Aalto University
School of Electrical
Engineering

ELEC-E5550 - Statistical Natural Language Processing

Automatic Speech Recognition

Tamás Grósz

Department of Information and Communications Engineering

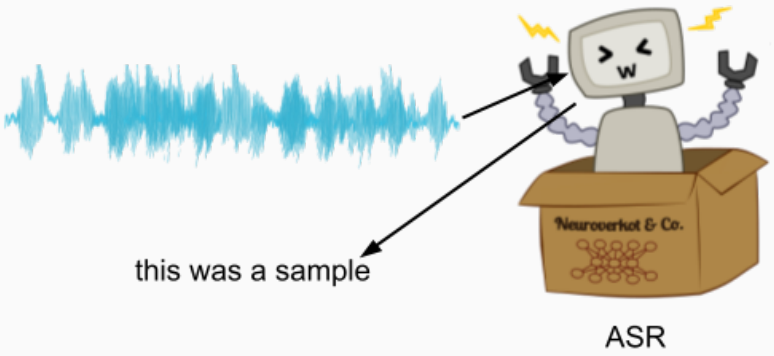
Outline

The goals of this lecture:

- to define what is automatic speech recognition
- to introduce commonly used techniques and models
- to give a general overview on traditional and end2end solutions
- to give a brief glance at the latest technologies

Introduction

Speech recognition



- Speech recognition also called Speech to Text
- 1952: Audrey developed at Bell Labs (single digit recognition)
- Nowadays many popular services (Alexa, Google Assistant, Siri, . . .)

Speech recognition challenges

Ice-breaker
Form groups of 3 and discuss about:

- have you used speech recognition services before?
- what are the challenges in speech recognition in your opinion?

Some examples from the lecturer:

- Noise
- Accent
- Speaking rate
- Disfluency

Speech recognition

We can categorize ASR systems in various ways.

- Isolated word recognition
- Voice command
- Keyword recognition
- Voice search
- Dictation systems
- Continuous/spontaneous speech recognition

From the data processing perspective:

- Offline (the system waits until the entire speech is recorded)
- Online (we see words appear while we speak)

Speech recognition

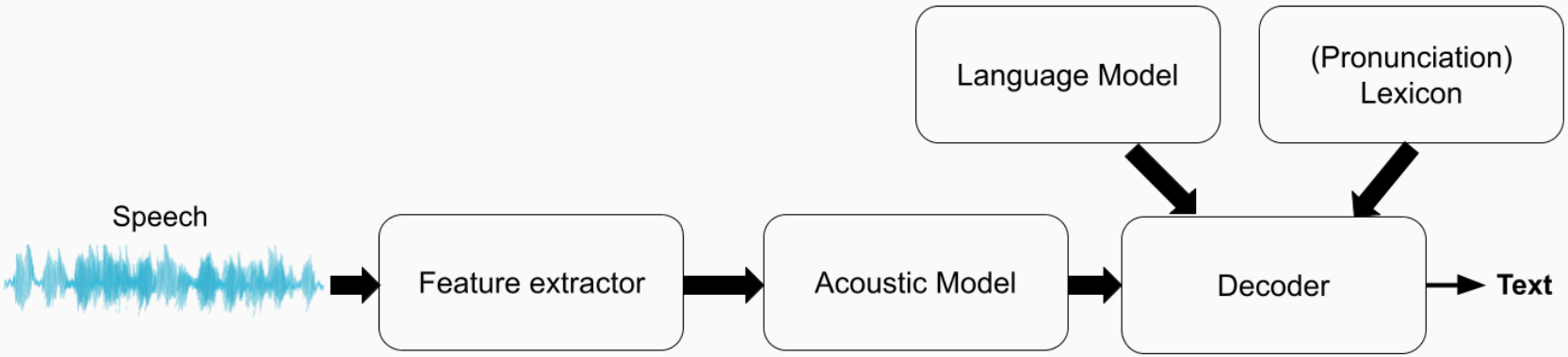
Some other categorizations:

- Speaker independent or dependent
- Read speech, planned speech, conversational speech
- Space and distance to the microphone: close-talk, near-field, far-field
- Single or multi microphone
- Noise robustness (what kind of environmental noises can it handle)
- Vocabulary: closed (only recognizes known words) or open

Traditional systems

Traditional ASR systems

Traditional ASR systems consists of several specialized modules.



Traditional ASR systems

Task: Find the most likely word sequence given the observations (speech) and the models for acoustics and language.

Formally:

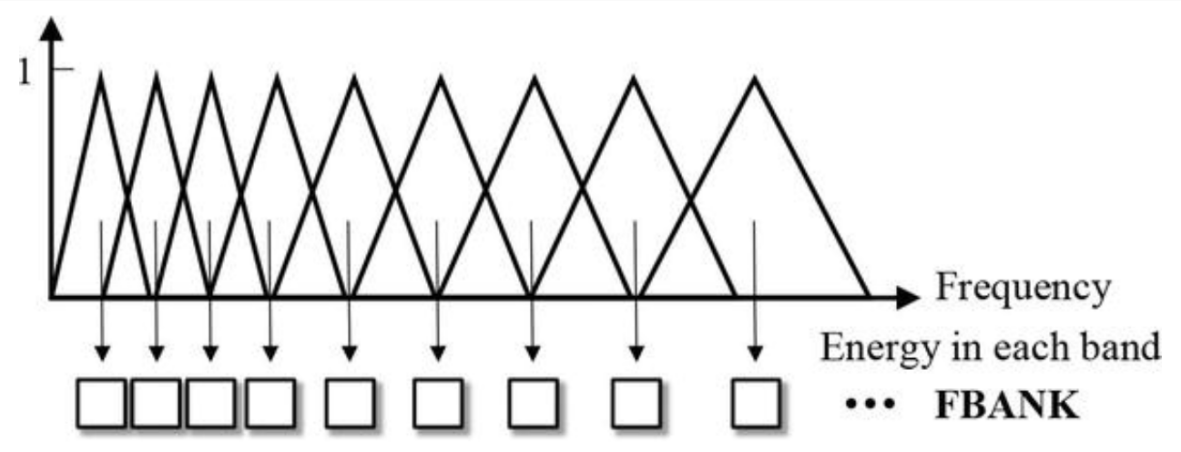
$$\hat{W} = \mathrm{argmax}_W p(\underbrace{W}_{\text{Words}} | \overbrace{O}^{\text{Speech/observations}})$$

$$= \mathrm{argmax}_W \underbrace{p(O|W)}_{\text{AcousticModel}} * \underbrace{p(W)}_{\text{LanguageModel}}$$

Features

- Raw speech is hard for ML models (noise, 16k or 44k samples/sec)
- Let's use some feature extraction!
- Feature design has been based on the knowledge of human hearing and psycho-acoustics. Typical features:
  - Mel-Frequency Cepstral Coefficients (MFCCs)
  - Perceptual Linear Prediction (PLP)
  - Logarithmic Mel-Filterbank Energies
- Common traits:

  non-linear frequency warping and energy compression



Acoustic Model

Task

An acoustic model is used in automatic speech recognition to represent the relationship between an audio signal and the **phonemes** or **other linguistic units** that make up speech.

- Different units can have different durations
- Solution: phoneme/unit classifier and a Hidden Markov Model (HMM) to deal with temporal information
- HMM states correspond to basic recognition units
- originally Gausian Mixture Models were used as phoneme classifiers, since 2013, DNNs replaced them

Acoustic Model (DNN)

Using HMM/DNN acoustic models is not trivial!

We need:

- a suitable loss function
  - Initially, standard Cross Entropy was used
  - which needed time-aligned labels mostly produced by an HMM/GMM
  - sequence discriminative losses (sMBR, MMI, MWER, CTC) offer an alternative but are difficult to implement and train.
- Large amount of data

Language Model

Task

A language model is a probability distribution over sequences of words. Given any sequence of words, a language model predicts the probability of the next word.

- N-gram model: built by counting how often word sequences occur in corpus text and then estimating the probabilities.
- Neural LM: trained to classify the next word, to many output options, tricks like Negative Sampling is used to alleviate it
- Large, pre-trained LMs: BERT, GPT,...

Negative sampling

1. Select a few non-target words (negative samples).
2. Pretend that the target word and the negative samples represent the entire vocabulary.
3. Update only these output units.

Aalto University
School of Electrical
Engineering

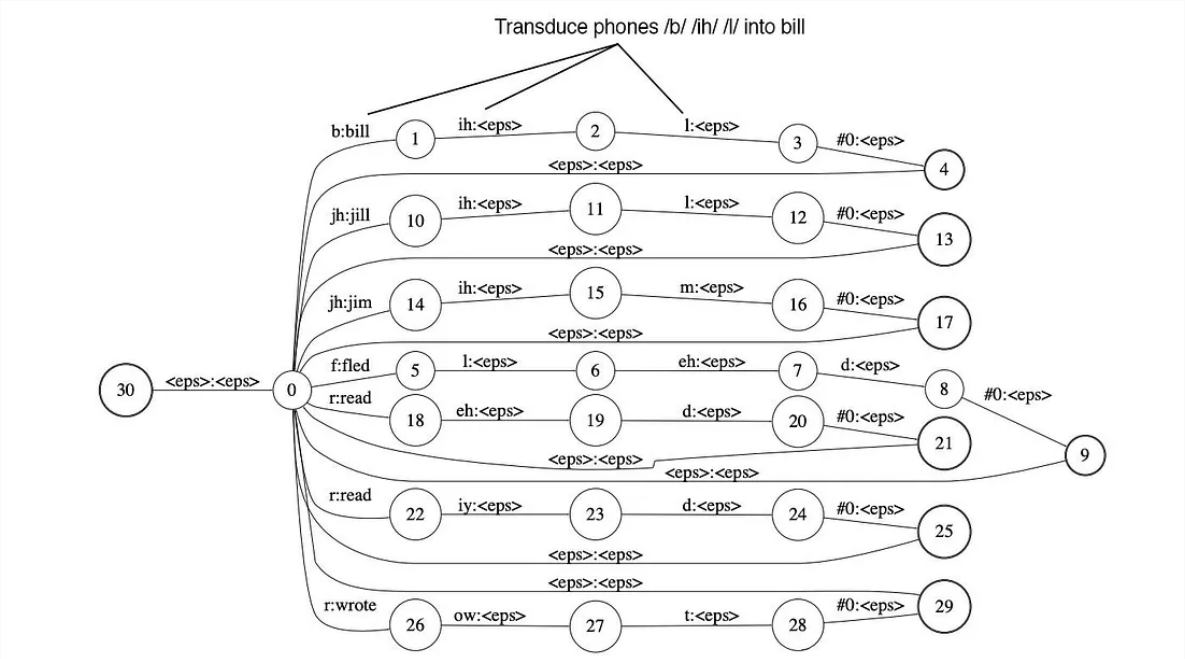
Decoder

The Pronunciation Lexicon contains mapping between words and acoustic units.

preferred $\longrightarrow$ *P R AX F ER DD* **or** *P R IX F ER DD*

The **Decoder** combines all components (AM, LM, lexicon) together to perform the speech to text task.

Decoders of traditional ASR systems are often Weighted Finite-State Transducers that can be used to search for the most probable sequences.



ASR performance

- How do we measure the performance of ASR?
- We can have a dedicated test set of recordings, and compare the human transcript to the ASR.
- How do we compare two texts? $\longrightarrow$ Levenshtein distance

Levenshtein distance

The Levenshtein distance between two sentences is the minimum number of word edits (substitution/deletion/insertion) required to change one word into the other. In ASR, we call it Word Error Rate (WER), or Character Error Rate (CER) if we calculate it with characters as base units.

$WER(ref, hyp) = \frac{S+D+I}{C+S+D}$, where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct words.

Aalto University
School of Electrical
Engineering

Exercise

Calculate the WER and CER metrics by comparing the ASR hyp to the human transcript!

ASR hyp:	he then appeared in the episode smackdown
Human transcript:	he then appeared on an episode of smackdown

Which metric measures the true accuracy better in your opinion and why?

Don't forget to submit you answer in MyCourse!

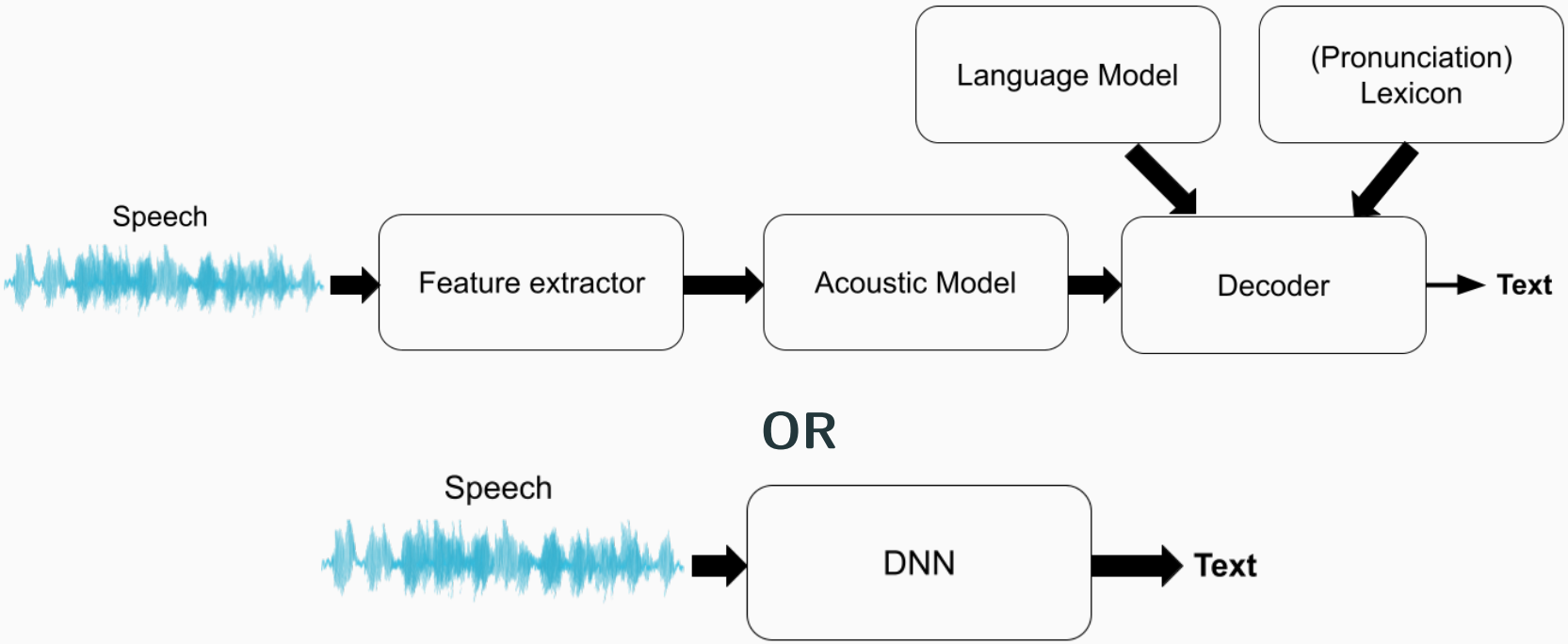
Break

15 min break

End-to-end ASR

New paradigm

Which type of system would you prefer?



Basic concept

- Instead of multiple, specialized subsystems we can use a single neural network.
- Easier to implement
- The decoding procedure gets much easier
- Harder to train!
- The model needs to learn all the tasks simultaneously
- We need specialized training methods
- Considerable amount of data is needed

Training methods

To train end-to-end models we need new methods.

CTC

Connectionist temporal classification is a very popular algorithm for training end-to-end systems.

- An optional *Blank* label is added to the output units, inserted between all units (character, sub-word or word)
- It computes the scores of all possible alignments of the GT text
- The goal is to maximize the summed score of the alignments
- The limitation of CTC loss is the input sequence must be longer than the output, and the longer the input sequence, the harder to train

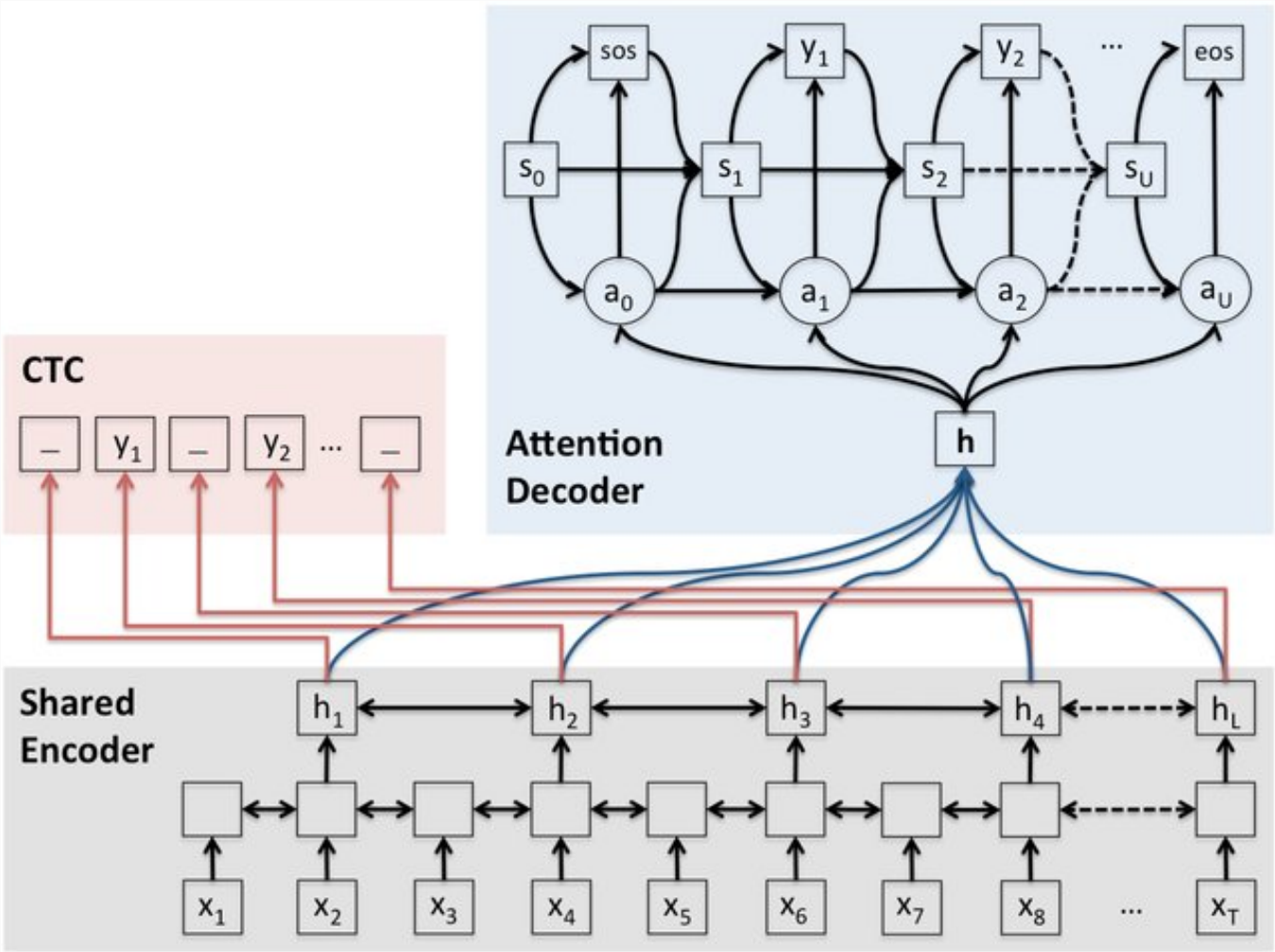
[0] A nice tutorial is available: https://voidful.medium.com/understanding-ctc-loss-for-speech-recognition-a16a3ef4da92

Model Types

Now the question is what kind of models are suitable for end-to-end ASR?

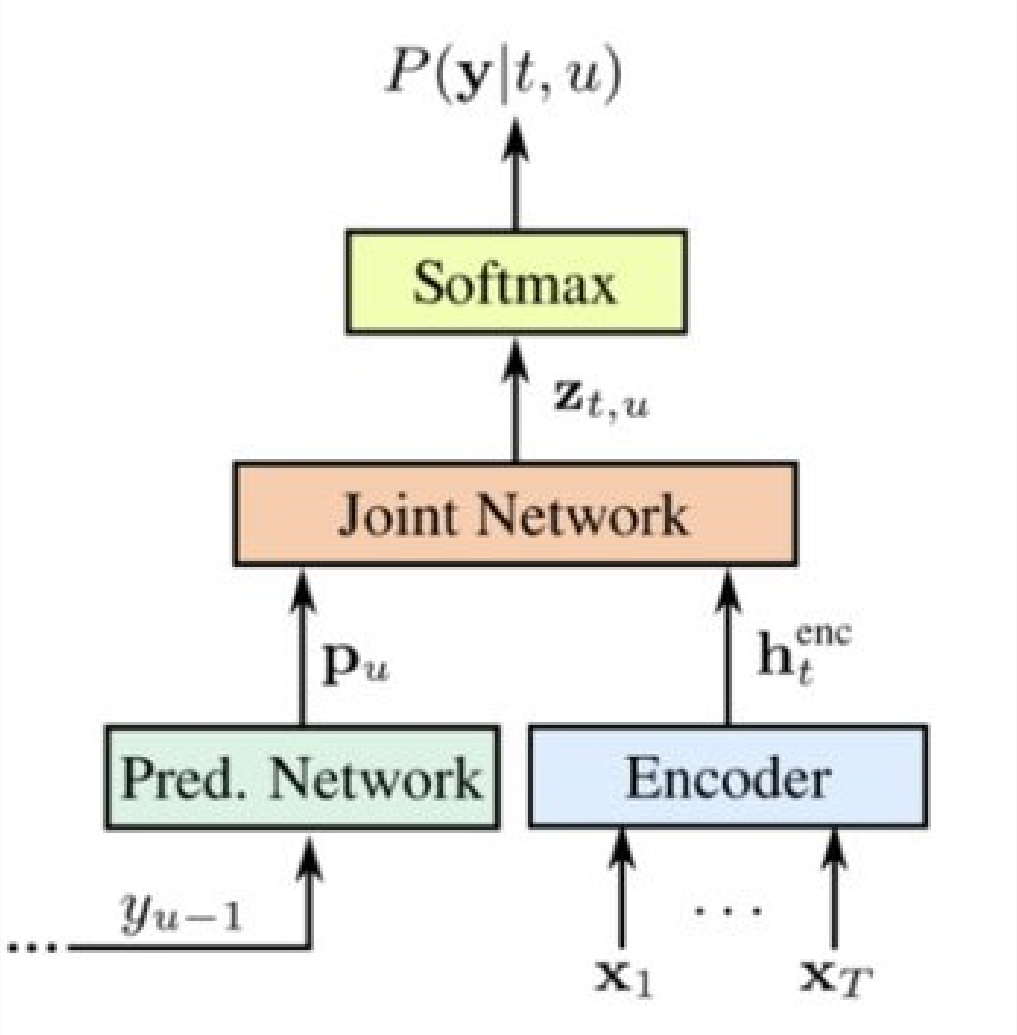
- Recurrent or attention components are necessary
- Convolution is often used to transform the input (raw audio or filterbank)
- Many models work best with short audios (reasons: CTC training and computation)

Model Types, Attention-based Encoder Decoder



Aalto University
School of Electrical
Engineering

Model Types, RNN transducer

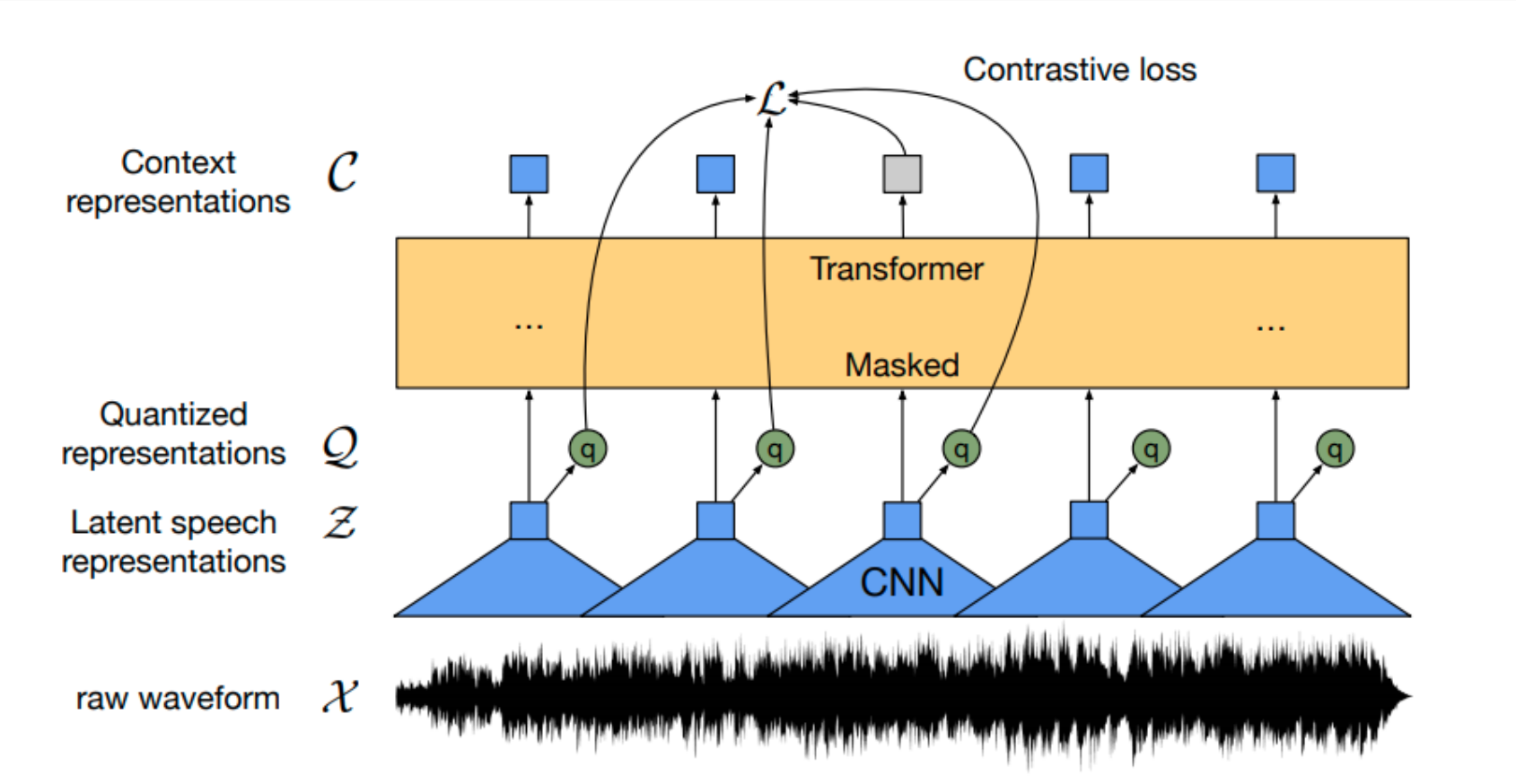


Pre-trained models

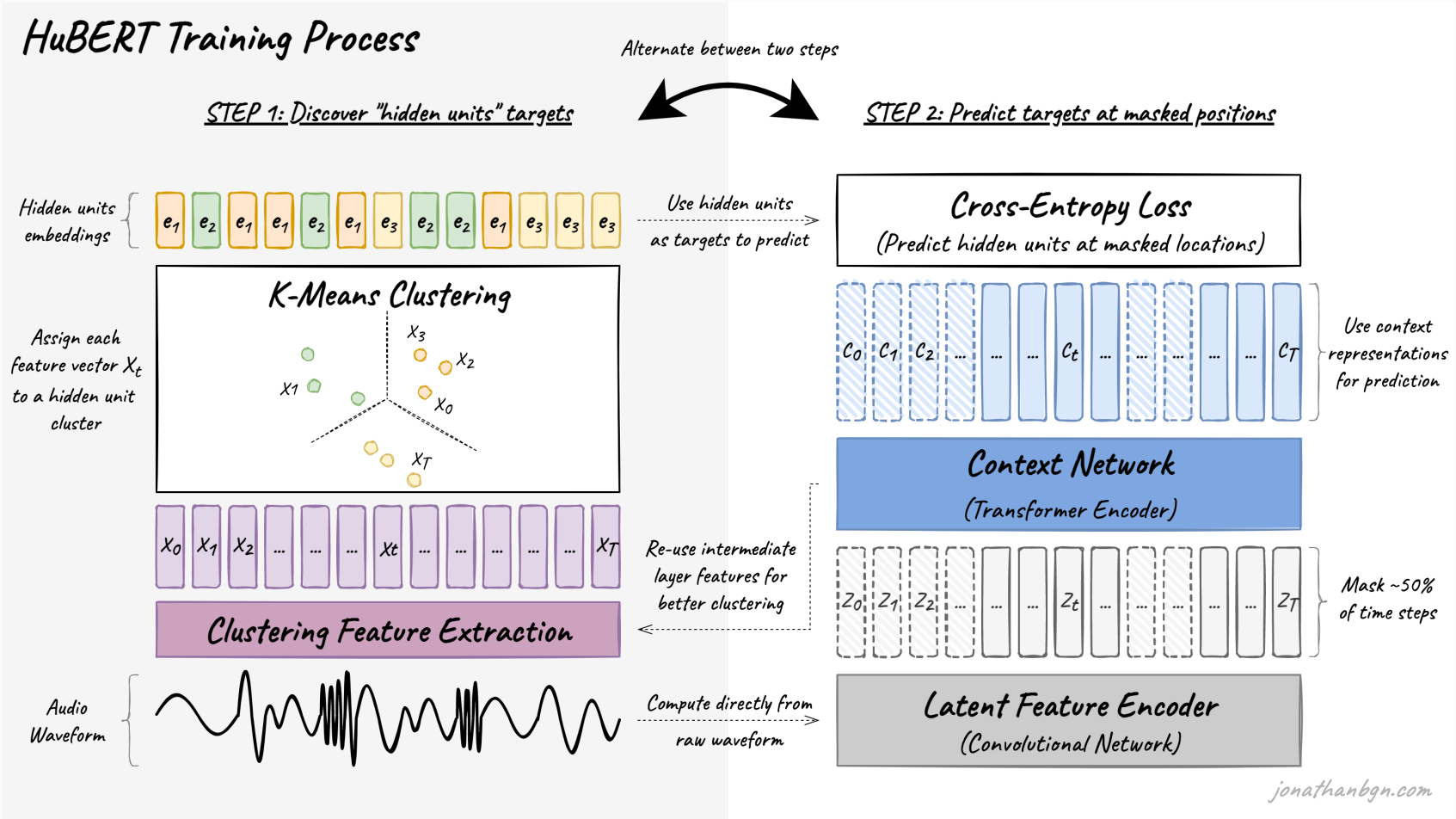
- We have large amounts of speech data
- but annotation is expensive!
- Can we still use the audios without transcripts? Yes, self-supervised learning offers a convenient solution.
- Most self-supervised pre-training methods rely on some smart way of clustering the data to discover acoustic and linguistic units.
- Self-supervised models still need a supervised fine-tuning with annotated data[1]
- Fine-tuning often uses CTC

[1] Except the completely self-supervised models often using GAN

Wav2Vec


Contrastive loss
$\mathcal{L}$
Context representations
$\mathcal{C}$
Transformer
Masked
...
...
Quantized representations
$\mathcal{Q}$
q
q
q
q
q
Latent speech representations
$\mathcal{Z}$
CNN
raw waveform
$\mathcal{X}$

HuBERT



[1] picture from https://jonathanbgn.com/2021/10/30/hubert-visually-explained.htmls

Other models

There are several other alternatives:

1. Whisper
2. wavLM
3. HuBERT

You can find several pre-trained and finetuned models on Hugging-Face, search for models with the automatic-speech-recognition tag

Summary

Summary

- We saw the definition of automatic speech recognition
- Commonly used techniques and models were introduced
- Overview on hybrid/traditional vs end2end solutions
- Brief glance at the latest technologies