Automatic Speech Recognition

Acoustic Modelling

Decoding

Applications

Janne Pylkkönen

Speechly

8.2.2022

Automatic Speech Recognition (ASR)

Speechly

- Lecture goals: To understand…
  - … what is automatic speech recognition
  - … how statistical models are used to recognise speech
  - … what are the fundamentals of modelling speech acoustics
  - … how deep neural networks are used in speech recognition
  - … how different applications use speech recognition
- In some forms, automatic speech recognition has existed already for over 50 years
- In the past decade, the use of speech recognition in consumer devices has exploded

Speech Recognition Tasks

Speechly

- Typical automatic speech recognition (ASR) tasks:
  - Keyword detection
  - Command-and-control
  - Search by speech
  - Dictation
  - Conversational interaction

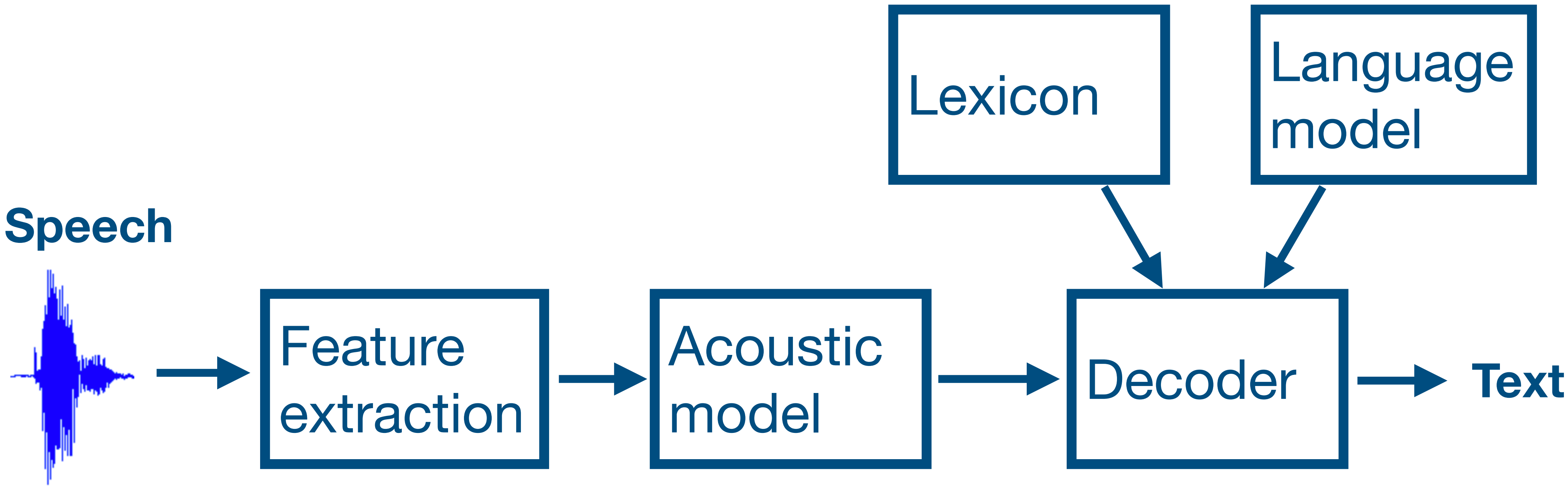
Easier

Harder

- Speech characteristics relating to the recognition task:
  - Isolated words vs. continuous speech
  - Speaker dependent vs. independent
  - Vocabulary size
  - Read speech, planned speech, conversational speech
  - Environmental noise
  - Space and distance to the microphone: close-talk, near-field, far-field
- Recognising everyday speech around us is challenging because it is speaker independent, conversational, large vocabulary, continuous speech, mixed with various environmental noises!

Components of a Traditional ASR System

Speechly



- Task of the automatic speech recognition: Find the most likely word sequence given the observations (speech) and the models for acoustics and language
- Speech acoustics are matched with a statistical model
- Language model is either a statistical model (n-gram, RNN), a fixed grammar, or in simple tasks just a vocabulary
- Lexicon ties together the units of acoustic and language model

The Fundamental Equation of ASR

Speechly

- Find the most likely word sequence given the observations and the models for acoustics and language:

Acoustic model:
Likelihood of the observations O, given the word sequence W

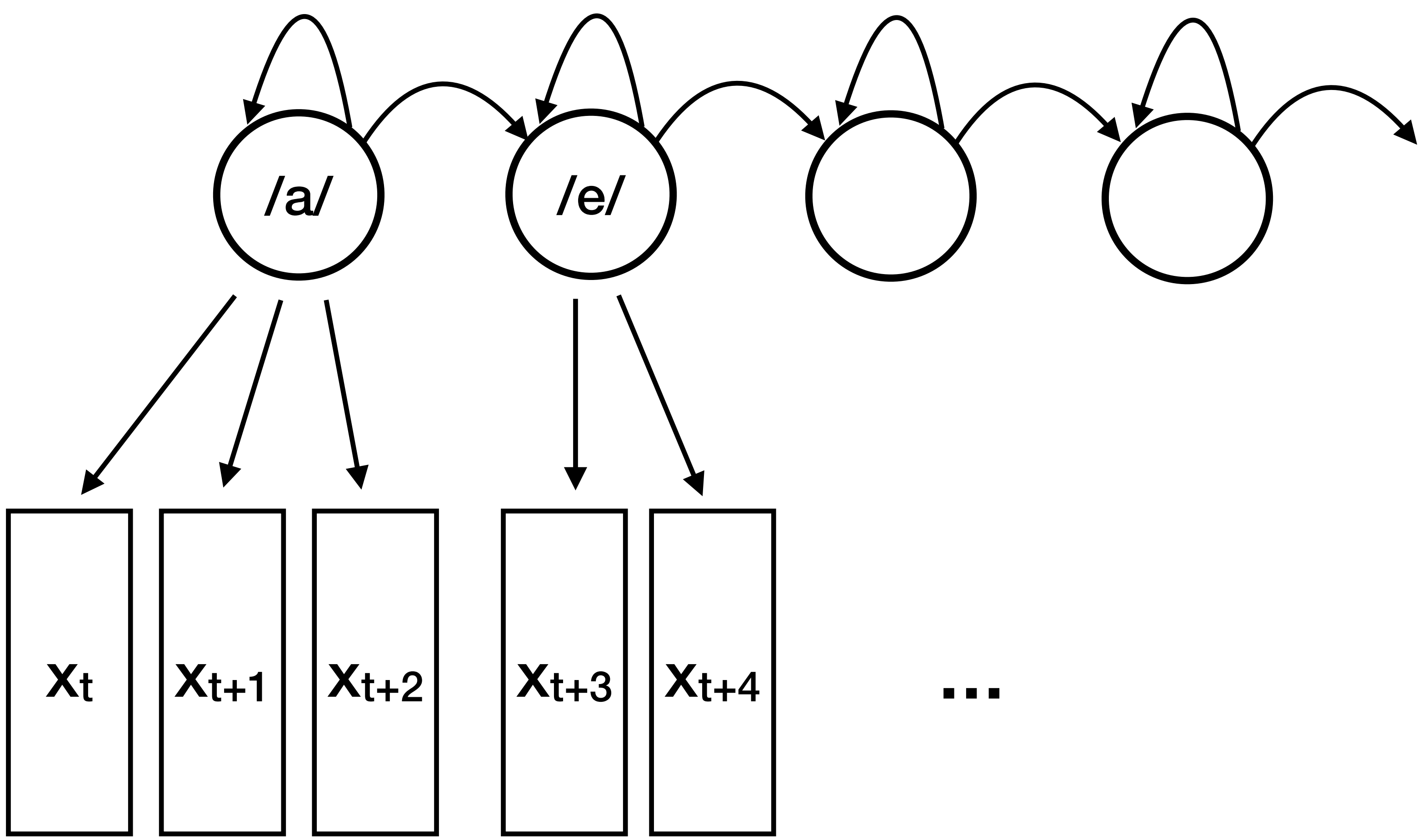
Language model:
Probability of the word sequence W

$$\hat{W} = \arg\max_{W} p(W \mid \boldsymbol{O}) = \arg\max_{W} p(\boldsymbol{O} \mid W)\, p(W)$$

Decoder:
Find the most likely word sequence W

Acoustic Model

- The information in speech signal is encoded in its time-varying properties
- The traditional model for the temporally varying speech signal is Hidden Markov Model (HMM): a sequence of states, each coupled with a specific emission probability model for the distribution of the observations
  - Nowadays emission distributions are modelled with neural networks (so called hybrid models), older systems used Gaussian mixture models
- HMM states correspond to basic recognition units, e.g. phones or senones



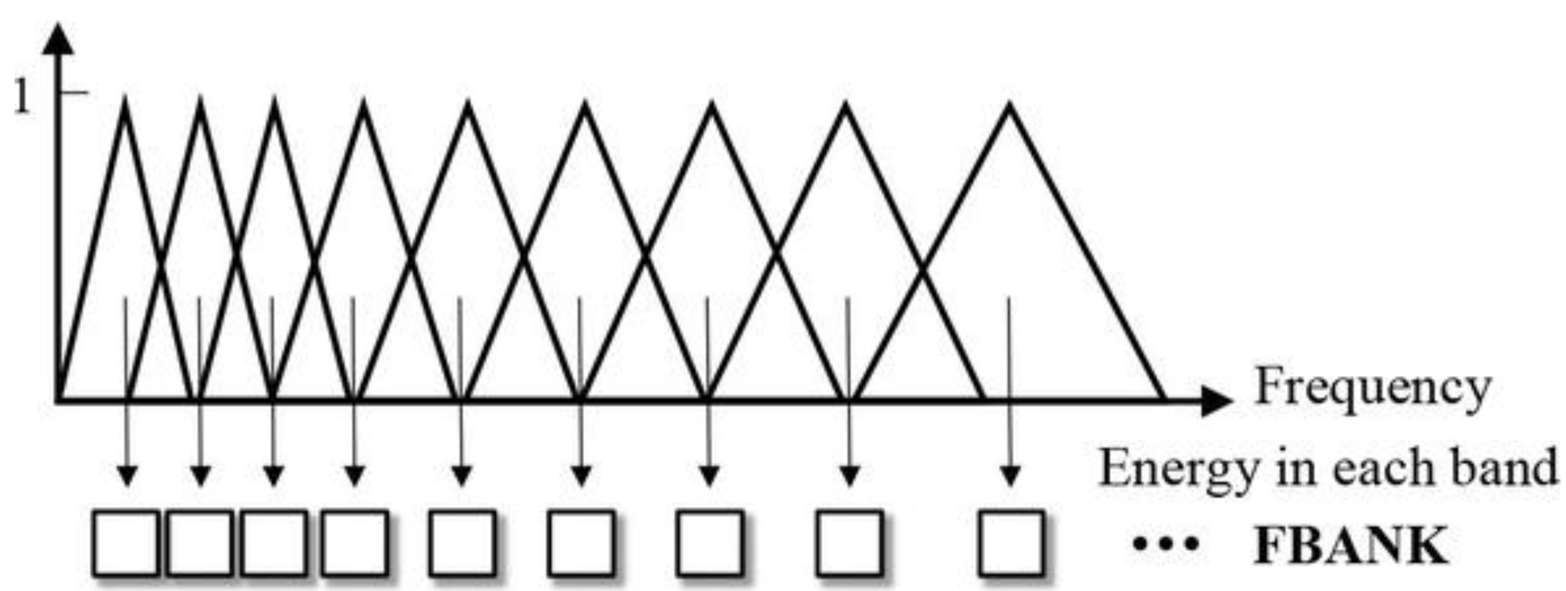
Phonemes, phones, triphones, senones

Speechly

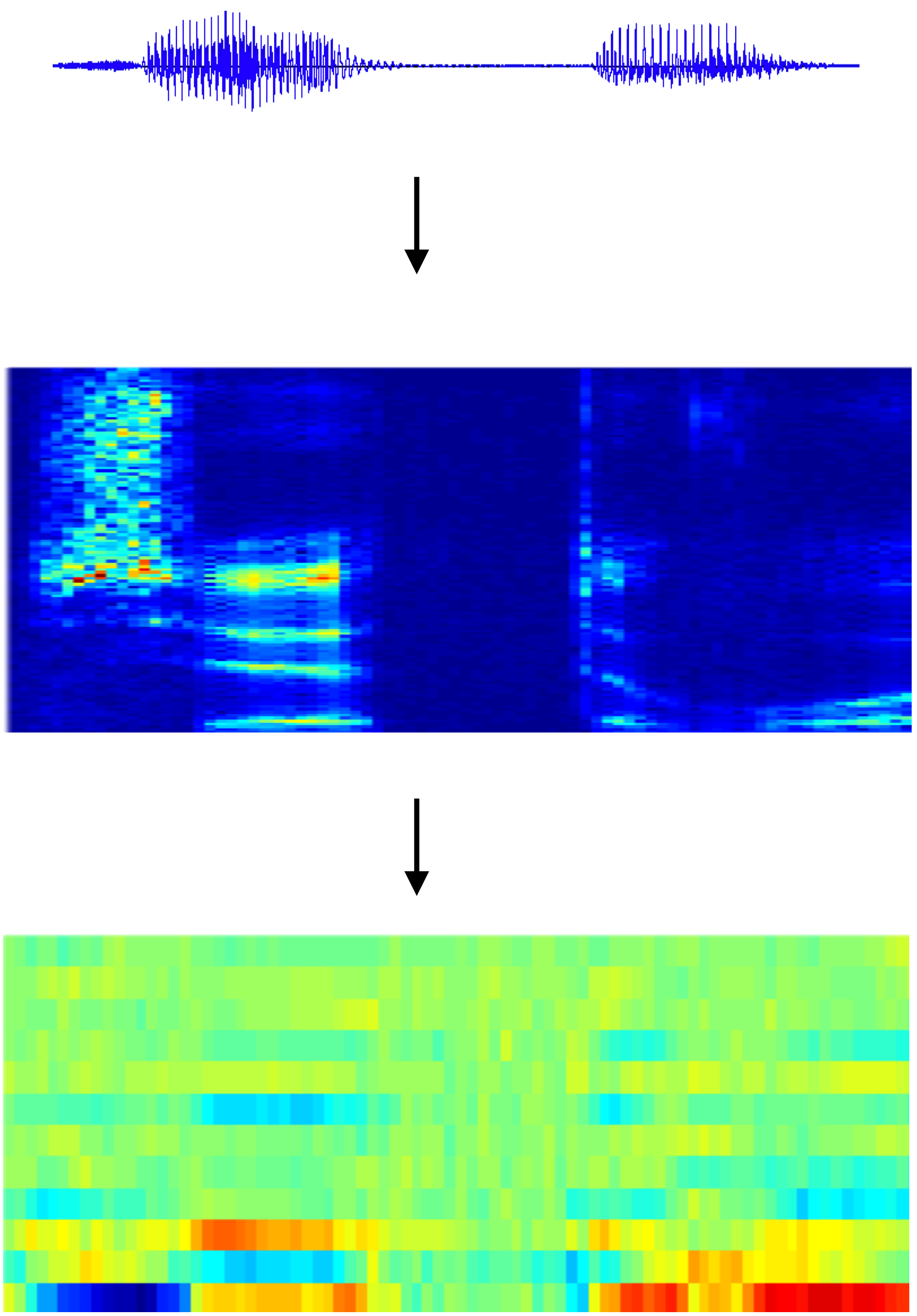
- ***Phoneme*** - The basic unit in spoken language, analogous to a letter in the written text
- ***Phone*** - Spoken realisation of a phoneme
- ***Lexicon*** - Mapping between words and phoneme sequences
- ***Context-dependent phone*** - A phone model which takes the surrounding phonemes into account
  - A large proportion of the acoustic variation of phones is due to this phoneme context
- ***Triphone*** - Context dependent phone which considers both the previous and the next phone, i.e. the left and the right context
  - Notation: **t-a+s** means phone ***/a/*** occurring between ***/t/*** and ***/s/***
- ***Senone*** - Part of a phone. Traditionally ASR systems have used 3 HMM states for modelling a single triphone. One state is then called a senone.

Features for Speech Acoustics

- To model the time varying speech signal with HMM-based acoustic models, the signal has to be converted into a sequence of short-time features
- The features need to retain the relevant information for the phone identities, while inhibiting unwanted variation (e.g. due to the speaker or environment)
- 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 characteristics of these features are non-linear frequency warping and energy compression



Example: MFCC feature extraction



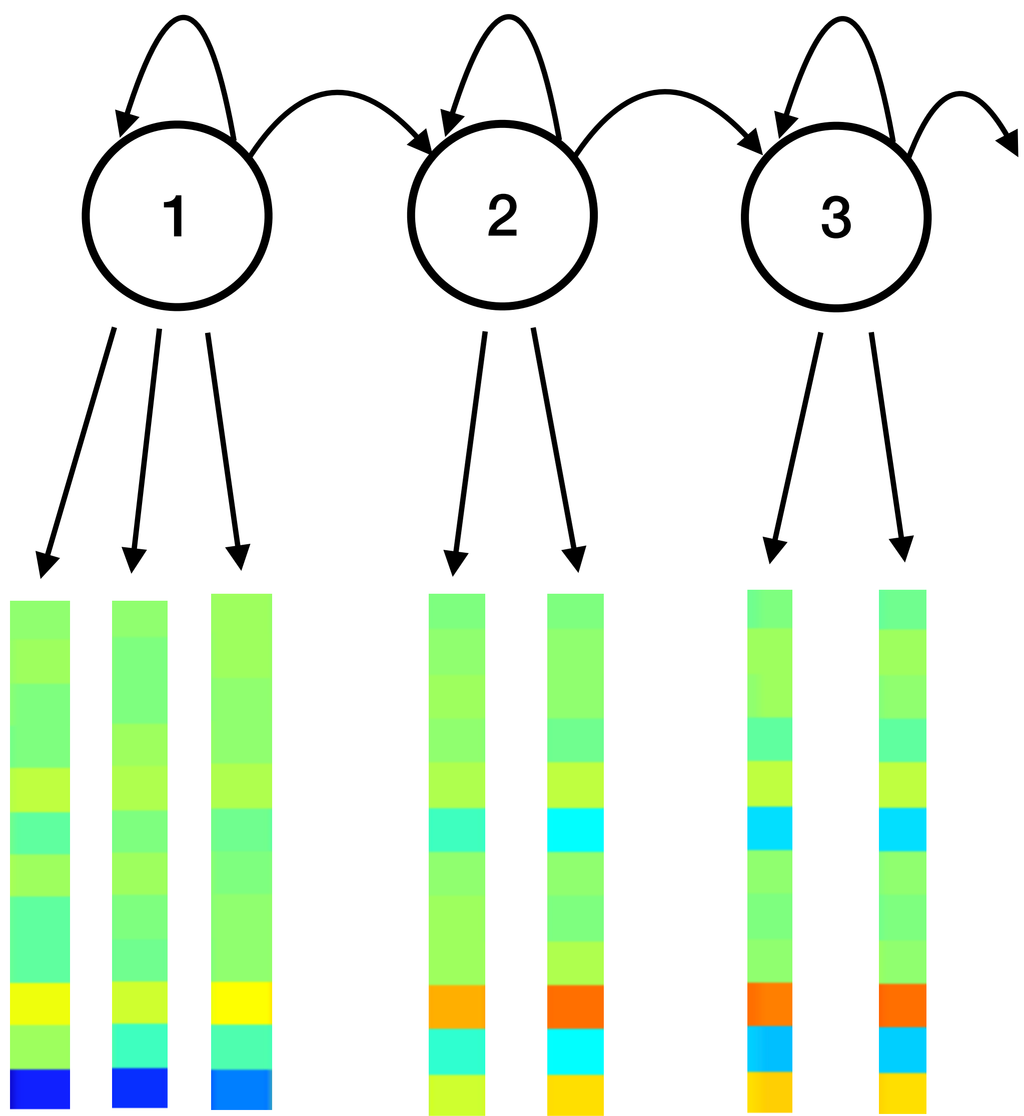
Typical properties of Mel-Frequency Cepstral Coefficient (MFCC) features in classical ASR:

- Feature vectors are 13 dimensional
- Each feature vector is extracted from a 25ms spectral analysis window
- Windows overlap such that the feature extraction generates 100 feature vectors per second

$$\longrightarrow \begin{pmatrix} 2.3 \\ -4.2 \\ 0.8 \\ \vdots \\ 1.3 \end{pmatrix} \begin{pmatrix} 1.7 \\ -3.4 \\ 2.1 \\ \vdots \\ 0.2 \end{pmatrix} \cdots \begin{pmatrix} 0.9 \\ 1.4 \\ -1.5 \\ \vdots \\ -2.6 \end{pmatrix}$$

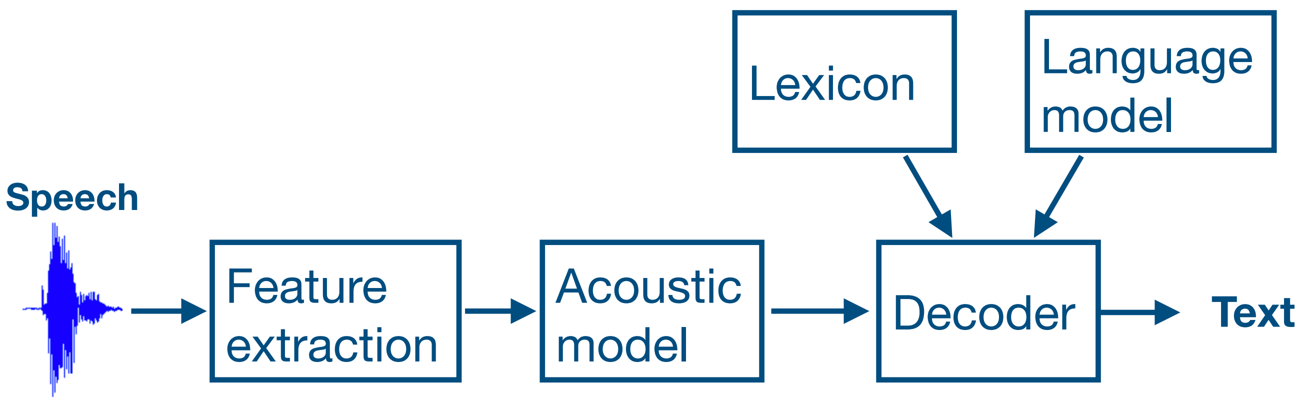
Estimating Emission Distribution Models

Speechly



- If we know the alignment of the feature vectors to the HMM states, it is possible to estimate the emission distribution model to represent the distribution of the observations in that state
- Typically this alignment is NOT known, and instead the alignment and emission distributions are estimated iteratively using the Expectation-Maximization (EM) algorithm
- The alignment over the HMM states can be obtained using the Viterbi algorithm
- The emission distribution is then estimated/ trained with the obtained alignment, and the process is repeated until convergence

Language Model and Decoder



- The output of the acoustic model is a sequence of probabilities of e.g. phones or senones
- That sequence needs to be decoded in order to find the most likely output message
- The phone sequences are converted to potential word sequences (hypotheses), using the information from the lexicon
- Language model (LM) defines the allowed words and gives probabilities for their sequences, effectively defining the decoder search space
- The decoder then finds the most probable output text for the input signal, combining the probabilities from the acoustic and language models

Example on scoring alternative hypotheses:

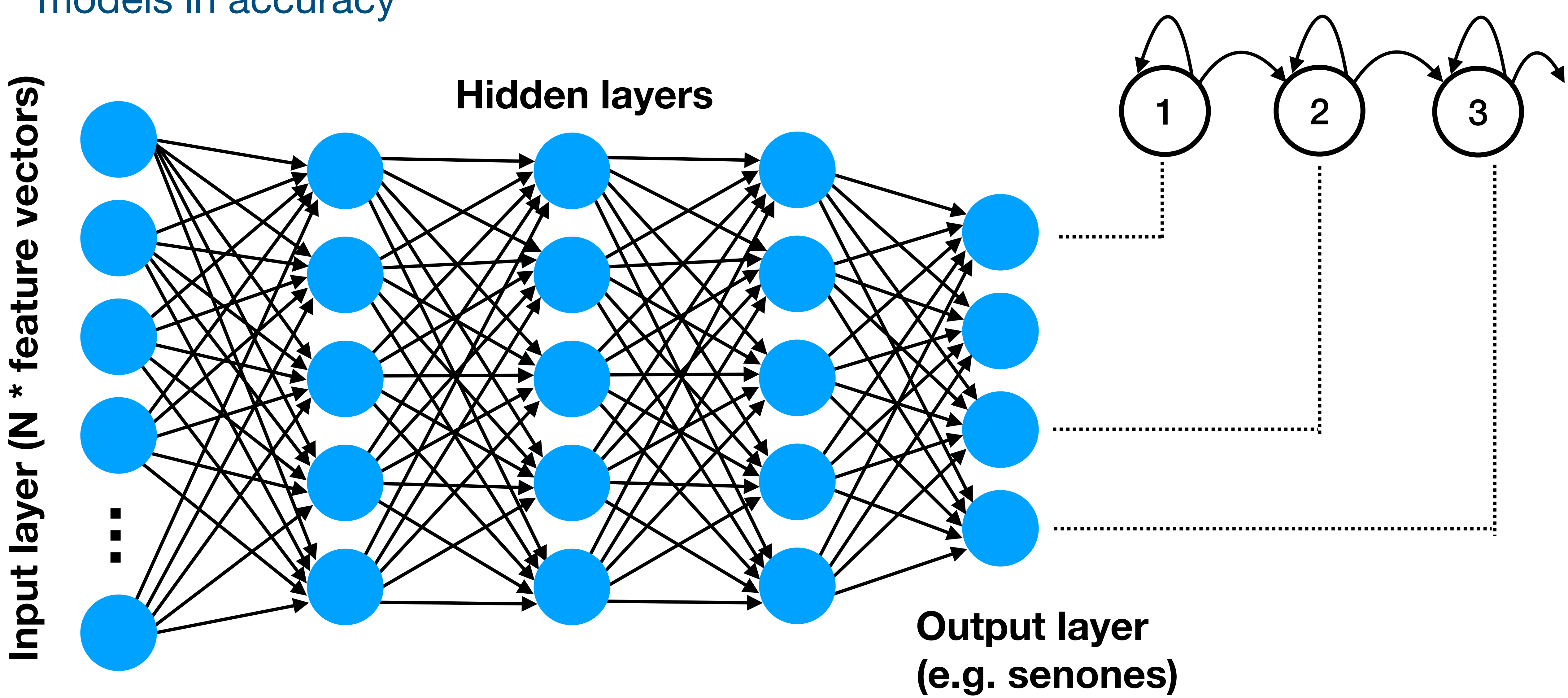
Wreck a nice beach → High AM score/Low LM score

Recognize speech → High AM score/High LM score → **Best hypothesis!**

Read the news → Low AM score/High LM score

Neural Networks for Acoustic Modelling

- Classical HMM-based ASR systems used generative Gaussian Mixture Models (GMMs) as the emission distribution models. Nowadays Deep Neural Networks (DNNs) are used instead.
- Such a combination of HMMs with DNNs is called a Hybrid ASR system
- Neural networks are discriminative models, which can outperform generative models in accuracy



Neural Network Acoustic Models

Speechly

- Neural network acoustic models come in many flavours:
  - Feed-forward networks
  - Recurrent networks (RNNs, LSTMs, GRUs)
  - Convolutional input layers
  - Attention-based models
- Some hybrid systems still use HMM/GMM models for initialisation and to define the DNN output layer
- Decoding relies on the acoustic model to produce likelihoods **p(o|s)**:

$$\hat{\mathrm{W}} = \arg\max_{\mathrm{W}} p(\boldsymbol{O} \mid \mathrm{W})\, p(\mathrm{W}) = \arg\max_{\mathrm{W}} \sum_{s_{1:T} \in W} \left(\prod_{t=1}^{T} p(o_t \mid s_t)\, p(s_t \mid s_{t-1}) \right) p(W)$$

- However, discriminative DNNs produce posterior probabilities **P(S|O)**
- Solution to the mismatch: Apply Bayes rule to convert posteriors to "pseudo-likelihoods", using state priors:

$$p(o \mid s) = \frac{P(s \mid o)\, p(o)}{P(s)} \propto \frac{P(s \mid o)}{P(s)}$$

End-to-End Models for ASR

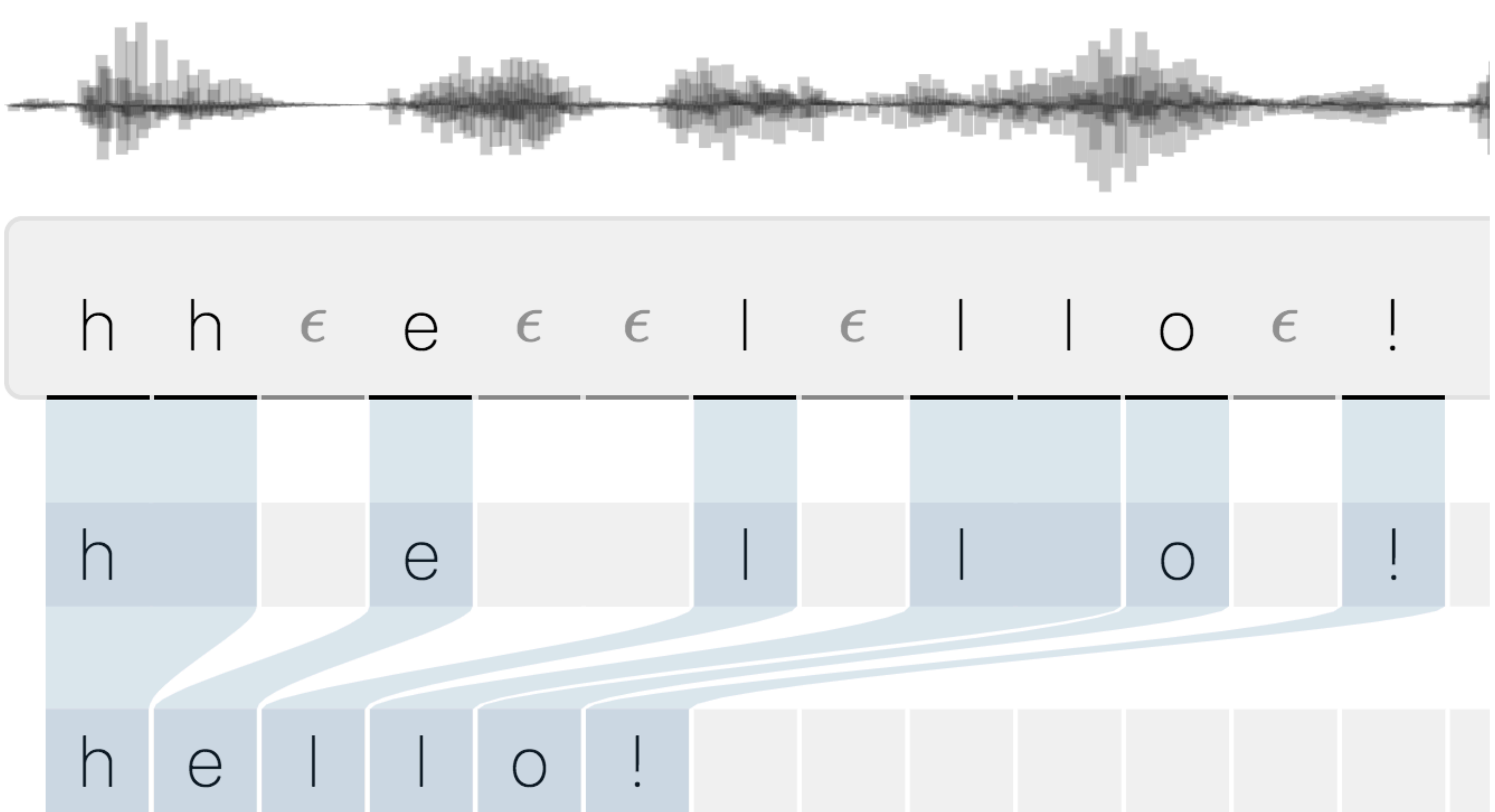
Speechly

- A growing trend in automatic speech recognition is to simplify the statistical modelling and decoding by using end-to-end models, more generally known as sequence-to-sequence classifiers
- End-to-end models take speech features as input, and produce text as output
- Benefits:
  - Simpler training procedure (just one model to train)
  - Possibility for more accurate models than with separated AM, LM, and lexicon
  - Decoding is significantly simpler and faster than with traditional ASR models
- Downsides:
  - Requires a lot of training data
  - Difficulty to adapt to new domains
  - Typically some language model is still needed for the best results, which complicates decoding

Connectionist Temporal Classification

Speechly

- A relatively simple method for alignment-free sequence modelling is called **CTC = Connectionist Temporal Classification**
  - Introduces a special blank symbol ε, which allows (potential) direct usage of the network output as a recognition result
- CTC refers to an output encoding scheme and a loss function for sequence classification problems
- Typically CTC is applied for training deep neural networks with recurrent layers (RNNS, LSTMs)



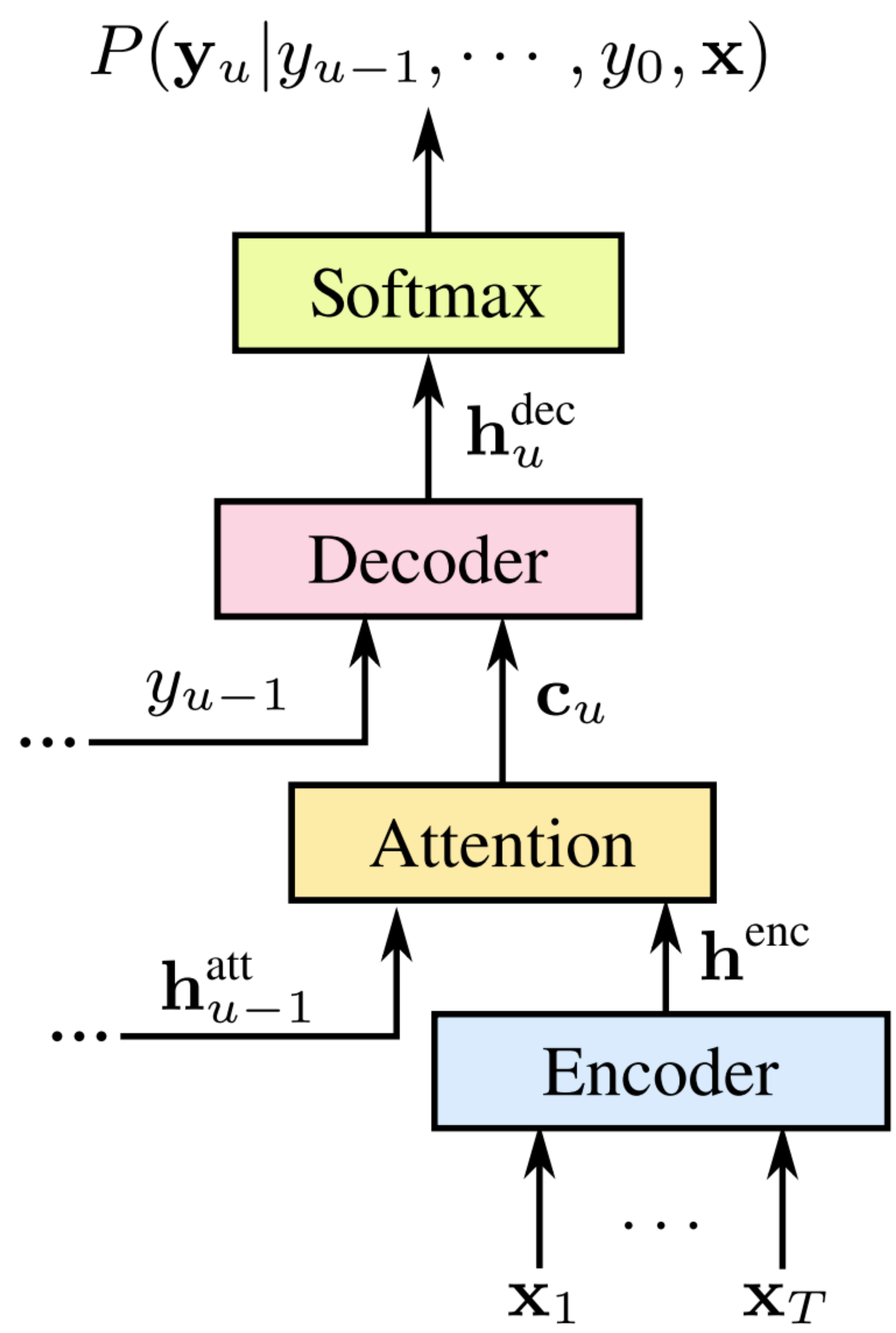
- CTC network outputs phones or letters (graphemes), instead of context-dependent senones
- In practice, a LM and a decoder is still needed

https://distill.pub/2017/ctc/

Encoder, Attention, Decoder

Speechly

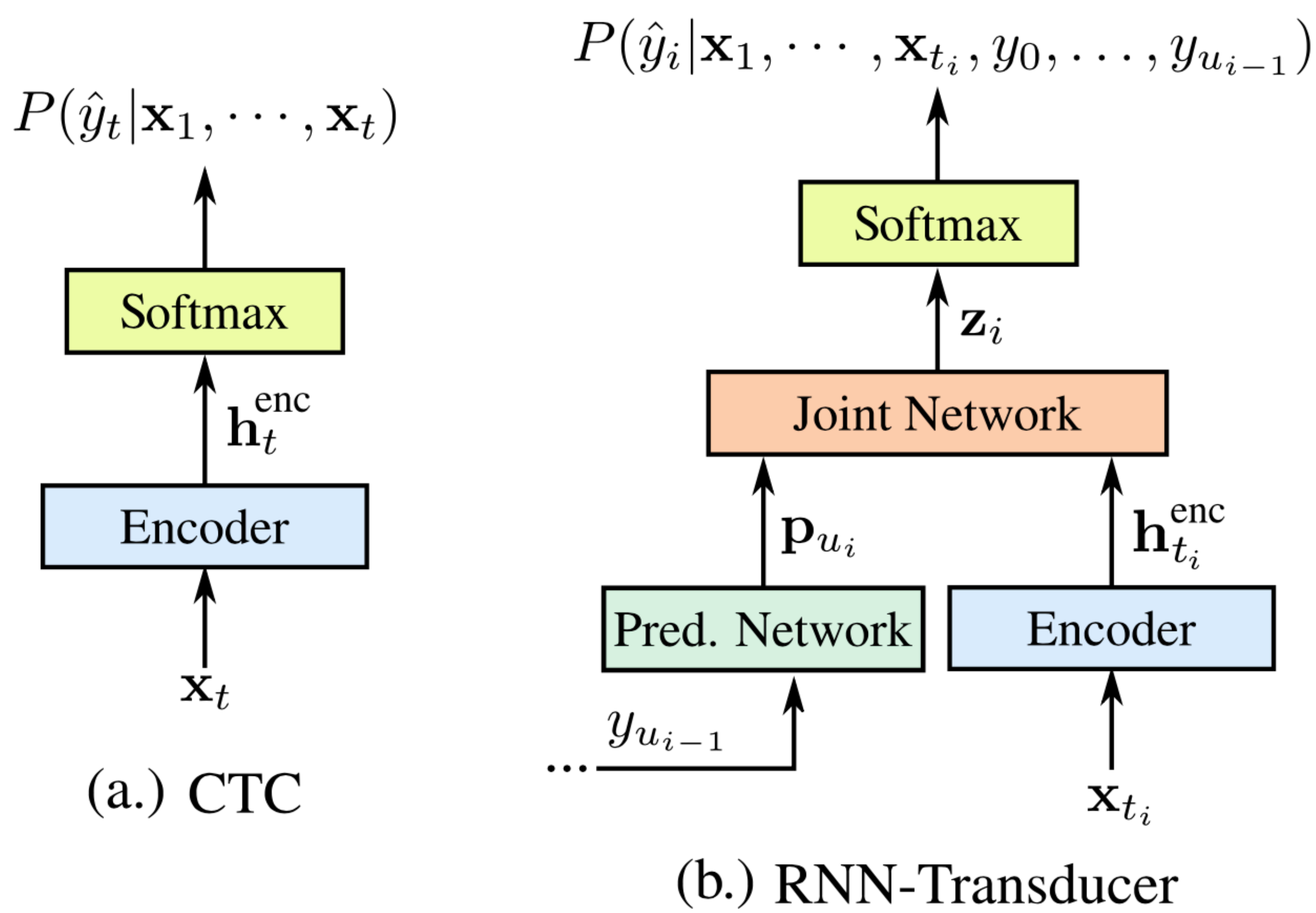
- Another type of end-to-end model uses an encoder-decoder approach with attention mechanism
- A complex neural network can output graphemes (letters) directly, without an explicit lexicon
- **Listen, Attend and Spell (LAS)** by Google consists of:
  - **Encoder** (Listener) resembles traditional acoustic model
  - **Attention** mechanism resolves alignment between input frames and output symbols
  - **Decoder** (Speller) acts as a language model and constructs the output
  - All the model blocks are optimised jointly
  - An additional LM can still improve the accuracy



Prabhavalkar *et al*.: A Comparison of Sequence-to-Sequence Models for Speech Recognition (Interspeech 2017)

RNN Transducer

- A recurrent architecture that has recently gained popularity for real-time ASR
- Originally proposed by Graves et al. (2012, 2013), later used e.g. by Google for their on-device (mobile phone) ASR
- The network output is either letters or words/word-pieces
- The encoder in RNN-T is similar to the CTC models
- RNN-T architecture introduces a predictor, which is an integrated neural language model with a recurrent input signal
- The joint network combines the outputs of the encoder and the predictor to produce the probability distribution of the next letter or word



(a.) CTC

(b.) RNN-Transducer

He *et al.*: Streaming End-to-End Speech Recognition for Mobile Devices (ICASSP 2018)

Applications of Speech Recognition

- The improvements in automatic speech recognition accuracy have led to the adoption of ASR in various applications
- ASR is often used as a more efficient and convenient replacement to typing, but speech recognition can also enable completely new kinds of interactions and levels of automation
- Typical uses of ASR include
  - Command-and-control applications
  - Dictation
  - Automatic call center operation
  - Generating transcriptions and TV subtitles
- Smart speakers such as Amazon Echo and Google Home have popularised using speech to control simple tasks
  - Home automation can be controlled with speech, even if the devices themselves don’t have ASR capabilities: It is enough that they can communicate with the smart speaker.

Challenges in ASR

- Although automatic speech recognition accuracy already matches human performance in many practical tasks, there are still challenges that need constant attention:
  - Out-of-vocabulary words are difficult to recognise correctly
  - Varying environmental noises impair recognition accuracy
  - Overlapping speech and “cocktail party” situations are especially problematic
  - Accented speech doesn’t work as well as native speech
  - Recognising child speech, or people with speech production disabilities, may perform poorly
- Often the key to a successful model is to obtain enough realistic, in-domain training data. Some data can be simulated if necessary.
- Many DNN-based models require huge amounts of data for training, in the order of thousands of hours. End-to-end models may need up to 100,000h of speech for the best performance!

ASR in Smart Speakers

Speechly

- Smart speakers are always on, waiting for a dedicated wake word
- Once the wake word is detected, the speech is streamed to the cloud, where speech recognition, natural language understanding, and response generation takes place
- Special challenges:
  - Robust wake word detection on the device
  - Far-field speech recognition, possibly with a lot of background noise
  - Low-latency cloud-based ASR
  - Personalisation to match user's needs and habits, like recognising songs from a personal playlist
- ASR solutions:
  - Noise-robust feature extraction with beam-forming and acoustic echo cancellation
  - Complex DNN-based ASR models in the cloud, trained from tens of thousands of hours of speech, using real or simulated room acoustics

Group Discussion

Speechly

- Think about an application where ASR would be useful, but where it is not yet commonly used. How would ASR change the user experience, or enable a new service? What are the biggest challenges for ASR in that use case?

Speechly

https://www.speechly.com/