從傳統語音特徵邁向深度學習自督導式語音特徵

F11944019 電信所博一 林子權

Outline

- 傳統語音特徵
- 語音自督導式模型（Speech Self-Supervised Models）
- 傳統語音特徵和語音自督導式特徵之表現比較
- 語音自督導式模型之特性
- 語音自督導式模型所面臨的挑戰

傳統語音特徵

- Spectrogram
- Log-Mel Spectrogram
- MFCC

Spectrogram

- 頻率隨著時間變化的一種視覺表示
- 生成方法：對信號進行短時傅立葉轉換（STFT）

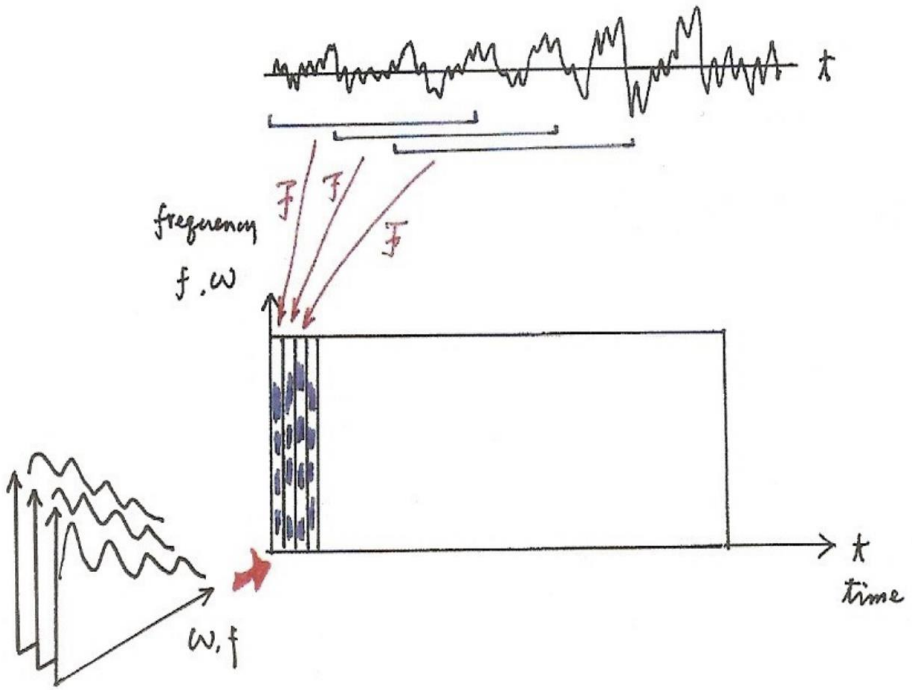


Image Reference: 李琳山老師數位語音處理概論

Spectrogram

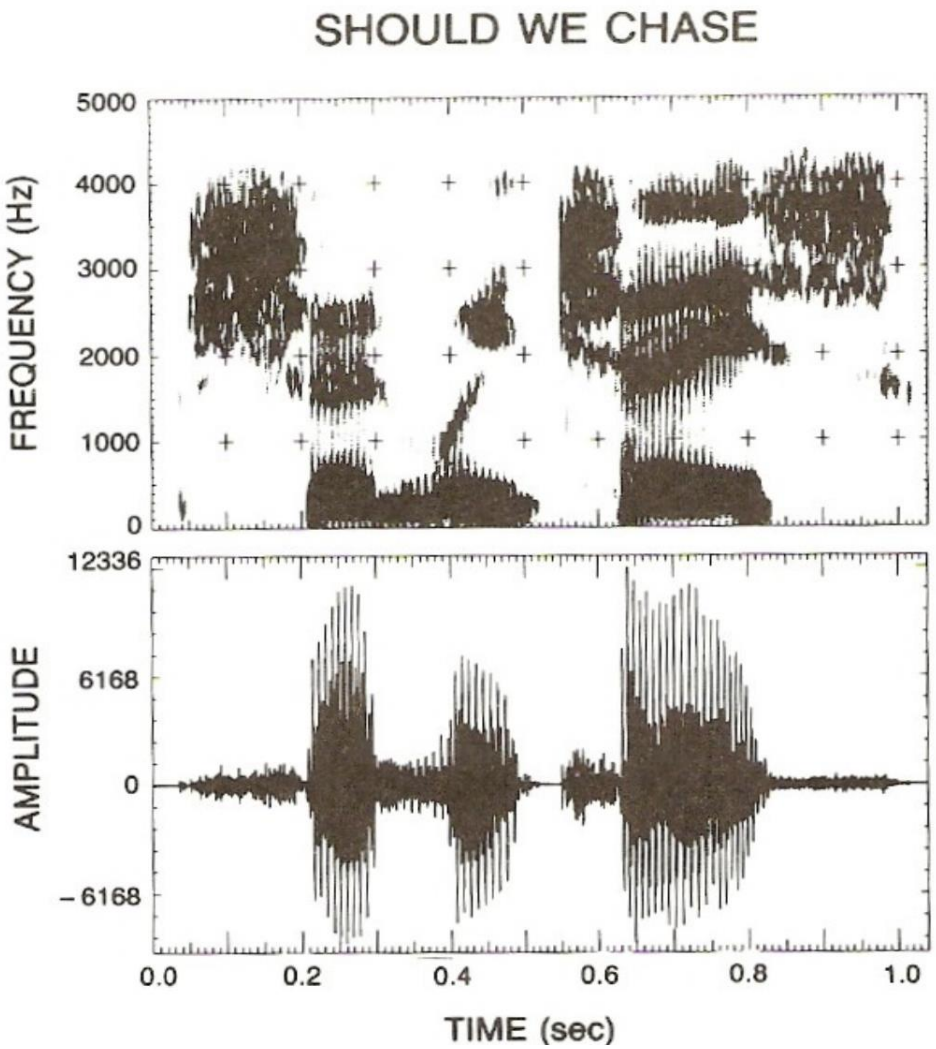


Image Reference: 李琳山老師數位語音處理概論

Log-Mel Spectrogram

- Motivation：人類對於低頻信號很敏感，但相對區分不清楚高頻的信號。

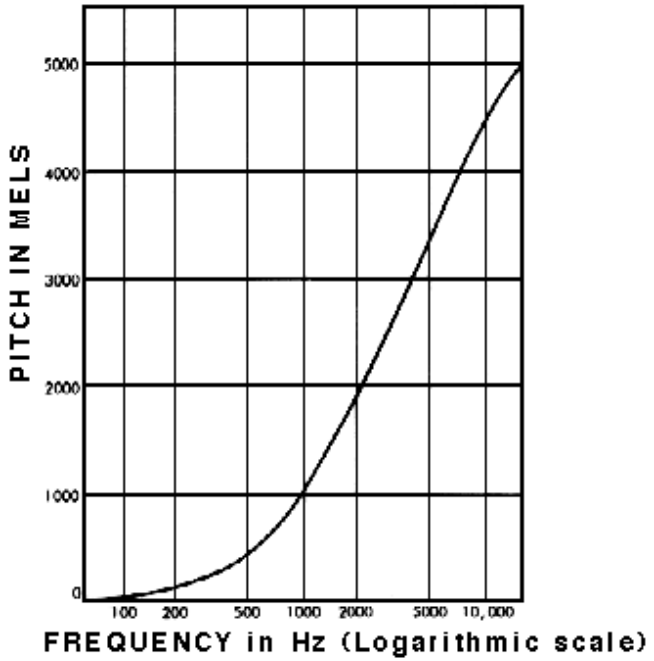
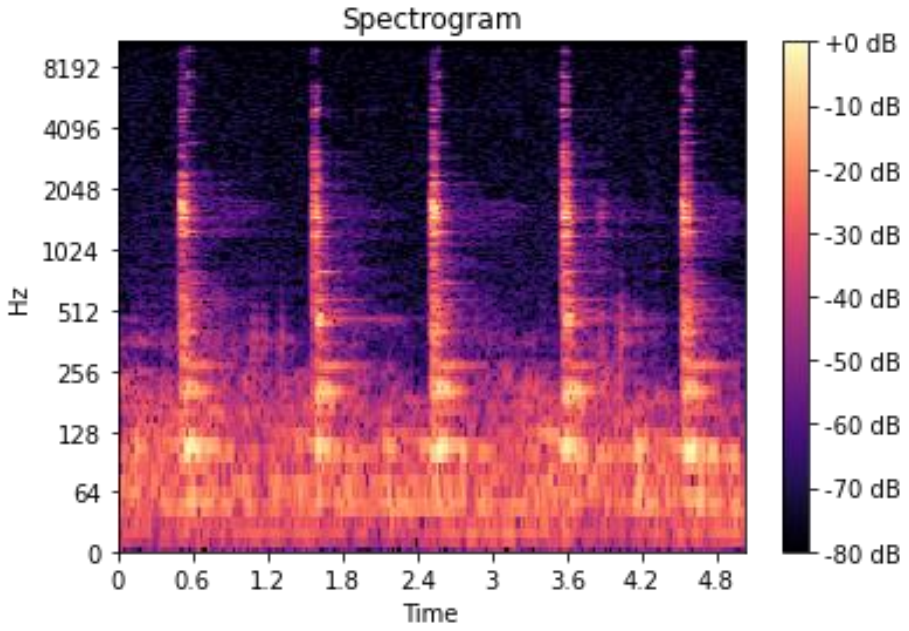


Image Reference: CSDN博客 - 理解梅尔频谱（mel spectrogram）

Log-Mel Spectrogram



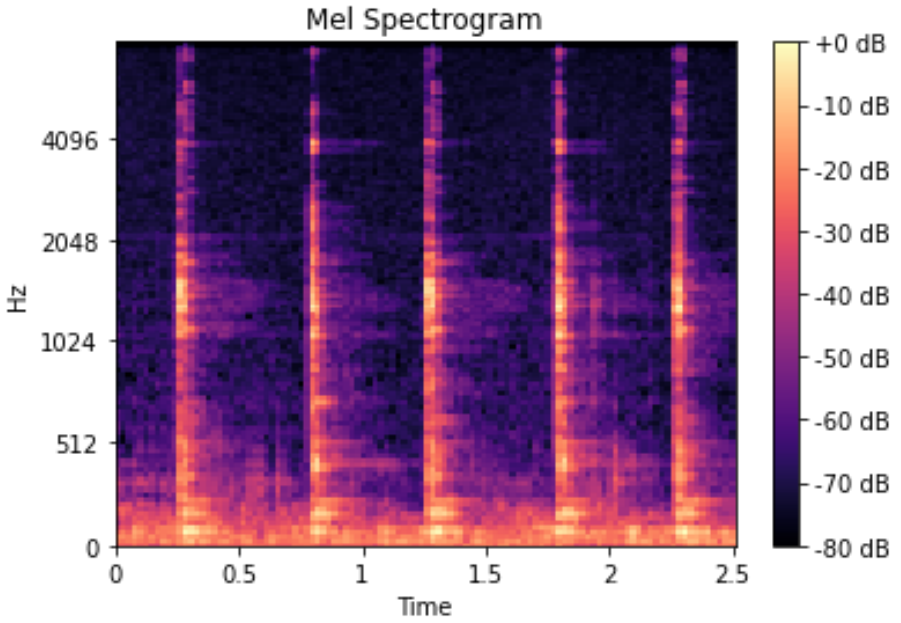


Image Reference: [CSDN博客 - 理解梅尔频谱（mel spectrogram）]

MFCC

- 在語音辨識上被廣泛地使用

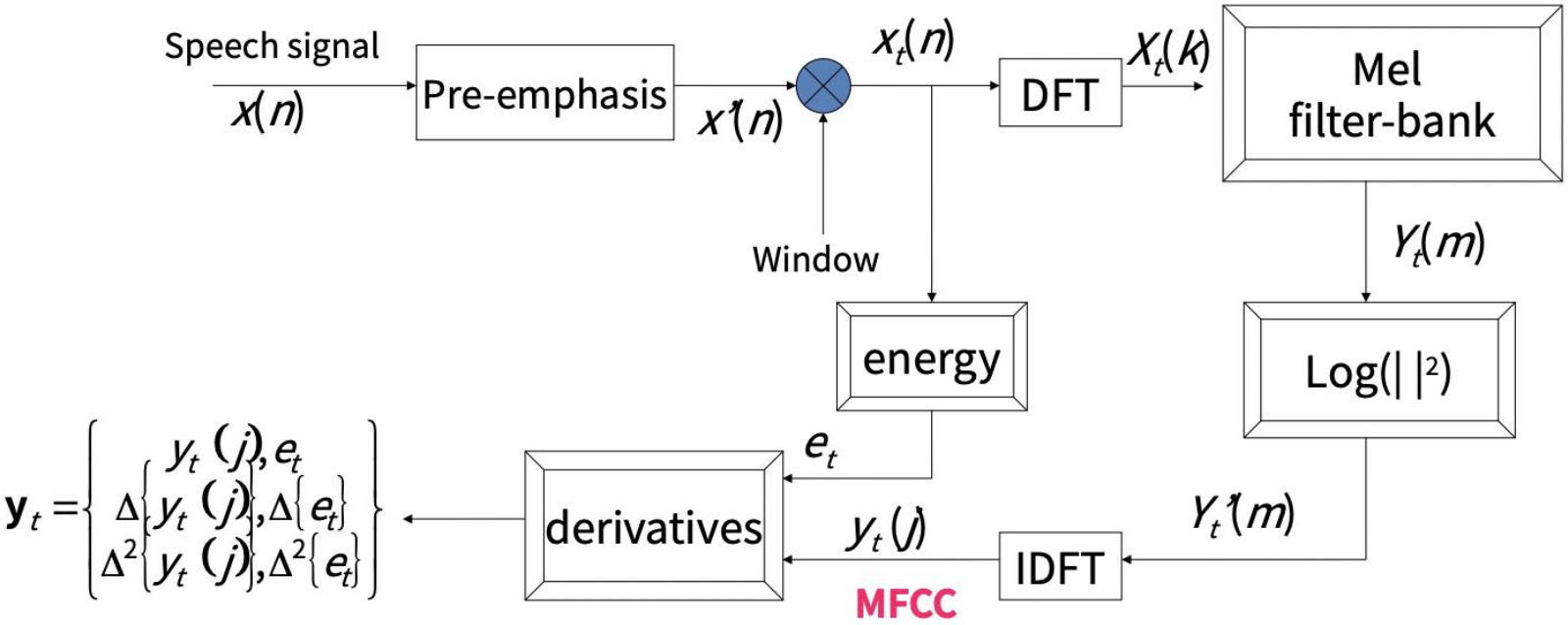


Image Reference: 李琳山老師數位語音處理概論

語音自督導式模型

- 一種能夠抽取語音特徵的深度學習模型

語音自督導式模型的類型

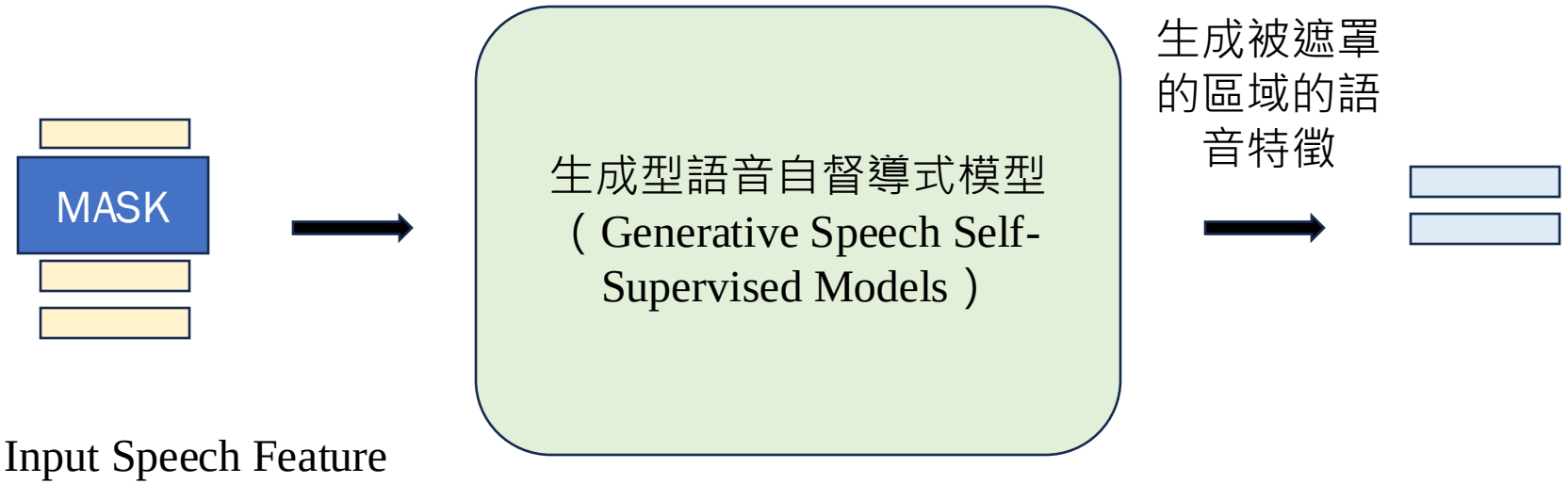
根據預訓練演算法，大致上可以歸納成三類

1. 生成型（Generative）
2. 對比型（Contrastive）
3. 預測型（Predictive）

雖然預訓練方法略有不同，但大致上目標都是「讓深度學習模型在學習預訓練任務的同時，逐漸理解並且抽取語音當中的脈絡信息」

生成型語音自督導式模型

1. 生成型（Generative）：預訓練時，讓模型學習生成輸入當中被遮罩（Masking）的區塊



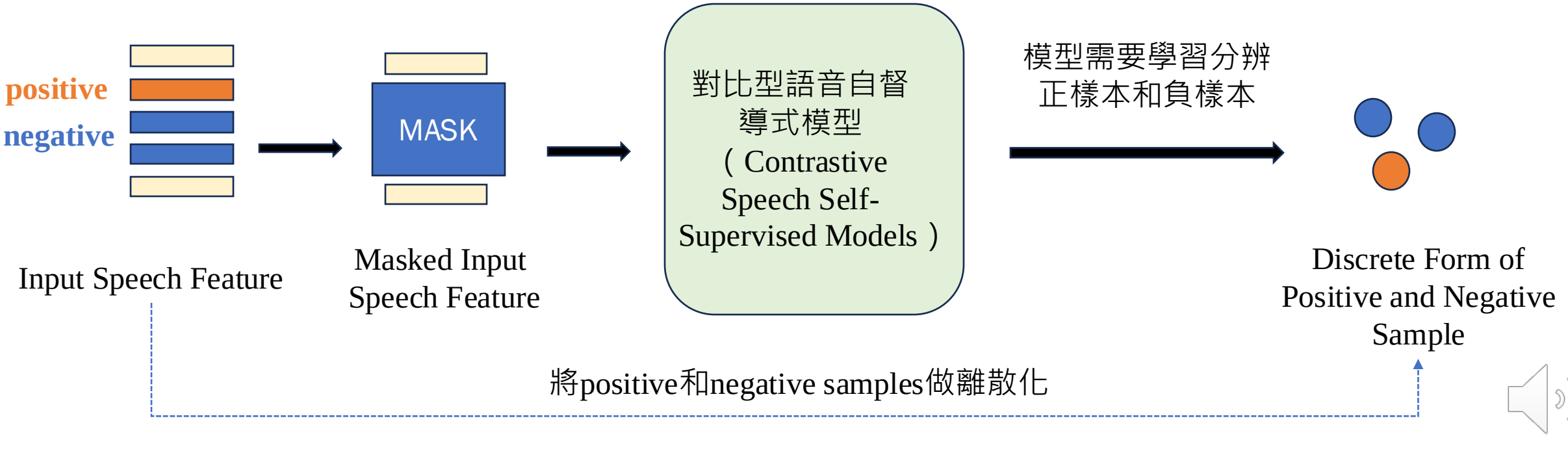
生成型語音自督導式模型

1. 生成型（Generative）：有名的例子是Mockingjay [1], TERA [2], APC [3]等等

- Mockingjay：訓練時對時間軸做隨機遮罩，生成遮罩區塊的輸入語音特徵
- TERA：為Mockingjay的延伸，同時對頻率軸以及時間軸做隨機遮罩
- APC：讓模型預測下一個時間序的輸入語音特徵

對比型語音自督導式模型

2. 對比型（Contrastive）：預訓練時，讓模型學習分辨輸入當中的正樣本（Positive Sample）和負樣本（Negative Sample）



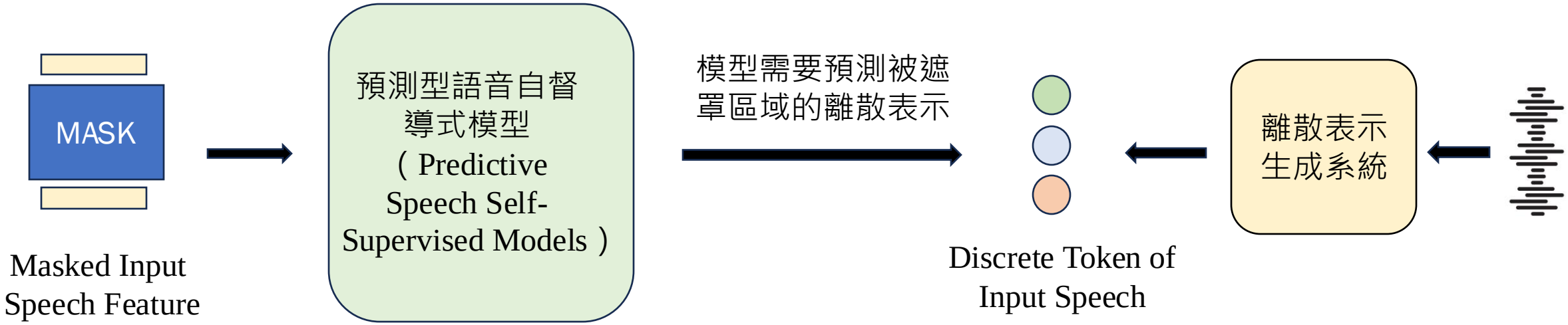
對比型語音自督導式模型

2. 對比型（Contrastive）：有名的例子是Wav2vec [4], Wav2vec 2.0 [5] 等等

- Wav2vec：沒有使用遮罩，正樣本為當前時間序，副樣本為未來時間序
- Wav2vec 2.0：使用遮罩，正樣本為當前時間序，副樣本從其他遮罩時間序做隨機採樣

預測型語音自督導式模型

3. 預測型（Predictive）：預訓練時，讓模型學習預測被遮罩區域的離散表示（Discrete Token）



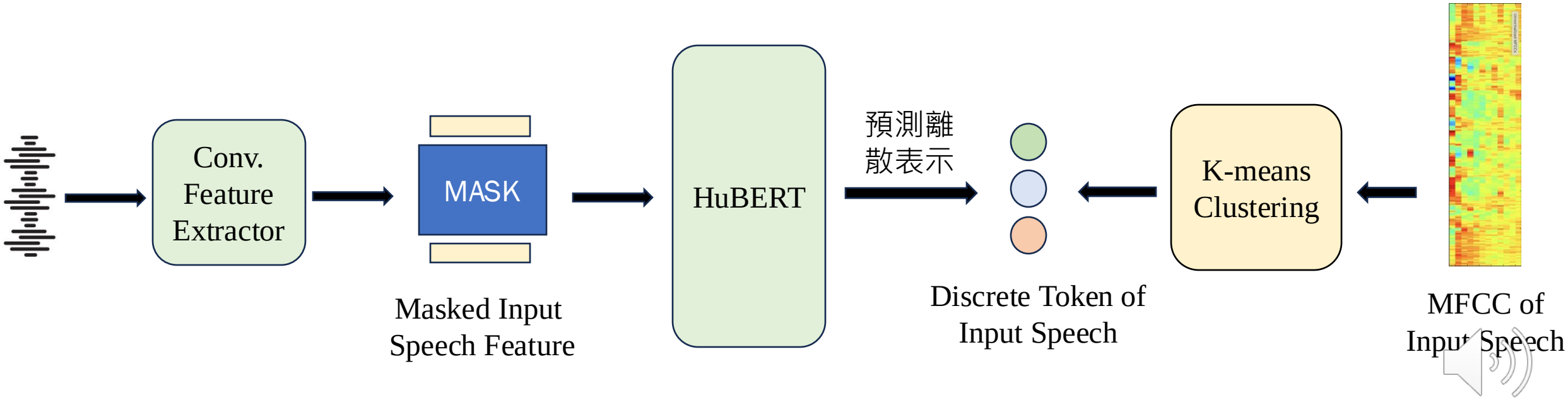
預測型語音自督導式模型

3. 預測型（Predictive）：有名的例子是HuBERT [6]、WavLM [7]等等

- HuBERT：離散表示生成系統為對MFCC特徵做k-means clustering
- WavLM：基於HuBERT做優化，在輸入加上類似data augmentation的東西

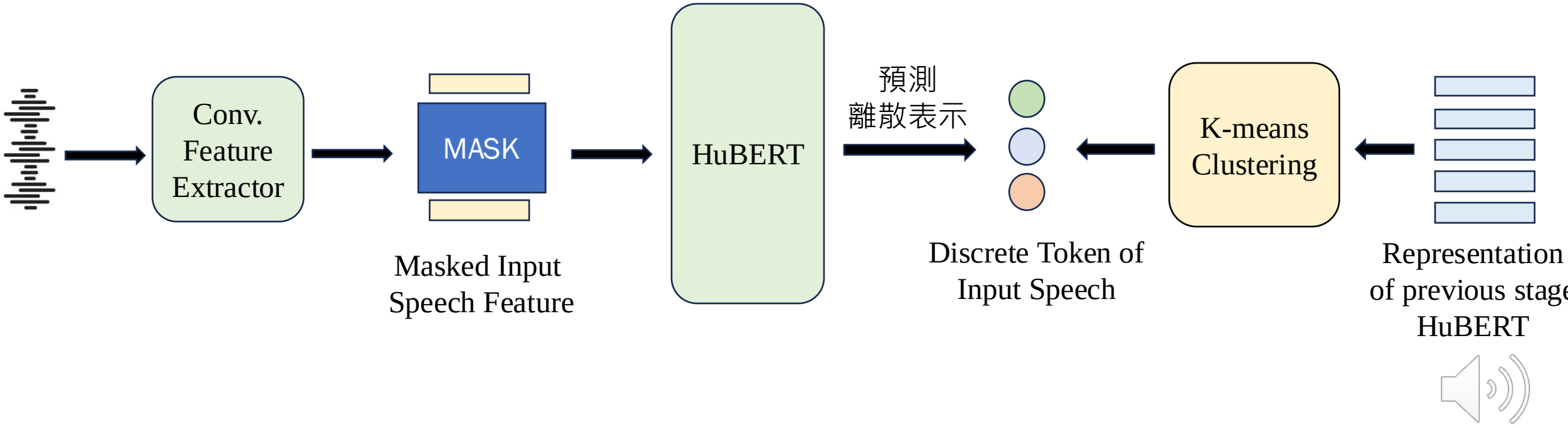
HuBERT

- 因其模型設計簡單明瞭且有效，為現在最主流使用的語音自督導式模型之一
- 以MFCC的離散表示作為目標，進行類似Masked Language Modeling的學習，可以得到比MFCC更好的特徵



HuBERT

- 可以做多階段的學習，學習的目標從第一階段的MFCC的離散表示，替換「前一階段自己所抽出的特徵的離散表示」



使用自督導式特徵來訓練下游任務

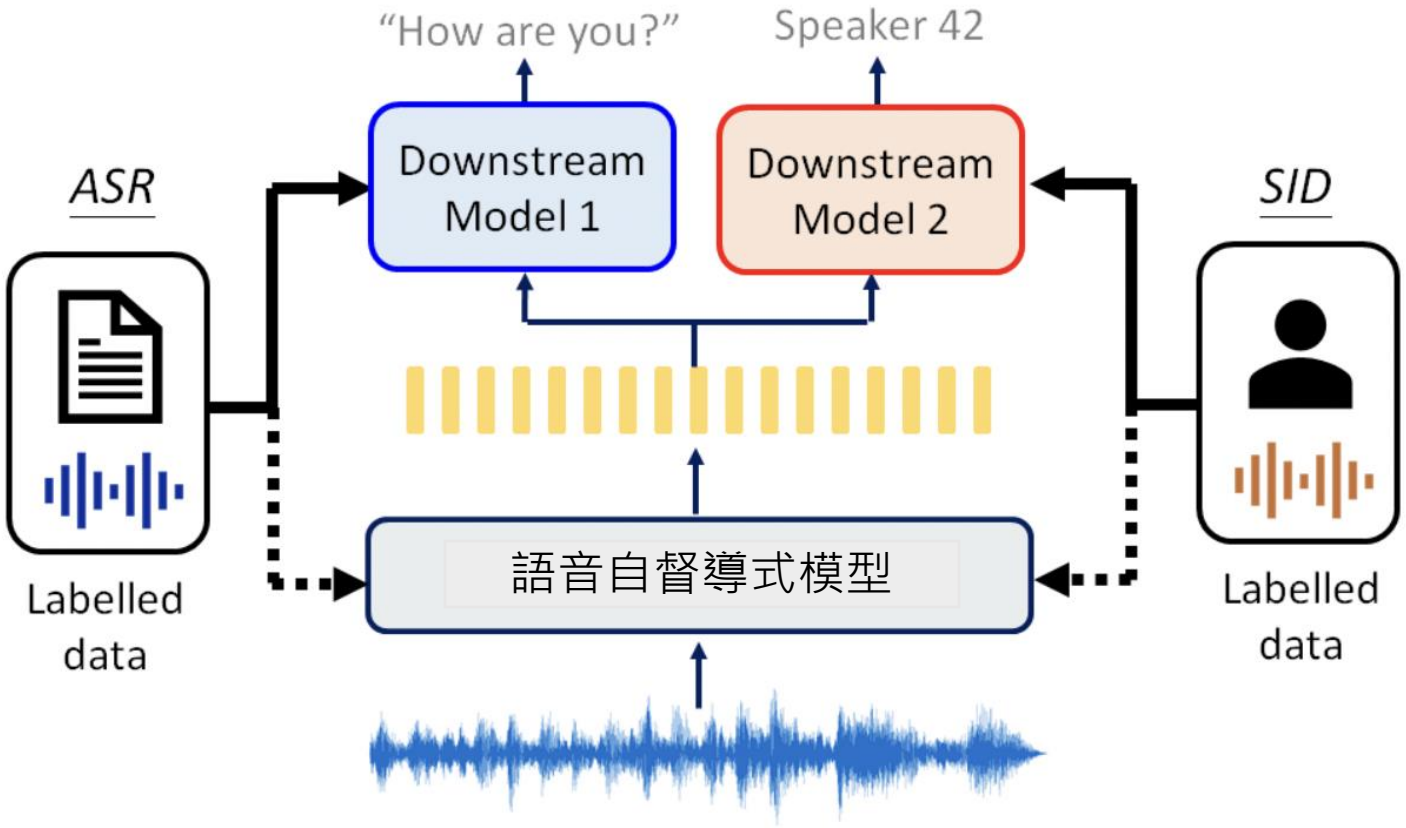


Image Reference: Self-Supervised Speech Representation Learning: A Review

語音特徵表現衡量基準

- 語音特徵表現衡量基準（SUPERB Benchmark）[8]：將自督導式語音特徵作為輸入來訓練下游任務模型，觀察不同任務的表現

Content
Phoneme Recognition
Automatic Speech Recognition
Query by Example
Keyword Spotting

Speaker
Speaker Identification
Speaker Verification
Speaker Diarization

Paralinguistic
Emotion Recognition

Semantic
Intent Classification
Slot Filling

語音特徴表現衡量基準

	PR	KS	IC	SID	ER	ASR (WER)		QbE	SF		ASV	SD
	PER ↓	Acc ↑	Acc ↑	Acc ↑	Acc ↑	w/o ↓	w/ LM ↓	MTWV ↑	F1 ↑	CER ↓	EER ↓	DER ↓
FBANK	82.01	8.63	9.10	8.5E-4	35.39	23.18	15.21	0.0058	69.64	52.94	9.56	10.05
PASE+ [16]	58.87	82.54	29.82	37.99	57.86	25.11	16.62	0.0072	62.14	60.17	11.61	8.68
APC [7]	41.98	91.01	74.69	60.42	59.33	21.28	14.74	0.0310	70.46	50.89	8.56	10.53
VQ-APC [32]	41.08	91.11	74.48	60.15	59.66	21.20	15.21	0.0251	68.53	52.91	8.72	10.45
NPC [33]	43.81	88.96	69.44	55.92	59.08	20.20	13.91	0.0246	72.79	48.44	9.4	9.34
Mockingjay [8]	70.19	83.67	34.33	32.29	50.28	22.82	15.48	6.6E-04	61.59	58.89	11.66	10.54
TERA [9]	49.17	89.48	58.42	57.57	56.27	18.17	12.16	0.0013	67.50	54.17	15.89	9.96
DeCoAR 2.0 [10]	14.93	94.48	90.80	74.42	62.47	13.02	9.07	0.0406	83.28	34.73	7.16	6.59
modified CPC [34]	42.54	91.88	64.09	39.63	60.96	20.18	13.53	0.0326	71.19	49.91	12.86	10.38
wav2vec [12]	31.58	95.59	84.92	56.56	59.79	15.86	11.00	0.0485	76.37	43.71	7.99	9.9
vq-wav2vec [13]	33.48	93.38	85.68	38.80	58.24	17.71	12.80	0.0410	77.68	41.54	10.38	9.93
wav2vec 2.0 Base [14]	5.74	96.23	92.35	75.18	63.43	6.43	4.79	0.0233	88.30	24.77	6.02	6.08
wav2vec 2.0 Large [14]	4.75	**96.66**	95.28	86.14	65.64	3.75	3.10	0.0489	87.11	27.31	5.65	**5.62**
HuBERT Base [35]	5.41	96.30	98.34	81.42	64.92	6.42	4.79	**0.0736**	88.53	25.20	**5.11**	5.88
HuBERT Large [35]	**3.53**	95.29	**98.76**	**90.33**	**67.62**	**3.62**	**2.94**	0.0353	**89.81**	**21.76**	5.98	5.75

Table Reference: SUPERB: Speech processing Universal PERformance Benchmark

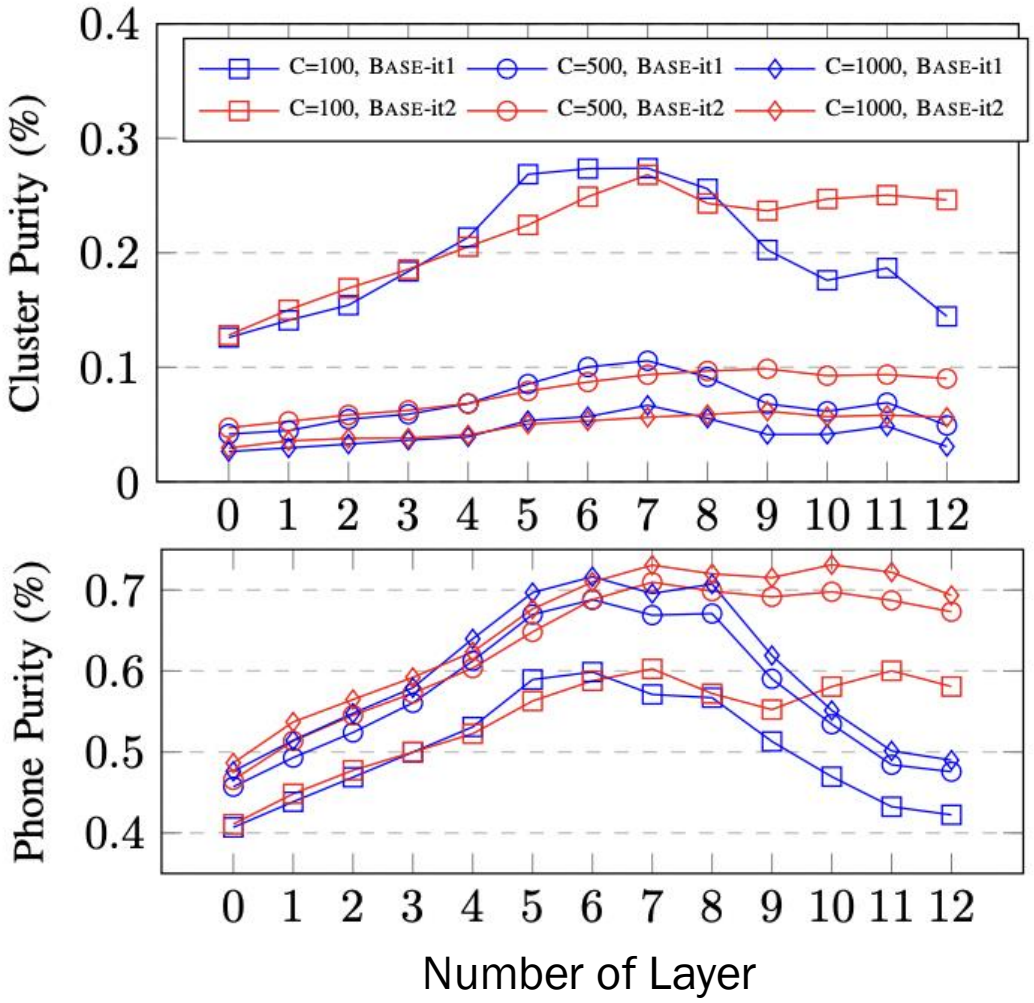
語音特徵表現衡量基準

SUPERB系列還有很多續作

- SUPERB-SG [9]：新增一些生成類型的語音下游任務，例如語音降噪、音源分離和語音翻譯
- SUPERB@SLT 2022 [10]：衡量不同語音自督導模型之模型參數量、運算量，和其表現之間的趨勢變化
- ML-SUPERB [11]：新增多語言的語音下游任務
- AV-SUPERB [12]：新增語音-圖像之多模態下游任務

語音自督導式特徵和音素的關聯性

將語音自督導特徵離散化之後，和音素標籤計算關聯性。Phone purity可以粗略的理解成離散語音自督導特徵和phone label之間的「關聯性」，越接近1代表越相似



語音自督導特徵之語者資訊

語者（Speaker）相關的資訊，大多也存在語音自督導模型較前面層的隱藏特徵裡

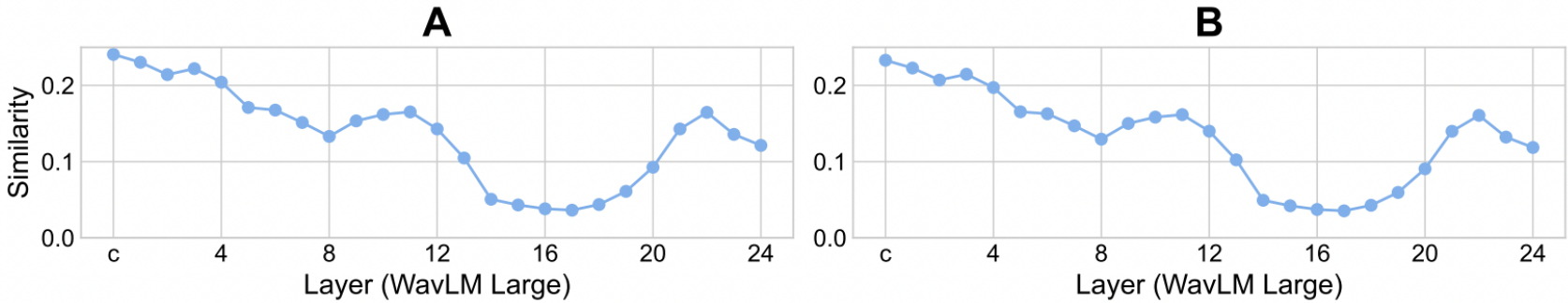


Fig. 6. Similarity results between each layer of WavLM LARGE and last layer of (A) DINO and (B) supervised speaker model.

Image Reference: [What Do Self-Supervised Speech and Speaker Models Learn? New Findings From a Cross Model Layer-Wise Analysis]

語音自督導特徵之語調資訊

語調（Prosody）相關的資訊，大多存在語音自督導模型較前面層的隱藏特徵裡

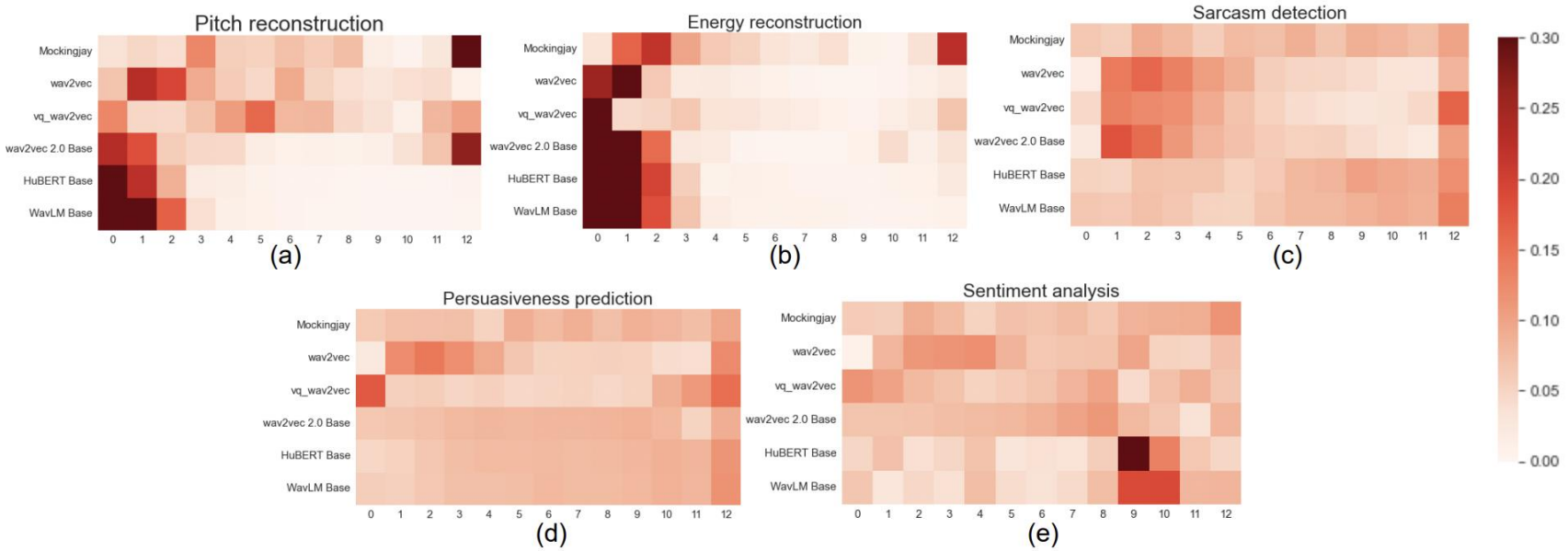
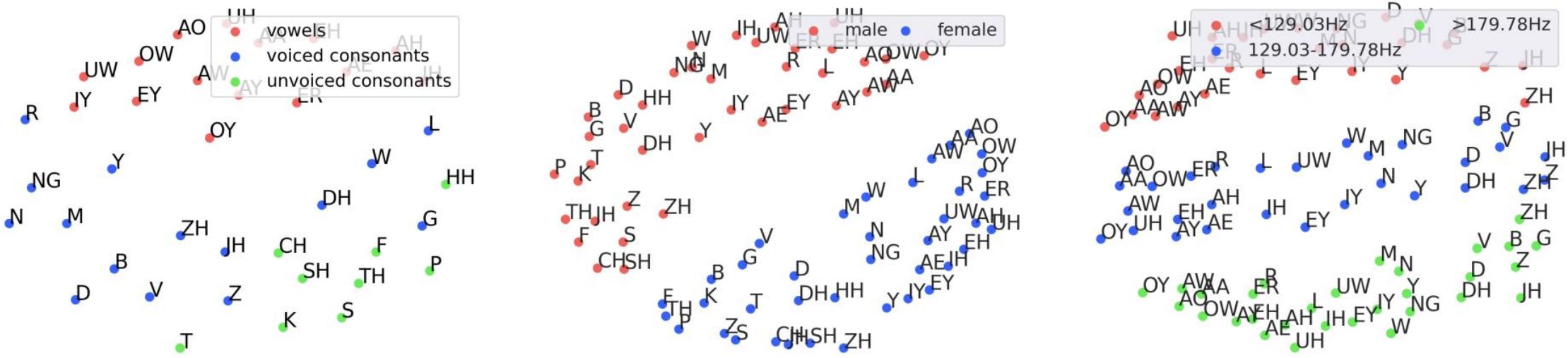


Image Reference: ON THE UTILITY OF SELF-SUPERVISED MODELS FOR PROSODY-RELATED TASKS

語音自督導式模型之可能的運作原理

- 語音自督導式模型當中，會有特定的神經元負責捕捉音素、語者和音高等語音相關之特性



(a) Phone classes (b) Gender (c) Pitch

Image Reference: [Property Neurons in Self-Supervised Speech Transformer]

語音自督導式模型所面臨的問題

- 面對噪音時的強健性（Robustness）不足
- 參數量過大，常見的HuBERT和WavLM高達90M個參數，無法在低運算資源的設備上使用
- Etc.

語音自督導式模型之噪音強健性不足

- 當輸入有噪音時，其抽取出來的特徵在下游任務上的表現下降許多
- 解決方法：Domain Adversarial Training（DAT）。訓練一個domain classifier，目標是分辨語音自督導式特徵是來自乾淨或是噪音的語音。語音自督導式模型的目標則是騙過domain classifier，希望抽出domain-invariant的語音特徵 [13]

語音自督導式模型之噪音強健性不足

- 使用DAT可以提升語音自督導式模型之噪音強健性

	IC (ACC%)		ER (ACC%)		KS (ACC%)	
	clean	noisy	clean	noisy	clean	noisy
Baseline	99.47	96.94	63.96	57.33	97.14	93.38
Baseline+DAT	99.66	99.45	69.95	66.64	96.85	95.42

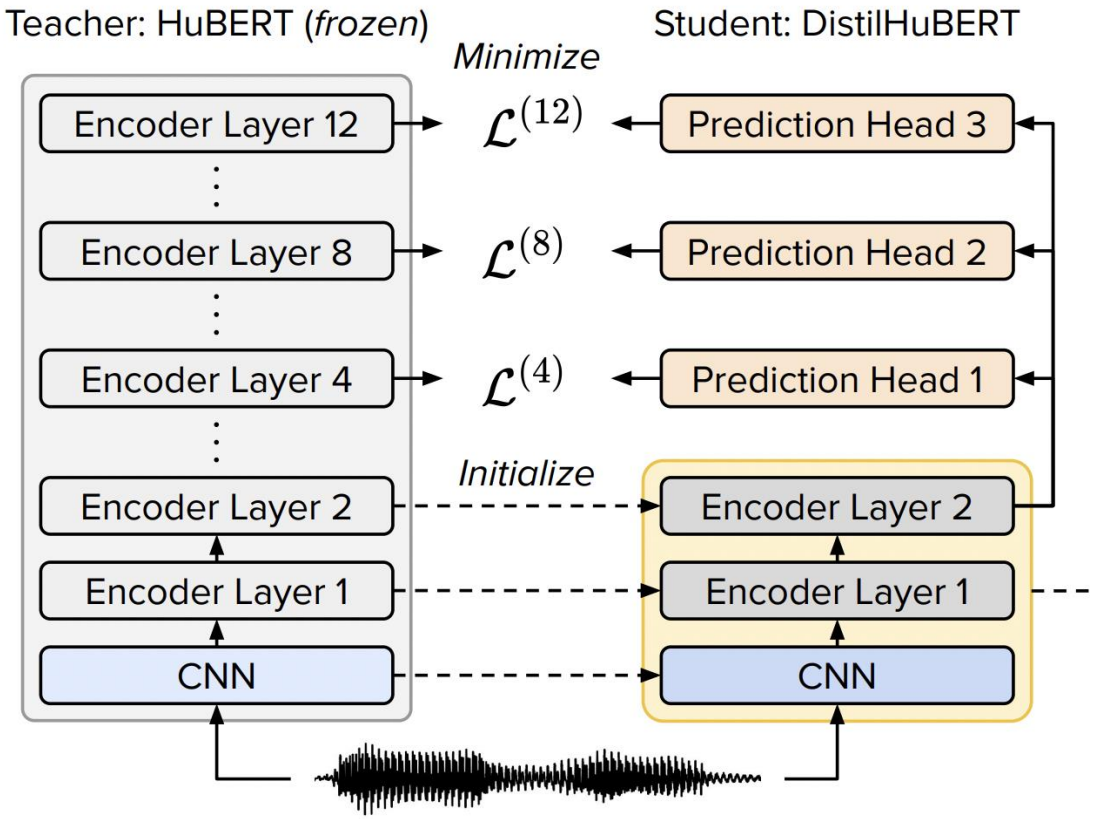
Table Reference: Improving Distortion Robustness of Self-supervised Speech Processing Tasks with Domain Adaptation

語音自督導式模型之參數量過多

- 語音自督導式模型的參數量過大（通常都大於90M），限制了其在低資源設備上之應用
- 解決方法：知識蒸餾、參數裁剪、參數量化等模型壓縮技術

知識蒸餾

- 訓練一個Student語音自督導式模型來「模仿」Teacher語音自督導式模型所抽出之語音特徵 [14]

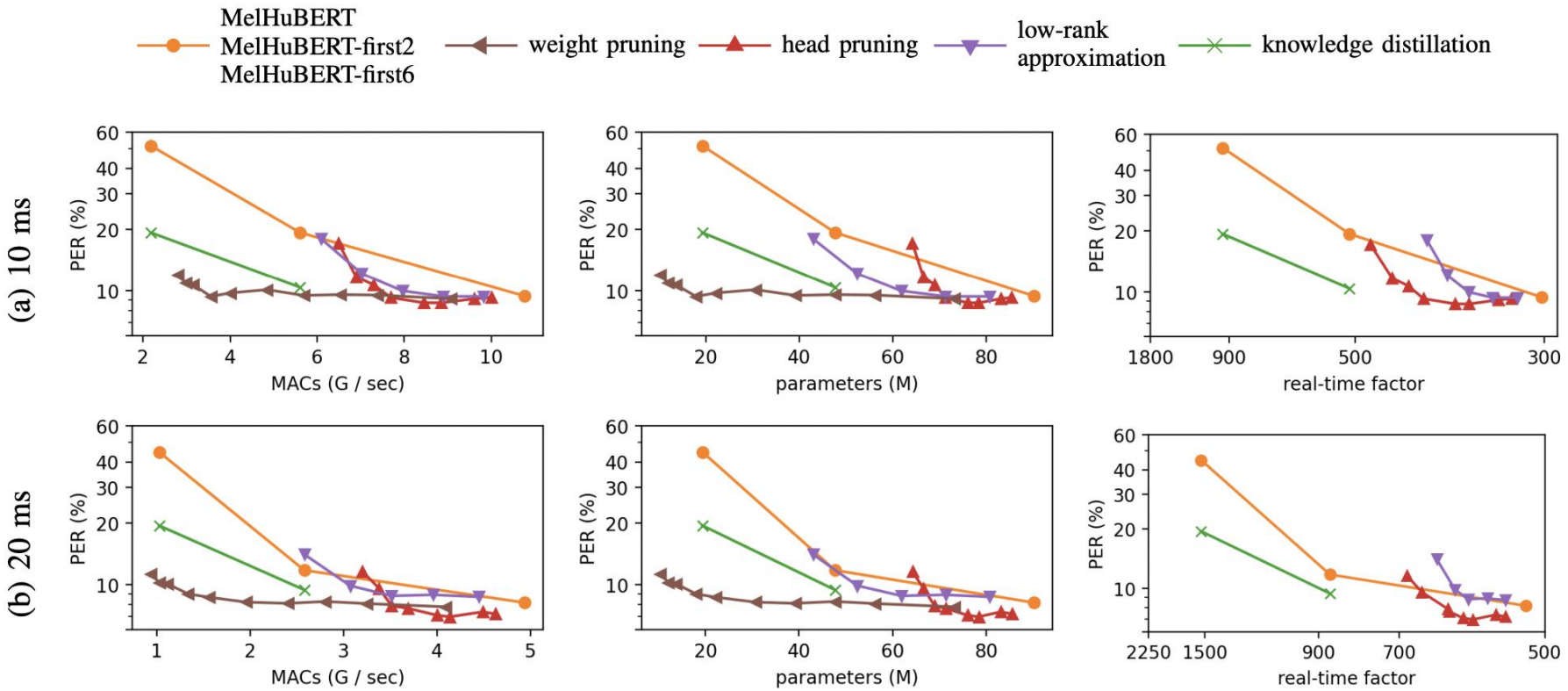


參數裁剪

- 透過L1-范數、Gradient等指標來衡量每個參數的重要性，將相對不重要的參數給設置成零，意即將其「裁剪」掉
- 根據裁剪的單位不同，又可分成權重裁減（Weight Pruning）、注意力頭裁剪（Head Pruning）、前饋層裁剪（FFN Pruning）等等 [15]

參數裁剪

- 不同的裁剪演算法，在不一樣的模型壓縮指標上具有不同的優勢 [15]



總結

- 語音自督導式特徵和傳統的語音特徵相比，其儲存的語音資訊，更容易被下游任務模型給提出，這個特性反映在SUPERB的分數上
- 和傳統語音特徵相比，語音自督導式模型的主要問題在於其龐大的參數量和運算量，大幅地限制了其在低運算資源設備上的應用。即使是如今最好的模型壓縮技術，壓縮率都仍然不夠解決這個問題

實驗模擬程式碼連結

下面的程式碼能夠模擬這個投影片有提到的其中幾個實驗結果

- Property Neurons in Self-Supervised Speech Transformers這篇論文的實驗模擬程式碼連結：https://github.com/nervjack2/PropertyNeurons
- Compressing transformer-based self-supervised models for speech processing這篇論文的實驗模擬程式碼連結：https://github.com/nervjack2/Speech-SSL-Compression

參考資料

[1] Liu, Andy T., et al. "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[2] Liu, Andy T., Shang-Wen Li, and Hung-yi Lee. "Tera: Self-supervised learning of transformer encoder representation for speech." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021): 2351-2366.

[3] Chung, Yu-An, et al. "An unsupervised autoregressive model for speech representation learning." *INTERSPEECH* (2019).

[4] Schneider, Steffen, et al. "wav2vec: Unsupervised pre-training for speech recognition." *arXiv preprint arXiv:1904.05862* (2019).

[5] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in neural information processing systems* 33 (2020): 12449-12460.

[6] Hsu, Wei-Ning, et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." *IEEE/ACM transactions on audio, speech, and language processing* 29 (2021): 3451-3460.

參考資料

[7] Chen, Sanyuan, et al. "Wavlm: Large-scale self-supervised pre-training for full stack speech processing." *IEEE Journal of Selected Topics in Signal Processing* 16.6 (2022): 1505-1518.

[8] Yang, Shu-wen, et al. "Superb: Speech processing universal performance benchmark." *arXiv preprint arXiv:2105.01051* (2021).

[9] Tsai, Hsiang-Sheng, et al. "SUPERB-SG: Enhanced speech processing universal performance benchmark for semantic and generative capabilities." *arXiv preprint arXiv:2203.06849* (2022).

[10] Feng, Tzu-hsun, et al. "Superb@ slt 2022: Challenge on generalization and efficiency of self-supervised speech representation learning." *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023.

[11] Shi, Jiatong, et al. "Ml-superb: Multilingual speech universal performance benchmark." *arXiv preprint arXiv:2305.10615* (2023).

[12] Tseng, Yuan, et al. "Av-superb: A multi-task evaluation benchmark for audio-visual representation models." *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

[13] Huang, Kuan Po, et al. "Improving distortion robustness of self-supervised speech processing tasks with domain adaptation." *arXiv preprint arXiv:2203.16104* (2022).

參考資料

[14] Chang, Heng-Jui, Shu-wen Yang, and Hung-yi Lee. "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert." *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

[15] Lin, Tzu-Quan, et al. "Compressing transformer-based self-supervised models for speech processing." *arXiv preprint arXiv:2211.09949* (2022).