HW2 - SSL

Team 5
B09901116 陳守仁、B09602017 白宗民、B08901207 呂俐君

Track 1

1. Run an **“upstream and downstream” pair** of your choice **(1 point)**
2. Read the description of the downstream task and prepare data (link below).
3. Run the experiment and report your results / observations.
4. An unique *“upstream and downstream” pair* counts for 1 point, you can do different pairs to get more points.
5. Setting different hyperparameters for an existing pair also counts for 1 point.

Upstream Models and Downstream Tasks

- Downstream Tasks
    - Keyword Spotting
    - Phoneme Recognition

- Upstream Models
    - Fbank
    - Wav2vec 2.0 - wav2vec 2.0 base - Wav2Vec2-Base-960h
        - 94.4M parameters
        - 960 hours of Librispeech on 16kHz sampled speech audio.
    - Hubert - hubert_base - hubert_base_ls960
        - 95M parameters
        - 960 hours of LibriSpeech audio with a batch size of at most 87.5 seconds of audio per GPU.
    - WavLM - WavLM Base+
        - 94.7M parameters
        - 60,000 hours of Libri-Light/10,000 hours of GigaSpeech/24,000 hours of VoxPopuli

Models Introduction

Upstream Model 1 - Fbank

Acoustic Feature Upstreams

We also provide classic acoustic features as baselines. For each upstream with `Name`, you can configure their options (available by their `Backend`) in `s3prl/upstream/baseline/Name.yaml`.

Feature	Name	Default Dim	Stride	Window	Backend
Spectrogram	spectrogram	257	10ms	25ms	torchaudio-kaldi
FBANK	fbank	80 + delta1 + delta2	10ms	25ms	torchaudio-kaldi
MFCC	mfcc	13 + delta1 + delta2	10ms	25ms	torchaudio-kaldi
Mel	mel	80	10ms	25ms	torchaudio
Linear	linear	201	10ms	25ms	torchaudio

Not pre-trained

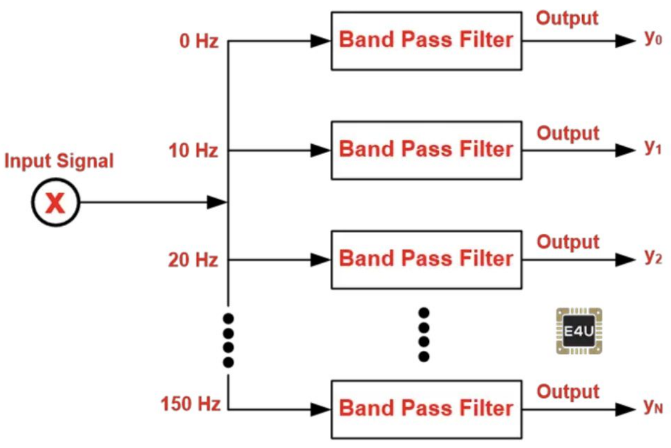
torchaudio.compliance.kaldi.fbank

Fbank Explained

TORCHAUDIO.COMPLIANCE.KALDI.FBANK

```
torchaudio.compliance.kaldi.fbank(
    waveform: Tensor,
    blackman_coeff: float = 0.42,
    channel: int = -1,
    dither: float = 0.0,
    energy_floor: float = 1.0,
    frame_length: float = 25.0,
    frame_shift: float = 10.0,
    high_freq: float = 0.0,
    htk_compat: bool = False,
    low_freq: float = 20.0,
    min_duration: float = 0.0,
    num_mel_bins: int = 23,
    preemphasis_coefficient: float = 0.97,
    raw_energy: bool = True,
    remove_dc_offset: bool = True,
    round_to_power_of_two: bool = True,
    sample_frequency: float = 16000.0,
    snip_edges: bool = True,
    subtract_mean: bool = False,
    use_energy: bool = False,
    use_log_fbank: bool = True,
    use_power: bool = True,
    vtln_high: float = -500.0,
    vtln_low: float = 100.0,
    vtln_warp: float = 1.0,
    window_type: str = 'povey'
) → Tensor    [SOURCE]
```

- Fbank (filter bank)
  - feature extraction technique that is commonly used in audio processing



Structure of a Filter Bank

- Output: Features(tensor)

Upstream Model 1 - Fbank on Different Tasks

- **Keyword Spotting**
  - test accuracy : 0.08114
  - Task type: Classification
  - Steps: 50,000

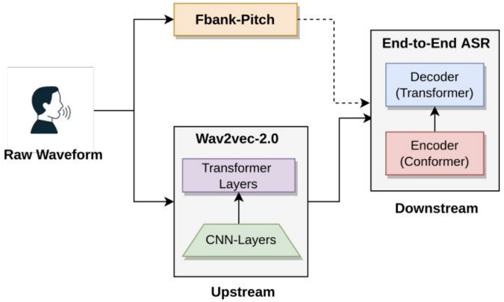
- **Phoneme Recognition**
  - test accuracy : 0.1332
  - Task type: Transcription
  - Steps: 50,000

Upstream Model 1 - Fbank Baseline Model

```
kaldi:
  feat_type: fbank
  fbank:
    num_mel_bins: 80
    frame_length: 25.0
    frame_shift: 10.0
    use_log_fbank: True

delta:
  order: 2
  win_length: 5

cmvn:
  use_cmvn: True
```

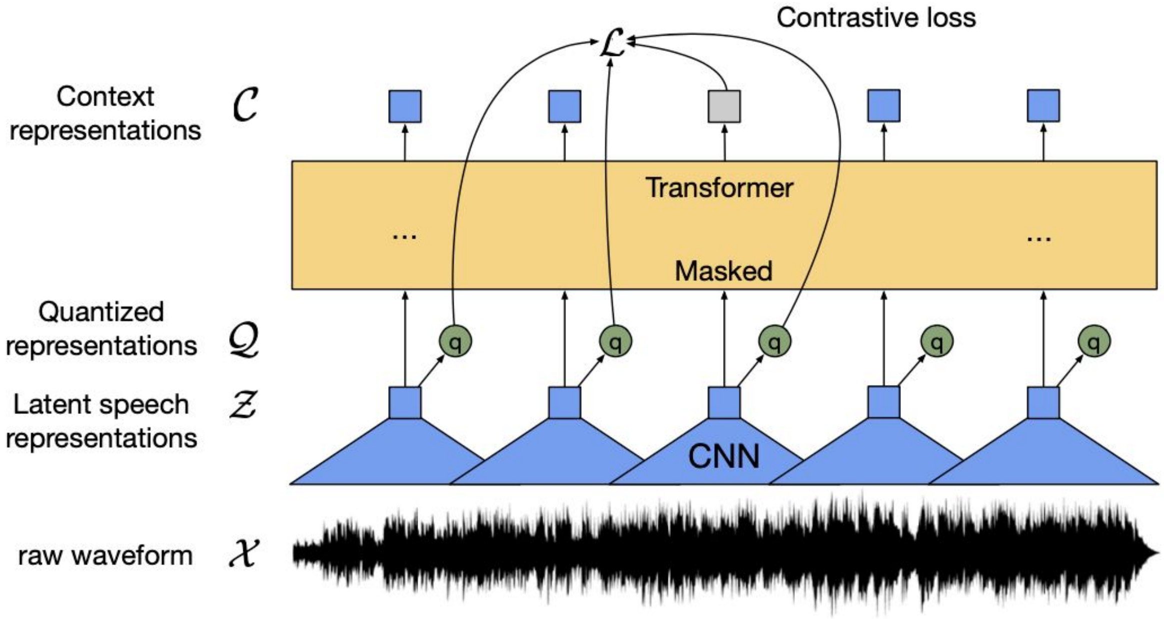


Fbank features are simply a representation of the spectral characteristics of the speech audio.

- Other models: Contain representations that is useful for understanding the **meaning of words/sounds**
- Fbank: They **do not contain** any information about the meaning of the words.

<u>It is not pre-trained</u>

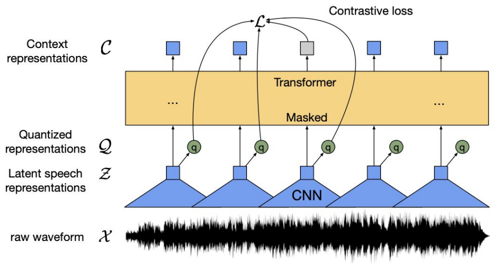
Upstream Model 2 - Wav2vec2



- **Convolutional layers:** process the raw waveform input to get latent representation - Z

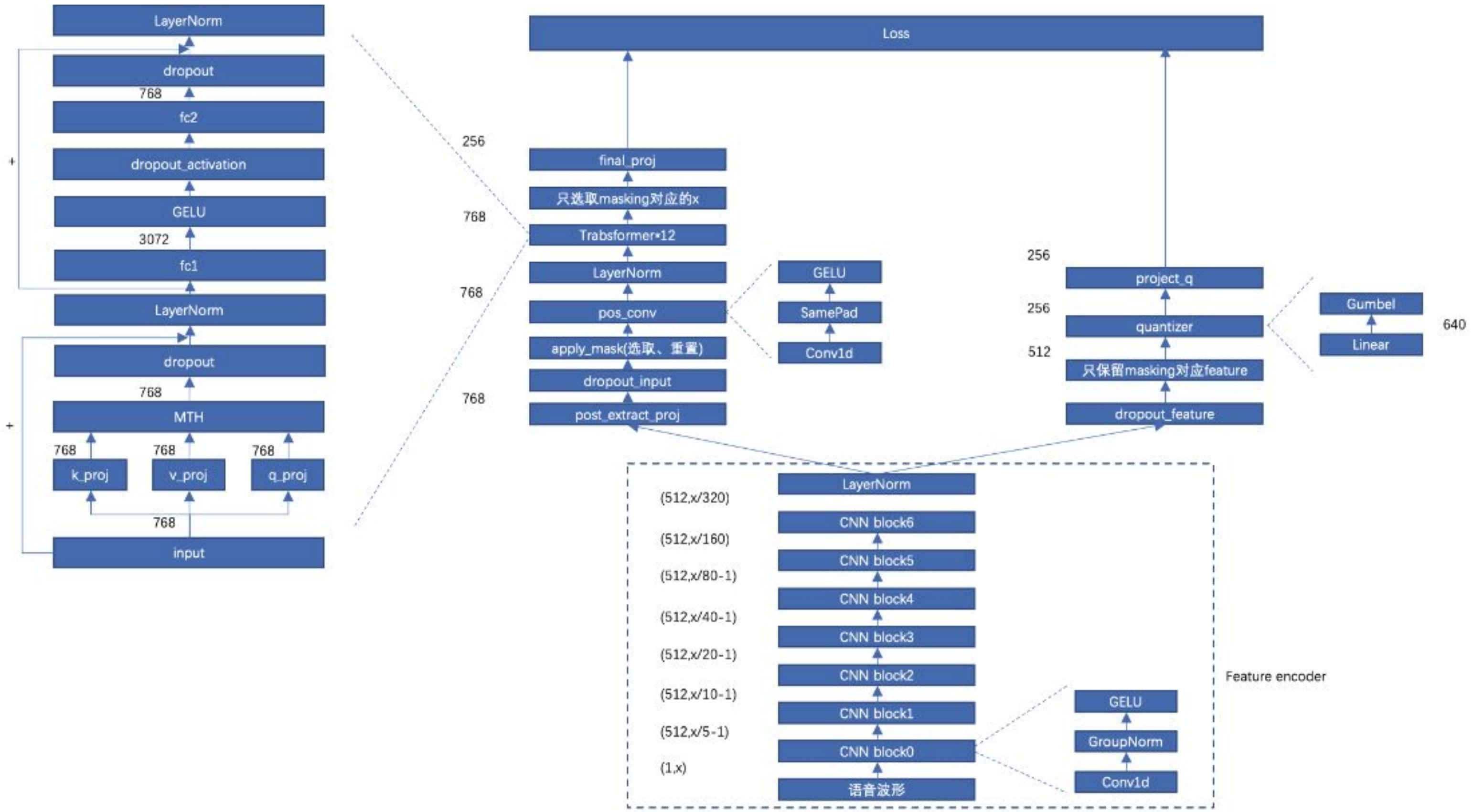
- **Transformer layers:** creating contextualised representation - C

Upstream Model 2 - Wav2vec2



- Encodes speech audio via a multi-layer CNN → preprocess raw waveform
- Masks spans of the resulting latent speech representations
- The latent representations are fed to a Transformer network to build contextualized representations → enhance the speech representation with context (since it's self-attention model)
- Trained via a contrastive task where the true latent is to be distinguished from distractors
- Learn discrete speech units via a gumbel softmax to represent the latent representations in the contrastive task
- Fine-tuned on labeled data with a Connectionist Temporal Classification (CTC) loss

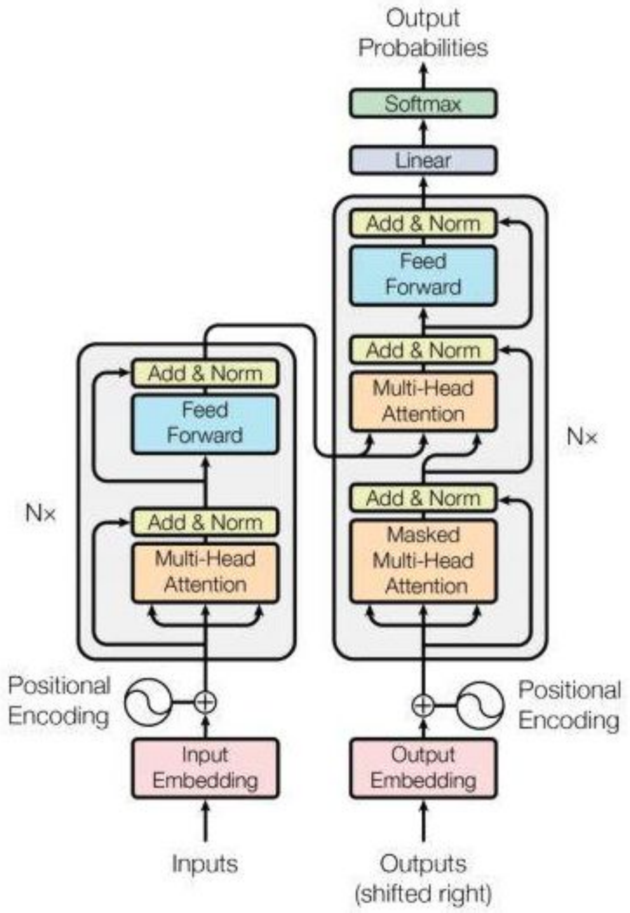
Upstream Model 2 - Wav2vec2



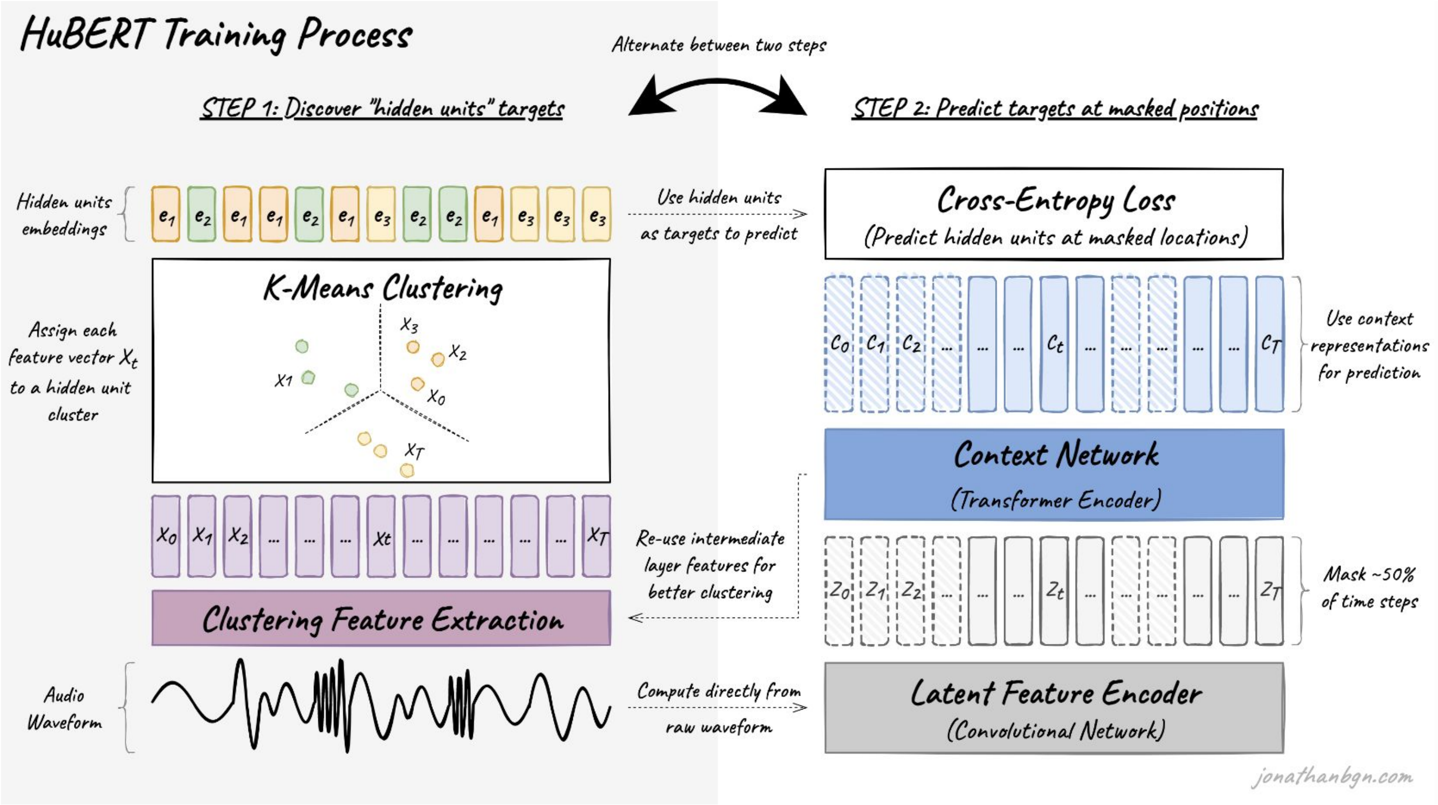
Upstream Model 2 - Wav2vec2

- **Keyword Spotting**
  - test accuracy : 0.96332
  - Task type: Classification
  - Advantage: Transformer's capacity to gauge the dependability of contextual information

- **Phoneme Recognition**
  - test accuracy : 0.93657
  - Task type: Transcription
  - Advantage: Transformer's ability to discern the trustworthiness of long-term sequences



Upstream Model 3 - HuBERT



Upstream Model 3 - HuBERT

1. **Clustering step** - create pseudo-targets
   a. extract the **h**idden **u**nits from audio → *K* clusters
   b. hidden unit –mapped→ *embedding vector* (for predictions)
   c. Clustering features decided by "Mel-Frequency Cepstral Coefficients (MFCCs)"
   d. Combine multiple clustering with a different number of clusters (optional)

2. **Prediction step:** guess these targets at masked positions.
   a. mask 50% of transformer encoder(BERT) input features
   b. predict the targets
   c. calculate the cosine similarity between
      [transformer outputs (projected to a lower dimension)] &
      [each hidden unit embedding from all possible hidden units] → give prediction logits.

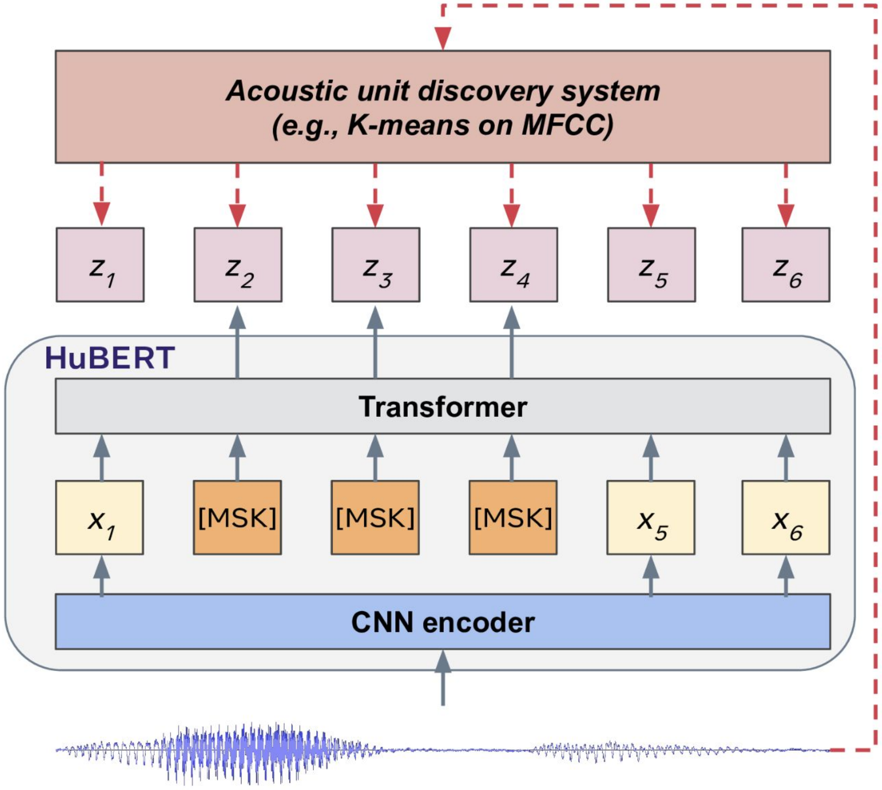
Upstream Model 3 - HuBERT

	HuBERT	Wav2vec2	Advantages
LOSS	cross-entropy loss	contrastive loss + diversity loss	easier and more stable training
builds targets & Clustering	seperately	simultaneously	Simplier
intermediate layers	re-uses embeddings from intermediate layers of encoder	only uses the CNN output for quantization.	better targets quality

Experiment 3 - HuBERT on different tasks

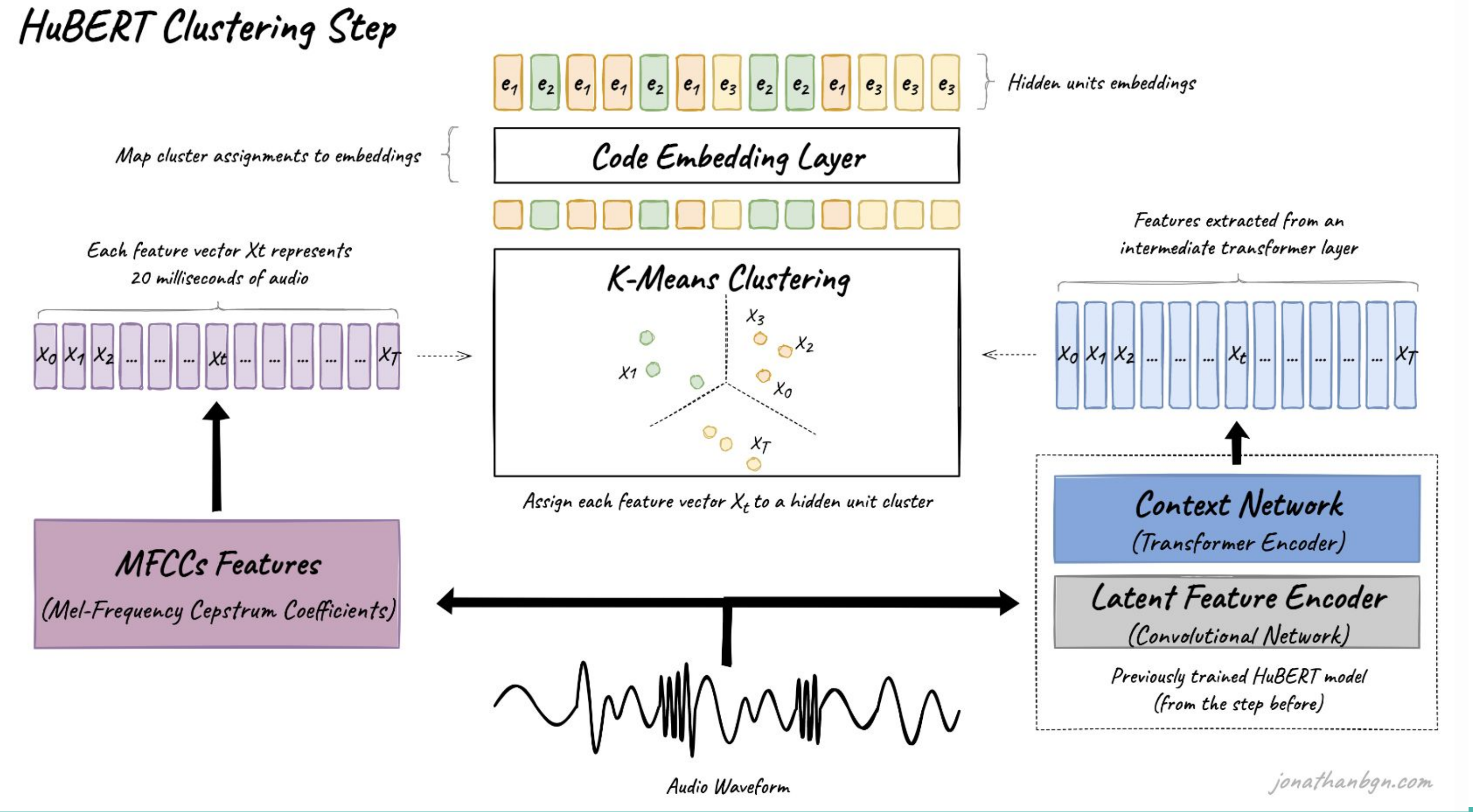
- Keyword Spotting
  - test accuracy : 0.96527
  - Task type: Classification
  - Steps: 72,000

- Phoneme Recognition
  - test accuracy : 0.9401
  - Task type: Transcription
  - Steps: 51,700

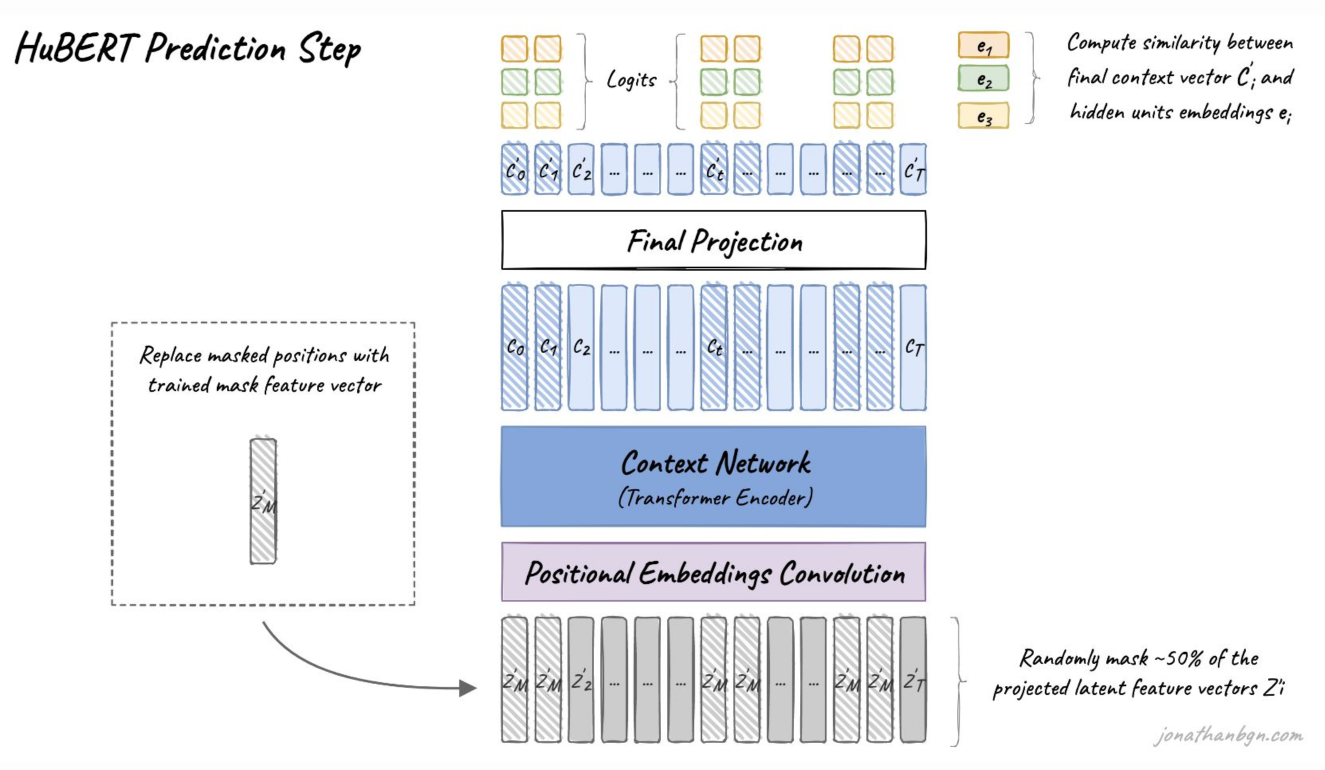


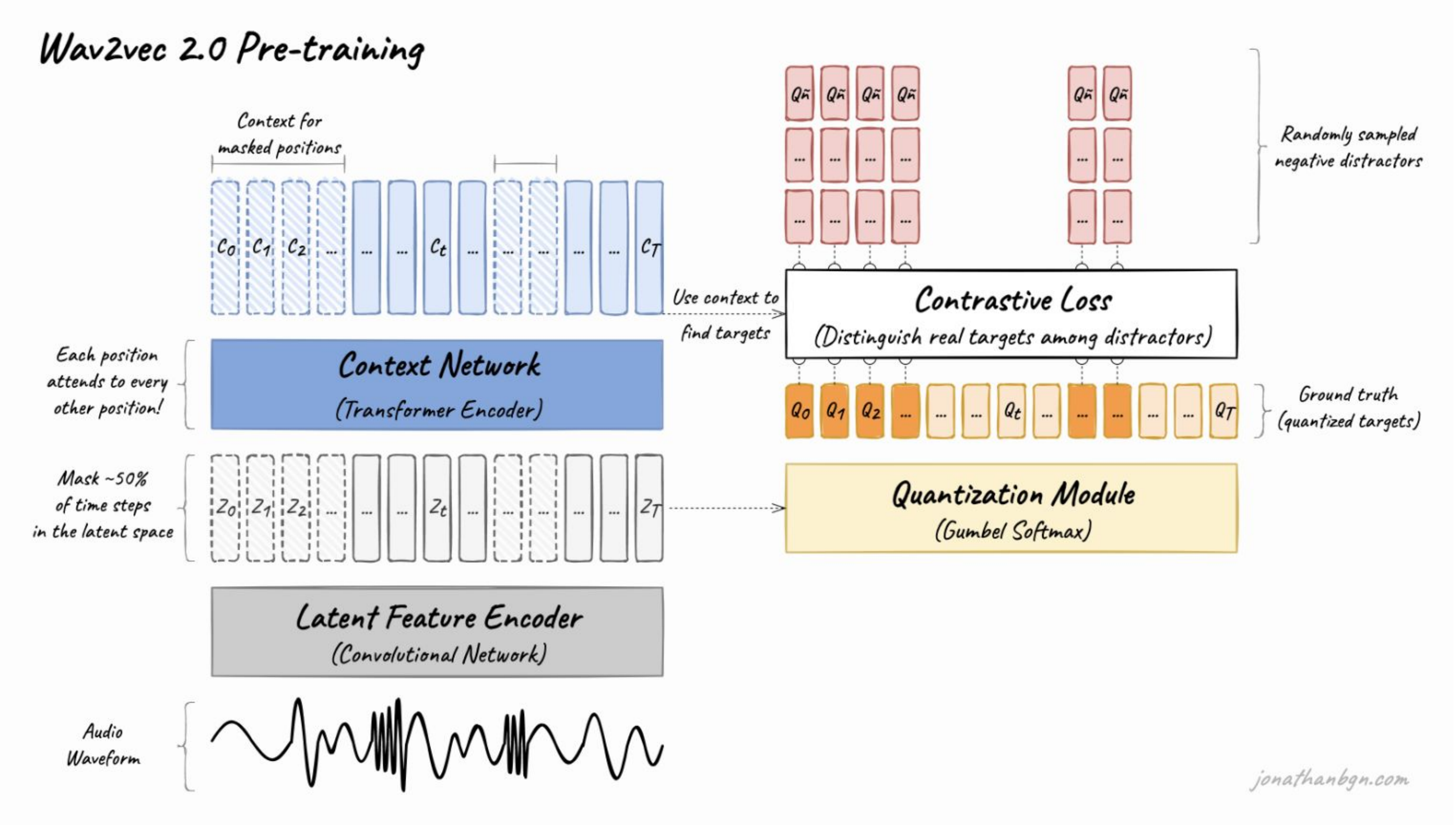
Experiment 3 - Hubert on different tasks

- Hu



Experiment 3 - Hubert on different tasks





Wav2vec 2.0 Pre-training
Context for masked positions
c_0 c_1 c_2 ... ... c_t ... ... c_T
Use context to find targets
Contrastive Loss
(Distinguish real targets among distractors)
Randomly sampled negative distractors
$Q_{\tilde{n}}$
Each position attends to every other position!
Context Network
(Transformer Encoder)
Ground truth (quantized targets)
Q_0 Q_1 Q_2 ... ... Q_t ... ... Q_T
Mask ~50% of time steps in the latent space
z_0 z_1 z_2 ... ... z_t ... ... z_T
Quantization Module
(Gumbel Softmax)
Latent Feature Encoder
(Convolutional Network)
Audio Waveform
jonathanbgn.com

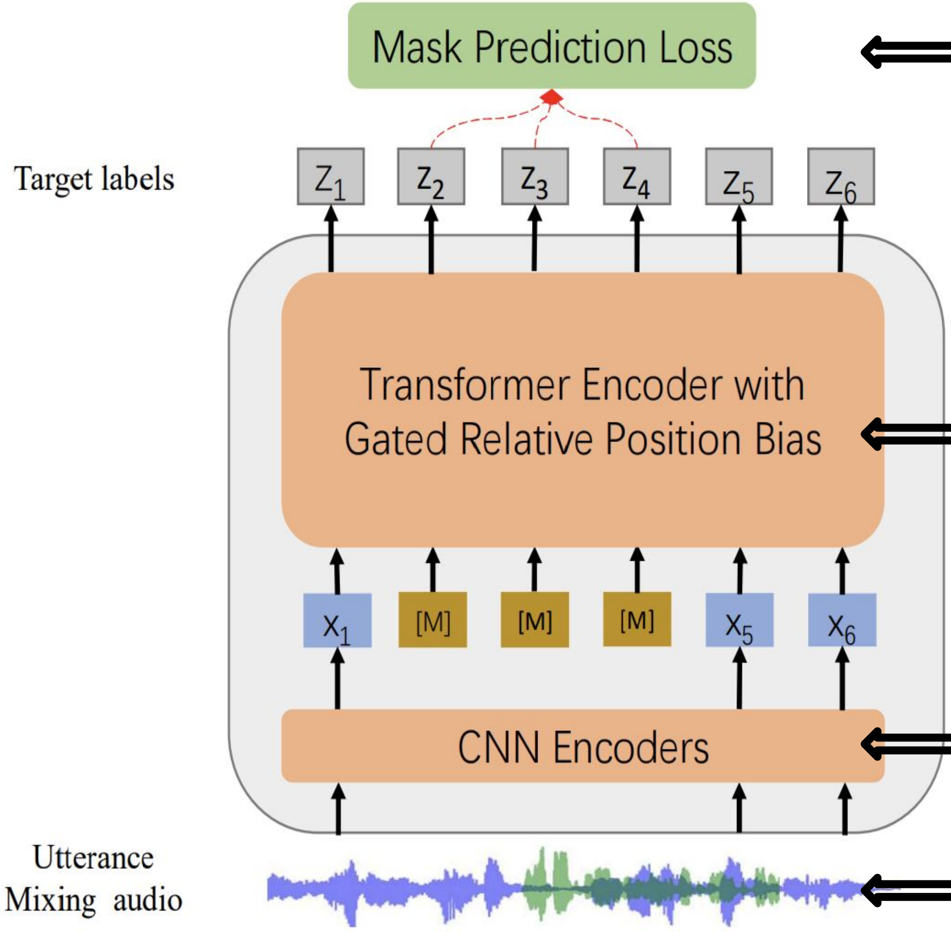
Upstream Model 4 - WavLM Structure

https://arxiv.org/pdf/2110.13900.pdf

Drawbacks of Existing Models

1. **Limited Multi-Speaker Performance:**
   a. Struggle to handle tasks involving multiple speakers, like identifying who's speaking or separating voices in mixed audio. They don't do a great job at this because they <u>weren't designed to tell speakers apart effectively</u> during their initial training.

2. **Over-Reliance on Audiobooks:**
   a. Depend heavily on using a huge amount of audio data, mainly from audiobooks. However, audiobooks are quite different from real-life scenarios, and using them exclusively makes the models less effective when dealing with real-world audio tasks.



Pseudo-labels generated using k-means clustering by sending in clean speech utterance at the last layer. ← HuBERT

GREP - encoded based on the offset between "key" and "query" in the Transformers self-attention mechanism

- Better understand the relative positional relationship between sounds
- Sequential ordering

1. 7 blocks of **temporal convolution**
   a. 512 channels of strides (5,2,2,2,2,2,2)
   b. kernel widths (10,3,3,3,3,2,2)
2. **Normalization** Layer
3. **GELU activation** Layer

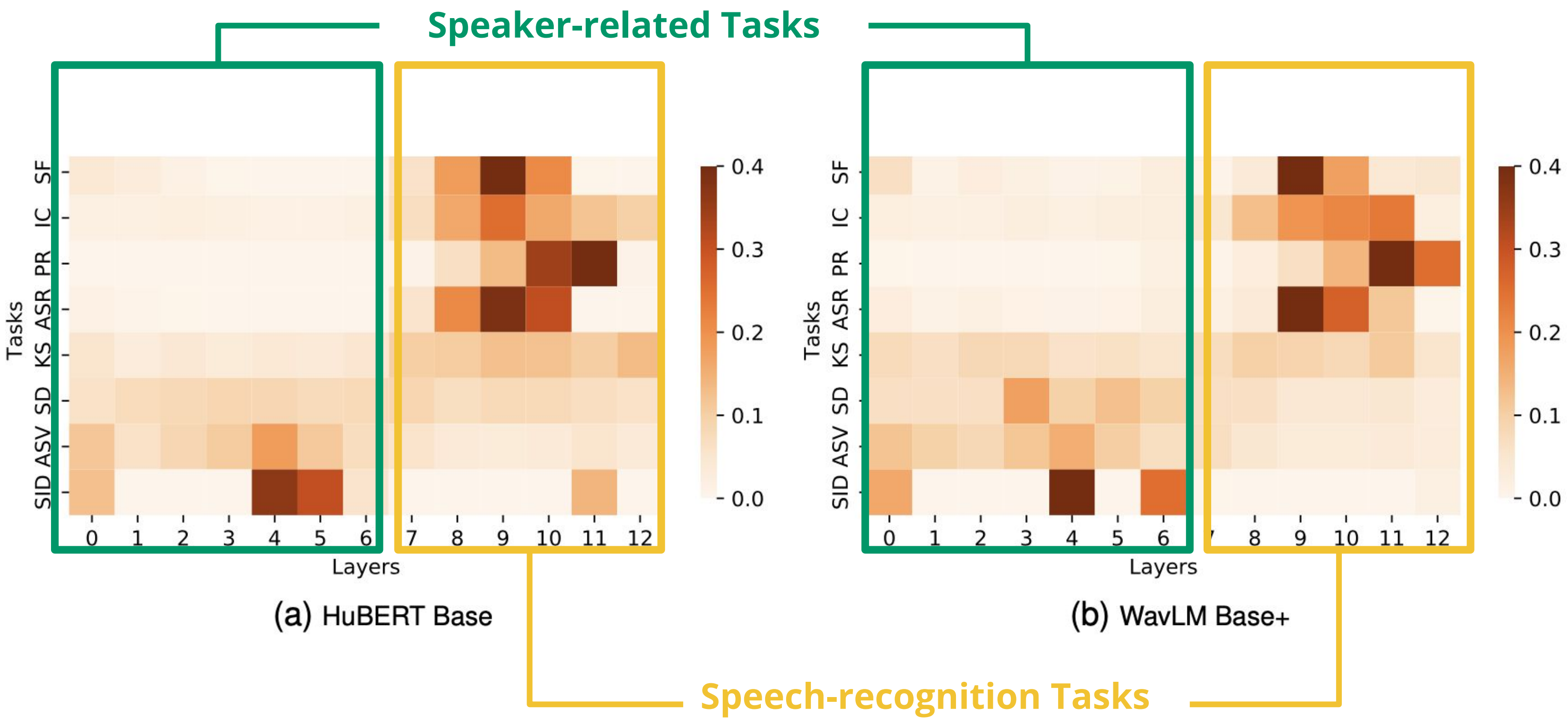
Simulates noisy/overlapped speech by taking clean speech utterances and **mixing them** with random noise or secondary utterances.

Upstream Model 4 - WavLM Structure

New in pre-training

- **masked speech denoising**
    - some inputs are simulated noisy/overlapped speech with masks
- **prediction framework**
    - target is to predict the pseudo-label of the original speech on the masked region like HuBERT

Models - Performances of Layers for Different Tasks



Experiment 4 - WavLM on different tasks

- **Keyword Spotting**
  - test accuracy : **0.969**
  - Task type: Classification
  - Steps: 50000

- **Phoneme Recognition**
  - test accuracy : **0.9536**
  - Task type: Transcription
  - Steps: 50000

In addition, we optimize the model structure and training data of HuBERT and wav2vec 2.0. We add **gated relative position bias (grep)** [15] to the Transformer structure as the backbone, which improves model performance for ASR and keeps almost the same parameter number and training speed. Compared with the convolutional relative position embedding used in wav2vec 2.0 and HuBERT, the gates **allow the relative position bias to be adjusted adaptively by conditioning on the current speech content**. To further improve the model robustness and alleviate the data mismatch, we **scale up unlabeled pre-training data to 94k hours of public audios**. The dataset consists of **60k hours of Libri-Light**, **10k hours of GigaSpeech** [16], and **24k hours of VoxPopuli** [17]. The new dataset consists of training instances from different scenarios, such as podcasts, YouTube, and European Parliament (EP) event recordings

Results of Upstream Models to PR & KS

Experiment 5 - PR with different models

- Phoneme Recognition is a transcribing task which tries to transcribe spoken words into phonemes.

Configurations

- Batch size = 16
- Dropout = 0.2
- Model = RNN(bidirectional, layernorm = true)/Wav2Letter
- Optimizer = Adam
- Module = LSTM
- Dim = 1024
- Steps = 50000

Dataset

Test : test-clean

Train: test-clean-100

Dev: dev-clean

Experiment 5 - Phoneme Recognition Evaluation Result

	Steps	test-Loss (torch.nn.CTCLoss)	test-WER/Accuracy
Fbank	50000	2.8204	0.8668/0.1332
Hubert	51700	0.26769	0.0599/0.9401
Wav2Vec2	50800	0.29295	0.0648/0.9352
WavLM	50000	0.21890	0.046388/0.9536

Fbank - Baseline Model

Fbank（FilterBank）：The general steps to obtain Fbank features from a speech signal are: pre-emphasis, framing, windowing, short-time Fourier transform (STFT), normalization, and Mel filtering, among others.

```
kaldi:
  feat_type: fbank
  fbank:
    num_mel_bins: 80
    frame_length: 25.0
    frame_shift: 10.0
    use_log_fbank: True

delta:
  order: 2
  win_length: 5

cmvn:
  use_cmvn: True
```

100% of testing result has WER > 0.5

Experiment 5 - Fbank to PR Results

I get tired of seeing men and horses going up and down, up and down.

Ground Truth: AY1 G EH1 T T AY1 ER0 D AH1 V S IY1 IH0 NG M EH0 N AE1 N D HH AO1 R S IH0 Z G OW1 IH0 NG AH1 P AE1 N D D AW1 N AH1 P AE1 N D D AW1 N

Fbank Output: D D T S NG S Z D

Mary(Mery) sighed

Ground Truth: M ER0 IY1 S AY1 D

Fbank Output: M R IY0 S D

Best Performance

WER : 0.5

Experiment 5 - Hubert to PR Results

Worse Performance

Robin Fitzooth

Ground Truth: R AA1 B IH0 N F IH0 T UW2 TH

Hubert Output: R AA1 P IH0 AH0 N F IH1 T S Y UW1 TH

WER : 0.6

Suppose it's a friend

Ground Truth: S AH0 P OW1 Z IH1 T S AH0 F R EH1 N D

Hubert Output: SH AH0 P OW1 Z IH0 T S T AH1 V B R AE1 IH1 N D

WER : 0.5714

Stephanos Dedalos

Ground Truth: S T EH0 F AA1 N OW0 S D EY0 D AA1 L OW0 Z

Hubert Output: S T F N ER0 S D T L AO1 S

WER : 0.5333

Experiment 5 - Wav2Vec2 to PR Results

Worse Performance

Stephanos Dedalos

Ground Truth:	S T EH1 F AH0 N S EH1 D L S	**WER : 0.5333**
Wav2Vec2 Output:	S T EH0 F AA1 N OW0 S D EY0 D AA1 L OW0 Z	

Ay Me

Ground Truth:	EY1 M	**WER : 0.5000**
Wav2Vec2 Output:	M	

Fine Glorious

Ground Truth:	F AY1 N G L AO1 R IY0 AH0 S	**WER : 0.5000**
Wav2Vec2 Output:	F F AY1 N AO1 AO1 R IY0 S Z S	

Result Comparison

Hubert

7176-92135-0030 - Line 2441 - WER: 0.3000

8555-284447-0016 - Line 2601 - WER: 0.3000

1089-134686-0003 - Line 2349 - WER: 0.3043

4970-29093-0016 - Line 2570 - WER: 0.3077

3729-6852-0027 - Line 2543 - WER: 0.3333

908-31957-0010 - Line 2606 - WER: 0.3333

3729-6852-0043 - Line 2261 - WER: 0.3600

237-134500-0025 - Line 2619 - WER: 0.4000

121-123852-0001 - Line 2620 - WER: 0.5000

1089-134691-0024 - Line 2546 - WER: 0.5333

8555-284447-0011 - Line 2562 - WER: 0.5714

61-70968-0038 - Line 2603 - WER: 0.6000

Wav2Vec2

8463-294828-0005 - Line 2433 - WER: 0.3000

4970-29093-0016 - Line 2570 - WER: 0.3077

7127-75947-0033 - Line 975 - WER: 0.3151

260-123286-0014 - Line 2245 - WER: 0.3200

1995-1837-0002 - Line 2492 - WER: 0.3333

908-31957-0010 - Line 2606 - WER: 0.3333

237-134500-0001 - Line 2616 - WER: 0.3333

61-70968-0034 - Line 2004 - WER: 0.3636

1995-1826-0025 - Line 2350 - WER: 0.3913

3729-6852-0043 - Line 2261 - WER: 0.4000

3729-6852-0027 - Line 2543 - WER: 0.4000

61-70968-0038 - Line 2603 - WER: 0.4000

2830-3980-0026 - Line 2617 - WER: 0.4000

8555-292519-0002 - Line 2618 - WER: 0.4000

237-134500-0025 - Line 2619 - WER: 0.4000

8555-284447-0011 - Line 2562 - WER: 0.4286

1995-1826-0014 - Line 2595 - WER: 0.4545

8555-284447-0016 - Line 2601 - WER: 0.5000

121-123852-0001 - Line 2620 - WER: 0.5000

1089-134691-0024 - Line 2546 - WER: 0.5333

Experiment 5 - WavLM to PR Result

8555-284447-0016 - Line 2601 - WER: 0.3000

4970-29093-0016 - Line 2570 - WER: 0.3077

260-123286-0014 - Line 2245 - WER: 0.3200

237-134500-0001 - Line 2616 - WER: 0.3333

3729-6852-0027 - Line 2543 - WER: 0.4000

61-70968-0038 - Line 2603 - WER: 0.4000

2830-3980-0026 - Line 2617 - WER: 0.4000

237-134500-0025 - Line 2619 - WER: 0.4000

121-123852-0001 - Line 2620 - WER: 0.5000

1089-134691-0024 - Line 2546 - WER: 0.5333

Fewer WER > 0.3

Still Perform Worst in Similar Sound Tracks

Experiment 6 - KS with different models

- Keyword spotting is a task which aims to detect a specific set of spoken words.
- Use the classification report from sklearn.metrics to do the further examination
- batch size = 32, lr = 2.5e-4, optimizer = AdamW, pooling = MeanPooling, steps = 50000
- Format of the datasets:

```
no-d7467392_nohash_0.wav no
no-1b4c9b89_nohash_4.wav no
no-b83c1acf_nohash_2.wav no
up-f6af2457_nohash_1.wav up
up-7e1054e7_nohash_0.wav up
up-9a69672b_nohash_0.wav up
```

(data of test_truth.txt)

Experiment 6 - KS with different models

Classification Reports:

wav2vec2:

	precision	recall	f1-score	support
stop	1.00	0.95	0.97	257
yes	0.94	0.95	0.94	257
down	0.98	0.96	0.97	253
on	0.92	0.96	0.94	251
right	0.94	0.98	0.96	267
go	0.94	0.96	0.95	252
silence	0.97	0.95	0.96	262
left	0.98	0.98	0.98	246
unknown	0.99	0.96	0.97	259
off	0.96	0.98	0.97	249
no	0.95	0.97	0.96	272
up	1.00	0.96	0.98	256
accuracy			0.96	3081
macro avg	0.96	0.96	0.96	3081
weighted avg	0.96	0.96	0.96	3081

hubert:

	precision	recall	f1-score	support
stop	1.00	0.93	0.96	257
yes	0.97	0.95	0.96	257
down	0.98	0.96	0.97	253
on	0.95	0.96	0.95	251
right	0.95	0.99	0.97	267
go	0.98	0.96	0.97	252
silence	0.93	0.97	0.95	262
left	0.99	0.96	0.98	246
unknown	0.96	0.98	0.97	259
off	0.98	0.98	0.98	249
no	0.96	0.97	0.96	272
up	0.94	0.98	0.96	256
accuracy			0.96	3081
macro avg	0.97	0.96	0.97	3081
weighted avg	0.97	0.96	0.96	3081

wavlm_base+:

	precision	recall	f1-score	support
down	1.00	0.96	0.98	257
left	0.94	0.98	0.96	257
on	0.95	0.97	0.96	253
up	0.99	0.96	0.97	251
off	0.97	0.98	0.97	267
silence	1.00	0.94	0.97	252
stop	0.92	0.98	0.95	262
right	0.98	0.98	0.98	246
yes	1.00	0.95	0.97	259
go	1.00	0.99	0.99	249
unknown	0.99	0.97	0.98	272
no	0.91	0.97	0.94	256
accuracy			0.97	3081
macro avg	0.97	0.97	0.97	3081
weighted avg	0.97	0.97	0.97	3081

- We can find out that all the indicators are pretty high

Experiment 6 - KS with different models

fbank:

	precision	recall	f1-score	support
off	0.00	0.00	0.00	257
no	0.00	0.00	0.00	257
left	0.00	0.00	0.00	253
stop	0.00	0.00	0.00	251
down	0.00	0.00	0.00	267
up	0.00	0.00	0.00	252
on	0.00	0.00	0.00	262
silence	0.00	0.00	0.00	246
go	0.00	0.00	0.00	259
right	0.00	0.00	0.00	249
unknown	0.00	0.00	0.00	272
yes	0.08	1.00	0.15	256
accuracy			0.08	3081
macro avg	0.01	0.08	0.01	3081
weighted avg	0.01	0.08	0.01	3081

```
for label_1, label_2 in zip(pred_labels, true_labels):
    total += 1
    if label_1 == label_2 and label_1 != "yes":
        print(label_1, label_2)
        num += 1

    if label_1 != "yes":
        print(label_1)
        count += 1
        print(total)
```

```
stop
519
up
1844
on
2230
off
2370
down
2371
on
2391
off
2505
_unknown_
2587
on
2614
_unknown_
2624
there's 10 labels not being yes
```

- We can find out that almost all the lpredicted labels are “yes”
- And there’s no matched labels except for “yes”
- The performance sucks

SUPERB Challenge

Method	Name	Description	URL	Params ↓	MACs ↓	(1) ↓	(2) ↓	(3) ↓	(4) ↓	Rank ↑	Score ↑	KS ↑	IC ↑	PR ↓	ASR ↓	ER ↑	QbE ↑
WavLM Large	Microsoft	M-P + VQ …		3.166e+8	4.326e+12	3.…	6.…	1.…	2.…	29.7	1145	97.86	99.31	3.06	3.44	70.62	8.86
CoBERT Base	ByteDanc…	Code Repr…		9.435e+7	1.660e+12	1.…	2.…	4.…	8.…	20.7	894	96.36	98.87	3.08	4.74	65.32	5.07
HuBERT Large	paper	M-P + VQ		3.166e+8	4.324e+12	3.…	6.…	1.…	2.…	21.75	919	95.29	98.76	3.53	3.62	67.62	3.53
data2vec Large	CI Tang	Masked G…		3.143e+8	4.306e+12	3.…	6.…	1.…	2.…	23.8	949	96.75	98.31	3.6	3.36	66.31	6.28
WavLM Base+	Microsoft	M-P + VQ …		9.470e+7	1.670e+12	1.…	2.…	4.…	8.…	27.95	1106	97.37	99	3.92	5.59	68.65	9.88
data2vec-aqc Base	Speech La…	Masked G…		9.384e+7	1.657e+12	1.…	2.…	4.…	8.…	22.05	935	96.36	98.92	4.11	5.39	67.59	6.65
LightHuBERT Sta…	LightHuBE…	Once-for-A…		9.500e+7	-	-	-	-	-	23.6	959	96.82	98.5	4.15	5.71	66.25	7.37
WavLM Base+	Lawrance	WavLM Ba…	-	9.470e+7	1.600e+2	1.…	1.…	1.…	1.…	5.9	-	96.92	-	4.64	-	-	-
data2vec base	CI Tang	Masked G…		9.375e+7	1.657e+12	1.…	2.…	4.…	8.…	19.55	884	96.56	97.63	4.69	4.94	66.27	5.76
wav2vec 2.0 Large	paper	M-C + VQ		3.174e+8	4.326e+12	3.…	6.…	1.…	2.…	20.5	914	96.66	95.28	4.75	3.75	65.64	4.89
WavLM Base	Microsoft	M-P + VQ …		9.470e+7	1.670e+12	1.…	2.…	4.…	8.…	24.55	1019	96.79	98.63	4.84	6.21	65.94	8.7
HuBERT Base	paper	M-P + VQ		9.470e+7	1.669e+12	1.…	2.…	4.…	8.…	20.65	941	96.3	98.34	5.41	6.42	64.92	7.36
wav2vec 2.0 Base	paper	M-C + VQ		9.504e+7	1.669e+12	1.…	2.…	4.…	8.…	15	818	96.23	92.35	5.74	6.43	63.43	2.33
ccc-wav2vec 2.0 …	Speech La…	M-C + VQ		9.504e+7	1.670e+12	1.…	2.…	4.…	8.…	20.25	940	96.72	96.47	5.95	6.3	64.17	6.73
Hubert Base	Lawrance	PR and KS	-	9.440e+7	1.600e+2	1.…	1.…	1.…	1.…	4.5	-	96.53	-	5.99	-	-	-
b0990106x	陳亭瑋	wav2vec2-…	-	1.600e+2	1.600e+2	1.…	1.…	1.…	1.…	2.1	-	-	-	6.28	-	-	-
Wav2Vec 2.0	Lawrance	Wav2Vec2…	-	9.500e+7	1.600e+2	1.…	1.…	1.…	1.…	3.9	-	96.33	-	6.49	-	-	-

Method	Name	Description	URL	Params ↓	MACs ↓	(1) ↓	(2) ↓	(3) ↓	(4) ↓	Rank ↑	Score ↑	KS ↑	IC ↑	PR ↓	ASR ↓	ER ↑	QbE ↑
WavLM Large	Microsoft	M-P + VQ ...	🔗	3.166e+8	4.326e+12	3....	6....	1....	2....	29.7	1145	97.86	99.31	3.06	3.44	70.62	8.86
WavLM Base+	Microsoft	M-P + VQ ...	🔗	9.470e+7	1.670e+12	1....	2....	4....	8....	27.95	1106	97.37	99	3.92	5.59	68.65	9.88
FaST-VGS+	Puyuan P...	FaST-VGS...	-	2.172e+8	-	-	-	-	-	17.05	809	97.27	98.97	7.76	8.83	62.71	5.62
WavLM Base+	Lawrance	WavLM Ba...	-	9.470e+7	1.600e+2	1....	1....	1....	1....	5.9	-	96.92	-	4.64	-	-	-
LightHuBERT Sta...	LightHuBE...	Once-for-A...	🔗	9.500e+7	-	-	-	-	-	23.6	959	96.82	98.5	4.15	5.71	66.25	7.37
WavLM Base	Microsoft	M-P + VQ ...	🔗	9.470e+7	1.670e+12	1....	2....	4....	8....	24.55	1019	96.79	98.63	4.84	6.21	65.94	8.7
data2vec Large	Cl Tang	Masked G...	🔗	3.143e+8	4.306e+12	3....	6....	1....	2....	23.8	949	96.75	98.31	3.6	3.36	66.31	6.28
ccc-wav2vec 2.0 ...	Speech La...	M-C + VQ	🔗	9.504e+7	1.670e+12	1....	2....	4....	8....	20.25	940	96.72	96.47	5.95	6.3	64.17	6.73
wav2vec 2.0 Large	paper	M-C + VQ	🔗	3.174e+8	4.326e+12	3....	6....	1....	2....	20.5	914	96.66	95.28	4.75	3.75	65.64	4.89
data2vec base	Cl Tang	Masked G...	🔗	9.375e+7	1.657e+12	1....	2....	4....	8....	19.55	884	96.56	97.63	4.69	4.94	66.27	5.76
Hubert Base	Lawrance	PR and KS	-	9.440e+7	1.600e+2	1....	1....	1....	1....	4.5	-	96.53	-	5.99	-	-	-
CoBERT Base	ByteDanc...	Code Repr...	🔗	9.435e+7	1.660e+12	1....	2....	4....	8....	20.7	894	96.36	98.87	3.08	4.74	65.32	5.07
data2vec-aqc Base	Speech La...	Masked G...	🔗	9.384e+7	1.657e+12	1....	2....	4....	8....	22.05	935	96.36	98.92	4.11	5.39	67.59	6.65
DPHuBERT	Yifan Peng	DPHuBER...	🔗	2.359e+7	6.541e+11	5....	1....	1....	3....	16.6	866	96.36	97.92	9.67	10.47	63.16	6.93
Wav2Vec 2.0	Lawrance	Wav2Vec2...	-	9.500e+7	1.600e+2	1....	1....	1....	1....	3.9	-	96.33	-	6.49	-	-	-
HuBERT Base	paper	M-P + VQ	🔗	9.470e+7	1.669e+12	1....	2....	4....	8....	20.65	941	96.3	98.34	5.41	6.42	64.92	7.36

Try Something FUN

Test PR - give random sentences to each upstream model

Our Test Data - Wav2Vec 2.0

We love 李宏毅.

WER : 0.3333

Ground Truth: W IY1 L AH1 V L IY1 HH AH1 NG W AY1

Wav2Vec2: W IY1 L AH1 V L IY0 HH AA1 NG

This presentation is so fun.

WER : 0.2381

Ground Truth: DH IH0 S P R EH1 Z AH0 N T EY1 SH AH0 N IH1 Z S OW0 F AH1 N

Wav2Vec2: DH IH0 S P R IY0 Z N EY1 SH AH0 N IH1 Z S S OW0 F AA1 N

Trust me.

WER : 0.1667

Ground Truth: T R AH1 S T M

Wav2Vec2: T R R AH1 S T M

Our Test Data - HuBERT

We love 李宏毅. **WER : 0.3333**

Ground Truth: W IY1 L AH1 V L IY1 HH AH1 NG W AY1

Hubert: W IY1 L AH0 V L IY0 HH AH1 NG Y

This presentation is so fun. **WER : 0.1905**

Ground Truth: DH IH0 S P R EH1 Z AH0 N T EY1 SH AH0 N IH1 Z S OW0 F AH1 N

Hubert: DH IH0 S P R EH2 Z EH0 N T EY1 SH AH0 N IH1 Z S OW0 F AA1 N D

Trust me. **WER : 0.000**

Ground Truth: T R AH1 S T M

Hubert: T R AH1 S T M

Our Test Data - WavLM

We love 李宏毅. **WER : 0.3333**

Ground Truth: W IY1 L AH1 V L IY1 HH AH1 NG W AY1

Wavlm: W IY1 L AH1 V L IY0 HH N NG Y IY1

This presentation is so fun. **WER : 0.1429**

Ground Truth: DH IH0 S P R EH1 Z AH0 N T EY1 SH AH0 N IH1 Z S OW0 F AH1 N

Wavlm: DH IH0 S P R EH2 Z EH0 N T EY1 SH AH0 N IH1 Z S OW0 F AA1 N

Trust me. **WER : 0.0000**

Ground Truth: T R AH1 S T M

Wavlm: T R AH1 S T M

Our Test Data - Overall Analysis

	Wav2Vec2	HuBERT	WavLM
We love Lee	0.3333	0.3333	0.3333
Presentation fun	0.2381	0.1905	0.1429
Trust me	0.1667	0.000	0.0000
Overall WER	0.256410	0.205128	0.179487

Better

Our Test Data - Overall Analysis

	Wav2Vec2	HuBERT	WavLM
We love Lee	0.3333	0.3333	0.3333

Ground Truth: W IY1 L AH1 V L IY1 HH AH1 NG W AY1
Wav2Vec2: W IY1 L AH1 V L IY0 HH AA1 NG
Hubert: W IY1 L AH0 V L IY0 HH AH1 NG Y
Wavlm: W IY1 L AH1 V L IY0 HH N NG Y IY1

Hypothesis: "Yi" is not common in English pronunciation

Our Test Data - Overall Analysis

	Wav2Vec2	HuBERT	WavLM
We love Lee	0.3333	0.3333	0.3333
[illegible]	[illegible]	[illegible]	[illegible]
Trust me	0.1667	0.000	0.0000
Overall WER	0.256410	0.205128	0.179487

Hypothesis 1: The shortest sentence → least prob. having error

Hypothesis 2: The words themselves are relatively easier for models

Further Testings

Test short sentences individually and collectively

$$A \longrightarrow WER,\ B \longrightarrow WER,\ C \longrightarrow WER$$
$$A + B + C \longrightarrow WER$$

Our Test Data - WavLM

WER : 0.0000 | WER : 0.0000 | WER : 0.0000

	Trust me.	Close the door.	I love you.
Full-sentence G.Truth	T R AH1 S T M	K L OW1 Z TH D AO1 R	AY1 L AH1 V Y UW1
Ground Truth	T R AH1 S T M	K L OW1 Z TH D AO1 R	AY1 L AH1 V Y UW1
Wavlm:	T R AH1 S T M	K L OW1 Z TH D AO1 R	AY1 L AH1 V Y UW1
Full-sentence Wavlm:	T R AH1 S T M	K L OW1 Z TH D AO1 R	AY1 L AH1 V Y UW1

WER : 0.0000

ALL CORRECT

Our Test Data - WavLM

	WER : 0.3333	WER : 0.2222	WER : 0.0000
	Dont go.	**He likes fish.**	**Jennie is pretty.**
Full-sentence G.Truth	D AA1 N T G OW1	HH IY1 L AY1 K S F IH1 SH	JH EH1 N IY0 IH1 Z P R IH1 T IY0
Ground Truth	D AA1 N T G OW1	HH IY1 L AY1 K S F IH1 SH	JH EH1 N IY0 IH1 Z P R IH1 T IY0
Wavlm:	D OW1 N G OW1	HH IY1 L AY1 K S F IH0 R	JH EH1 N IY0 IH1 Z P R IH1 T IY0
Full-sentence Wavlm:	D OW1 N T G OW1	HH IY1 L AY1 K S F IH1 SH	JH EH1 N IY0 IH1 Z P R IH1 T IY0

0.0385

Our Test Data - WavLM

	Beautiful girl.	Over the world.	Time flies.
	WER : 0.1818	WER : 1.00	WER : 0.1667
Full-sentence G.Truth	B Y UW1 T AH0 F AH0 L G ER1 L	OW2 V ER0 TH W ER1 L D	T AY1 M F L AY1 Z
Ground Truth	B Y UW1 T AH0 F AH0 L G ER1 L	OW2 V ER0 TH W ER1 L D	T AY1 M F L AY1
Wavlm:	B Y UW1 T AH0 F AH0 L G R AY1	EH1 V V R IY0 TH IH0 N AY1	T AY1 M F L AY1 Z
Full-sentence Wavlm:	B Y UW1 T AH0 F AH0 L G ER1 AH0 L	OW2 V ER0 TH W ER1 L D	T AY1 M F L AY1 Z

WER : 0.0385

Our Test Data - HuBERT

	WER : 0.0000	WER : 0.0000	WER : 0.0000
	Trust me.	**Close the door.**	**I love you.**
Full-sentence G.Truth	T R AH1 S T M	K L OW1 Z TH D AO1 R	AY1 L AH1 V Y UW1
Ground Truth	T R AH1 S T M	K L OW1 Z TH D AO1 R	AY1 L AH1 V Y UW1
HuBERT:	T R AH1 S T M	K L OW1 Z TH D AO1 R	AY1 L AH1 V Y UW1
Full-sentence HuBERT:	T R AH1 S T M K L OW1 Z TH D AO1 R AY1 L AH1 V Y UW1		

WER : 0.0000

ALL CORRECT

Our Test Data - HuBERT

	WER : 0.1667	WER : 0.1111	WER : 0.4545
	Dont go.	**He likes fish.**	**Jennie is pretty.**
Full-sentence G.Truth	D AA1 N T G OW1	HH IY1 L AY1 K S F IH1 SH	JH EH1 N IY0 IH1 Z P R IH1 T IY0
Ground Truth	D AA1 N T G OW1	HH IY1 L AY1 K S F IH1 SH	JH EH1 N IY0 IH1 Z P R IH1 T IY0
HuBERT:	D OW1 N T G OW1	HH IY1 L AY2 K S F IH1 SH	JH N S P R IH1 IY0
Full-sentence HuBERT:	D OW1 N T G OW1	HH IY1 L AY2 K S F IH1 SH	JH N IY0 IH1 Z P R IH1 T IY0

0.1154

Our Test Data - HuBERT

	WER : 0.2727	WER : 1.00	WER : 0.1667
	Beautiful girl.	**Over the world.**	**Time flies.**
Full-sentence G.Truth	B Y UW1 T AH0 F AH0 L G ER1 L	OW2 V ER0 TH W ER1 L D	T AY1 M F L AY1 Z
Ground Truth	B Y UW1 T AH0 F AH0 L G ER1 L	OW2 V ER0 TH W ER1 L D	T AY1 M F L AY1
HuBERT:	B IH1 UW1 T AH0 F AH0 L G R	AE1 N P P R EH1 N AY1	T AY1 M F L Z
Full-sentence HuBERT:	B Y UW1 T AH0 F AH0 L G ER1 L	OW2 V ER0 TH W ER1 L L D	T AY1 M F L AY1 Z

WER : 0.0385

Our Test Data - Wav2vec2

WER : 0.0000 WER : 0.0000 WER : 0.0000

	Trust me.	Close the door.	I love you.
Full-sentence G.Truth	T R AH1 S T M	K L OW1 Z TH D AO1 R	AY1 L AH1 V Y UW1
Ground Truth	T R AH1 S T M	K L OW1 Z TH D AO1 R	AY1 L AH1 V Y UW1
Wav2vec2:	T R AH1 S T M	K L OW1 Z TH D AO1 R	AY1 L AH1 V Y UW1
Full-sentence Wav2vec2:	T R AH1 S T M K L OW1 Z TH D AO1 R AY1 L AH1 V Y UW1		

WER : 0.0000

ALL CORRECT

Our Test Data - Wav2vec2

WER : 0.5000 | WER : 0.2222 | WER :0.1818

	Dont go.	He likes fish.	Jennie is pretty.
Full-sentence G.Truth	D AA1 N T G OW1	HH IY1 L AY1 K S F IH1 SH	JH EH1 N IY0 IH1 Z P R IH1 T IY0
Ground Truth	D AA1 N T G OW1	HH IY1 L AY1 K S F IH1 SH	JH EH1 N IY0 IH1 Z P R IH1 T IY0
Wav2vec2:	D OW1 N G	HH IY1 L AY1 IH1 K S F IH1 IH0 SH	JH EH1 N AH0 Z P R IH1 T IY0
Full-sentence Wav2vec2:	D OW1 N T G OW1	HH IY1 L AY1 K S F IH1 SH	JH EH1 N IH1 Z P R IH1 T IY0

0.0769

Our Test Data - Wav2vec2

WER : 0.0000 | WER : 1.00 | WER : 0.3333

	Beautiful girl.	Over the world.	Time flies.
Full-sentence G.Truth	B Y UW1 T AH0 F AH0 L G ER1 L	OW2 V ER0 TH W ER1 L D	T AY1 M F L AY1 Z
Ground Truth	B Y UW1 T AH0 F AH0 L G ER1 L	OW2 V ER0 TH W ER1 L D	T AY1 M F L AY1
Wav2vec2:	B Y UW1 T AH0 F AH0 L G ER1 L	V R IY0 G R T Z S	T AY1 M P F L Z
Full-sentence Wav2vec2:	B Y UW1 T F AH0 L G ER1 L	OW2 V ER0 TH W ER1 L D	T AY1 M F L Z

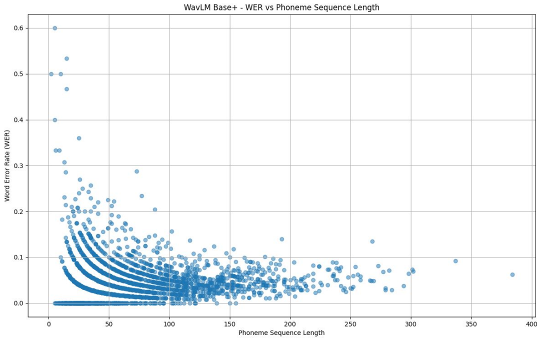
WER : 0.0769

	Trust me.	Close the door.	I love you.	Dont go.	He likes fish.	Jennie is pretty.	Beautiful girl.	Over the world.	Time flies.
Full-sentence G.Truth	T R AH1 S T M K L OW1 Z TH D AO1 R AY1 L AH1 V Y UW1			D AA1 N T G OW1 HH IY1 L AY1 K S F IH1 SH JH EH1 N IY0 IH1 Z P R IH1 T IY0			B Y UW1 T AH0 F AH0 L G ER1 L OW2 V ER0 TH W ER1 L D T AY1 M F L AY1 Z		
Ground Truth	T R AH1 S T M	K L OW1 Z TH D AO1 R	AY1 L AH1 V Y UW1	D AA1 N T G OW1	HH IY1 L AY1 K S F IH1 SH	JH EH1 N IY0 IH1 Z P R IH1 T IY0	B Y UW1 T AH0 F AH0 L G ER1 L	OW2 V ER0 TH W ER1 L D	T AY1 M F L AY1
Wav2vec2:	T R AH1 S T M	K L OW1 Z TH D AO1 R	AY1 L AH1 V Y UW1	D OW1 N G	HH IY1 L AY1 IH1 K S F IH1 IH0 SH	JH EH1 N AH0 Z P R IH1 T IY0	B Y UW1 T AH0 F AH0 L G ER1 L	V R IY0 G R T Z S	T AY1 M P F L Z
Full-sentence Wav2vec2:	T R AH1 S T M K L OW1 Z TH D AO1 R AY1 L AH1 V Y UW1			D OW1 N T G OW1 HH IY1 L AY1 K S F IH1 SH JH EH1 N IH1 Z P R IH1 T IY0			B Y UW1 T F AH0 L G ER1 L OW2 V ER0 TH W ER1 L D T AY1 M F L Z		
HuBERT:	T R AH1 S T M	K L OW1 Z TH D AO1 R	AY1 L AH1 V Y UW1	D OW1 N T G OW1	HH IY1 L AY2 K S F IH1 SH	JH N S P R IH1 IY0	B IH1 UW1 T AH0 F AH0 L G R	AE1 N P P R EH1 N AY1	T AY1 M F L Z
Full-sentence HuBERT:	T R AH1 S T M K L OW1 Z TH D AO1 R AY1 L AH1 V Y UW1			D OW1 N T G OW1 HH IY1 L AY2 K S F IH1 SH JH N IY0 IH1 Z P R IH1 T IY0			B Y UW1 T AH0 F AH0 L G ER1 L OW2 V ER0 TH W ER1 L L D T AY1 M F L AY1 Z		
Wavlm:	T R AH1 S T M	K L OW1 Z TH D AO1 R	AY1 L AH1 V Y UW1	D OW1 N G OW1	HH IY1 L AY1 K S F IH0 R	JH EH1 N IY0 IH1 Z P R IH1 T IY0	B Y UW1 T AH0 F AH0 L G R AY1	EH1 V V R IY0 TH IH0 N AY1	T AY1 M F L AY1 Z
Full-sentence Wavlm:	T R AH1 S T M K L OW1 Z TH D AO1 R AY1 L AH1 V Y UW1			D OW1 N T G OW1 HH IY1 L AY1 K S F IH1 SH JH EH1 N IY0 IH1 Z P R IH1 T IY0			B Y UW1 T AH0 F AH0 L G ER1 AH0 L OW2 V ER0 TH W ER1 L D T AY1 M F L AY1 Z		

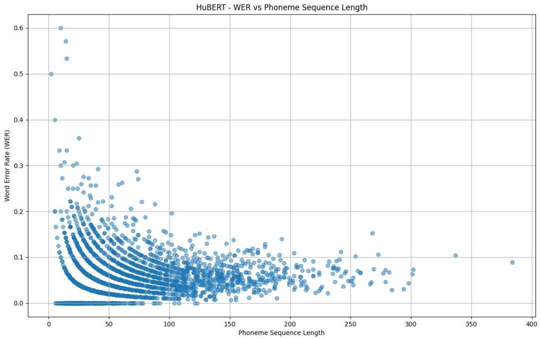
Test Result Analysis

Based on our test, we think

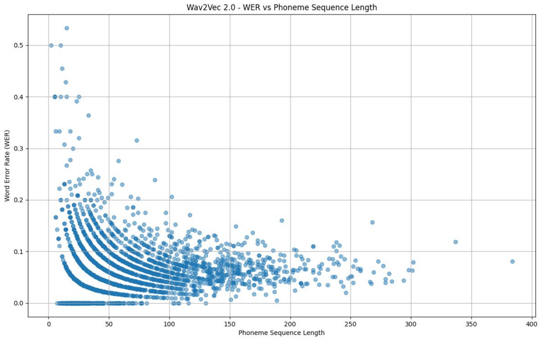
longer Sentences have better correctness rates



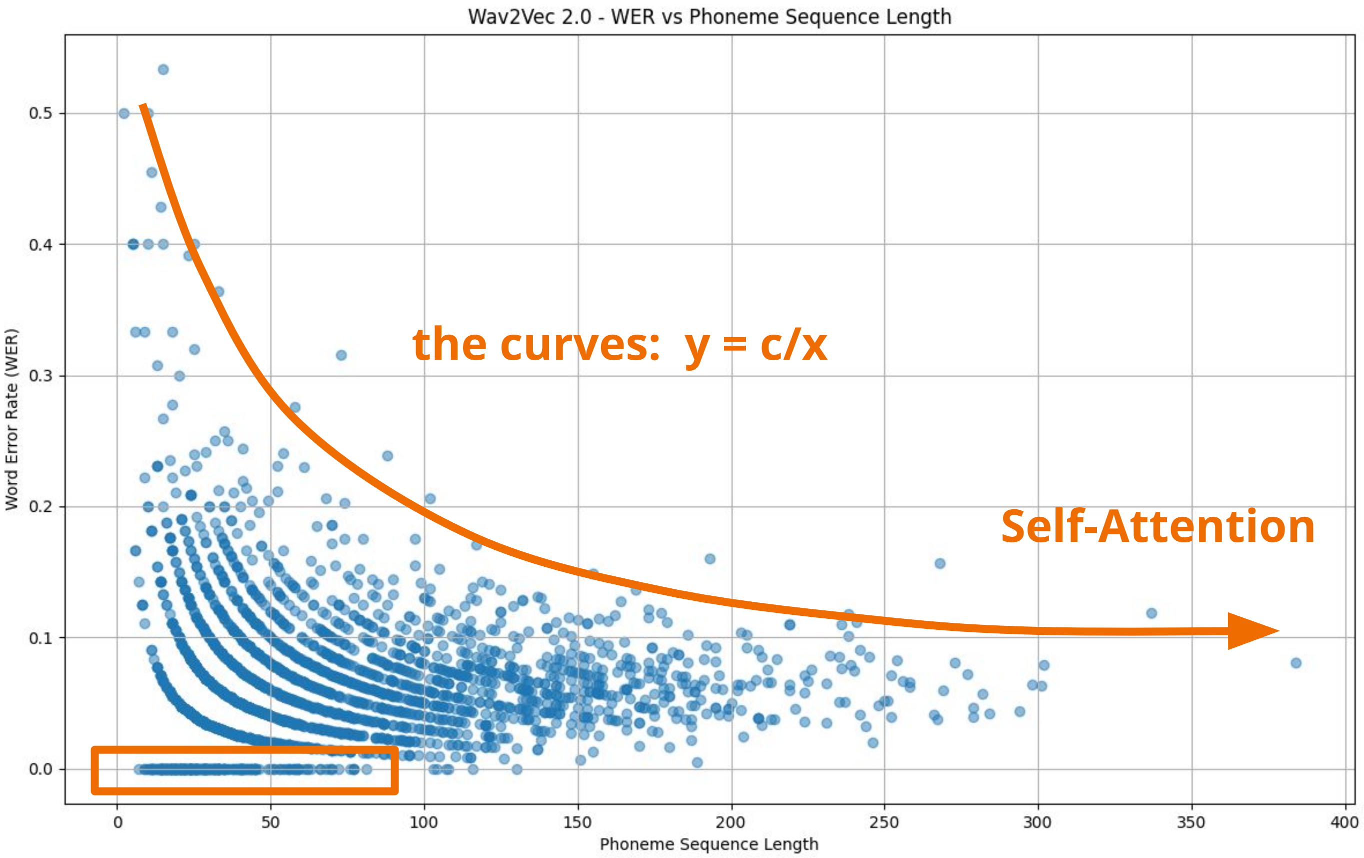
WavLM



HuBERT

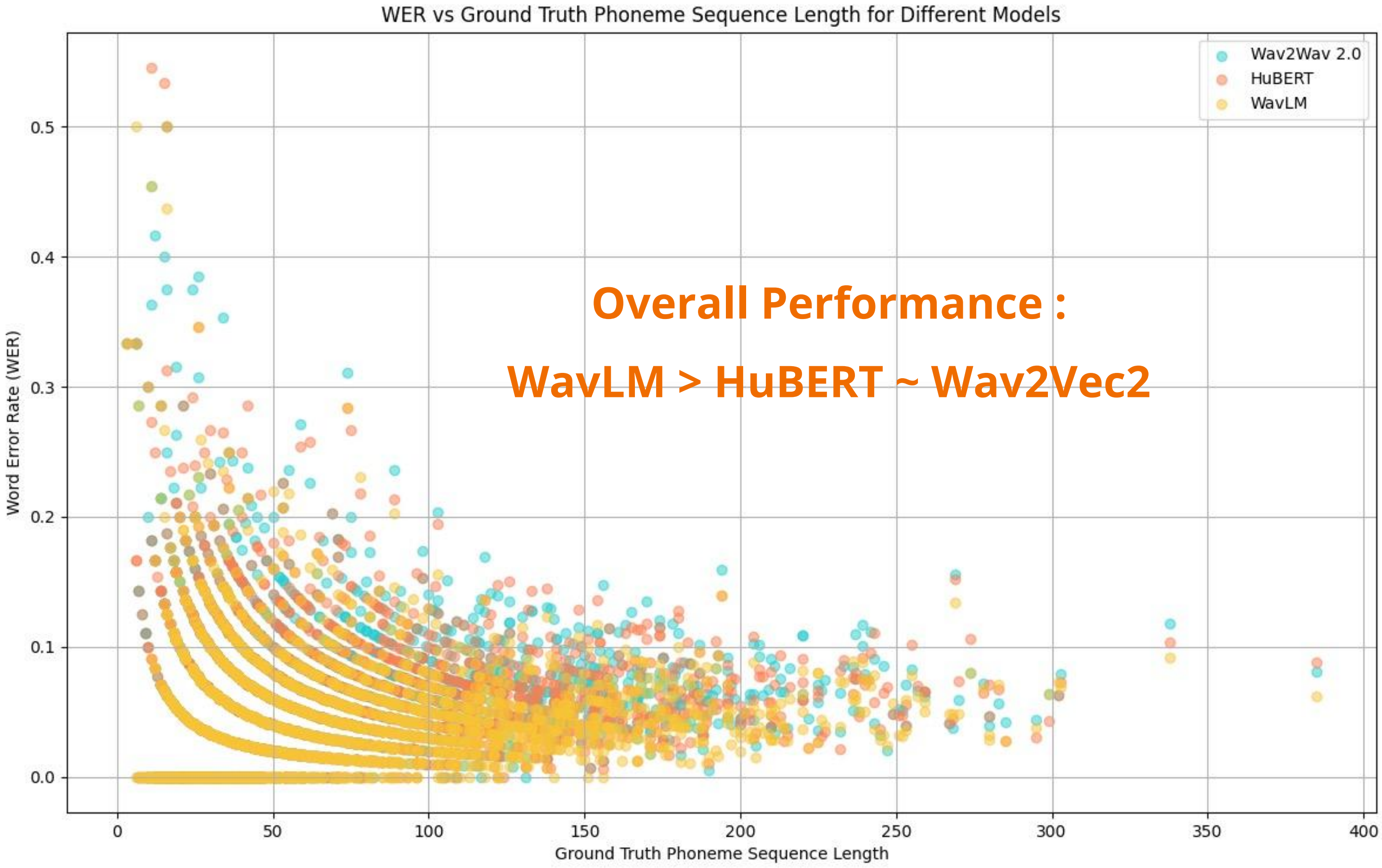


Wav2vec 2.0



Shorter sentence → totally correct

```
Length: 28,     WER: 0.0000,    File #: 61-70970-0005, Text: THE LAD HAD CHECKED HIM THEN
Length: 26,     WER: 0.0000,    File #: 61-70970-0006, Text: NEVER THAT SIR HE HAD SAID
Length: 14,     WER: 0.0000,    File #: 61-70970-0008, Text: NOW TO BED BOY
Length: 60,     WER: 0.0000,    File #: 61-70970-0014, Text: PRESENTLY HE CROSSED THE FLOOR OF HIS ROOM WITH DECIDED STEP
Length: 84,     WER: 0.0000,    File #: 61-70970-0016, Text: WE WILL GO OUT TOGETHER TO THE BOWER THERE IS A WAY DOWN TO THE COURT FROM MY WINDOW
Length: 34,     WER: 0.0000,    File #: 61-70970-0017, Text: REST AND BE STILL UNTIL I WARN YOU
Length: 66,     WER: 0.0000,    File #: 61-70970-0029, Text: FROM THE BLACKNESS BEHIND THE LIGHT THEY HEARD A VOICE WARRENTON'S
Length: 42,     WER: 0.0000,    File #: 61-70970-0030, Text: SAVE ME MASTERS BUT YOU STARTLED ME RARELY
Length: 51,     WER: 0.0000,    File #: 61-70968-0001, Text: GIVE NOT SO EARNEST A MIND TO THESE MUMMERIES CHILD
Length: 61,     WER: 0.0000,    File #: 61-70968-0003, Text: HE WAS LIKE UNTO MY FATHER IN A WAY AND YET WAS NOT MY FATHER
Length: 54,     WER: 0.0000,    File #: 61-70968-0004, Text: ALSO THERE WAS A STRIPLING PAGE WHO TURNED INTO A MAID
Length: 37,     WER: 0.0000,    File #: 61-70968-0007, Text: SISTER NELL DO YOU HEAR THESE MARVELS
Length: 63,     WER: 0.0000,    File #: 61-70968-0008, Text: TAKE YOUR PLACE AND LET US SEE WHAT THE CRYSTAL CAN SHOW TO YOU
Length: 139,    WER: 0.0000,    File #: 61-70968-0010, Text: FORTHWITH ALL RAN TO THE OPENING OF THE TENT TO SEE WHAT MIGHT BE AMISS BUT MASTER WILL WHO PEEPED OUT FIRST NEEDED NO MORE THAN ONE GLANCE
Length: 25,     WER: 0.0000,    File #: 61-70968-0018, Text: SO I DID PUSH THIS FELLOW
Length: 61,     WER: 0.0000,    File #: 61-70968-0037, Text: WHAT IS YOUR NAME LORDING ASKED THE LITTLE STROLLER PRESENTLY
Length: 44,     WER: 0.0000,    File #: 61-70968-0044, Text: IT WILL NOT BE SAFE FOR YOU TO STAY HERE NOW
Length: 47,     WER: 0.0000,    File #: 61-70968-0048, Text: AND HENRY MIGHT RETURN TO ENGLAND AT ANY MOMENT
Length: 23,     WER: 0.0000,    File #: 61-70968-0058, Text: WILL YOU FORGIVE ME NOW
Length: 51,     WER: 0.0000,    File #: 61-70968-0060, Text: NO THANKS I AM GLAD TO GIVE YOU SUCH EASY HAPPINESS
Length: 33,     WER: 0.0000,    File #: 5639-40744-0009, Text: MOTHER DEAR FATHER DO YOU HEAR ME
Length: 65,     WER: 0.0000,    File #: 5639-40744-0010, Text: IT IS THE ONLY AMENDS I ASK OF YOU FOR THE WRONG YOU HAVE DONE ME
Length: 115,    WER: 0.0000,    File #: 5639-40744-0029, Text: THIS TRUTH WHICH I HAVE LEARNED FROM HER LIPS IS CONFIRMED BY HIS FACE IN WHICH WE HAVE BOTH BEHELD THAT OF OUR SON
Length: 137,    WER: 0.0000,    File #: 5639-40744-0033, Text: HER BEARING WAS GRACEFUL AND ANIMATED SHE LED HER SON BY THE HAND AND BEFORE HER WALKED TWO MAIDS WITH WAX LIGHTS AND SILVER CANDLESTICKS
Length: 71,     WER: 0.0000,    File #: 6829-68769-0003, Text: IT WAS A DELIBERATE THEFT FROM HIS EMPLOYERS TO PROTECT A GIRL HE LOVED
Length: 81,     WER: 0.0000,    File #: 6829-68769-0009, Text: THEY WERE RECEIVED IN THE LITTLE OFFICE BY A MAN NAMED MARKHAM WHO WAS THE JAILER
Length: 46,     WER: 0.0000,    File #: 6829-68769-0010, Text: WE WISH TO TALK WITH HIM ANSWERED KENNETH TALK
Length: 55,     WER: 0.0000,    File #: 6829-68769-0014, Text: THEY FOLLOWED THE JAILER ALONG A SUCCESSION OF PASSAGES
Length: 72,     WER: 0.0000,    File #: 6829-68769-0033, Text: IT WAS BETTER FOR HIM TO THINK THE GIRL UNFEELING THAN TO KNOW THE TRUTH
Length: 31,     WER: 0.0000,    File #: 6829-68769-0041, Text: I'M NOT ELECTIONEERING JUST NOW
Length: 26,     WER: 0.0000,    File #: 6829-68769-0042, Text: OH WELL SIR WHAT ABOUT HIM
Length: 47,     WER: 0.0000,    File #: 6829-68769-0044, Text: IT HAS COST ME TWICE SIXTY DOLLARS IN ANNOYANCE
Length: 41,     WER: 0.0000,    File #: 6829-68769-0046, Text: YOU'RE FOOLISH WHY SHOULD YOU DO ALL THIS
Length: 54,     WER: 0.0000,    File #: 6829-68769-0051, Text: THERE WAS A GRIM SMILE OF AMUSEMENT ON HIS SHREWD FACE
Length: 93,     WER: 0.0000,    File #: 6829-68771-0006, Text: AND THIS WAS WHY KENNETH AND BETH DISCOVERED HIM CONVERSING WITH THE YOUNG WOMAN IN THE BUGGY
Length: 121,    WER: 0.0000,    File #: 6829-68771-0018, Text: FOR A MOMENT BETH STOOD STARING WHILE THE NEW MAID REGARDED HER WITH COMPOSURE AND A SLIGHT SMILE UPON HER BEAUTIFUL FACE
Length: 65,     WER: 0.0000,    File #: 6829-68771-0022, Text: I ATTEND TO THE HOUSEHOLD MENDING YOU KNOW AND CARE FOR THE LINEN
Length: 74,     WER: 0.0000,    File #: 6829-68771-0027, Text: THEY THEY EXCITE ME IN SOME WAY AND I I CAN'T BEAR THEM YOU MUST EXCUSE ME
Length: 60,     WER: 0.0000,    File #: 6829-68771-0028, Text: SHE EVEN SEEMED MILDLY AMUSED AT THE ATTENTION SHE ATTRACTED
Length: 39,     WER: 0.0000,    File #: 6829-68771-0034, Text: I WISH I KNEW MYSELF SHE CRIED FIERCELY
Length: 52,     WER: 0.0000,    File #: 908-157963-0016, Text: I PASS AWAY YET I COMPLAIN AND NO ONE HEARS MY VOICE
Length: 58,     WER: 0.0000,    File #: 908-157963-0022, Text: COME FORTH WORM AND THE SILENT VALLEY TO THY PENSIVE QUEEN
Length: 26,     WER: 0.0000,    File #: 908-31957-0000, Text: ALL IS SAID WITHOUT A WORD
Length: 54,     WER: 0.0000,    File #: 908-31957-0002, Text: I DID NOT WRONG MYSELF SO BUT I PLACED A WRONG ON THEE
Length: 190,    WER: 0.0000,    File #: 908-31957-0004, Text: SHALL I NEVER MISS HOME TALK AND BLESSING AND THE COMMON KISS THAT COMES TO EACH IN TURN NOR COUNT IT STRANGE WHEN I LOOK UP TO DROP ON A NEW RANGE OF
WALLS AND FLOORS ANOTHER HOME THAN THIS
Length: 17,     WER: 0.0000,    File #: 908-31957-0011, Text: AND LOVE BE FALSE
Length: 110,    WER: 0.0000,    File #: 908-31957-0019, Text: THOU CANST WAIT THROUGH SORROW AND SICKNESS TO BRING SOULS TO TOUCH AND THINK IT SOON WHEN OTHERS CRY TOO LATE
Length: 175,    WER: 0.0000,    File #: 908-31957-0025, Text: I LOVE THEE WITH A LOVE I SEEMED TO LOSE WITH MY LOST SAINTS I LOVE THEE WITH THE BREATH SMILES TEARS OF ALL MY LIFE AND IF GOD CHOOSE I SHALL BUT LOVE
THEE BETTER AFTER DEATH
Length: 45,     WER: 0.0000,    File #: 672-122797-0000, Text: OUT IN THE WOODS STOOD A NICE LITTLE FIR TREE
Length: 165,    WER: 0.0000,    File #: 672-122797-0001, Text: THE PLACE HE HAD WAS A VERY GOOD ONE THE SUN SHONE ON HIM AS TO FRESH AIR THERE WAS ENOUGH OF THAT AND ROUND HIM GREW MANY LARGE SIZED COMRADES PINES AS
WELL AS FIRS
Length: 187,    WER: 0.0000,    File #: 672-122797-0002, Text: HE DID NOT THINK OF THE WARM SUN AND OF THE FRESH AIR HE DID NOT CARE FOR THE LITTLE COTTAGE CHILDREN THAT RAN ABOUT AND PRATTLED WHEN THEY WERE IN THE
WOODS LOOKING FOR WILD STRAWBERRIES
Length: 49,     WER: 0.0000,    File #: 672-122797-0003, Text: BUT THIS WAS WHAT THE TREE COULD NOT BEAR TO HEAR
Length: 26,     WER: 0.0000,    File #: 672-122797-0011, Text: AND THEN WHAT HAPPENS THEN
Length: 87,     WER: 0.0000,    File #: 672-122797-0013, Text: I AM NOW TALL AND MY BRANCHES SPREAD LIKE THE OTHERS THAT WERE CARRIED OFF LAST YEAR OH
Length: 62,     WER: 0.0000,    File #: 672-122797-0015, Text: WERE I IN THE WARM ROOM WITH ALL THE SPLENDOR AND MAGNIFICENCE
Length: 61,     WER: 0.0000,    File #: 672-122797-0017, Text: SOMETHING BETTER SOMETHING STILL GRANDER MUST FOLLOW BUT WHAT
Length: 63,     WER: 0.0000,    File #: 672-122797-0021, Text: AND TOWARDS CHRISTMAS HE WAS ONE OF THE FIRST THAT WAS CUT DOWN
Length: 236,    WER: 0.0000,    File #: 672-122797-0022, Text: THE AXE STRUCK DEEP INTO THE VERY PITH THE TREE FELL TO THE EARTH WITH A SIGH HE FELT A PANG IT WAS LIKE A SWOON HE COULD NOT THINK OF HAPPINESS FOR HE
WAS SORROWFUL AT BEING SEPARATED FROM HIS HOME FROM THE PLACE WHERE HE HAD SPRUNG UP
Length: 38,     WER: 0.0000,    File #: 672-122797-0024, Text: THE DEPARTURE WAS NOT AT ALL AGREEABLE
Length: 53,     WER: 0.0000,    File #: 672-122797-0027, Text: THE SERVANTS AS WELL AS THE YOUNG LADIES DECORATED IT
Length: 26,     WER: 0.0000,    File #: 672-122797-0028, Text: THIS EVENING THEY ALL SAID
Length: 30,     WER: 0.0000,    File #: 672-122797-0029, Text: HOW IT WILL SHINE THIS EVENING
```



Overall Performance :

WavLM > HuBERT ~ Wav2Vec2

Short Conclusion

- The longer the sentence, the better the performance → self attention.
- Phoneme recognition often **relies on acoustic features of speech**.
  - Shorter sentences → less phonetic variation,

    → models hard to distinguish between similar phonemes → errors.

References

Hubert: https://jonathanbgn.com/2021/10/30/hubert-visually-explained.html

Wav2Vec2: https://jonathanbgn.com/2021/09/30/illustrated-wav2vec-2.html

Hubert Paper: https://arxiv.org/pdf/2106.07447.pdf

Fbank No Pre-Trained Paper: https://arxiv.org/pdf/2203.16973.pdf

PR across Languages Paper: https://arxiv.org/pdf/2206.12489.pdf

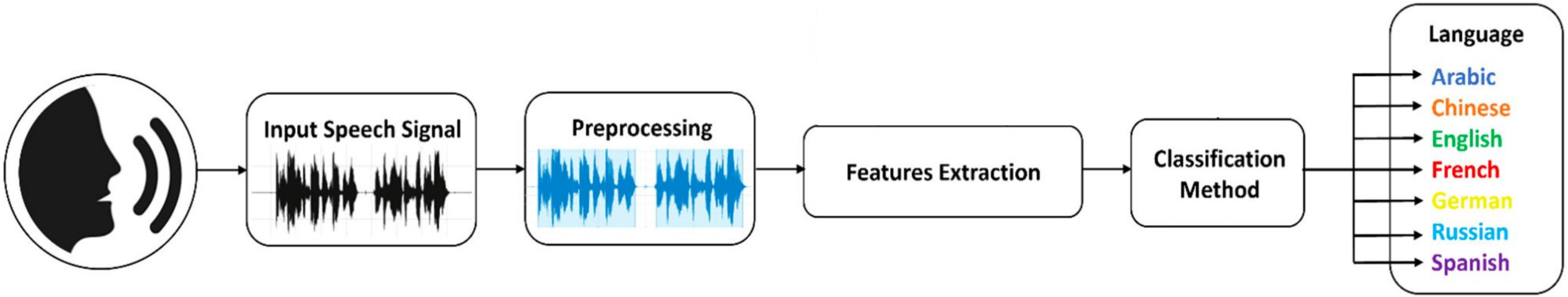
WavLM Paper: https://arxiv.org/pdf/2110.13900.pdf

Thanks

Track 4

OUR TASK - Spoken Language Identification (SLID)

To identify which language the speaker in the audio clip speaks.



Track 4 -

- **Downstream Outline**
- Find Suitable Dataset
- Compare & Analyze Upsteam Model

Reference We found

Hubert vs Wav2Vec2.0:

https://jonathanbgn.com/2021/10/30/hubert-visually-explained.html

https://neurosys.com/blog/wav2vec-2-0-framework

Phoneme 對照表

WavLM

KS 50,000 test acc: 0.9691658552418047

PR at 21000 -

test loss: 0.23664355278015137

test per: 0.054100164229794155

50000-

test loss: 0.21889734268188477

test per: 0.046388273248614144