Speech Processing

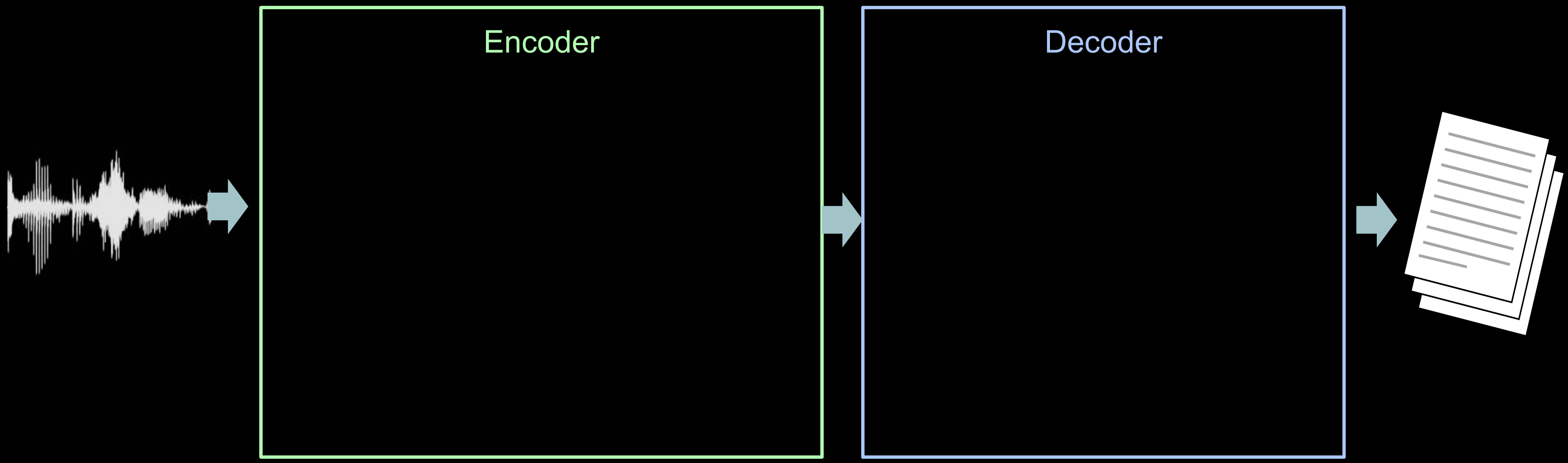
CSE538

Topics

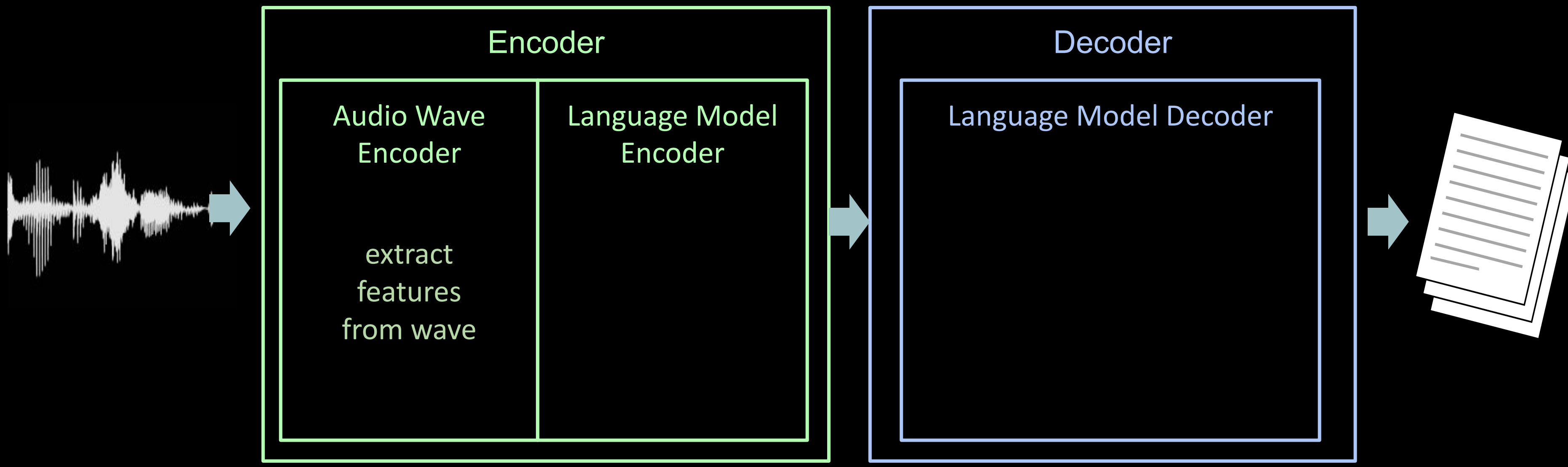
- Concept: Automatic Speech Recognition (ASR)
- Encoding Waves: Spectrograms
- Wave2Vec
- Whisper

ASR: Automatic Speech Recognition

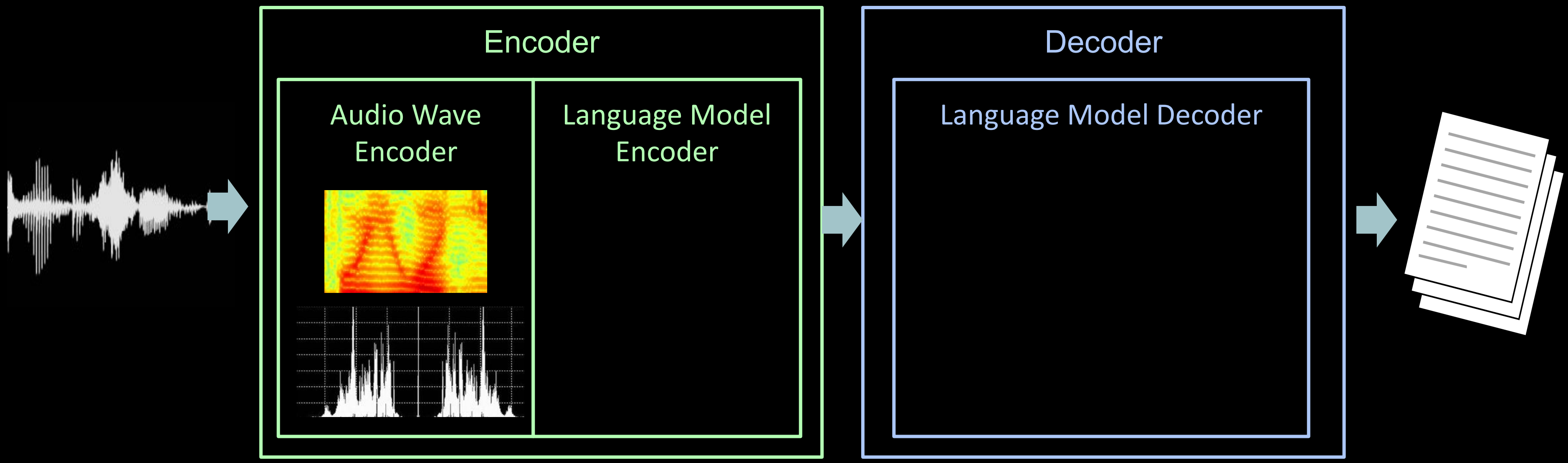
ASR: Automatic Speech Recognition



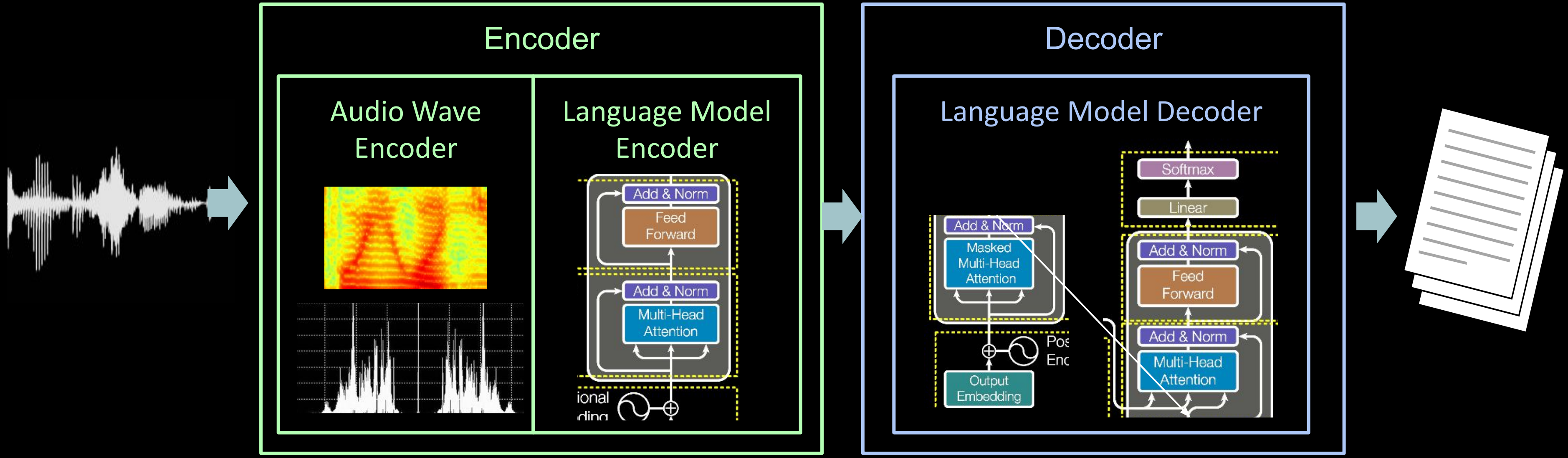
ASR: Automatic Speech Recognition



ASR: Automatic Speech Recognition



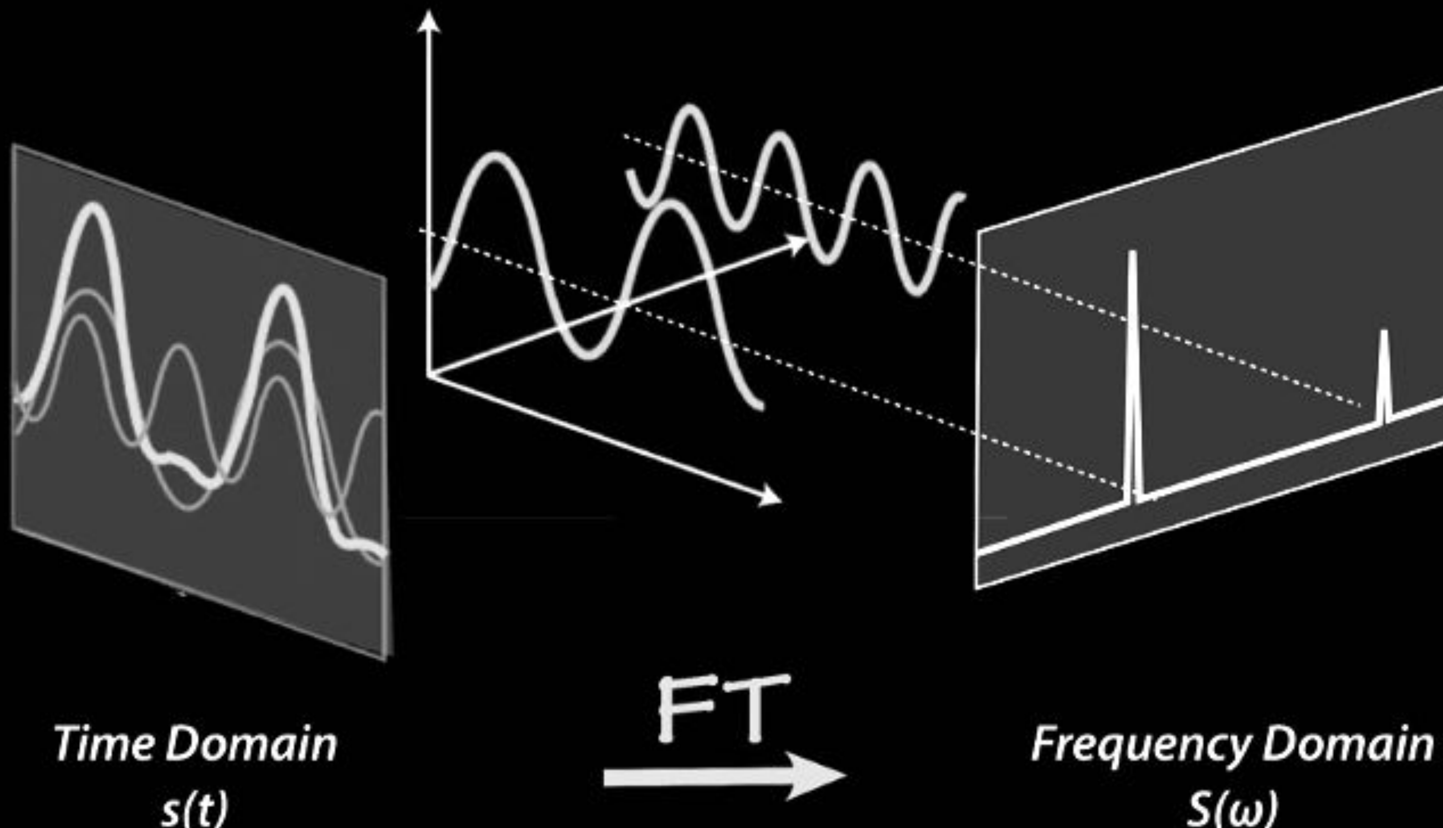
ASR: Automatic Speech Recognition



Topics

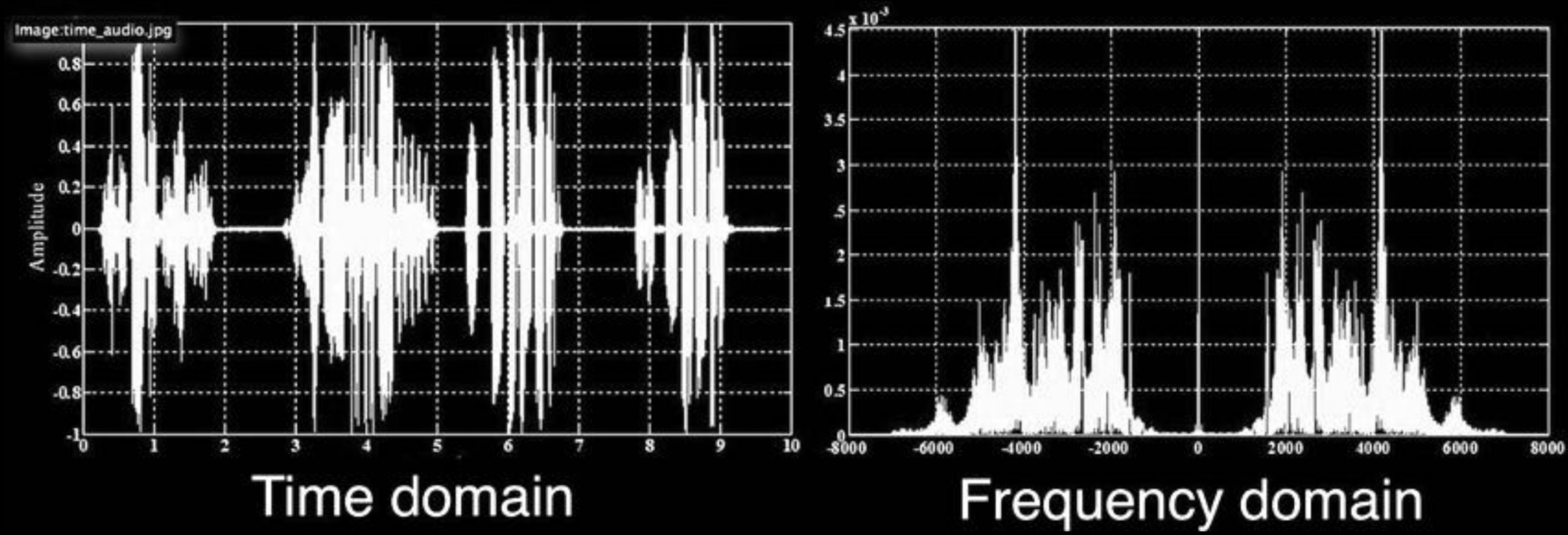
- Concept: Automatic Speech Recognition (ASR)
- **Encoding Waves: Spectrograms**
- Wave2Vec
- Whisper

Encoding Waves: Fourier Transform



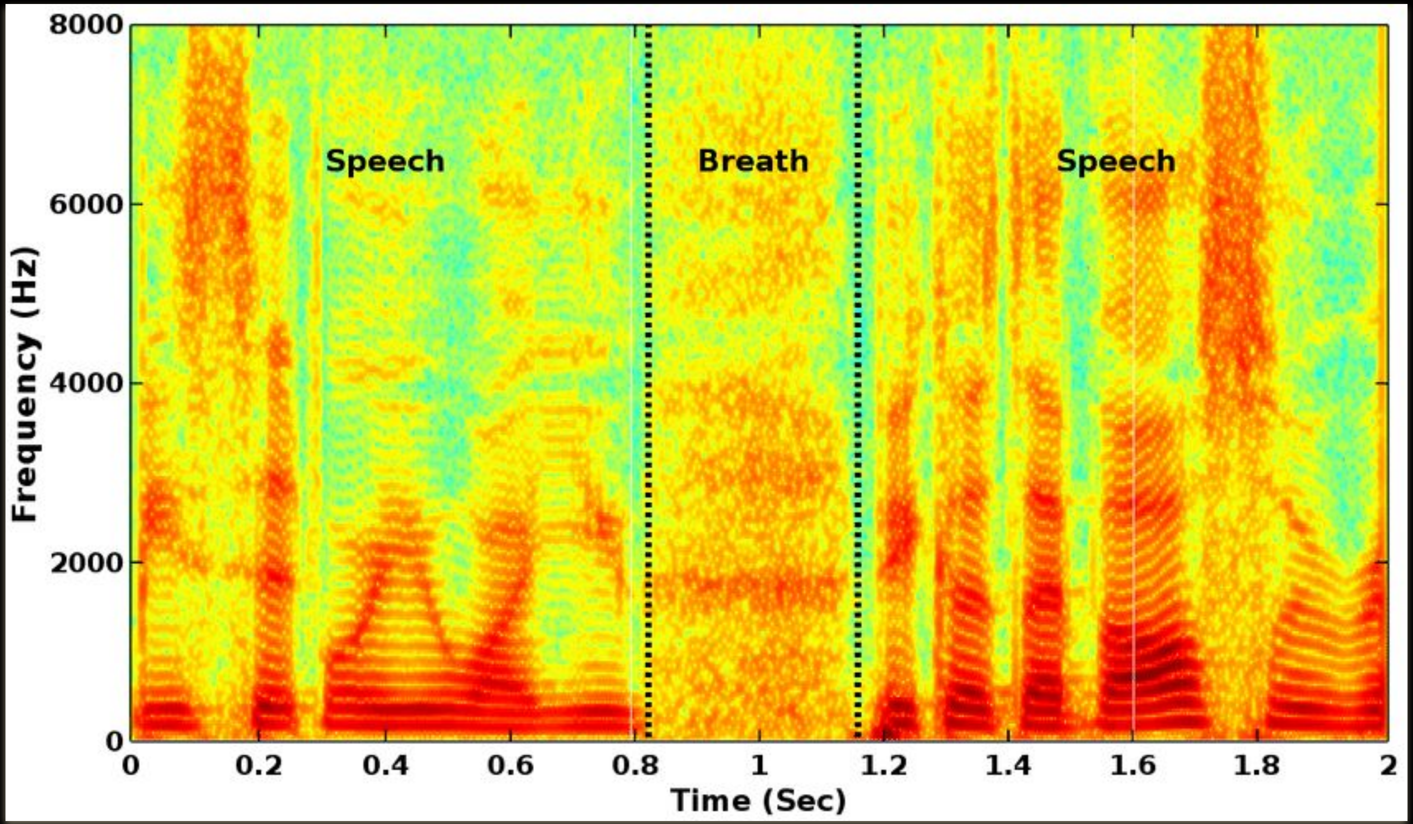
(Choudhary, 2020)

Encoding Waves: Fourier Transform



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

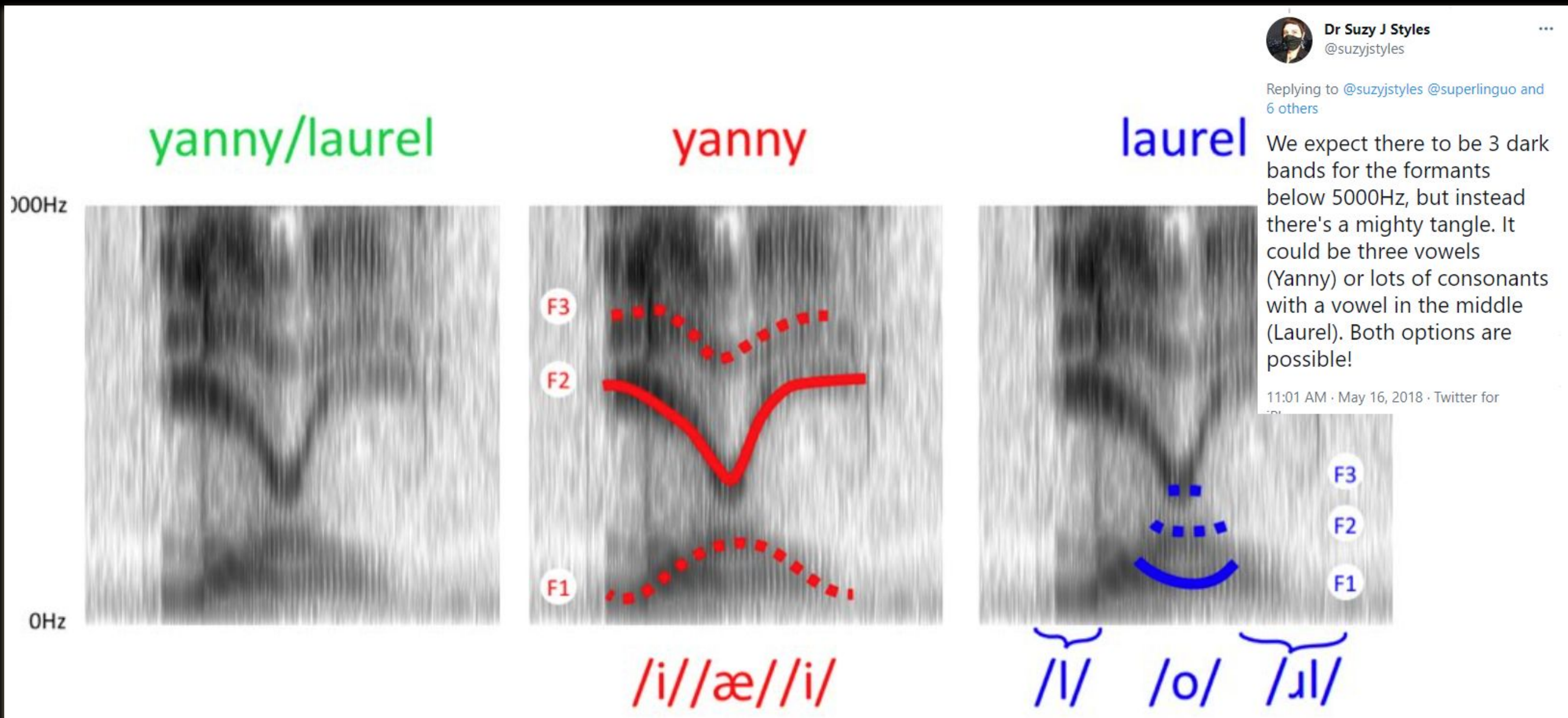
Spectrogram



(Dumpalla & Alluri, ICSC 2017)

Yanny Laurel

wiki/File:YannyLaurel.ogg



Spectrogram in Practice: The Mel Spectrum

Motivation: Hearing perception is logarithmic to frequency:

Less ability to distinguish 1 hertz change at higher frequencies

In music: Low A (A0) is 27.5*hrtz* versus A1 is 55hrtz;   A4 is 440 hrtz versus A5 is 880hrtz

$$mel(f) = 1127\ln\left(1 + \frac{f}{700}\right) \qquad (16.7)$$

(SLP3-16)

Spectrogram in Practice: The Mel Spectrum

Motivation: Hearing perception is logarithmic to frequency:

Less ability to distinguish 1 hertz change at higher frequencies

In music: Low A (A0) is 27.5*hrtz* versus A1 is 55hrtz;   A4 is 440 hrtz versus A5 is 880hrtz

$$mel(f) = 1127\ln(1 + \frac{f}{700}) \qquad (16.7)$$

(SLP3-16)

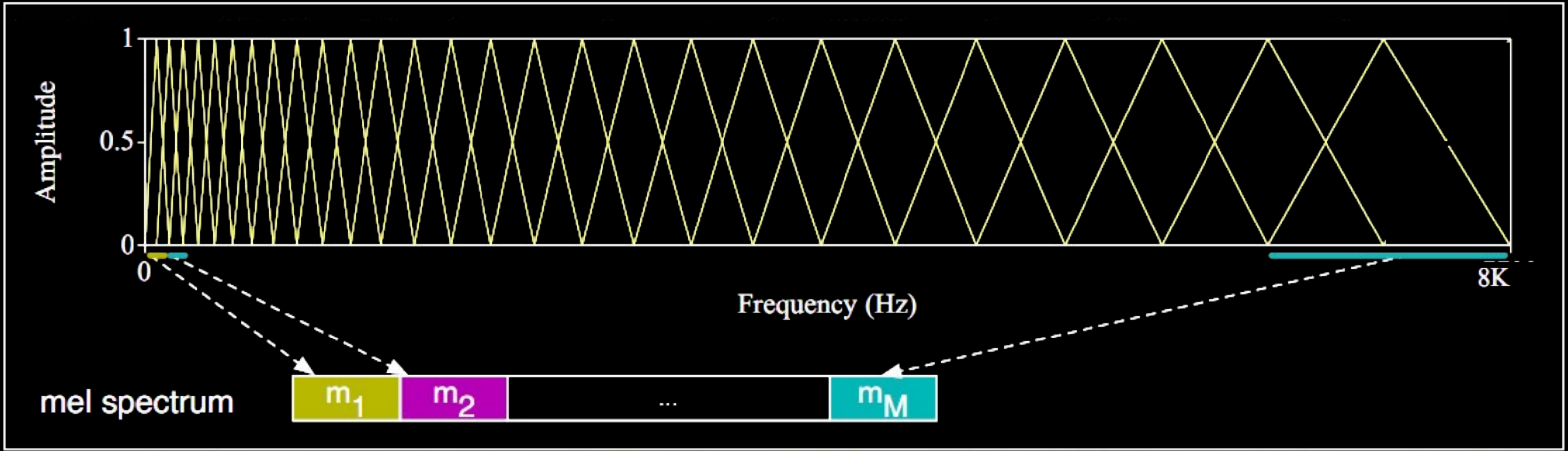


Figure 16.7 The mel filter bank (Davis and Mermelstein, 1980). Each triangular filter, spaced logarithmically along the mel scale, collects energy from a given frequency range.

Topics

- Concept: Automatic Speech Recognition (ASR)
- Encoding Waves: Spectrograms
- **Wave2Vec**
- Whisper

Wave2Vec Objective

Autoregressive future sample prediction

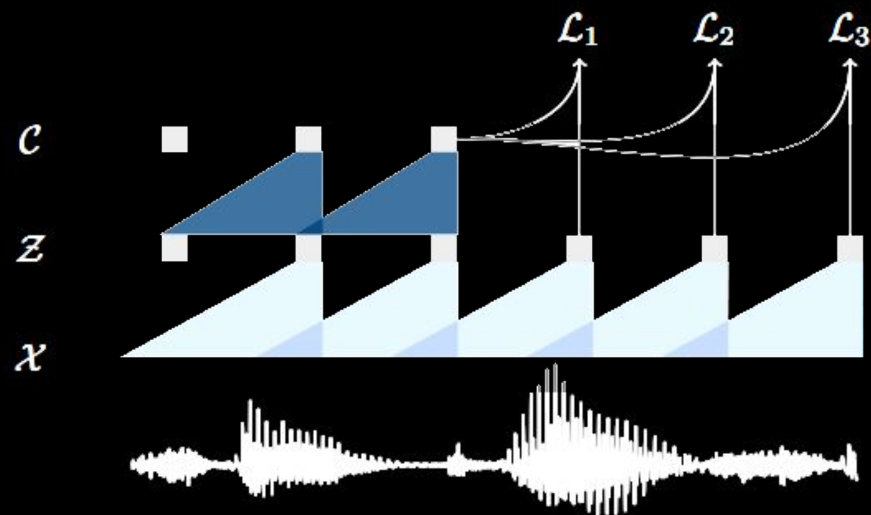


Figure 1: Illustration of pre-training from audio data $\mathcal{X}$ which is encoded with two convolutional neural networks that are stacked on top of each other. The model is optimized to solve a next time step prediction task.

(Steffen Schneider et al., 2019)

raw mel spectrogram

			nov93dev		nov92	
			LER	WER	LER	WER
Deep Speech 2 (12K h labeled speech; Amodei et al., 2016)			-	4.42	-	3.1
Trainable frontend (Zeghidour et al., 2018a)			-	6.8	-	3.5
Lattice-free MMI (Hadian et al., 2018)			-	5.66†	-	2.8†
Supervised transfer-learning (Ghahremani et al., 2017)			-	4.99†	-	2.53†
4-GRAM LM (Heafield et al., 2013)						
Baseline	–	–	3.32	8.57	2.19	5.64
wav2vec	Librispeech	80 h	3.71	9.11	2.17	5.55
wav2vec	Librispeech	960 h	2.85	7.40	1.76	4.57
wav2vec	Libri + WSJ	1,041 h	2.91	7.59	1.67	4.61
wav2vec large	Librispeech	960 h	2.73	6.96	1.57	4.32
WORD CONVLM (Zeghidour et al., 2018b)						
Baseline	–	–	2.57	6.27	1.51	3.60
wav2vec	Librispeech	960 h	2.22	5.39	1.25	2.87
wav2vec large	Librispeech	960 h	2.13	5.16	1.02	2.53
CHAR CONVLM (Likhomanenko et al., 2019)						
Baseline	–	–	2.77	6.67	1.53	3.46
wav2vec	Librispeech	960 h	2.14	5.31	1.15	2.78
wav2vec large	Librispeech	960 h	2.11	5.10	0.99	2.43

Table 1: Replacing log-mel filterbanks (Baseline) by pre-trained embeddings improves WSJ performance on test (nov92) and validation (nov93dev) in terms of both LER and WER. We evaluate pre-training on the acoustic data of part of clean and full Librispeech as well as the combination of all of them. † indicates results with phoneme-based models.

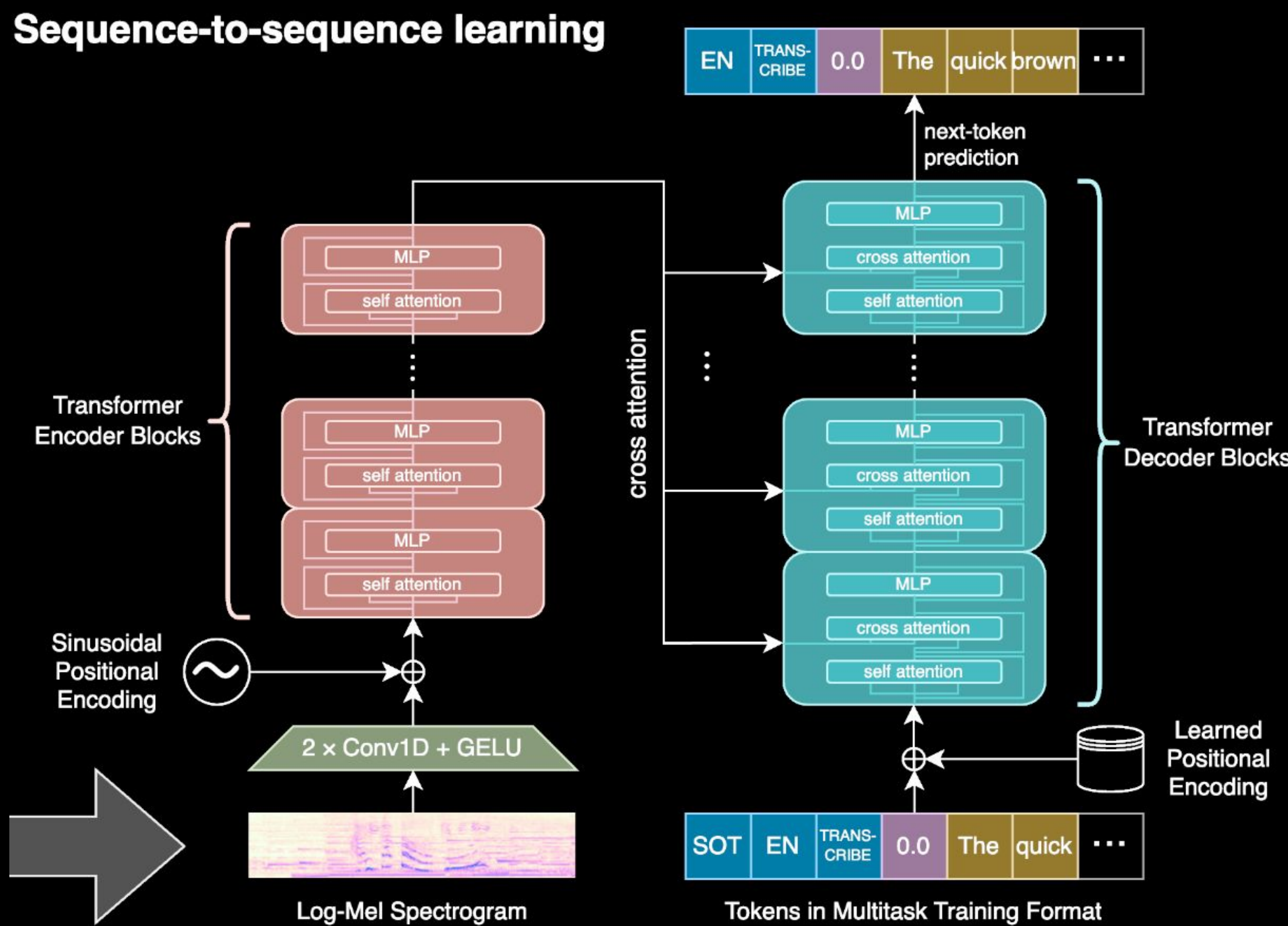
Wave2vec 2

https://ai.meta.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/

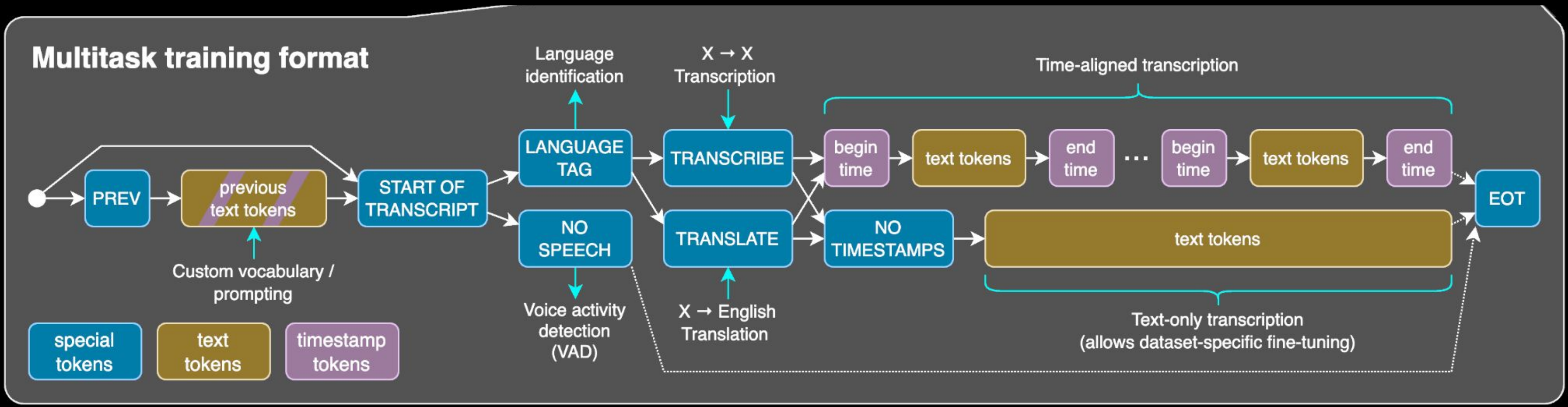
Topics

- Concept: Automatic Speech Recognition (ASR)
- Encoding Waves: Spectrograms
- Wave2Vec
- **Whisper**

Whisper


Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
...
next-token prediction
MLP
self attention
cross attention
cross attention
Transformer Encoder Blocks
Transformer Decoder Blocks
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Learned Positional Encoding
SOT
Log-Mel Spectrogram
Tokens in Multitask Training Format

Whisper


Multitask training format
Language identification
X → X Transcription
Time-aligned transcription
PREV
previous text tokens
START OF TRANSCRIPT
LANGUAGE TAG
NO SPEECH
TRANSCRIBE
TRANSLATE
begin time
text tokens
end time
...
begin time
text tokens
end time
NO TIMESTAMPS
text tokens
EOT
Custom vocabulary / prompting
Voice activity detection (VAD)
X → English Translation
Text-only transcription (allows dataset-specific fine-tuning)
special tokens
text tokens
timestamp tokens

Whisper

Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

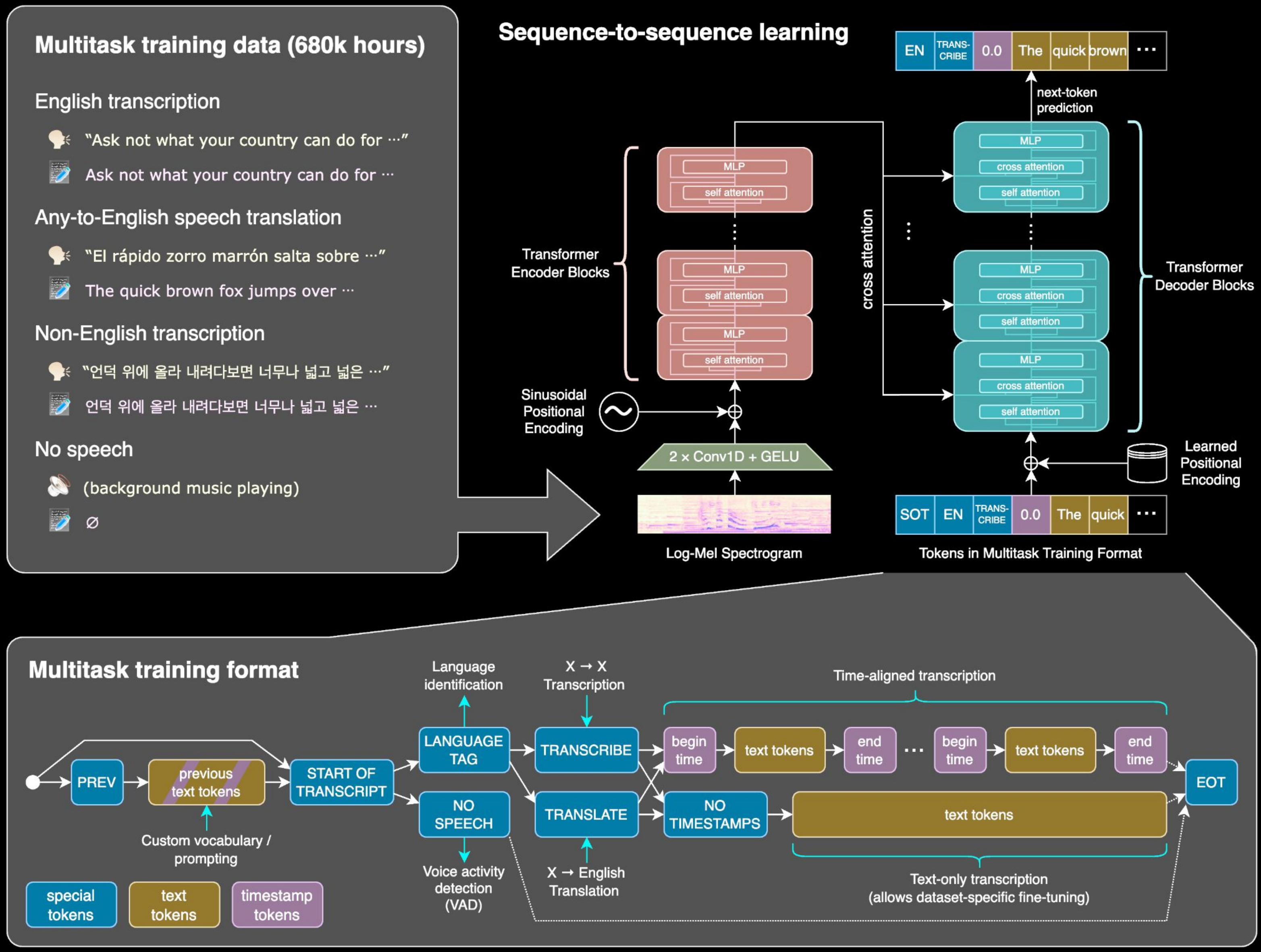
Table 1. Architecture details of the Whisper model family.

Dataset	wav2vec 2.0 Large (no LM)	Whisper Large V2	RER (%)
LibriSpeech Clean	**2.7**	**2.7**	0.0
Artie	24.5	**6.2**	74.7
Common Voice	29.9	**9.0**	69.9
Fleurs En	14.6	**4.4**	69.9
Tedlium	10.5	**4.0**	61.9
CHiME6	65.8	**25.5**	61.2
VoxPopuli En	17.9	**7.3**	59.2
CORAAL	35.6	**16.2**	54.5
AMI IHM	37.0	**16.9**	54.3
Switchboard	28.3	**13.8**	51.2
CallHome	34.8	**17.6**	49.4
WSJ	7.7	**3.9**	49.4
AMI SDM1	67.6	**36.4**	46.2
LibriSpeech Other	6.2	**5.2**	16.1
Average	29.3	**12.8**	55.2

Table 2. **Detailed comparison of effective robustness across various datasets.** Although both models perform within 0.1% of each other on LibriSpeech, a zero-shot Whisper model performs much better on other datasets than expected for its LibriSpeech performance and makes 55.2% less errors on average. Results reported in word error rate (WER) for both models after applying our text normalizer.

Current Challenges for ASR

- Live simultaneous transcription
- Single-channel multi-speaker transcription ("Cocktail room problem")
    - diarization: task of identifying speakers
- Multilingual transcription


Multitask training data (680k hours)
English transcription
"Ask not what your country can do for ⋯"
Ask not what your country can do for ⋯
Any-to-English speech translation
"El rápido zorro marrón salta sobre ⋯"
The quick brown fox jumps over ⋯
Non-English transcription
"언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯"
언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯
No speech
(background music playing)
∅
Sequence-to-sequence learning
EN TRANS-CRIBE 0.0 The quick brown ⋯
next-token prediction
MLP
cross attention
self attention
Transformer Encoder Blocks
cross attention
Transformer Decoder Blocks
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Learned Positional Encoding
Log-Mel Spectrogram
SOT EN TRANS-CRIBE 0.0 The quick ⋯
Tokens in Multitask Training Format
Multitask training format
Language identification
X → X Transcription
Time-aligned transcription
PREV
previous text tokens
START OF TRANSCRIPT
LANGUAGE TAG
NO SPEECH
TRANSCRIBE
TRANSLATE
begin time
text tokens
end time
NO TIMESTAMPS
EOT
Custom vocabulary / prompting
Voice activity detection (VAD)
X → English Translation
Text-only transcription (allows dataset-specific fine-tuning)
special tokens
text tokens
timestamp tokens