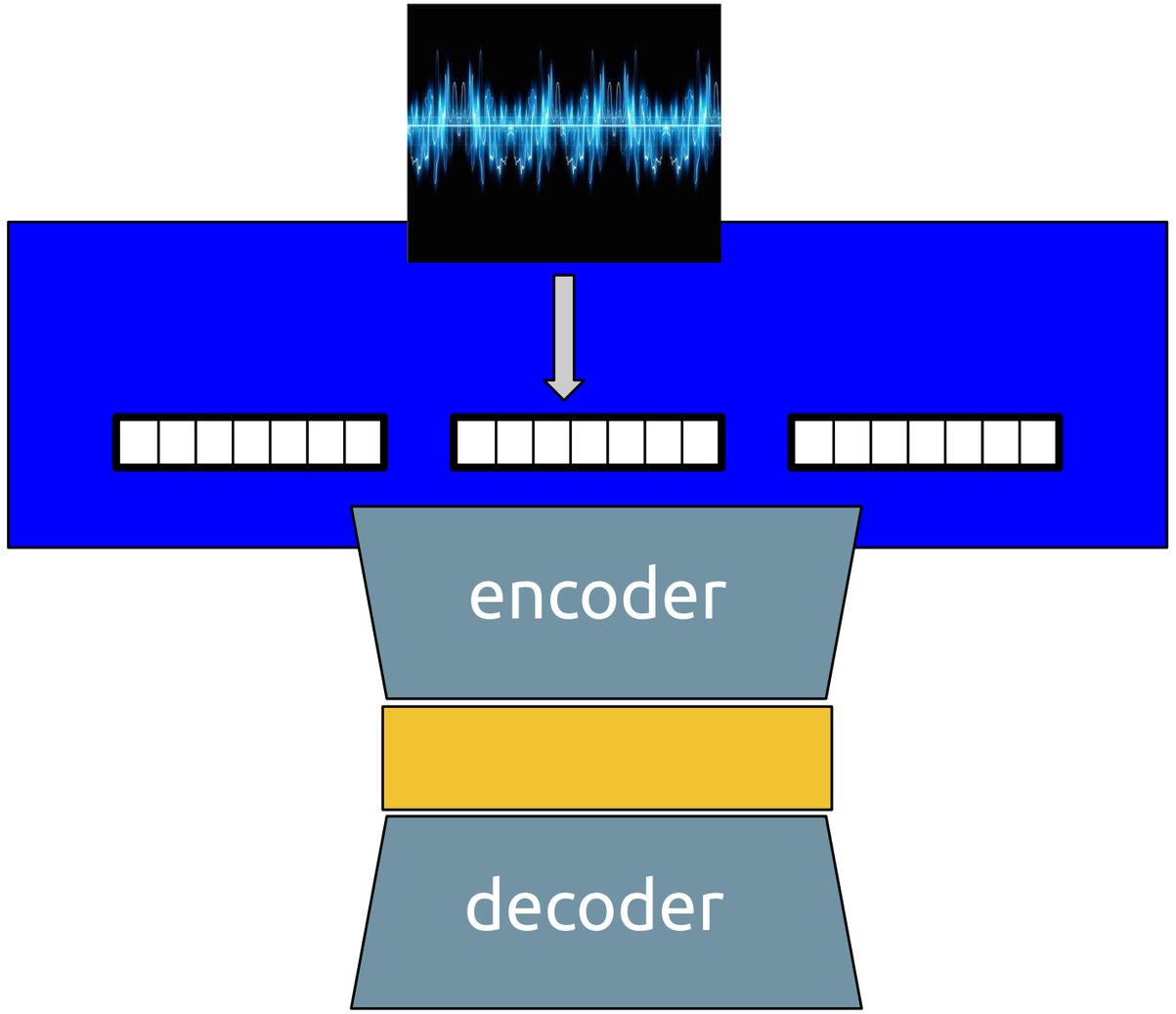
Sec 2.2

Input representations

From text translation to speech translation

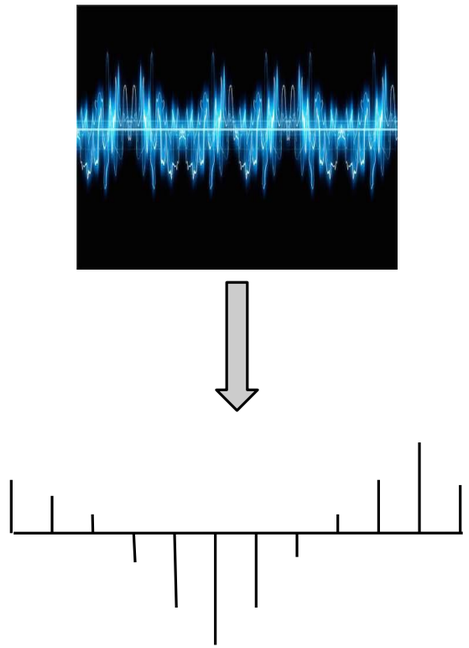
- Encoder-decoder models:
  - Can apply similar techniques

- Main differences to text translation
  - Input: Audio signal
    - Continuous
    - Longer



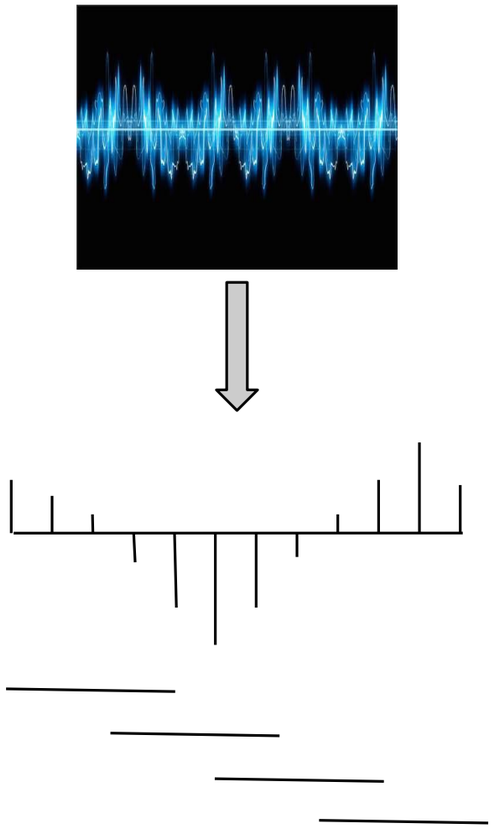
Audio representation

- Following best-practice from ASR
- Sampling
  - Measure Amplitude of signal at time t
  - Typically 16 kHz



Audio representation

- Following best-practice from ASR
- Sampling
  - Measure Amplitude of signal at time t
  - Typically 16 kHz
- Windowing
  - Split signal in different windows
    - Length: ~ 20-30 ms
    - Shift: ~ 10 ms
- Result:
  - One representation every 10 ms

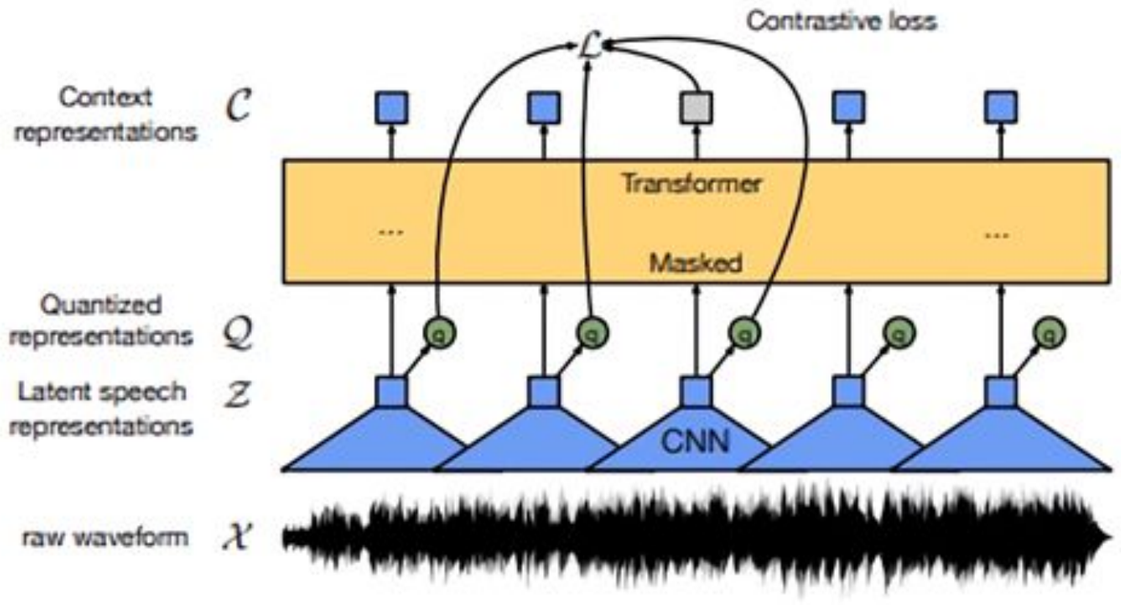


Audio representation

- Input features:
  - Signal processing:
    - Most common:
      - Mel-Frequency Cepstral Coefficients (MFCC)
      - Log mel-filterbank features (FBANK)
    - Idea:
      - Analyse frequencies of the signal
    - Steps:
      - Discrete Fourier Transformation
      - Mel filter-banks
      - Log scale
      - (Inverse Discrete Fourier Transformation)
    - Size:
      - 20-100 features per frame
  -

Audio representation

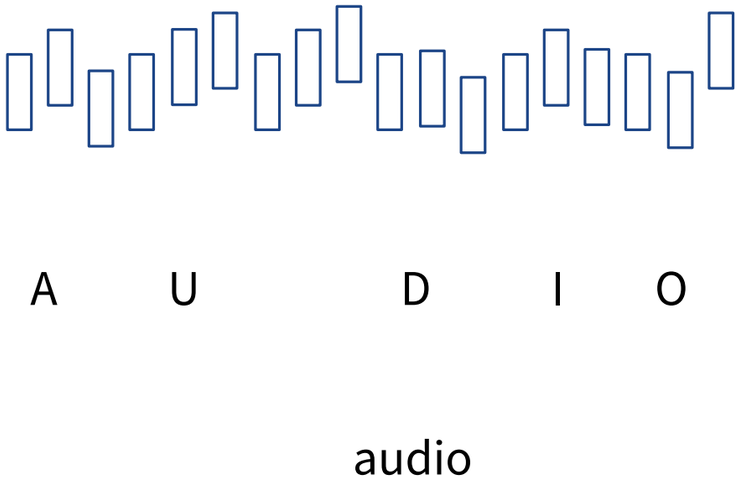
- Input features:
  - Signal processing:
  - Deep Learning:
    - Self-supervised Learning
      - Predict frame based on context
    - E.g. Wav2Vec 2.0 (Baevski et al., 2020)



Baevski et al. 2020

Challenges

- Variation
  - Many different ways to speech same sentence
  - Data augmentation
- Sequence Length
  - IWSLT test set 2020
    - Segments: 1804
    - Words: 32.795
    - Characters: 149.053
    - Features: 1.471.035
  - Architectural changes

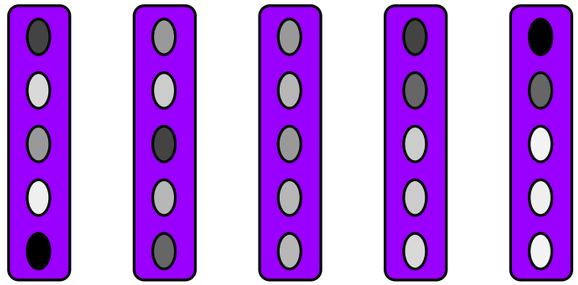


Data augmentation

- Limited training data
- Generate synthetic training data
- ASR investigated several possibilities
  - Noise injection (Hannun et al., 2014)
  - Speed perturbation (Ko et al., 2015)
- Successful technique in deep learning ASR
  - SpecAugment (Spark et al., 2019)
  - Also applied in ST (Bahar et al, 2019)

SpecAugment

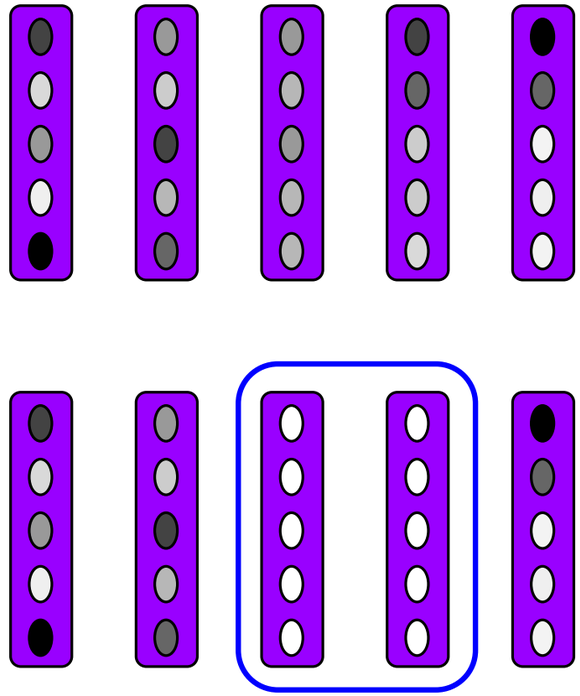
- Directly applied on audio features
- Idea:
  - Mask information



SpecAugment

- Directly applied on audio features
- Idea:
  - Mask information

- *Time masking*
  - Set several consecutive feature vector to zero



SpecAugment

- Directly applied on audio features
- Idea:
  - Mask information

- *Time masking*
  - Set several consecutive feature vector to zero

- *Frequency masking*
  - Mask consecutive frequency channels