How to Estimate Model Transferability of Pre-Trained Speech Models?

Zih-Ching Chen[1], Chao-Han Huck Yang[2*,3,2], Bo Li[2], Yu Zhang[2] Nanxin Chen[2], Shou-Yiin Chang[2], Rohit Prabhavalkar[2], Hung-yi Lee[1], Tara N. Sainath[2]

[1]National Taiwan University, Taiwan [2]Google, USA [3]Georgia Tech, USA

Overview

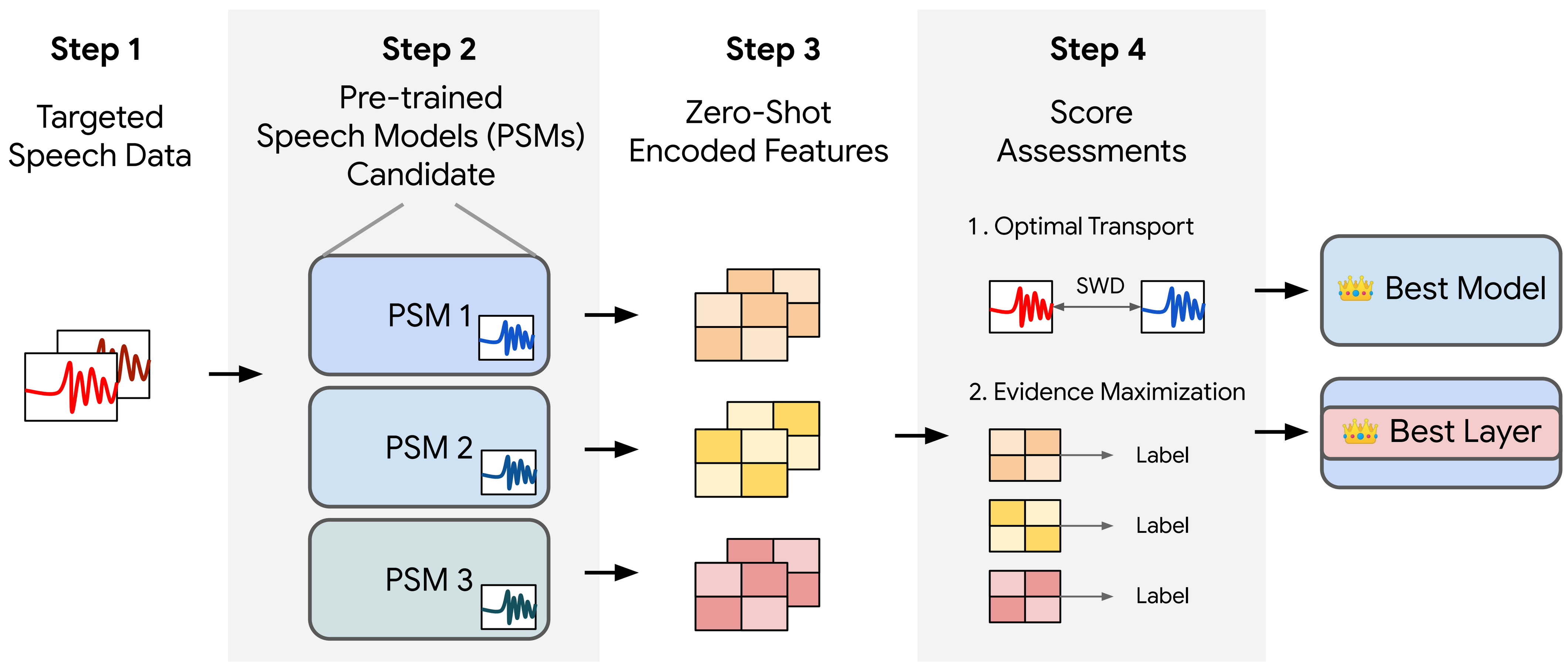


Figure 1:A step-by-step illustration of the proposed framework for providing scores for assessing the best pre-trained model and the best layer for transfer learning with speech data. Our score assessment method can choose the best model candidate and the best layer for fine-tuning.

- Introduce a "**score-based assessment**" framework for evaluating the **transferability** of pre-trained speech models (PSMs).
- Leverage Bayesian and optimal transport theories to **rank** PSM candidates.
- Our method could be used to selection the **best model** and the **best layer** for fine-tuning.
- Validates the efficiency and accuracy of our approach against actual fine-tuning results.

Transferability Estimation for Speech Models

We estimate the correlation between the label sequence and the features extracted from the input sequence

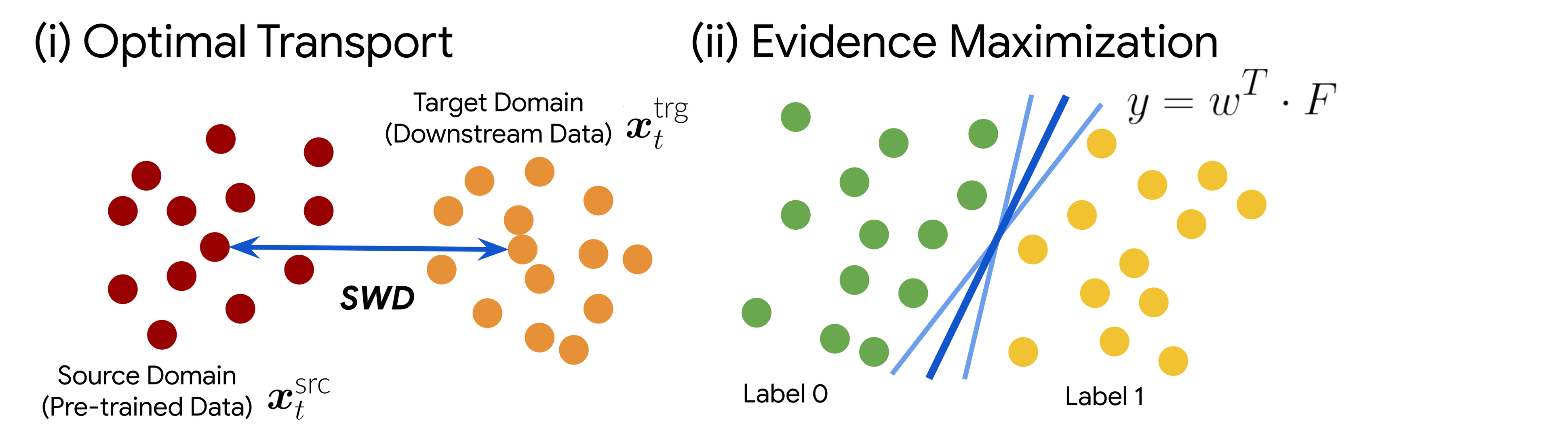
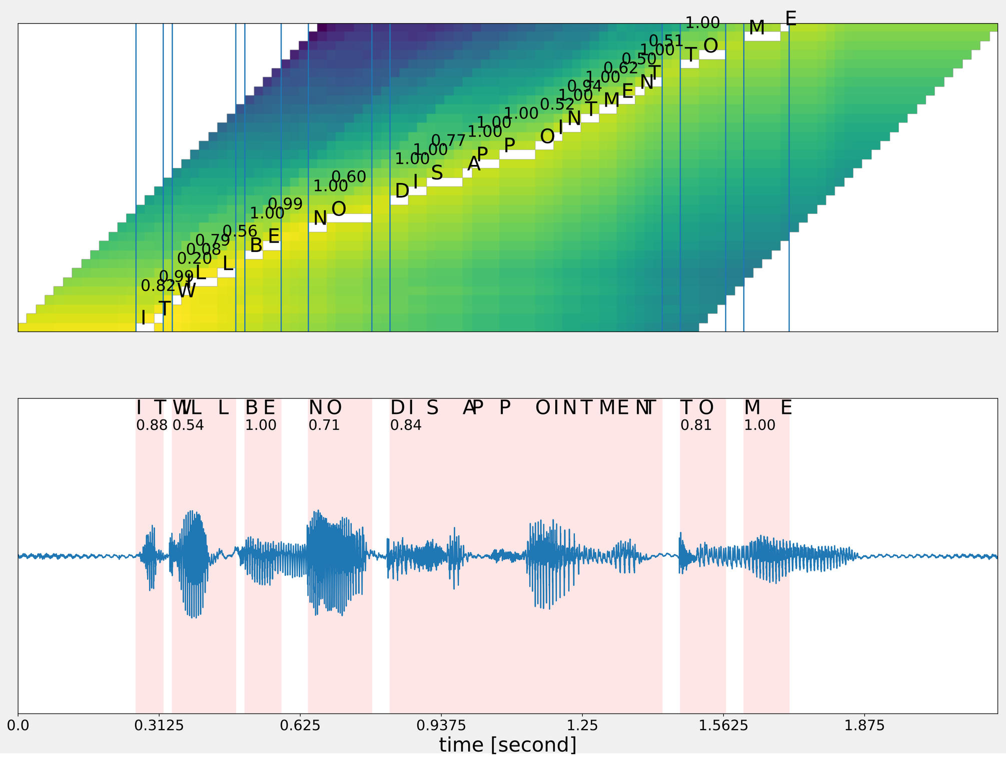


Figure 2:Illustration of the two approaches for estimating transferability in speech processing tasks. The transferability metric for optimial transport is SWD ($\mathcal{W}_p(\boldsymbol{\mu}_t^{\text{src}}, \boldsymbol{\mu}_t^{\text{trg}})$), while for evidence maximization, we use LogME ($\log p(\boldsymbol{y}|F)/n$), where $p(\boldsymbol{y}|F) = \int p(w)p(\boldsymbol{y}|F, w)dw$, to assess transferability.

How do we handle the sequence length mismatch in ASR tasks?

- **Force Alignment:** Corrects input-output misalignments.
- **CTC Alignment:** Utilizes CTC force alignment algorith for the transferability estimation.
- **Backtracking Integration:** Refines alignments by revisiting and adjusting previous frames.



Optimal Transport

Sliced Wasserstein Distance (SWD) [1]

- Sample random batches of pre-trained ($\boldsymbol{x}_t^{\text{src}}$) and downstream inputs ($\boldsymbol{x}_t^{\text{trg}}$) per time step.
- Sliced Wasserstein Distance (SWD) employed.
- Score = $\mathcal{W}_p(\boldsymbol{\mu}_t^{\text{src}}, \boldsymbol{\mu}_t^{\text{trg}})$.

Maximum Evidence

LogME [2] ($\log p(\boldsymbol{y}|F)$)

- LogME measures the suitability of the encoded features for predicting labels.
- It is estimated by mapping the extracted feature F to $\boldsymbol{y}$ using a linear transformation parameterized by w.

$$p(\boldsymbol{y}|F) = \int p(w)p(\boldsymbol{y}|F, w)\, dw$$

- The log-likelihood of the labels are summed to obtain the LogME score for the speech model.

References

[1] S. Kolouri et al. Generalized sliced wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

[2] K. You et al. Logme: Practical assessment of pre-trained models for transfer learning. In *Proc. of ICML*. PMLR, 2021.

Experimental Setup

Continuous ASR

- Pre-trained English-only Conformer model.
- Multilingual Librispeech with 7 languages.

Phoneme and Isolated Word Recognition

- Phoneme Recognition: LibriSpeech dataset.
- Isolated word recognition: Google Speech Commands dataset.

Experiments

Layerwise exploration

Cross-Lingual Adaptation of English-Only Conformer for Multilingual Librispeech

RNN-T Layer	WER (↓)	Rank_{FT}	$\text{Rank}_{\text{tSNE}}$	$\text{Rank}_{\text{LogME}}$	Rank_{SWD}
Conf-01	62.63	17	17	17	17
Conf-02	53.49	15	16	10	11
Conf-03	53.61	16	15	15	10
Conf-04	47.75	14	13	13	14
Conf-05	37.02	11	14	16	13
Conf-06	48.71	13	12	12	16
Conf-07	42.13	12	5	14	15
Conf-08	32.32	10	3	6	8
Conf-09	21.74	7	1	7	9
Conf-10	22.56	8	10	9	4
Conf-11	19.86	4	4	5	6
Conf-12	21.71	6	8	11	12
Conf-13	25.56	9	7	8	7
Conf-14	19.23	3	9	4	5
Conf-15	20.09	5	11	3	2
Conf-16	18.87	2	6	2	3
Conf-17	18.27	1	2	1	1
Spearman's rank correlation coefficient (↑)			0.69	0.87	0.81
p-value (↓)			1×10^{-3}	6×10^{-6}	7×10^{-5}

- **Layer Adaptation:** Superior performance observed when adapting the top layers.
- **Evaluation of Transferability Metrics:** LogME emerges as the most effective.

SSL model adaptation for Phoneme Recognition

HuBERT Layer	PER (↓)	Rank_{FT}	$\text{Rank}_{\text{LogME}}$
Layer-01	35.63	12	9
Layer-02	29.61	9	10
Layer-03	27.51	8	8
Layer-04	25.43	7	7
Layer-05	23.75	6	6
Layer-06	18.83	5	5
Layer-07	14.35	4	4
Layer-08	10.86	3	2
Layer-09	8.73	2	3
Layer-10	7.40	1	1
Layer-11	29.84	10	12
Layer-12	30.37	11	11
Spearman's rank correlation coefficient (↑)			0.94
p-value (↓)			3×10^{-6}

- **SSL Model Transferability:** tSNE and SWD are unsuitable due to the absence of source data.
- **HuBERT vs. RNN-T:** Top layer tuning in HuBERT doesn't always improve performance.

Model-wise exploration

Pre-trained speech model tuned on classification

Models	Para.	Acc. (↑)	Rank_{FT}	$\text{Rank}_{\text{LogME}}$	$\text{Rank}_{\text{tSNE}}$	Rank_{SWD}
†HuBERT	95M	95.94	2	2	1	2
†Wav2Vec2	95M	92.27	5	5	5	5
†DeCoAR2.0	90M	92.63	4	4	4	4
†Vggish	72M	96.78	1	1	2	1
Yamnet	4M	94.32	3	3	3	3
fsFCNN	20M	91.34	6	6	6	6
Time†			~ 0.61Day	24.09s	28.01s	10.86s

- Evaluated transferability scores for speech classification tasks.
- **LogME and SWD Effectiveness:** Accurately estimates transferability in speech classification in minimal time.