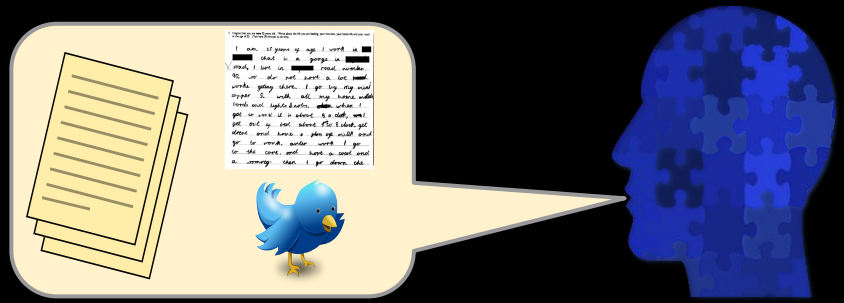
Automatic Speech Recognition

CSE354 - Spring 2021
Natural Language Processing



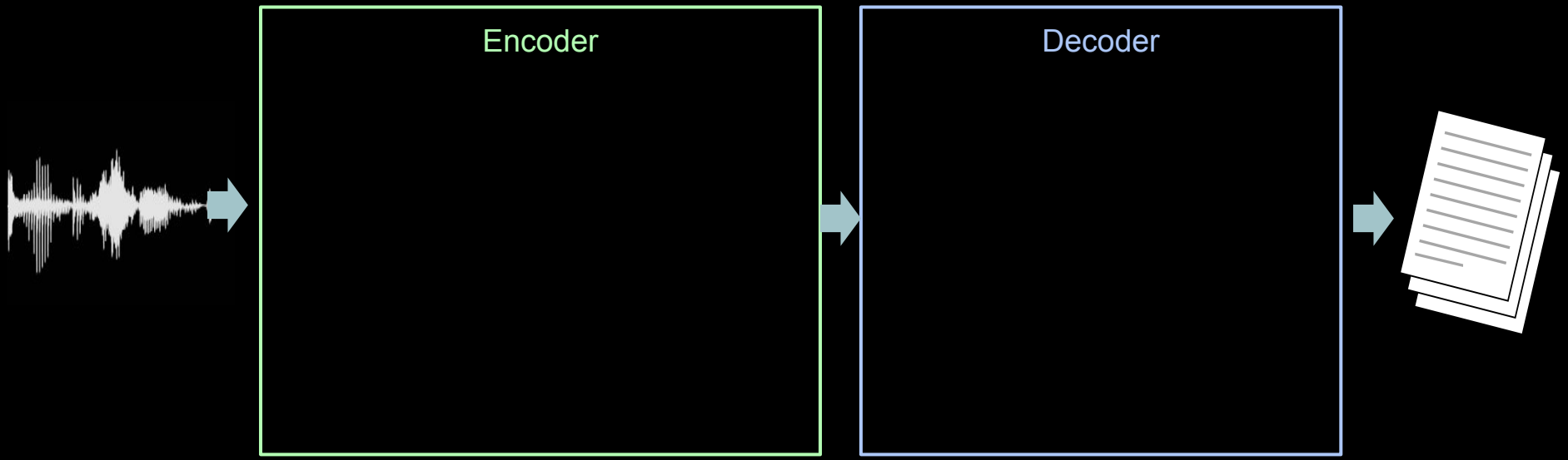
Most NLP Tasks. E.g.

- Automatic Speech Recognition

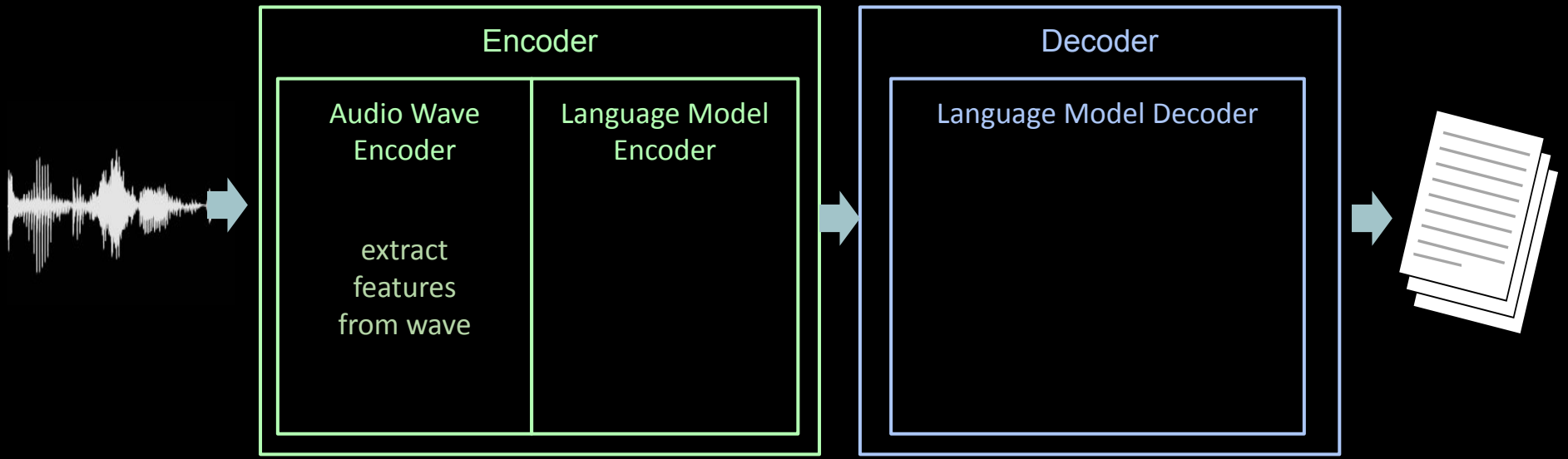
- Encoder-Decoders
  - Sequence Models
    - RNNs
    - Transformers
    - Pre-trained transformer networks (e.g. GPT)
  - Audio feature encoders

ASR: Automatic Speech Recognition

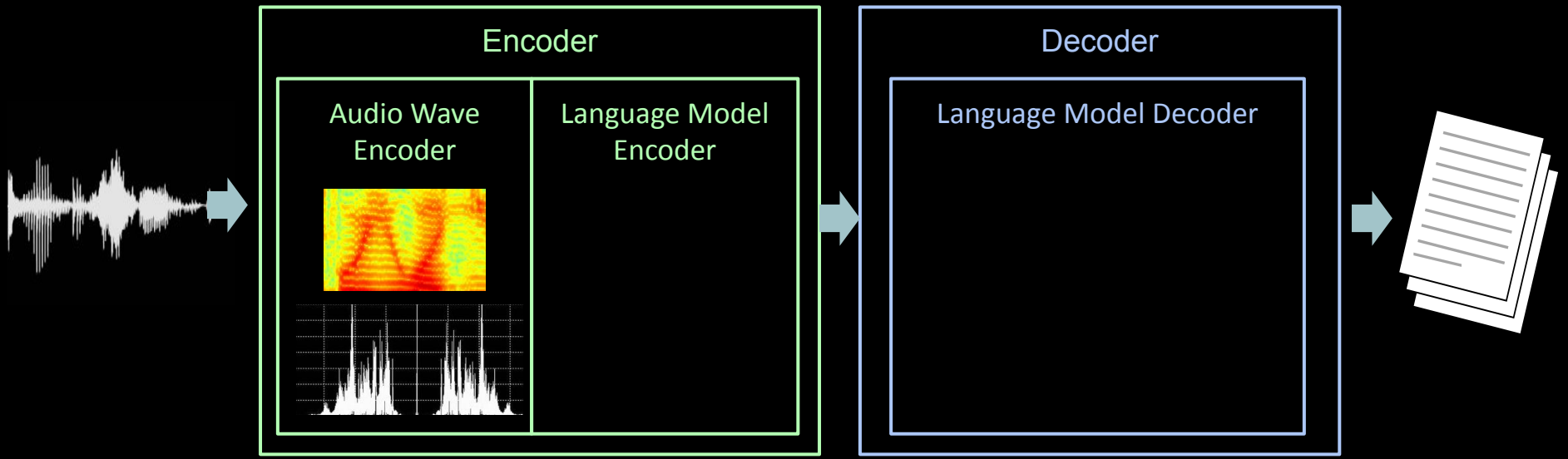
ASR: Automatic Speech Recognition



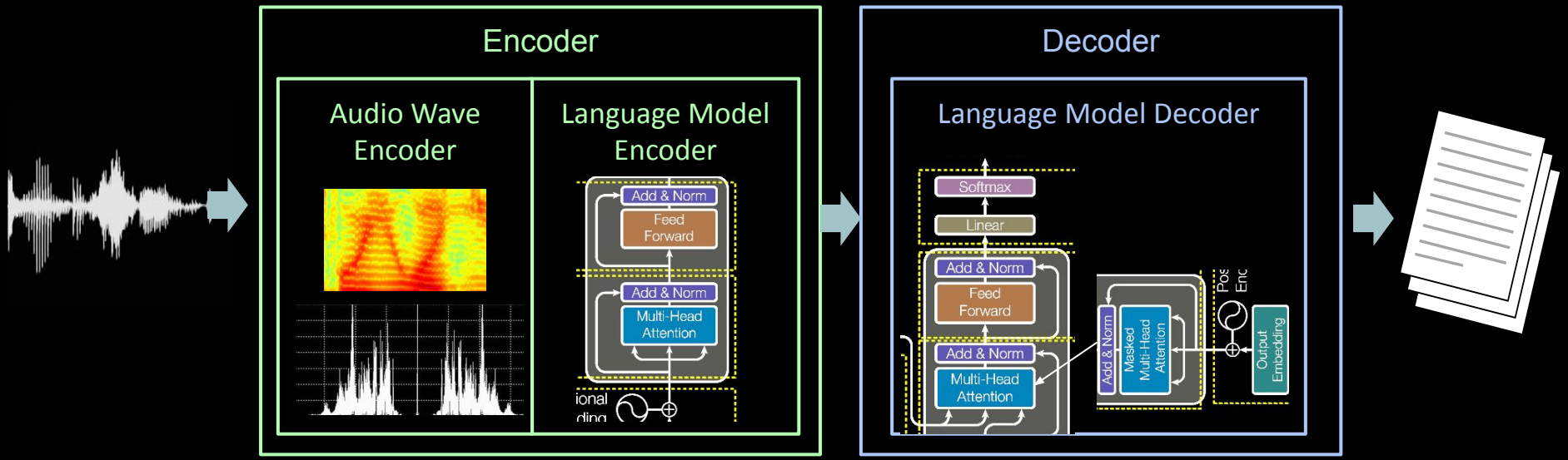
ASR: Automatic Speech Recognition



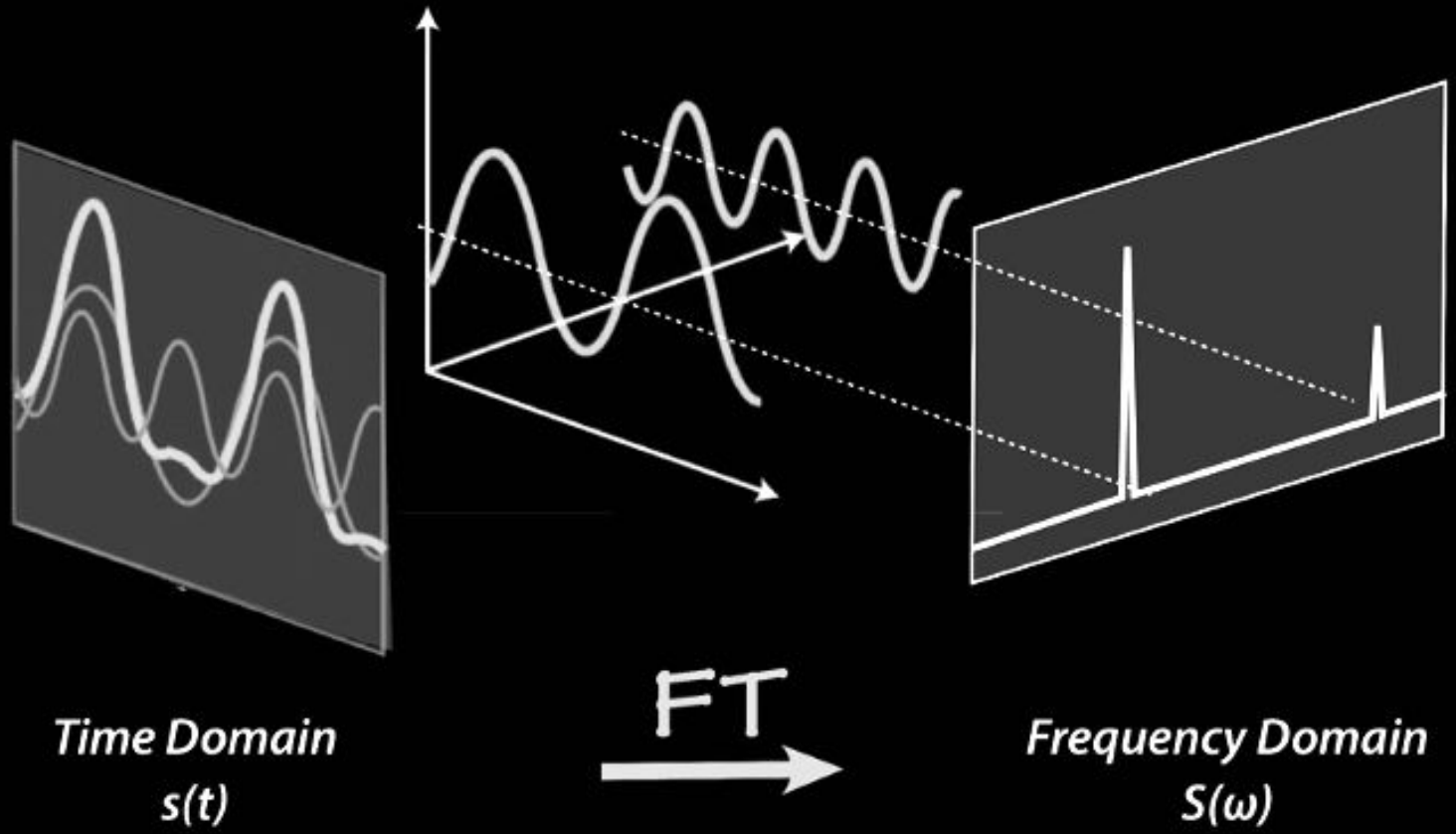
ASR: Automatic Speech Recognition



ASR: Automatic Speech Recognition

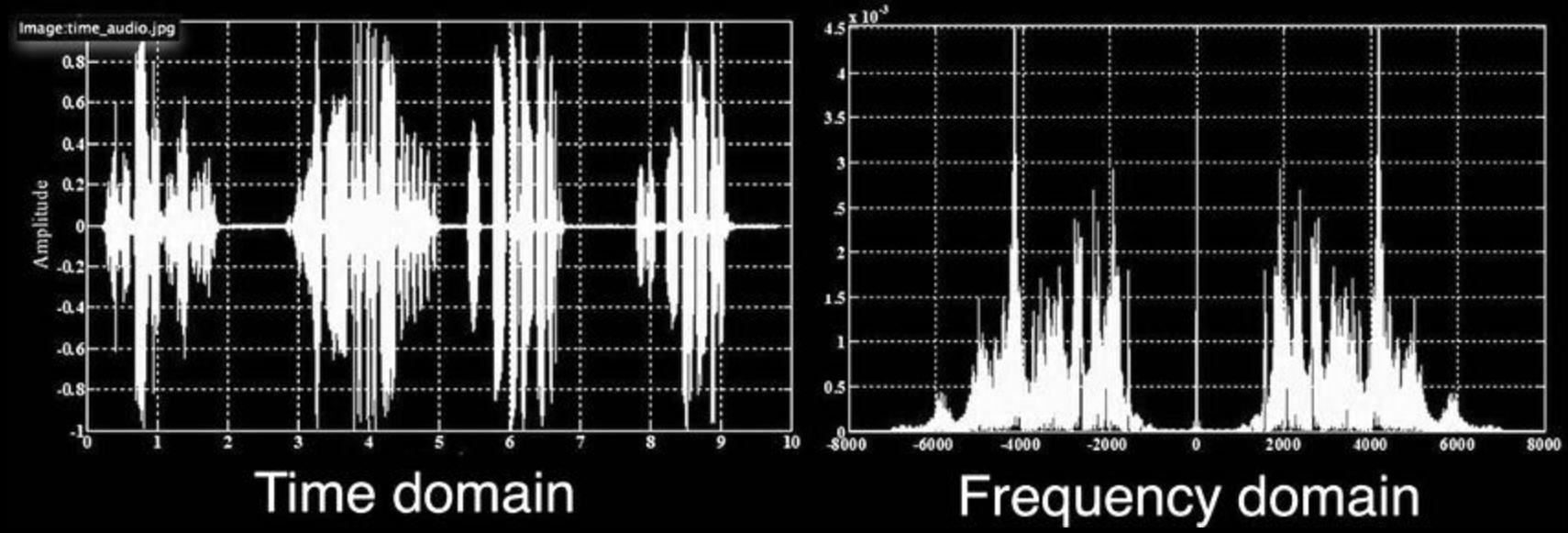


Encoding Waves: Fourier Transform



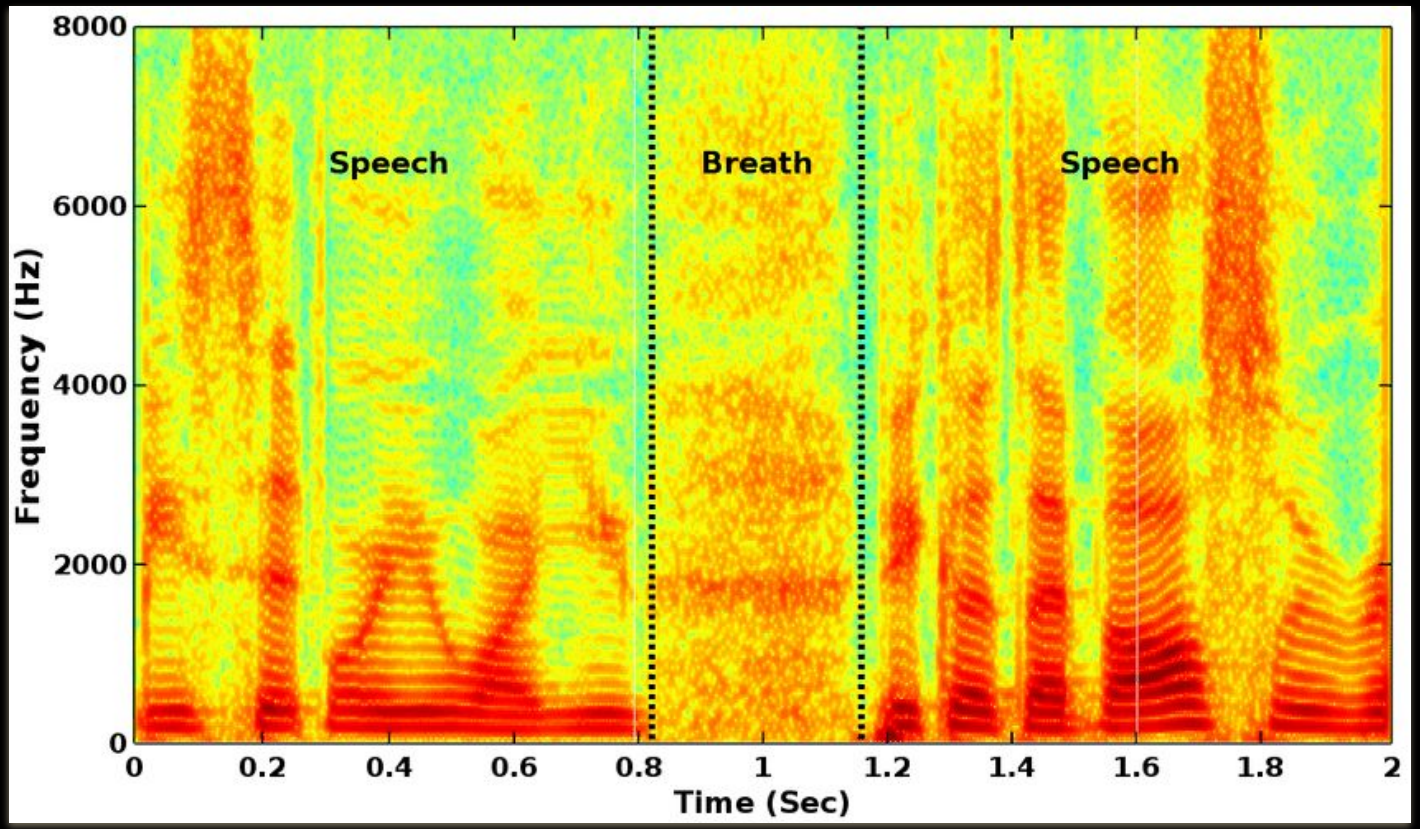
(Choudhary, 2020)

Encoding Waves: Fourier Transform



(Abdulsalam, Ayad. (2017). Audio Classification Based on Content Features.)

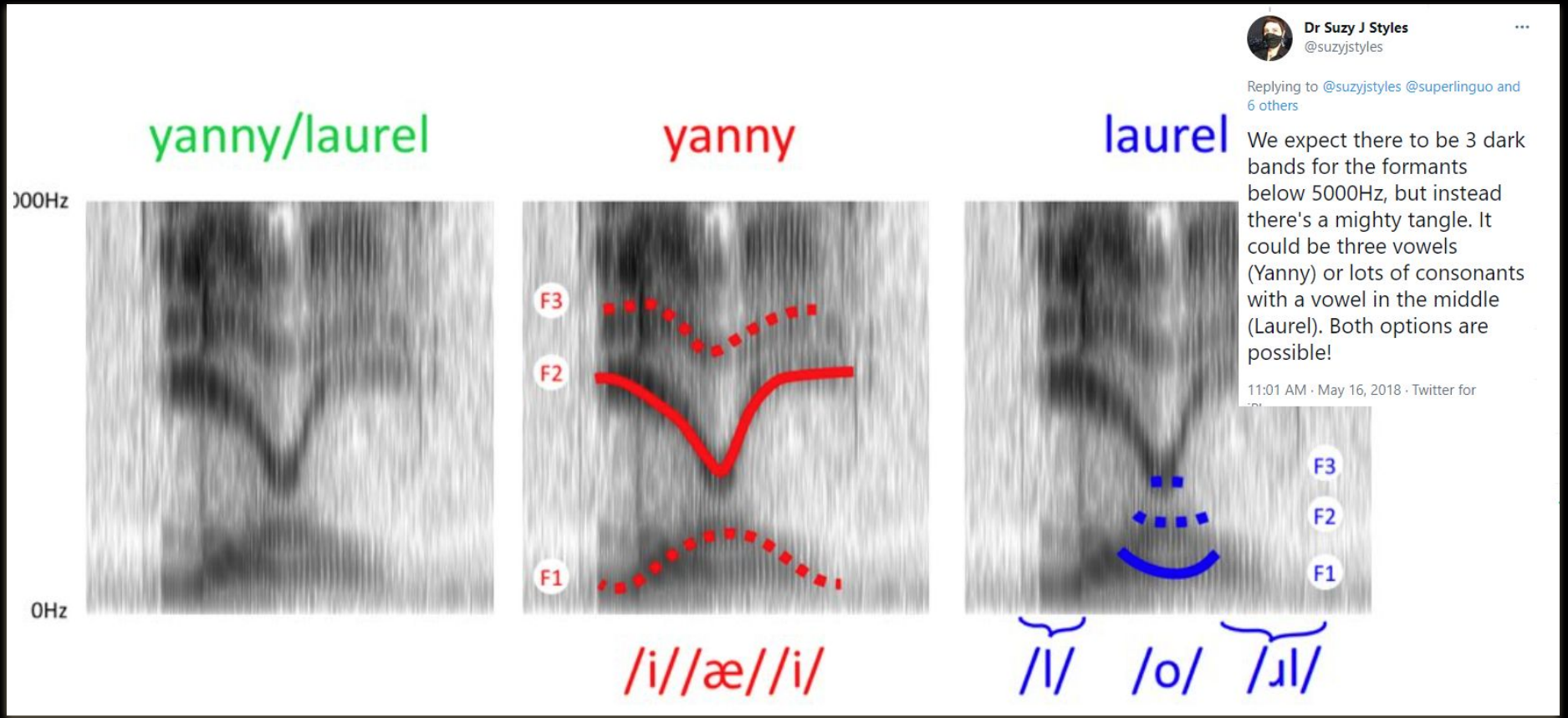
Spectrogram



(Dumpalla & Alluri, ICSC 2017)

Yanny Laurel

wiki/File:YannyLaurel.ogg



Dr Suzy J Styles
@suzyjstyles

Replying to @suzyjstyles @superlinguo and 6 others

We expect there to be 3 dark bands for the formants below 5000Hz, but instead there's a mighty tangle. It could be three vowels (Yanny) or lots of consonants with a vowel in the middle (Laurel). Both options are possible!

11:01 AM · May 16, 2018 · Twitter for

Masked Transformer for Encoding

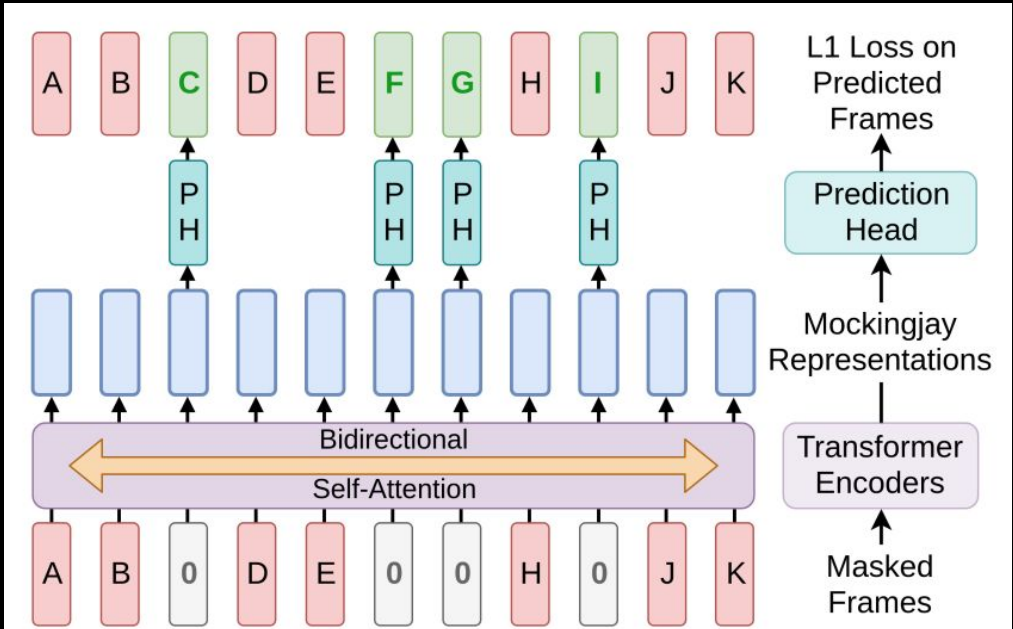
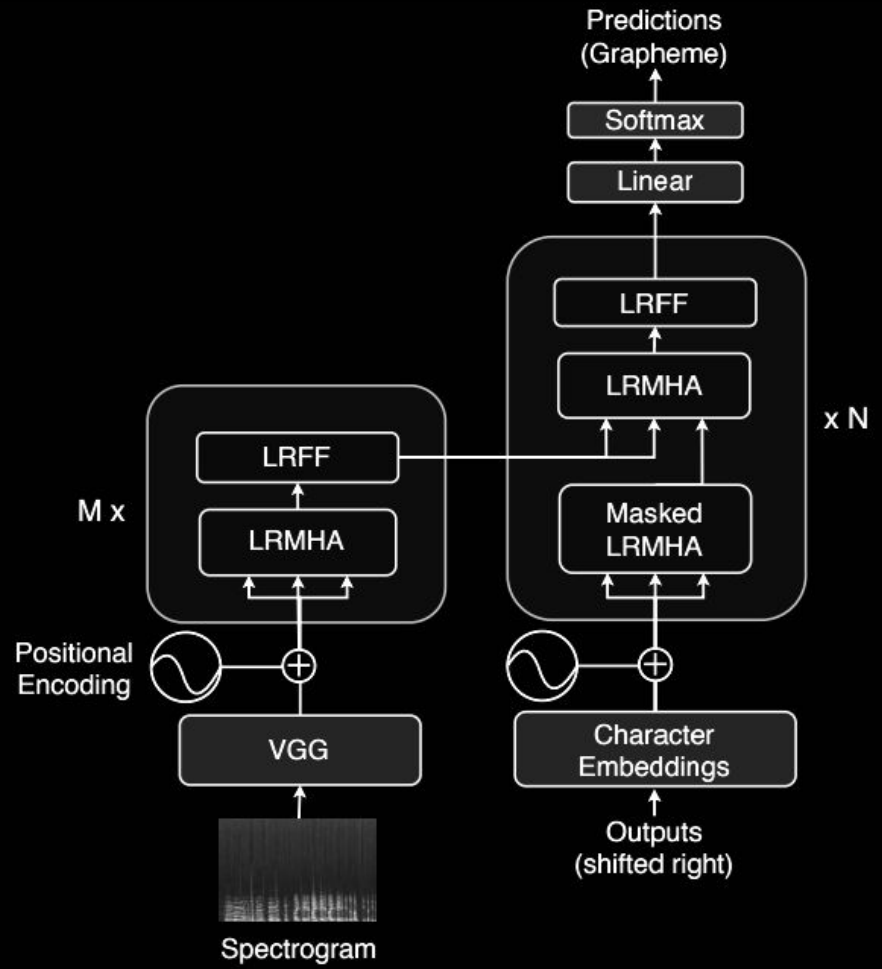


Fig. 1. The proposed Masked Acoustic Model pre-training task, 15% of input the frames are masked to zero at random.

(Liu et al., 2020)

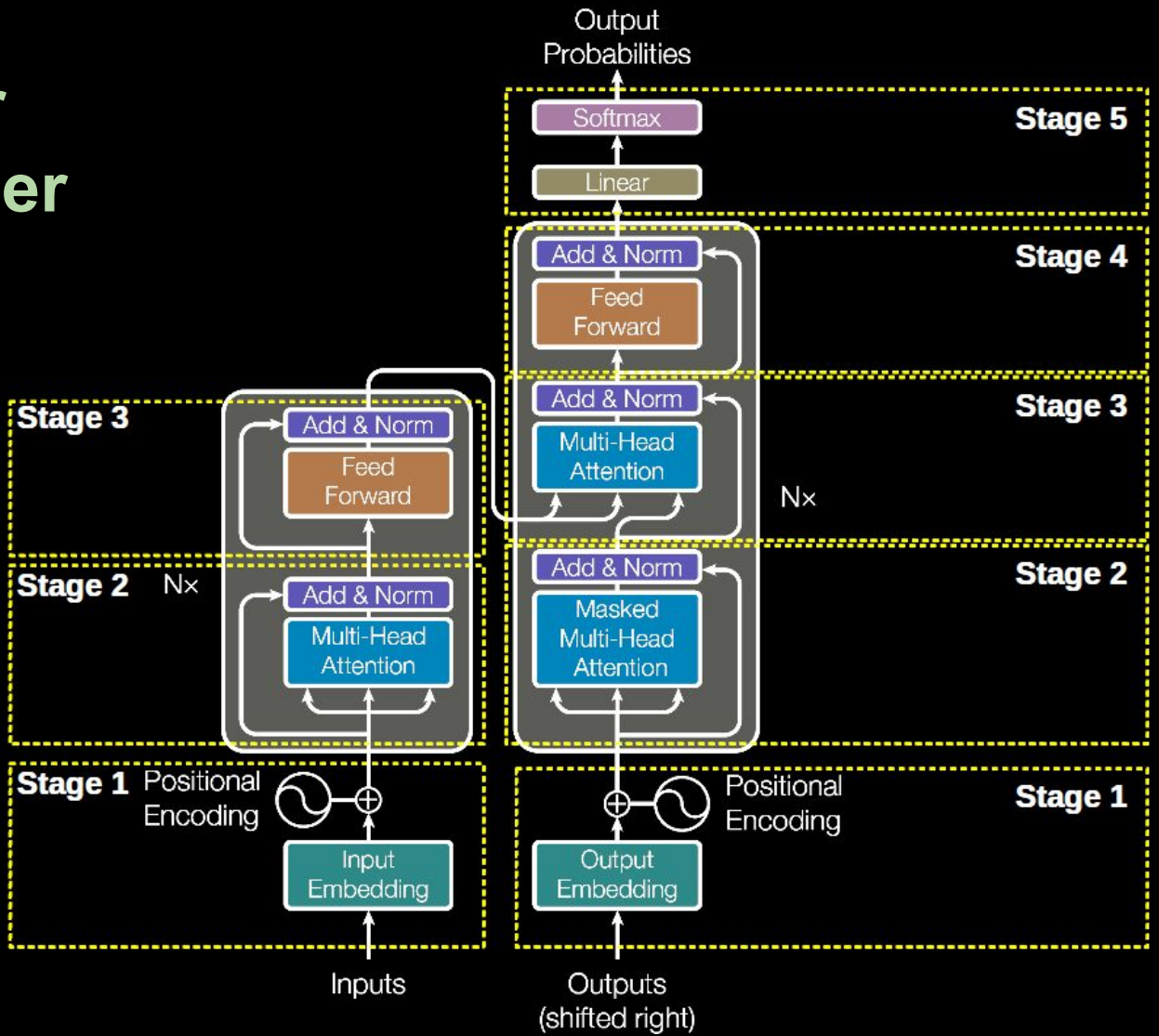
Simple Transformer Encoder-Decoder



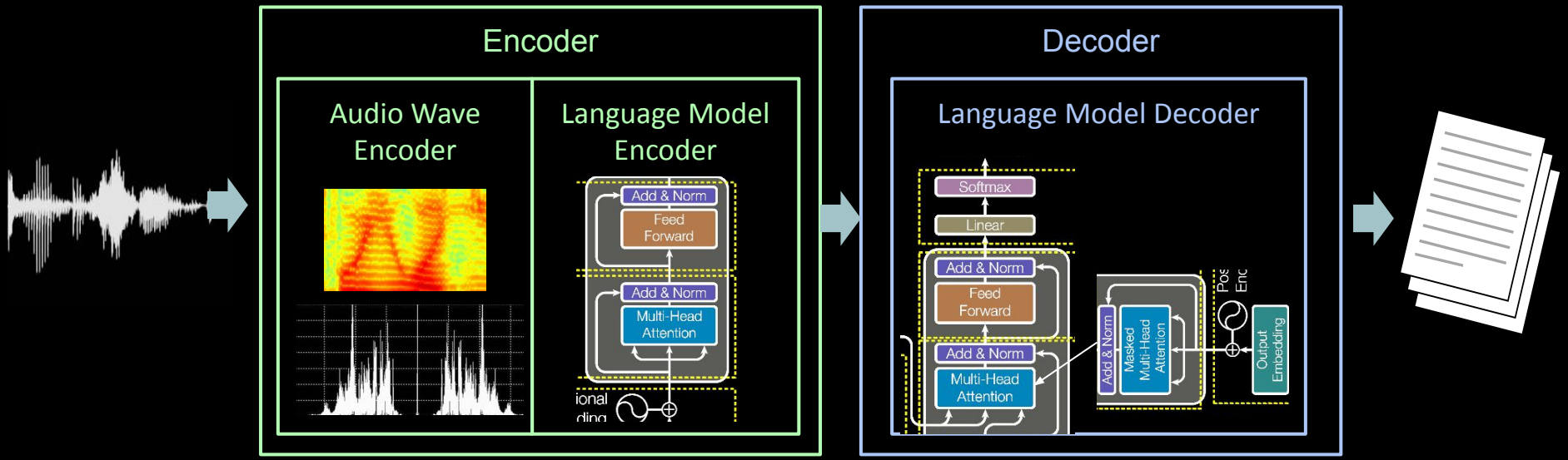
(Winata et al., 2020)

Fig. 1. Low-Rank Transformer Architecture.

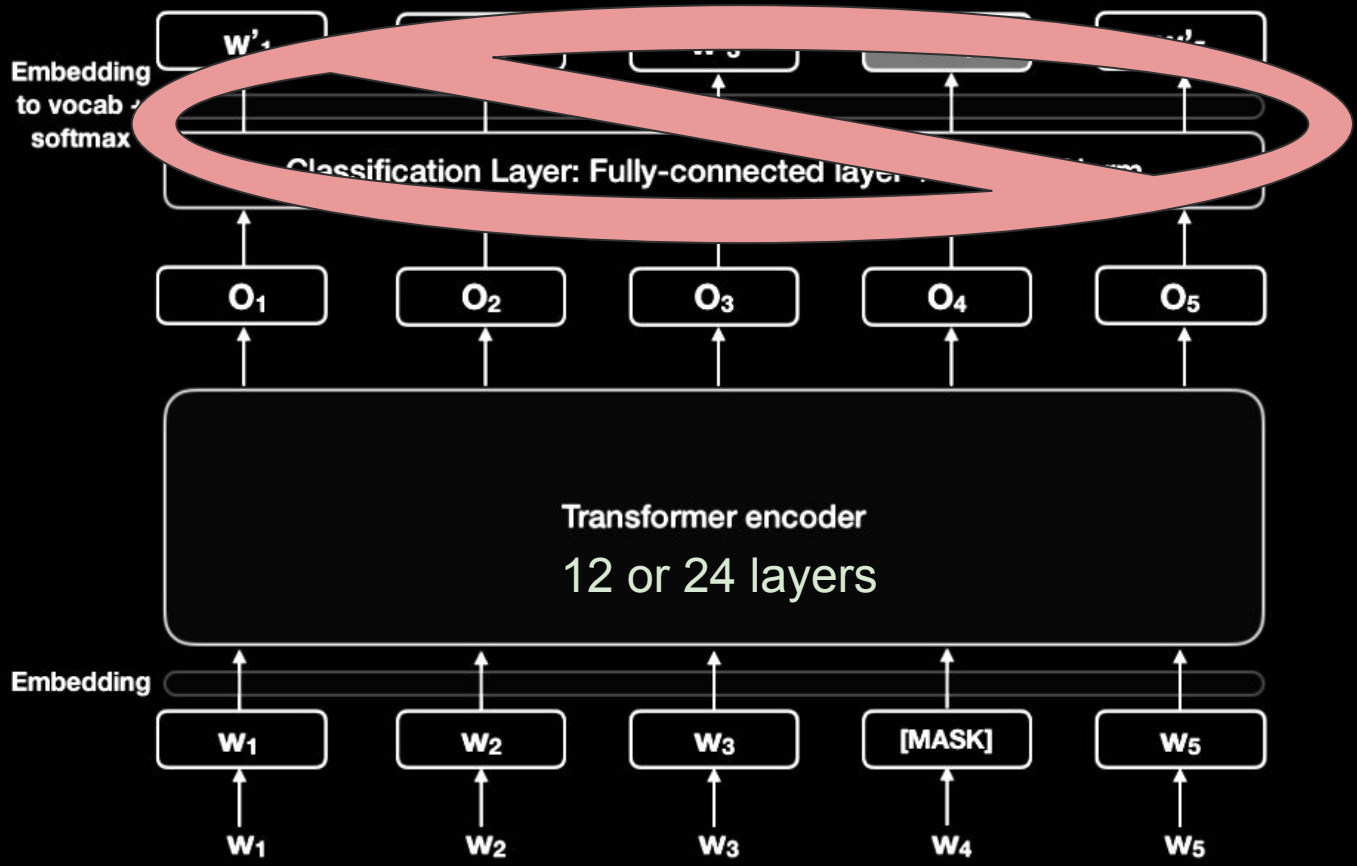
Transformer for Encoder-Decoder



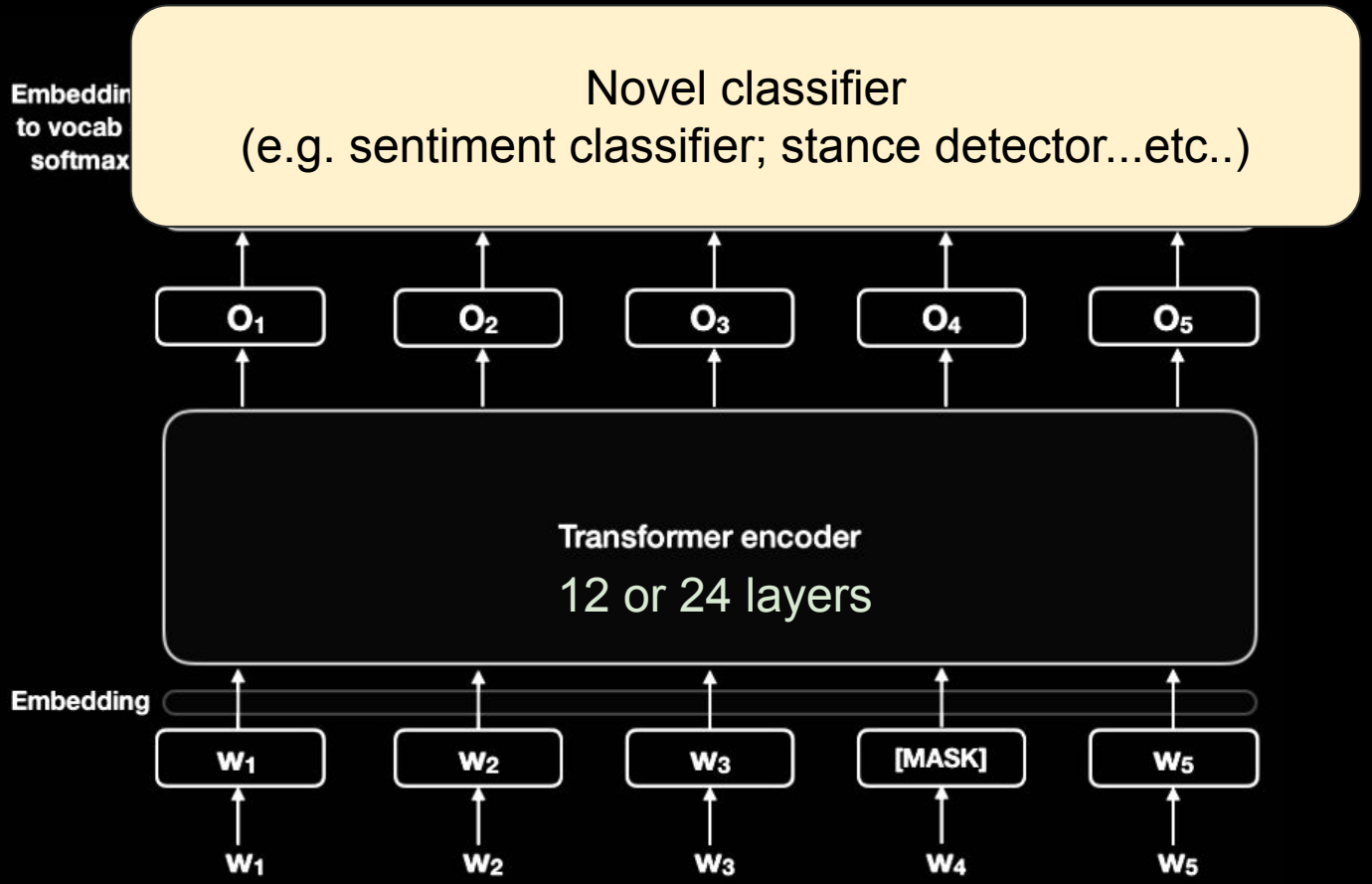
ASR: Automatic Speech Recognition



Pre-training; Fine-tuning



Pre-training; Fine-tuning



Current Architectures

Table 1. Performance in terms of average WER [%] on the **single-channel anechoic** wsj1-2mix corpus.

Model	dev	eval
RNN-based 1-channel Model [9]	24.90	20.43
Transformer-based 1-channel Model	**17.11**	**12.08**

Current Challenges for ASR

- Live simultaneous transcription
- Single-channel multi-speaker transcription ("Cocktail room problem")
- Multilingual transcription

Timeline: *Language Modeling* and *Vector Semantics*