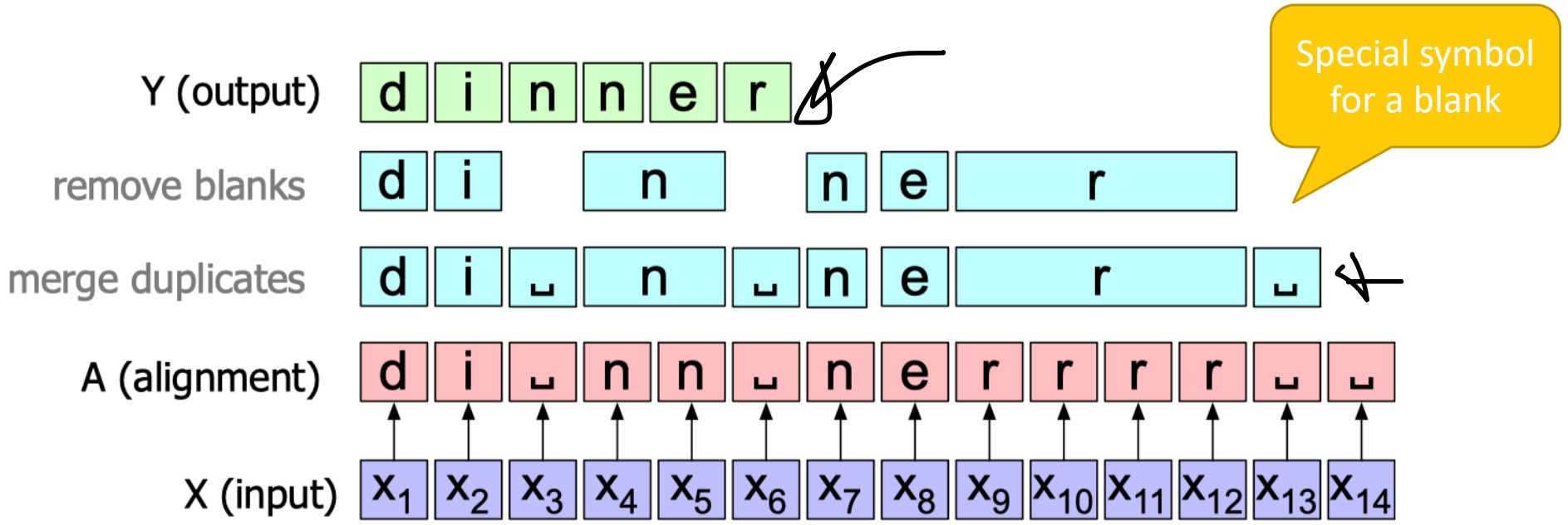
CTC

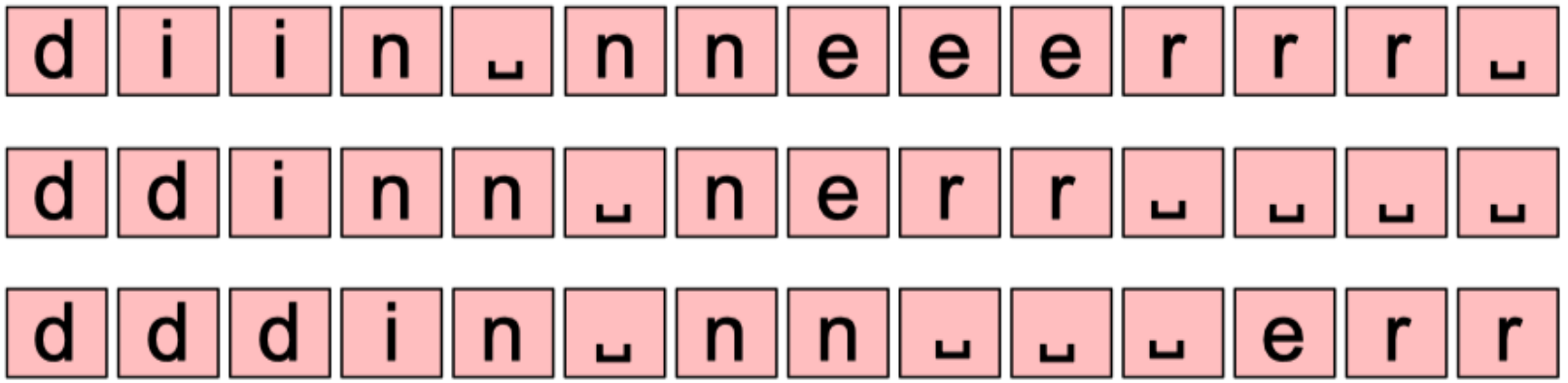
- Two properties in speech,
  - We have a very long acoustic input sequence X mapping to a much shorter sequence of letters Y,
  - It's hard to know exactly which part of X maps to which part of Y
- An alternative to encoder-decoder: an algorithm and loss function called Connectionist Temporal Classification (CTC).
- The intuition:
  - Output a single character for every frame of the input, so that the output is the same length as the input, and
  - Apply a collapsing function that combines sequences of identical letters, resulting in a shorter sequence

Naïve Alignment of Input with Letters

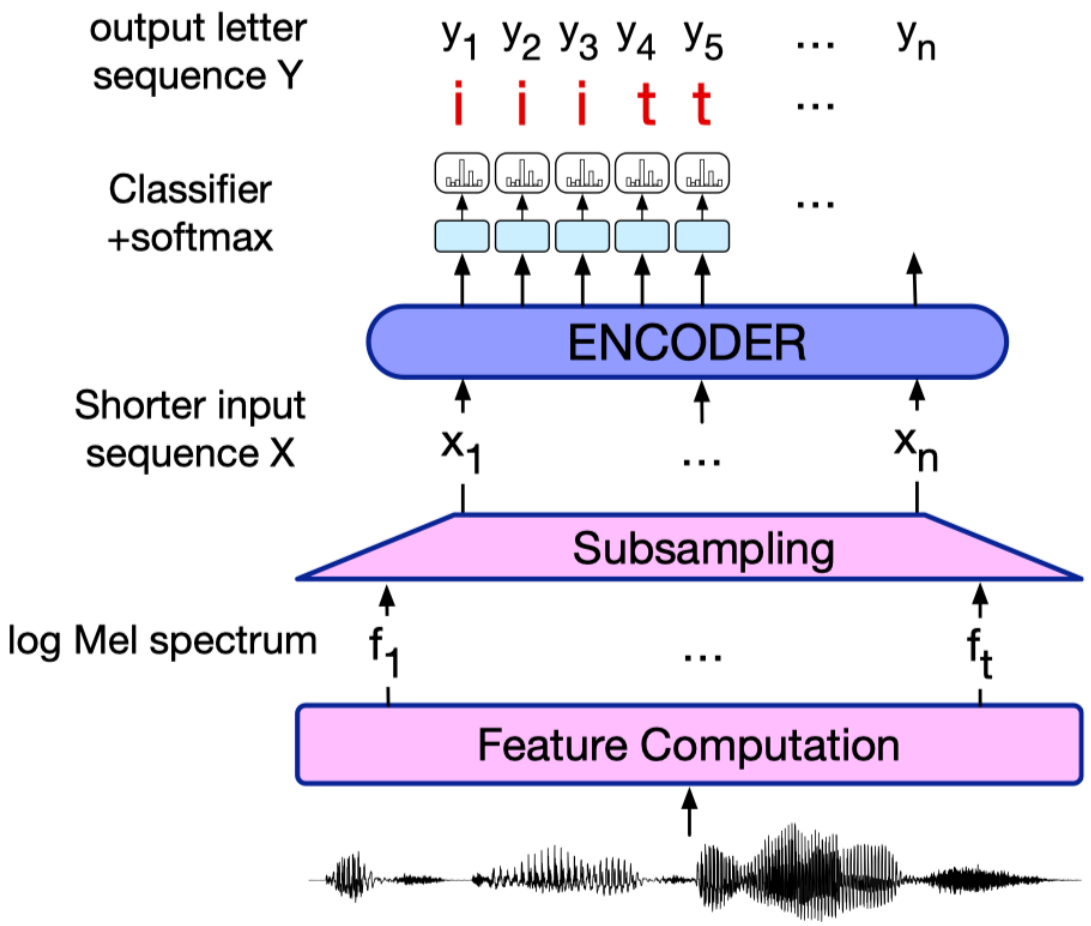


CTC Collapsing Function




d i i n ␣ n n e e e r r r ␣
d d i n n ␣ n e r r ␣ ␣ ␣ ␣
d d d i n ␣ n n ␣ ␣ ␣ e r r

CTC Inference

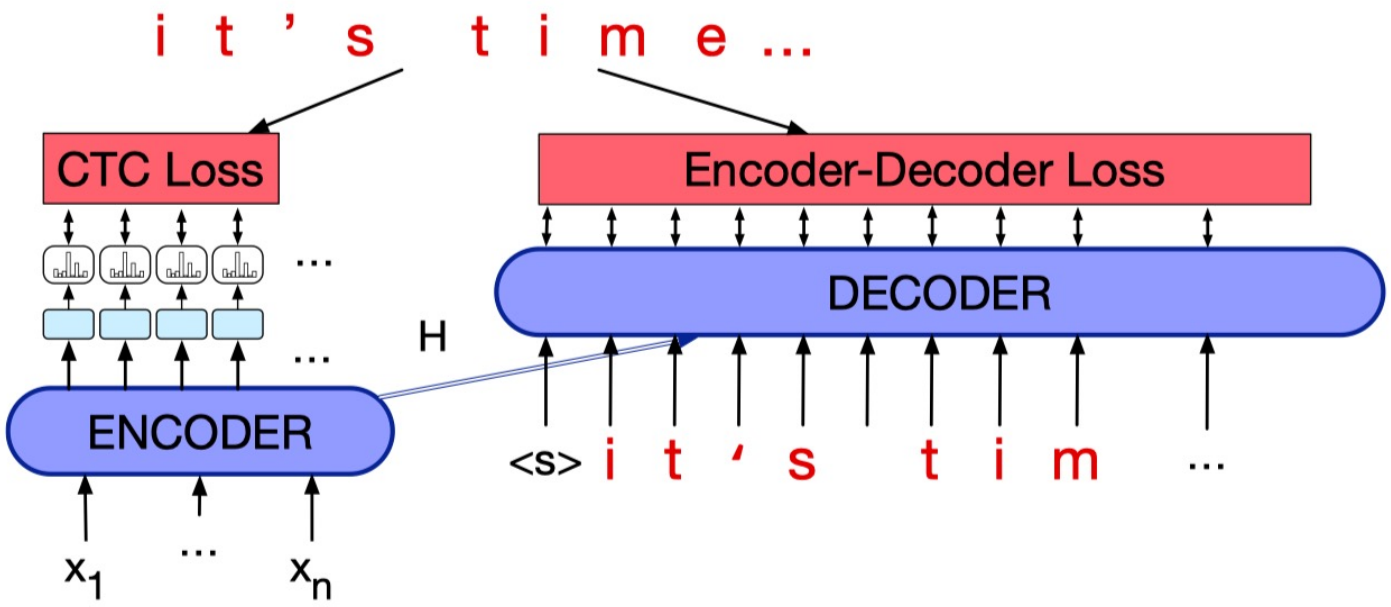


CTC Training

- Negative log-likelihood loss with a special CTC loss function.
- The loss for an entire dataset D is the sum of the negative log-likelihoods of the correct output Y for each input X.

- Naively summing over all possible alignments is not feasible.
  - Can be done using dynamic programming to merge alignments.

Combining CTC and Encoder-Decoder



ASR Evaluation

- Word Error Rate
  - How much the word string returned by the recognizer (the hypothesized word string) differs from a reference transcription.
- Minimum edit distance in words between the hypothesized and correct strings (word substitutions, insertions, and deletions).

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

Calculating Word Error Rate

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		

- Word Error Rate = 100 * (6 + 3 + 1) / 13 = 76.9%

Statistical Significance for ASR

- We need to know whether a particular improvement in word error rate is significant or not.
- Determining whether two WERs are different can be done with Matched-Pair Sentence Segment Word Error (MAPSSWE) test.
- Earlier works used McNemar's test for significance

Speech Synthesis (or) Text-to-Speech

The Task of Text-to-Speech

- Mapping from strings of letters to waveforms.
- TTS systems are generally based on the encoder-decoder architecture, either using LSTMs or Transformers.
- The default condition for ASR systems is to be speaker-independent.
  - trained on large corpora with thousands of hours of speech from many speakers
- It's less crucial to use multiple voices in TSS
  - The basic TTS systems are speaker-dependent.
  - Trained on much less data, but all from one speaker.

TTS

- Spectrogram Prediction
  - maps from strings of letters to mel spectrographs: sequences of mel spectral values over time.
- Vocoder
  - maps from mel spectrograms to waveforms

Preprocessing: Text Normalization

- Handle non-standard words: numbers, monetary amounts, dates, and other concepts that are verbalized differently than they are spelled
- The number 1750 can be spoken in at least four different ways, depending on the context:
  - 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”)

Non-Standard Words

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

- Normalization can be done by rule or by an encoder-decoder model.
- Rule-based normalization is done in two stages: tokenization and verbalization.
  - In the tokenization stage, we hand-write rules to detect non-standard words.
  - Regular expressions, like the following for detecting years:
  - /(1[89][0-9][0-9]) | (20[0-9][0-9]/
  - Verbalization uses more complex rule-system.
    - classifies and parses each input into a normal form
    - then produces text using a verbalization grammar

Normalization

- Encoder-decoder models have been shown to work better than rules for such tasks
- But they require expert-labeled training sets in which non-standard words have been replaced with the appropriate verbalization
  - such training sets for some languages are available

Mapping in Encoder-Decoder

They live at 224 Mission St.

They live at two twenty four Mission Street

45 minutes

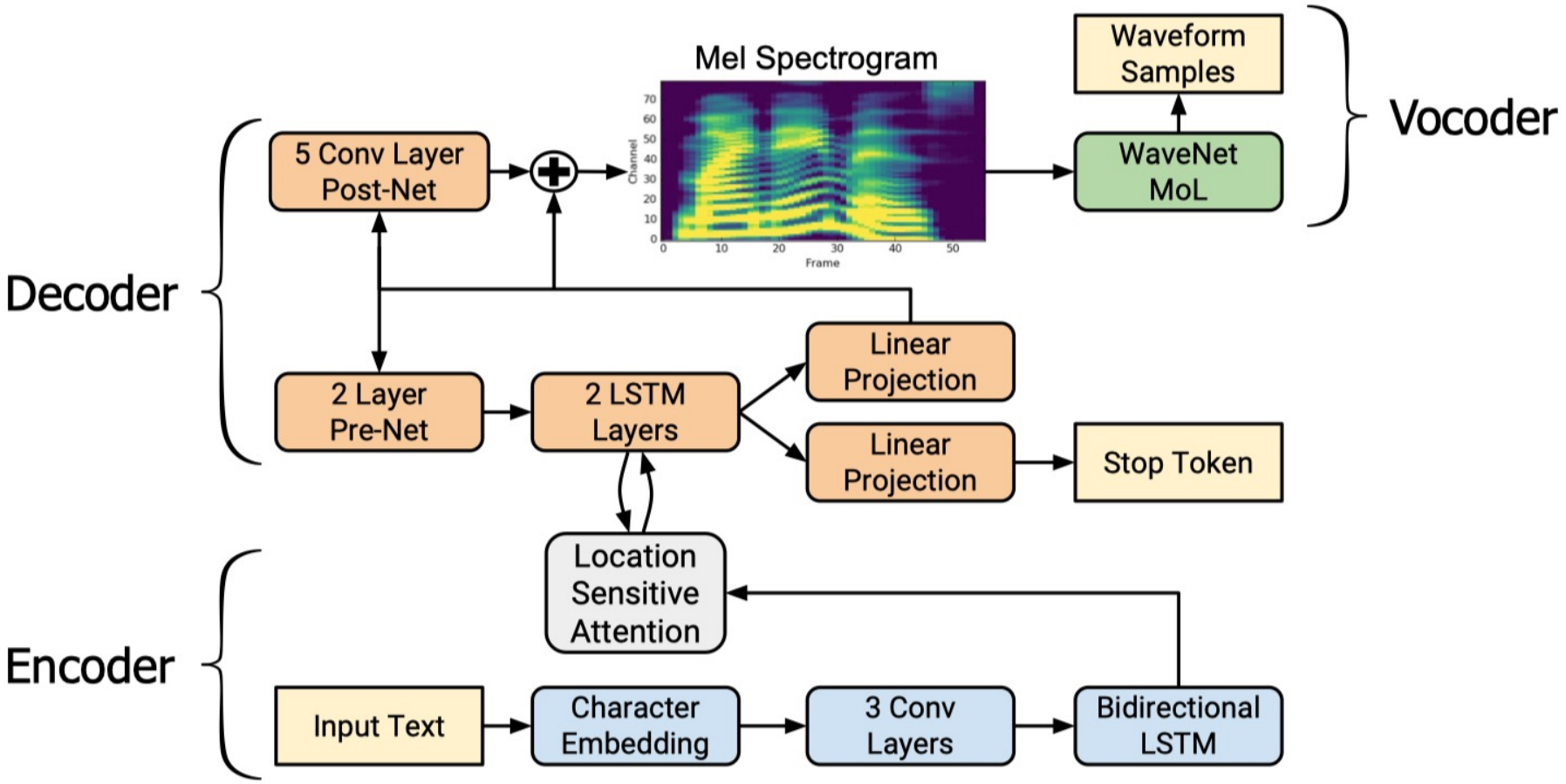
Forty five millimeters

TTS: Spectrogram Prediction

- The encoder-decoder with attention can be used for the first component of TTS.
- It maps from graphemes mel spectrograms, followed by a vocoder that maps to wavefiles.
- Encoder takes a sequence of letters and produce a hidden representation representing the letter sequence.
  - used by the attention mechanism in the decoder.
- In the decoder, the predicted mel spectrum from the prior time slot is passed through a small pre-net as a bottleneck.

Tacotron2 Architecture

Tacotron2 Architecture

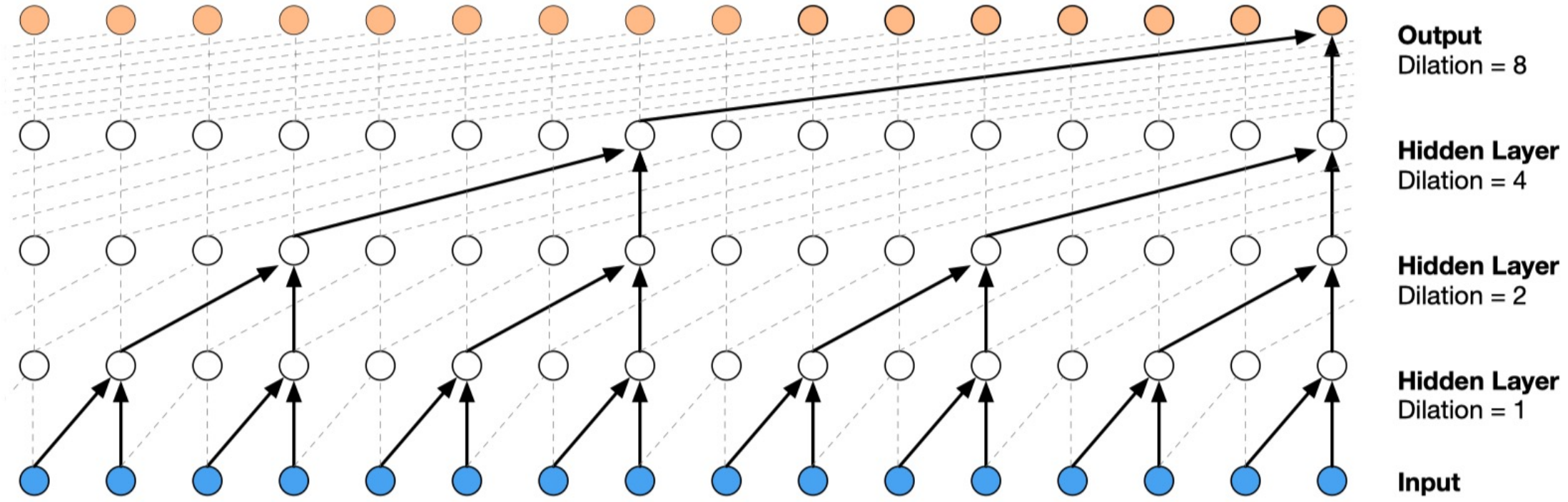


TTS: Vocoding

- Tacotron2's vocoder adapts from WaveNet's vocoder.
- Vocoding inverts a log mel spectrum representations back into a time-domain waveform representation.
- WaveNet is an autoregressive network.
  - It takes spectrograms as input and produces audio output represented as sequences of 8-bit mu-law.
  - The probability of a waveform $Y = y_1, y_2, \ldots, y_t$ given an intermediate input mel spectrogram h is computed.

$$p(Y) = \prod_{t=1}^{t} P(y_t | y_1, \ldots, y_{t-1}, h_1, \ldots, h_t)$$

Dilated Convolution



Training

- The spectrogram prediction encoder-decoder and the WaveNet vocoder are trained separately.
- After the spectrogram predictor is trained, the spectrogram is run in teacher-forcing mode.
- This sequence of ground truth-aligned features and gold audio output is then used to train the vocoder.

TTS Evaluation

- Speech synthesis systems are evaluated by human listeners.
- Mean Opinion Score (MOS)
- AB Tests