Fundamentals of Speech Recognition

Xiong XIAO

Principal Applied Scientist

Microsoft

March 28, 2023

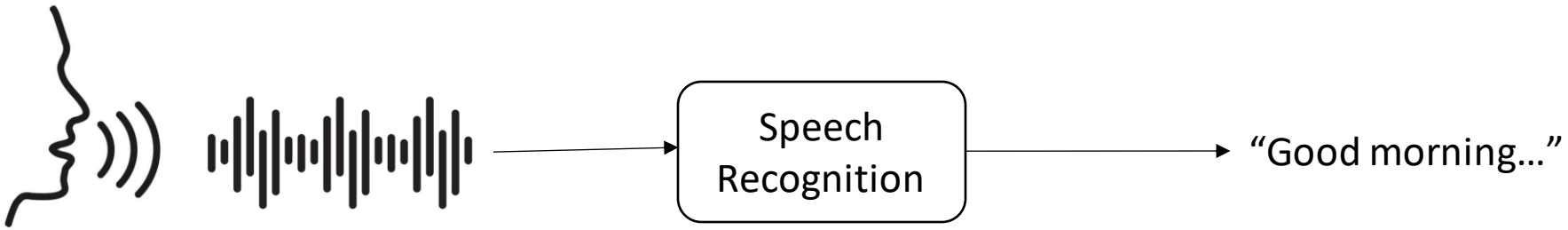
Outline

- Introduction and applications of ASR
- A naïve way of speech recognition
- Statistical modeling
- Deep learning
- Future of ASR
- Q&A

Introduction to ASR

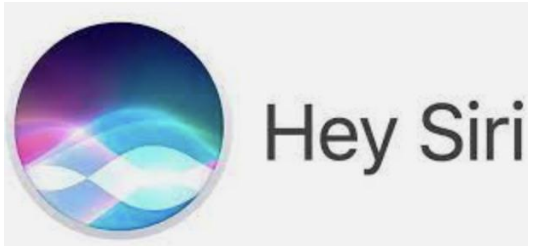
Speech Recognition Task

- Convert human speech waveform to human text.
- Also called automatic speech recognition (ASR) or speech-to-text (STT).
- ASR allows human to talk to machine in the most natural way.

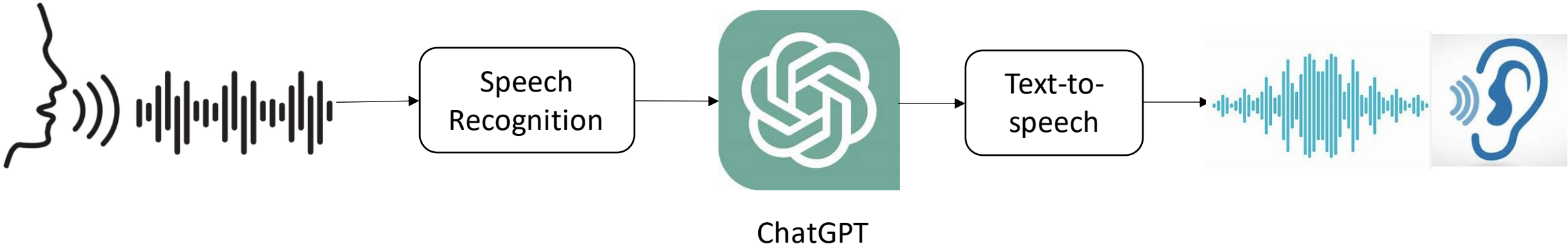


Applications of ASR

- Dictation
- Language learning
- Smart speakers (Alexa, Siri)
- Accessibility for hearing impaired
- Voice command
- Automatic captioning
- Audio indexing
- Machine translation
- Meeting understanding and summarization
- Call center analysis
- TV remote
- …



Enable ChatGPT with voice input/output

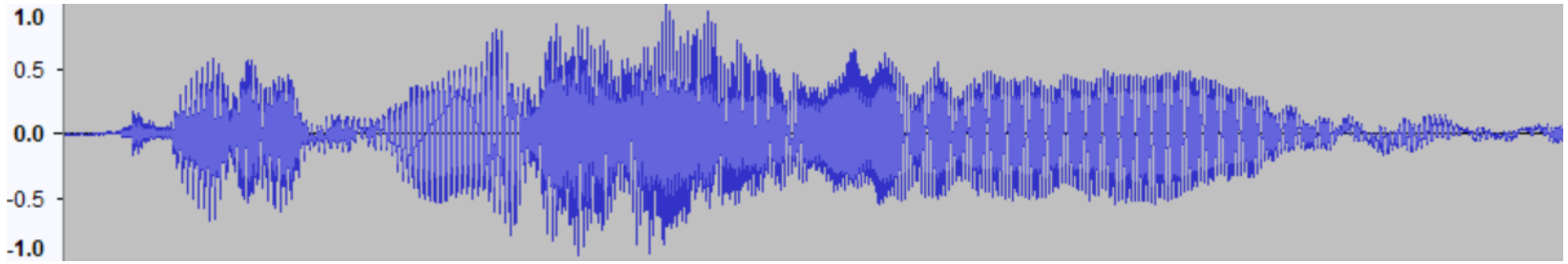


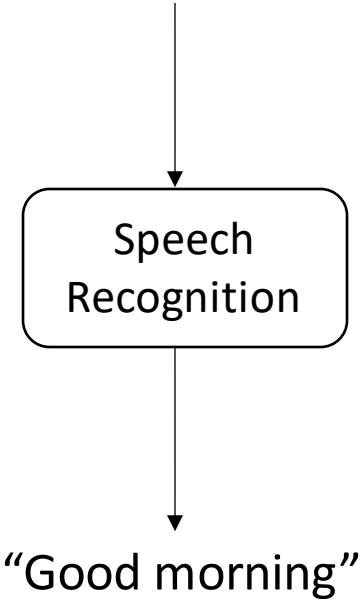
A naïve view of speech recognition

- Let’s first use a naïve and intuitive view to understand how to perform speech recognition.

Speech waveforms

Waveform:





ASR is a sequential pattern recognition task.

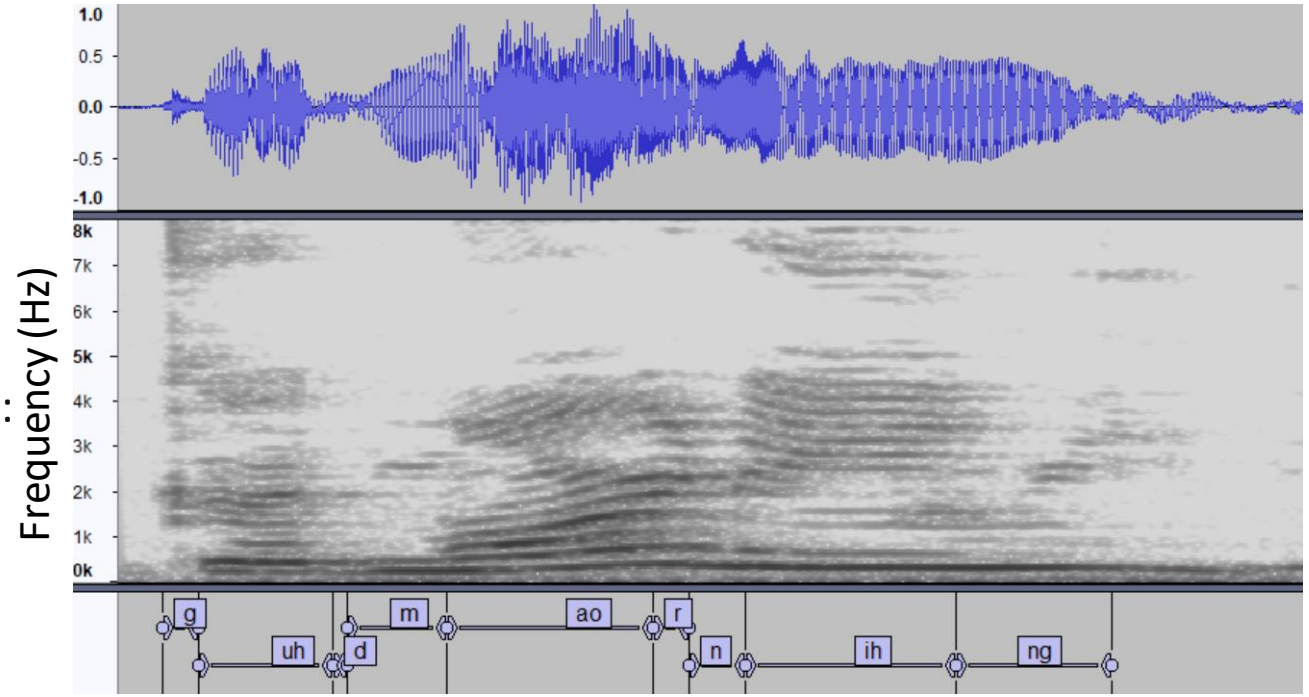
Subtasks:

- Where speech is? (VAD)
- Where is word beginning? (Segmentation)
- What are each word? (Classification)

Spectrogram and phonemes

Waveform:

Spectrogram:



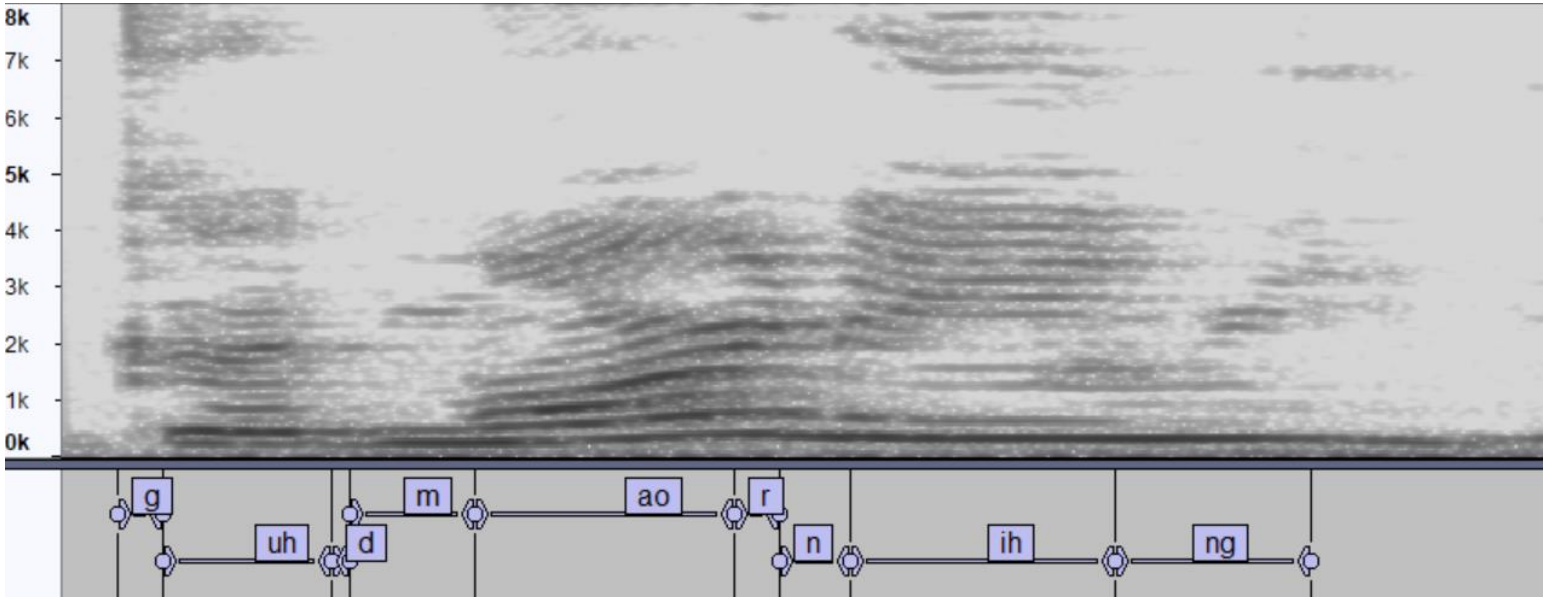
- Words can be divided into phonemes, only about 40 phonemes in English.
- Different phonemes have different “patterns”
  - Duration
  - Spectrum

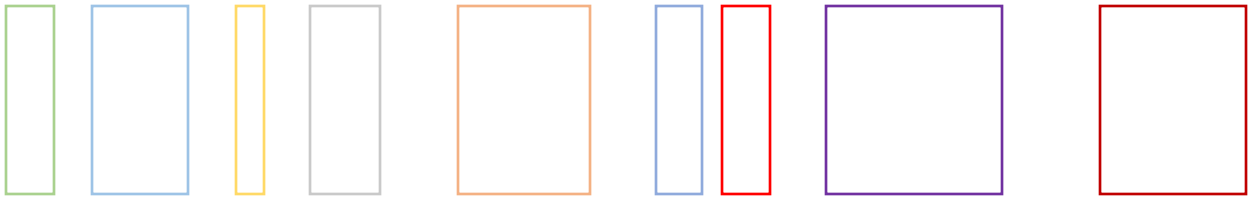
phonemes: /g/ /uh/ /d/ /m/ /ao/ /r/ /n/ /ih/ /ng/

IPA: /gʊd ˈmɔːnɪŋ/

Sentence: Good morning

Building templates for phonemes

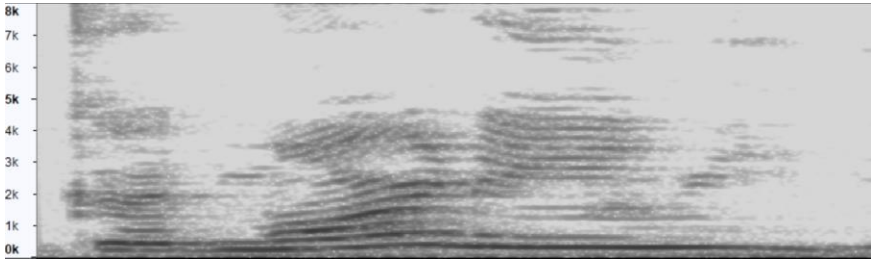


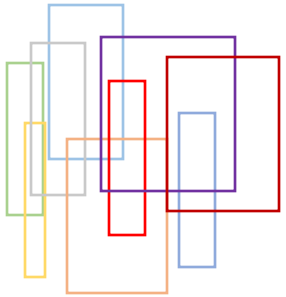


templates

- Each phoneme has some templates.
- A templates should capture the characteristics of a phoneme, including its duration and spectral distributions.
- Templates should be elastic to handle variations of phonemes.
- We will cover the actual techniques later.

Decoding





templates

Decoder

/g/ /uh/ /d/ /m/ /ao/ /r/ /n/ /ih/ /ng/

Find the most possible words

Good morning

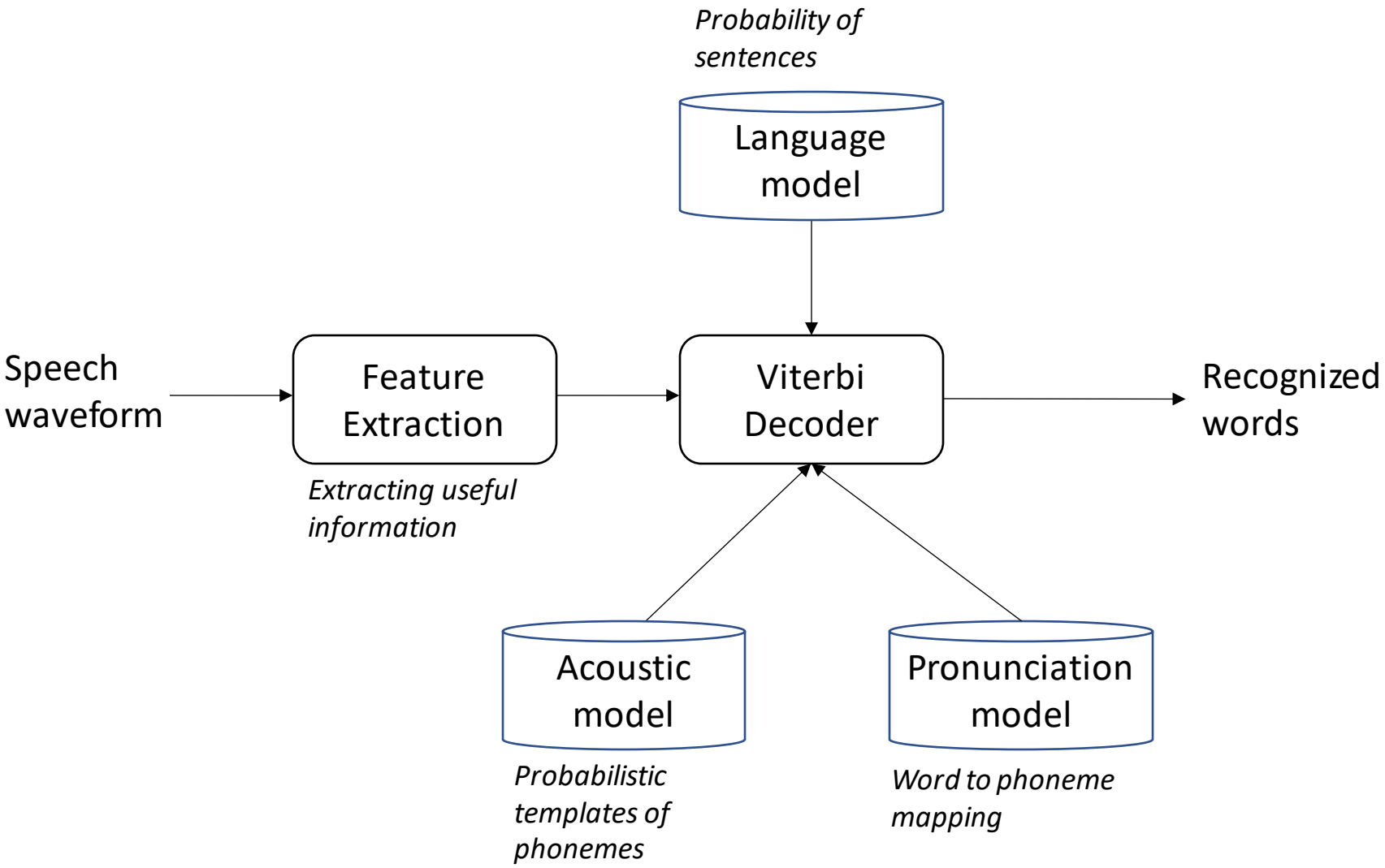
- Decoding is a search problem: find the most likely word sequence given the audio signal.
- Decoder needs to
  - detect start/end times of potential monophones
  - match them with the templates to find out the best template for each region

- Next, we are going to study a classic speech recognition framework, called HMM/GMM + n-gram ASR system. It is the mainstream ASR system from 1980s to 2000s.

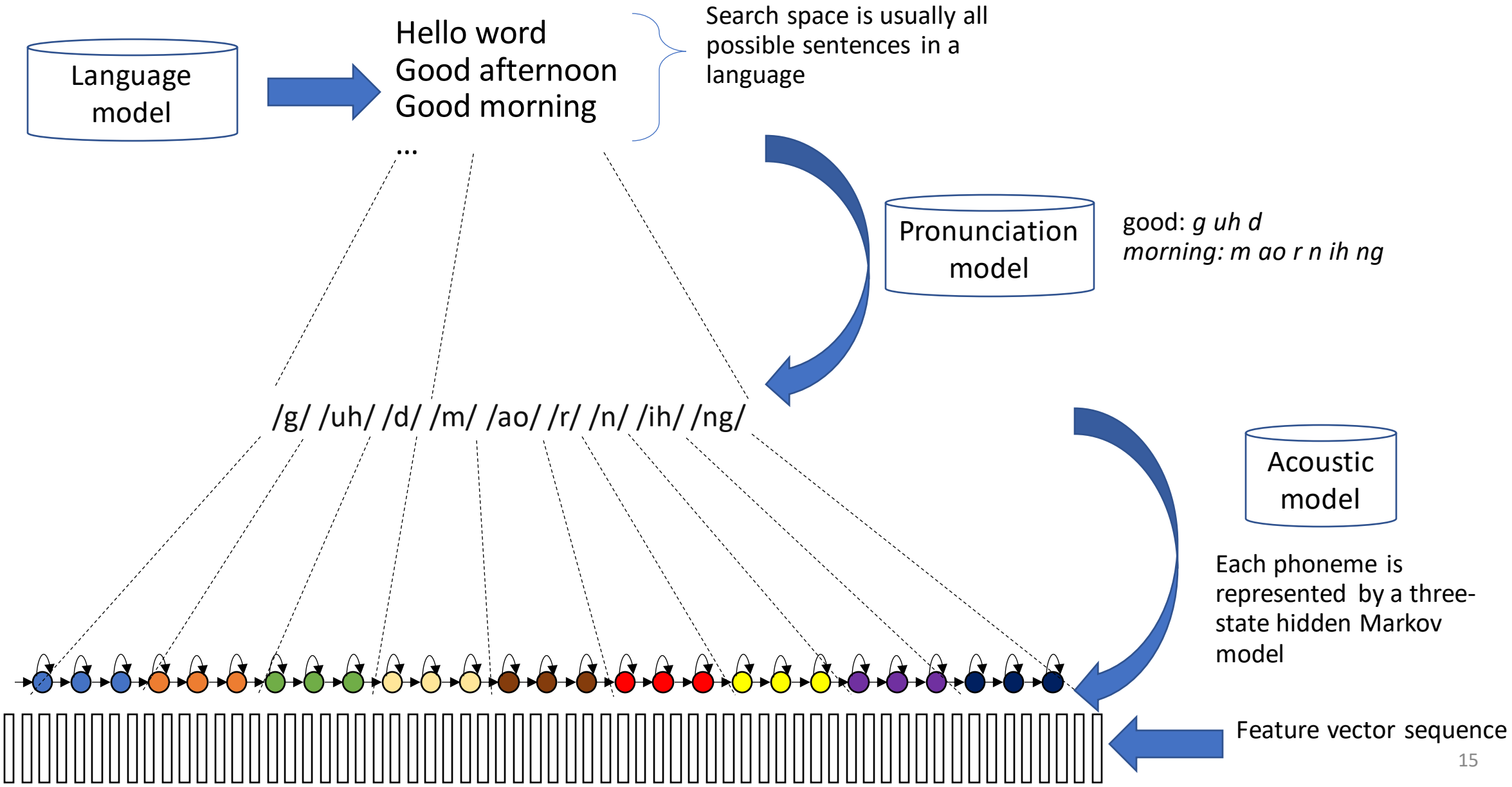
Statistical modeling for ASR

- Templates: acoustic model
- How to form words from phonemes: pronunciation model
- How to model word relationships: language model
- Decoding: Viterbi decoding and beam search

A more formal view of ASR



Layers of models



Language model

Captures the probability of a candidate sentence in the search space.

N-gram language models for a word sequence $W_1, W_2, \dots, W_{t-2}, W_{t-1}, W_t, \dots$

- Unigram LM uses unconditional word probabilities: $P(W_t)$
- Bigram LM uses: $P(W_t|W_{t-1})$
- Trigram LM uses: $P(W_t|W_{t-1}, W_{t-2})$
- …

Note that language model has a predefined vocabulary

The probability of the sentence "Good morning, how are you?" according to a bigram language model:

P(Good morning, how are you?) = P(Good | <sent-start>)
* P(morning| Good)
* P(how | morning)
* P(are | how)
* P(you | are)
* P(<sent-end> | you)

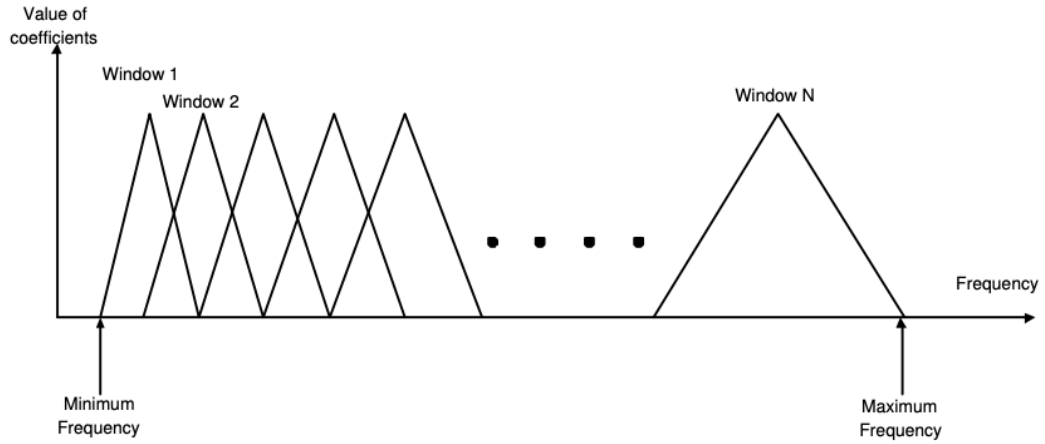
Challenges in language and pronunciation model

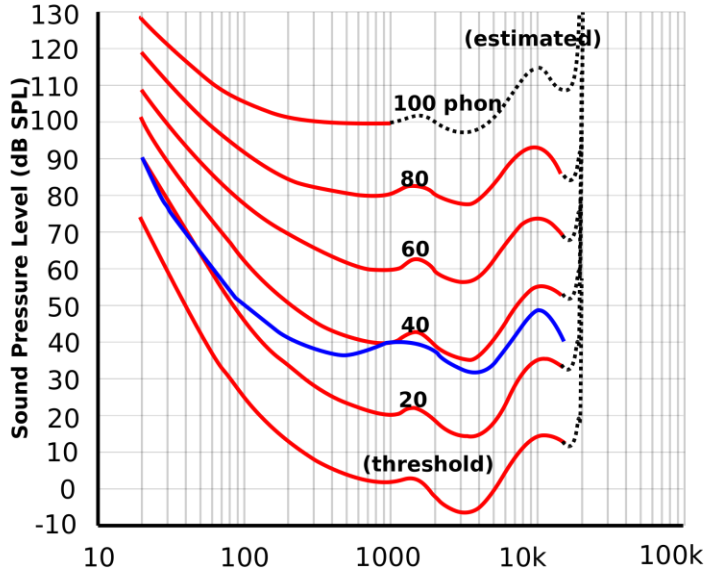
- Out-of-vocabulary (OOV) words, usually new name entities, long tail distribution
- Data sparsity: most high-order n-grams are unseen in training data
- N-gram model cannot capture long term dependency
- Variation of the pronunciation of words by different people and in different dialects

Acoustic features: Mel-frequency cepstral coefficients (MFCC)

- Features for acoustic modelling should
  - be compact for HMM-GMM modelling;
  - carry useful information to discriminate phoneme classes.
- MFCC is the most popular features for HMM-GMM acoustic model
- Motivated by human auditory system.
  - Mel-scale filterbanks: higher frequency resolution for low frequencies
  - Take logarithm of speech power: mimic equal loudness curve and convert sound intensity to perceived loudness

Illustration of Mel frequency filterbanks





Equal-loudness contours (red) (from ISO 226:2003 revisic
Original ISO standard shown (blue) for 40-phons

Davis, S. and Mermelstein, P. (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing, 28, 357-366.

MFCC extraction steps

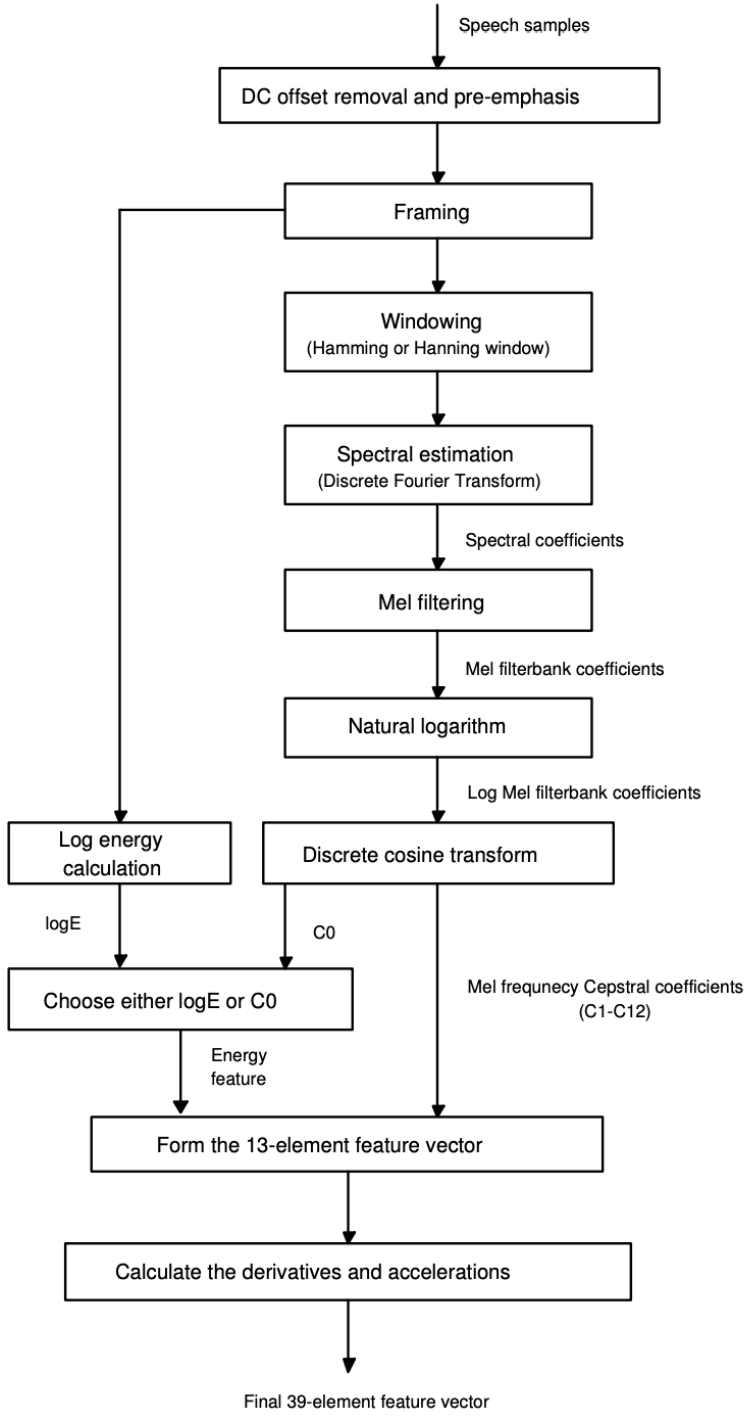


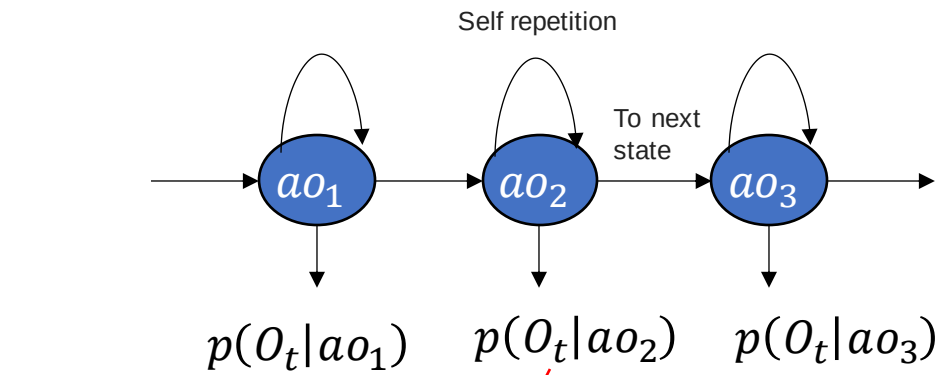
Figure adapted from Xiao X,

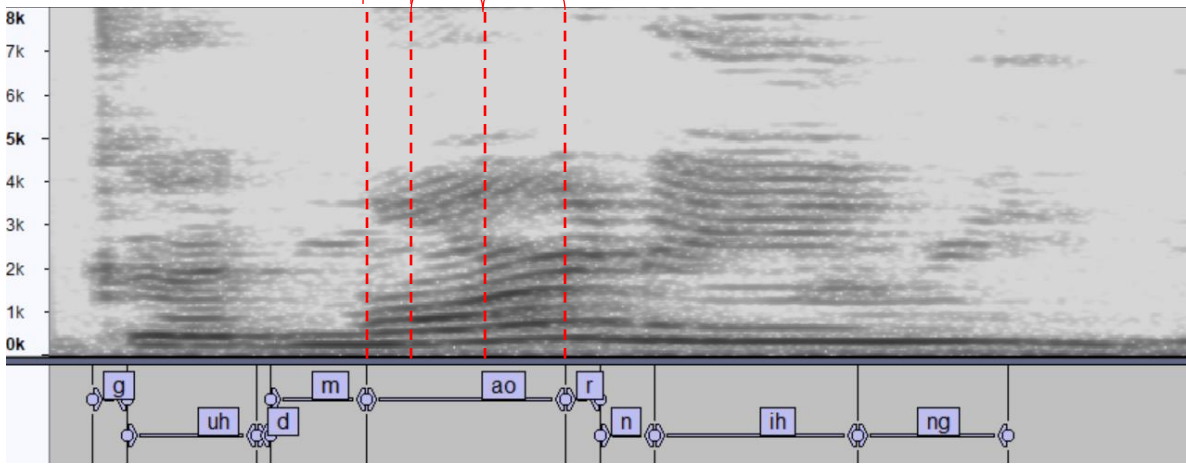
Acoustic model: Hidden Markov Model (HMM) for phonemes

- Each phoneme is typically represented by a 3-state HMM.
- Each state usually has a transition probability to the next state, e.g. from ao_1 to ao_2
- There is also a self-repetition probability: allow each state to "consume" variable number of frames → handling duration variations of phonemes.
- The distribution of the acoustic features at each state is assumed to be ***stationary***, usually represented by a Gaussian mixture model (GMM):

$$p(O_t|S_t = ao_1) = GMM(ao_1)$$

where O_t and S_t is the observed feature vector and hidden state at time index t, respectively.





Rabiner, L. and Juang, B., 1986. An introduction to hidden Markov models. *IEEE ASSP magazine*, *3*(1), pp.4-16.

Observation distribution for HMM states: Gaussian Mixture Model (GMM)

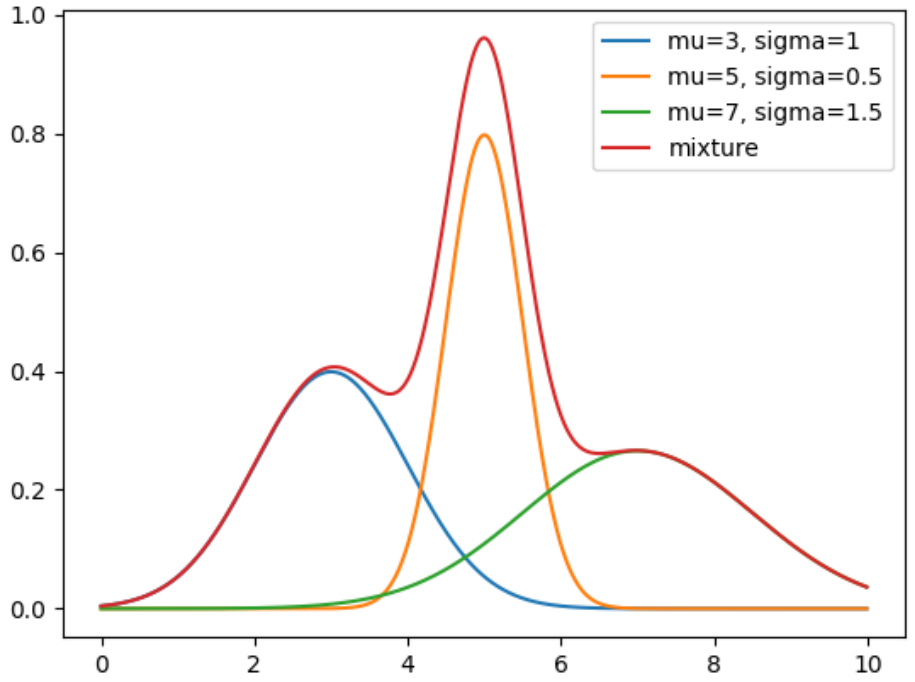
- GMM is a weighted sum of multiple Gaussian distributions.

$$p(O_t|\Lambda) = \sum_{m=0}^{M} w_m N(\mu_m, \Sigma_m)$$

where μ_m and Σ_m are mean and covariance of the mth Gaussian component.

- GMM can model any distribution given enough components.
- Number of components for each state varies depending on the amount of training data.

Example of 1-dimensional GMM with 3 components



Context dependent phonemes

- A phoneme may be pronounced differently in different contexts
  - “p” in “spin” and “pin” sounds differently.
  - “t” in “stop” and “top” sounds different.
- This is called co-articulation phenomenon.
- In large ASR systems, context dependent phonemes are used.
- Examples:
  - spin: /s/ /p/ /ih/ /n/ # context independent phoneme, also called monophone
  - spin: **/?-s+p/ /s-p+ih/ /p-ih+n/ /ih-n+?/** # context dependent triphones (left and right contexts, ? is determined by last/first of the preceding and following words)
- If there are 40 phonemes, there are in theory about 40*40*40= 64,000 triphones (although many of them does not exist in English)
- To reduce the number of classes, we cluster the triphone states to a predefined numbers, e.g. 10,000, for more efficient modeling.
- In the deep learning era, those clustered states are sometimes called senones.

Other choices for subword units

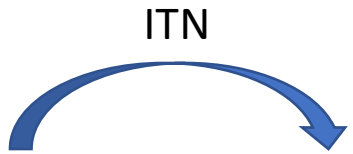
- Besides phonemes and context dependent phonemes, we can also use other subword units for acoustic modeling.
- Graphemes: letters or combination of letters.
- Morphemes: smallest meaning component of words
- Word pieces: variable-length units that are learned from the data, using techniques such as byte-pair encoding (BPE). More popular with End-to-end ASR. Handle OOV words better.

Training of the models

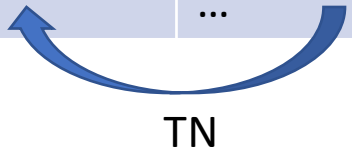
- HMM/GMM acoustic model: EM (expectation-maximization) algorithm. Forward/backward algorithm for efficiency
- Pronunciation model: usually defined by linguistics manually, expensive.
- Language model: mainly counting the ngrams and smoothing

Post processing: inverse text normalization

- When we train the language model, we usually apply text normalization (TN) to the training text to get clean word stream.
- In order to produce more readable recognized text, we need to apply inverse text normalization (ITN) to recognized words.



Categories	Text from ASR	After ITN
numbers	one hundred and twenty five	125
money	fifty dollars	$50
time	two thirty five	2:35
capitalization and punctuation	good morning	Good morning!
…	…	…



Evaluation Metrics

- Reference: The quick brown fox jumped over the lazy dog
- Hypothesis: The quick brown fox jumps over ---- lazy dog too

- Word error rate:
  - $WER = \frac{D+S+I}{N}$
  - D: number of deleted words
  - S: number of subsituted words
  - I: number of inserted words
  - N: number of reference words

- Readability: whether the recognized text is easy to read by human.

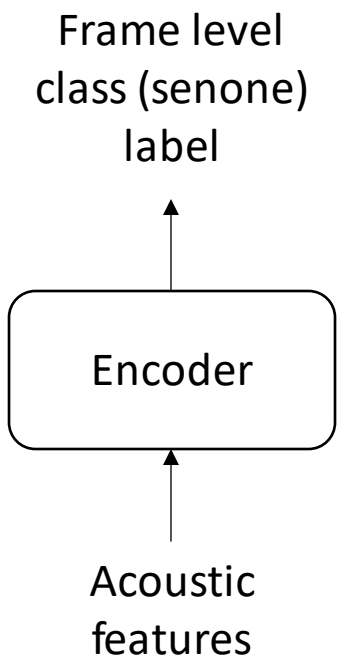
Deep learning

Deep learning for ASR

- Hybrid system: only replace HMM/GMM acoustic model with neural networks
- End-to-end ASR: replace the whole ASR system with neural works

Hybrid acoustic model

- Replace the generative HMM/GMM with a discriminative neural networks
- HMM/GMM models $p(o_t|s_t)$
- Hybrid models $p(s_t|o_t)$
- Common practices
  - Train an HMM/GMM first
  - Use it to align the label (senone sequences) to the feature sequence.
  - Train neural networks to predict frame level senone labels



Encoder Structures

- DNN
- CNN
- LSTM
- Transformer
- Or any combination of them

End-to-end ASR

- End-to-end ASR systems try to do ASR with a single model
- Three main approaches
  - Connectionist Temporal Classification
  - RNN Transducers
  - Sequence-to-Sequence

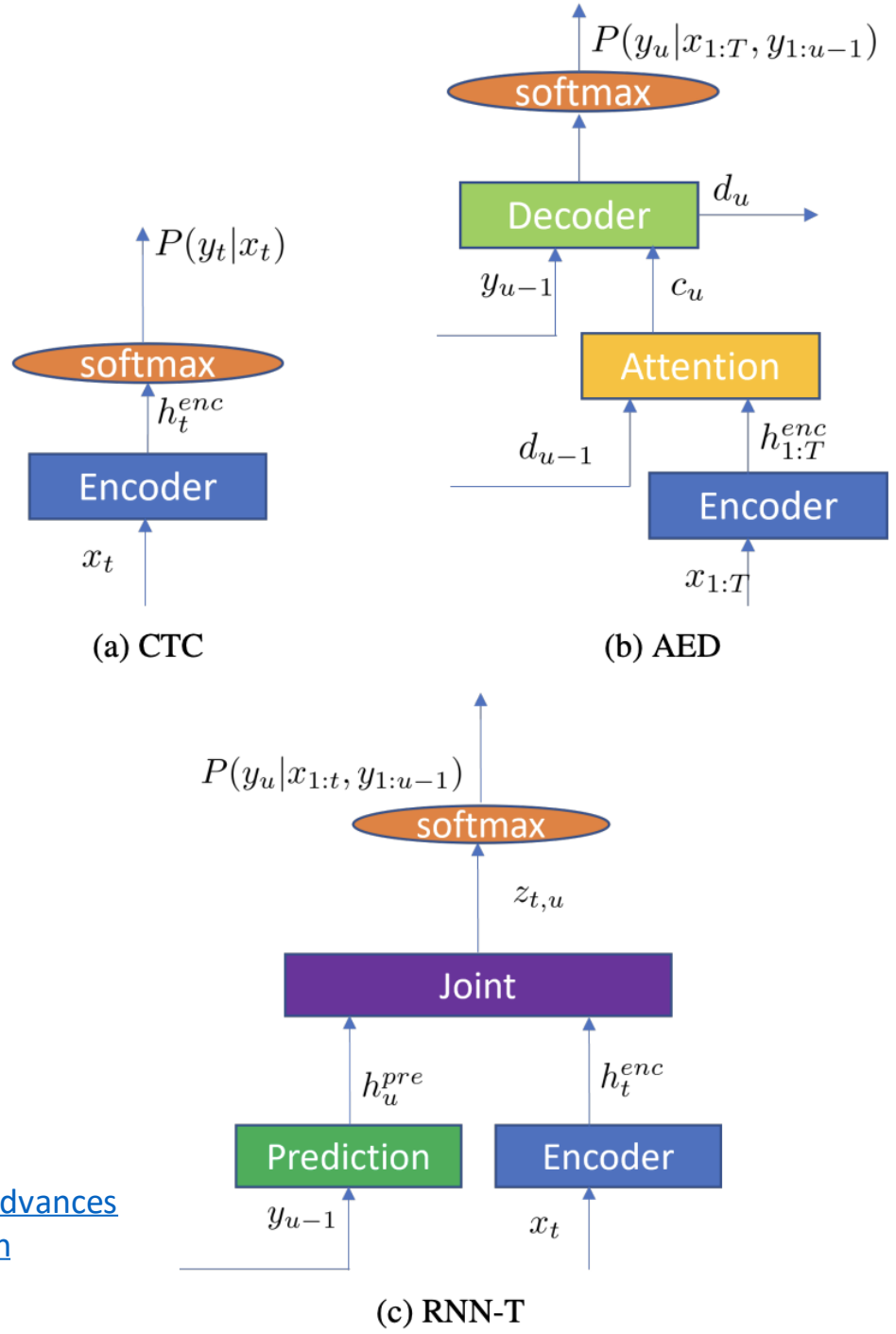


Figure adapted from [2111.01690] Recent Advances in End-to-End Automatic Speech Recognition (arxiv.org)

Connectionist Temporal Classification (CTC)

- As feature frames is usually much longer than label sequence, CTC introduces a blank label ”B” between two labels.
- Strength: No frame level alignment required, sum over all possible alignments
- Weakness: Independency assumption in output label. Still need language model to get good performance

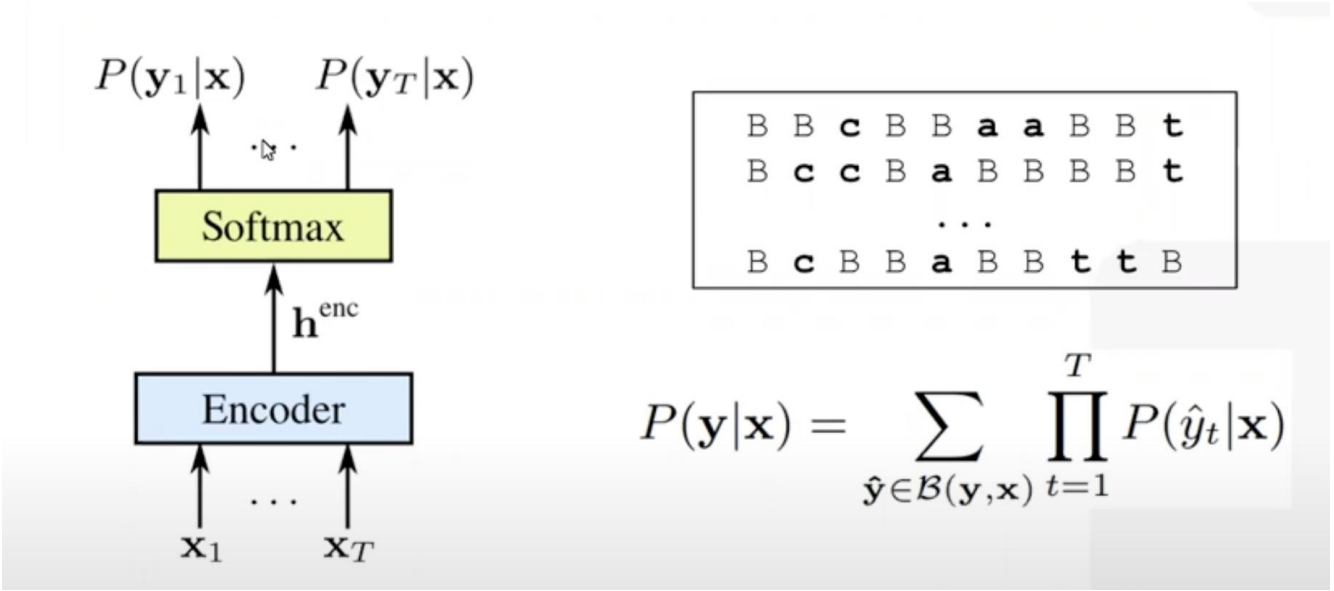


Figure adapted from “Tara Sainath - End-to-End Speech Recognition: The Journey from Research to Production – YouTube”

Graves, A. and Jaitly, N., 2014, June. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning* (pp. 1764-1772). PMLR.

Sequence-to-sequence (S2S)

- S2S is also called attention encoder decoder (AED)
- Encoder: similar to acoustic model
- Attention: alignment model
- Decoder: similar to pronunciation and language model
- Offline model

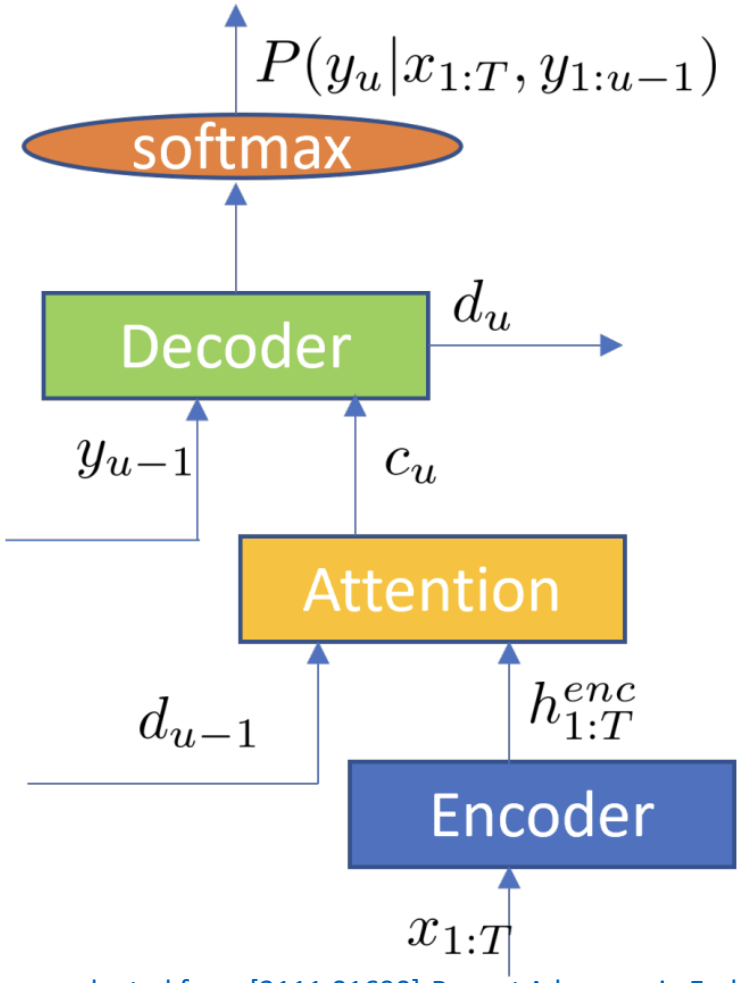


Figure adapted from [2111.01690] Recent Advances in End-to-End Automatic Speech Recognition (arxiv.org)

Chan, W., Jaitly, N., Le, Q.V. and Vinyals, O., 2015. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*.

Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K. and Bengio, Y., 2015. Attention-based models for speech recognition. *Advances in neural information processing systems*, *28*.

RNN Transducers (RNN-T)

- Called RNN-T because originally RNN is used as the encoder model structure.
- Newer models uses transformers or conformers as encoder
- A native streaming model

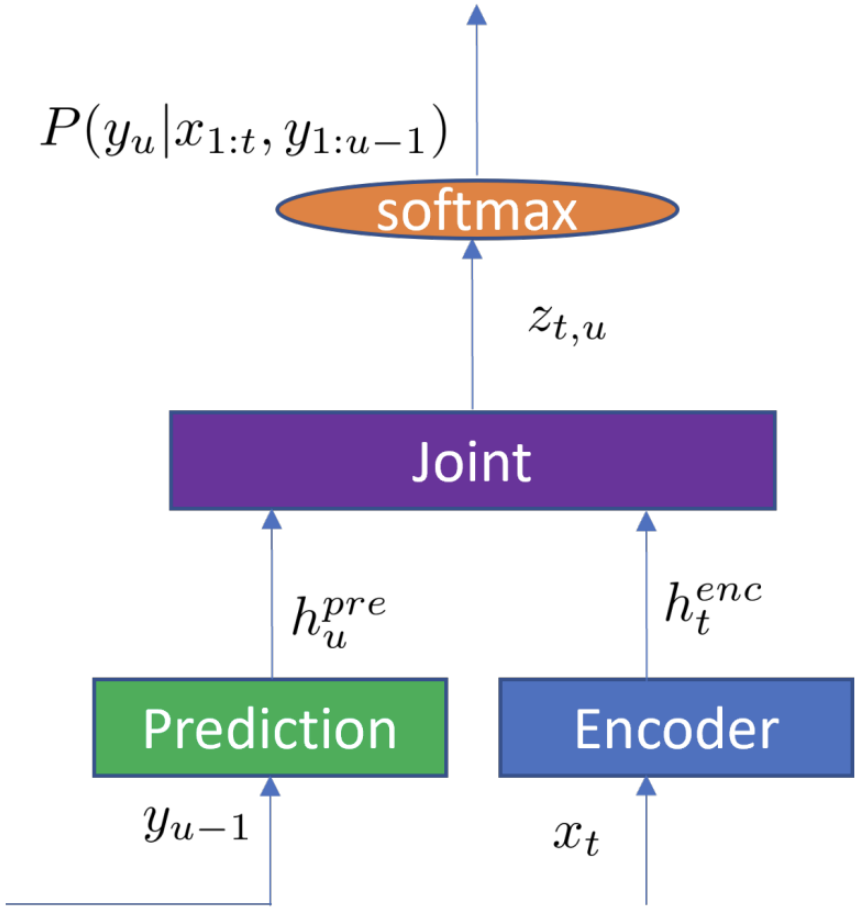


Figure adapted from [2111.01690] Recent Advances in End-to-End Automatic Speech Recognition (arxiv.org)

Graves, A., 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

More resources on E2E ASR

- Review papers
  - [2111.01690] Recent Advances in End-to-End Automatic Speech Recognition (arxiv.org)
- Videos
  - Dr. Jinyu Li, Microsoft, "Recent Advances in End-to-End Automatic Speech Recognition" - CSIP Seminar – YouTube
  - SANE2022 | Tara Sainath - End-to-End Speech Recognition: The Journey from Research to Production – YouTube

Challenges of ASR

- Robustness against various factors:
  - Acoustic environment: noise, reverb, transmission channel
  - Speaker variation
  - Accent
  - Domain
- Pronunciation variation, spontaneous speech, accent
- Overlapped speech
- Out of vocabulary (OOV) words, long tail effect
- Multilingual/Code-switching
- Low-resource languages
- Fast adaptation

Future of ASR

- Self-supervised training and the use of unlabeled data, wav2vec2, wavLM etc
- ASR for overlapped speech and speaker diarization
- Audio-visual ASR
- Directly output display format text: e.g. Whisper
- Multilingual modeling

Acknowledgement

- ChatGPT(GPT4) has been used to help determine the structure of this presentation, and also look up for the details of some concepts.

Q&A