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

Zih-Ching Chen[1], Chao-Han Huck Yang[2,3], 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

{r09942176, hungyilee}@ntu.edu.tw; huckiyang@gatech.edu; {boboli,ngyuzh}@google.com

Abstract

In this work, we introduce a "score-based assessment" framework for estimating the transferability of pre-trained speech models (PSMs) for fine-tuning target tasks. We leverage upon two representation theories, Bayesian likelihood estimation and optimal transport, to generate rank scores for the PSM candidates using the extracted representations. Our framework efficiently computes transferability scores without actual fine-tuning of candidate models or layers by making a temporal independent hypothesis. We evaluate some popular supervised speech models (e.g., Conformer RNN-Transducer) and self-supervised speech models (e.g., HuBERT) in cross-layer and cross-model settings using public data. Experimental results show a high Spearman's rank correlation and low p-value between our estimation framework and fine-tuning ground truth. Our proposed transferability framework requires less computational time and resources, making it a resource-saving and time-efficient approach for tuning speech foundation models.

Index Terms: Pre-trained speech models, transfer learning, model transferability, and foundation speech models

1. Introduction

In recent years, large-scale pre-trained neural networks, also known as Foundation Models [1, 2, 3, 4], have demonstrated numerous benefits in various fields. One of these benefits includes the application of supervised learning [5] or self-supervised learning (SSL) models [6, 7, 8] trained on few-shot adaptation [9] or continuous learning [10] from pre-training. FMs have shown strong generalization and zero-shot learning abilities [11, 12] by learning representations, making them a popular choice for speech processing tasks [2, 3, 13]. These representations can then be utilized by downstream models for specific task-related feature extraction [14]. However, fine-tuning separate FMs for multiple downstream tasks can be computationally expensive and resource-intensive due to the size of the FMs. Even in partial model tuning settings, determining where and how to insert parameter-efficient learning modules such as residual adapters [15, 16] and neural reprogramming [17] relies heavily on hand-crafted expertise.

In this work, we aim to address the aforementioned challenges of determining the best FM or candidate layer for tuning by *introducing a model transfer assessment framework for speech processing* applications. As illustrated in Figure 1, our process consists of four steps: step 1 and step 2 involve collecting target speech data and pre-trained speech models (PSMs), step 3 involves using frozen PSMs to extract features, and step 4 involves using score-based assessment methods to determine

---

*Work was mainly done during a research intern at Google.

the best model for the targeted task or the best layer for partial tuning or insertion of parameter-efficient modules.
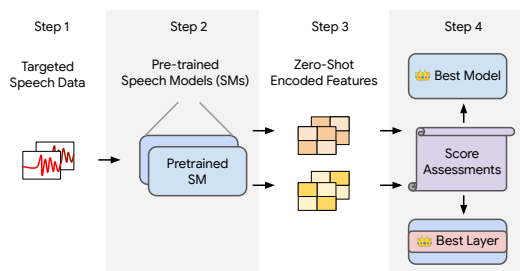

Figure 1: *A step-by-step illustration of the proposed framework for providing scores for assessing the best pre-trained model and the best layer for transfer learning with speech data.*

More specifically, our model assessment framework is based on two theoretical approaches: (i) optimal transport with latent space measurement [17] and (ii) maximum evidence [18] with Bayesian probabilistic graphical models (PGMs). Since both of these theoretical approaches have not been extensively explored in speech processing, we make a very first attempt to approximately model continuous speech recognition and isolated word recognition as classification problems by establishing a temporal independent hypothesis [19] (TIH). In the following paragraphs, we first introduce recent work on model transferability estimation in non-speech processing tasks and then explain how the TIH connects to the two theoretical backbones, even in a simplified and non-autoregressive decoding setting.

Related Pre-Trained Model Transferability Works: One of the earliest works on pre-trained model transferability estimation in computer vision tasks was the use of negative conditional entropy (NCE) [20]. NCE uses a pre-trained vision model and evaluates conditional entropy between target pseudo labels, which are the assigned label of the source model and real target labels. A subsequent work to NCE is the log expected empirical predictor (LEEP) [21], which modifies NCE by using soft predictions from the source model. LEEP calculates the log-likelihood between the target labels and the predictions to estimate model transferability, which allows the input data to come from arbitrarily different distributions. However, when evaluating the quality of model transferability estimation, both NCE and LEEP have been reported with correlation coefficient ≤ 0.6 with high p-value, as noted in [18].

On the other hand, a recent model transferability assessment solution for vision and language processing tasks is LogME [18]. LogME predicts accuracy on the target task by estimating the marginalized likelihood of labeled target exam-

ples, under the premise that a linear classifier is added on top of the pre-trained model. LogME can be applied to both supervised and unsupervised learning tasks for cross-task adaptation, as it considers zero-shot features encoded by pre-trained models.

In recent works [22, 23, 24], LogME has been shown to be an appropriate first step for encoder selection in *natural language processing* [22]. Preliminary attempts in [23] also showed high correlation results when using a modified version of LogME for ensemble selection of large language models. Meanwhile, to the best of the authors' knowledge, there are no studies of using the aforementioned log-likelihood methods to analyze speech processing-based tasks from large-scale pre-trained FM(s) to adaptation. While classification scores have been used as evaluation metrics to assess deep learning models [21, 18], they may not fully capture the performance of speech models due to the unique characteristics of speech signals. The sequential nature of speech signals requires modeling time-step dependencies, which cannot be directly assessed by classification metrics. Next, we review some related speech processing methods and empirical studies to highlight the differences between existing methods and the proposed model assessment framework in speech processing.

Pre-trained Speech Model (PSM) Selection Works: In the realm of neural network-based PSM for various speech application tasks, previous research has explored the possibility of selectively reusing established PSMs by only fine-tuning certain trainable components, such as the acoustic encoder or language decoder [14]. For instance, prior research [25] suggests that tuning only selected encoders of RNN-Transducer [26] closest to the *input layer* of PSM results in improved performance compared to fully fine-tuning. In contrast, *Pasad et al.* [27] discovered that re-initializing the transformer layers closest to the *prediction head* outperforms initializing all layers from a pre-trained SSL-based PSM. Given that different PSMs feature distinct encoder-decoder architectures that incorporate both acoustic and language information, the selection of an appropriate fine-tuning configuration can be task-dependent and model-specific, requiring heuristic search. In other words, training a model using every possible encoder or decoder has shown an extremely large sample complexity and a significant amount of time [28, 22, 29], making it difficult to reproduce for researchers with limited computational resources.

A preliminary attempt to estimate pre-trained model transferability was proposed in [17]. This attempt is based on optimal transport and measures the distance between a target domain and its source domain in terms of population risk, as its difficulty model transferability. However, their work was focused on *selecting* PSMs for *cross-modal* adaptation to classify sensor data, and its effectiveness in speech processing tasks has not been studied. In this work, we also advance the use of optimal transport [17, 30] for estimating PSMs and provide some initial explorations for speech processing, which is based on a proposed temporal independent condition discussed in Section 2.1.

The recent advent of FMs and their breakthroughs in speech processing tasks [31, 7, 32, 8, 2] has created a growing need for new transferability estimation techniques evaluating the strengths of FMs. But speech tasks have yet to be extensively investigated with transferability estimation techniques. In this work, we aim to answer the question of how well we can estimate the transferability of pre-trained speech models to specific speech recognition tasks. We establish baselines for layer-wise and model-wise evaluations. The proposed framework could serve as one first attempt to evaluate pre-trained speech models, which has some potential to enhance the development of model tuning for future studies.

Our contributions include:

- We connect model transferability estimation to speech tasks by leveraging a simplified hypothesis of temporal independence to relaxes the posterior nature of speech recognition.
- We advance two different perspectives on model transferability for speech processing: (i) optimal transport and (ii) evidence maximization, to provide interpretable scores for assessing PSMs.
- By conducting initial attempts in cross-layer and cross-model transfer learning setups, we evaluate that our framework [1] has a high rank correlation and low p-value.

2. Transferability for Speech Models

This study proposes two perspectives on evaluating model transferability in speech processing. We explore the use of optimal transport and evidence maximization as methods for generating interpretable scores to assess the performance of speech processing models (PSMs). To enable the evaluation of model transferability, we introduce a simplified hypothesis of temporal independence in Section 2.1, which relax the posterior nature of speech processing in this section and enables evaluating model transferability in sequential speech data. In Section 2.2, we incorporate TIH into the optimal transport sliced Wasserstein Distance, which is utilized for measuring the distance between source and target data distributions. In Section 2.3, we apply TIH in the likelihood aspect, LogME, a technique that models the relationship between extracted features of a speech signal and the output labels.

2.1. Modeling Temporal Independent Hypothesis (TIH)

To estimate the transferability for speech processing asks, we estimate the correlation between the label sequence and the features extracted from the input sequence. Since in speech processing tasks, the length of the input sequence and the label are not aligned, a temporal independent hypothesis was introduced in the connectionist temporal classification (CTC) mechanism [19] for automatic speech recognition (ASR) modeling. It is based on a simplified condition that ignores posterior information during the loss computation. This simplified condition assumes that a speech model can learn sequential information through its neural network encoders, such as attention or recurrent networks.

Let $f_i \in \mathbb{R}^D$ be the i-th feature extracted with D dimensions by a pre-trained model ϕ for the input x, and let $y_i \in \mathbb{Z}^+$ be the scalar label. The collection of all features is represented by a matrix $F \in \mathbb{R}^{D \times n}$, and the collection of all labels is represented by the vector $y \in \mathbb{Z}^n$:

$$p(\boldsymbol{y}|F) = \sum_{A \in A_F} \prod_{t=1}^{T} p_t(a_t|F), \qquad (1)$$

where A is defined as the set of all valid alignments between the input x and the output y, and $P(\boldsymbol{a} \mid \boldsymbol{x})$ is the probability of a specific alignment $\boldsymbol{a}$ given the input $\boldsymbol{x}$. The probability density $p(\boldsymbol{y}|F)$ measures the compatibility between the feature matrix F and the label sequence $\boldsymbol{y}$. It is calculated by summing

---

[1]Code is available at https://github.com/virginiakm1988/LogME-CTC

over all possible alignments A_F between the input and output sequences, using Equation 1. This equation multiplies the probability of the alignment at each time step t, given the feature matrix F and the label sequence $\boldsymbol{y}$.

Forced Alignment: In the case when the input sequence and the output label are not aligned, we estimate the transferability of pre-trained models using the CTC forced alignment algorithm by aligning the model's output with the reference transcriptions of the target task. Our forced alignment process includes backtracking , which enables the correction of misaligned frames by considering previous frames and making necessary adjustments to the current alignment.
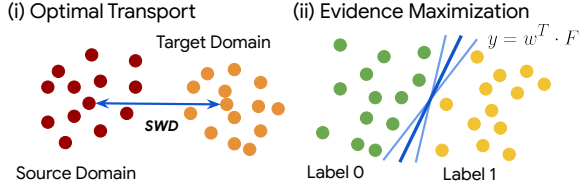

Figure 2: *Illustration of the two approaches for estimating transferability in speech processing tasks: optimal transport (Section 2.2) and maximum evidence (Section 2.3). The transferability metric in 2.2 is SWD [17], while in 2.3, we use LogME* $(\log p(\boldsymbol{y}|F)/n)$, *where* $p(\boldsymbol{y}|F) = \int p(w)p(\boldsymbol{y}|F,w)dw$, *to assess transferability.*

2.2. Optimal Transport by Sliced Wasserstein Distance

In speech processing tasks, the output label and input sequence may not align, directly calculating the median distance of the data distribution is infeasible. So here, we introduce optimal transport with TIH. Building upon Eq. (1), we conduct latent space measurement by sampling a random batch of source ($\boldsymbol{x}_t^{\text{src}}$) and target inputs ($\boldsymbol{x}_t^{\text{trg}}$) for time step t. We model character-level prediction using the TIH and the latent distribution of $\boldsymbol{\mu}_t^{\text{src}}$ and $\boldsymbol{\mu}_t^{\text{trg}}$, which represent the zero-shot latent representation. We measure the distance between these two probability distributions, $\boldsymbol{\mu}_t^{\text{src}}$ and $\boldsymbol{\mu}_t^{\text{trg}}$, using the sliced Wasserstein distance [33] (SWD), $\mathcal{W}_p(\boldsymbol{\mu}_t^{\text{src}}, \boldsymbol{\mu}_t^{\text{trg}})$, under L_p-norm, where $p = 1$ is used to obtain a closed form solution based on the theory presented in [17]. A larger SWD value (or score) indicates a greater difficulty in aligning the latent representation between the two selected domains and serves as a measure of the difficulty of model adaptation. In practice, we calculate the median SWD value over the total number of time steps, T, and report this value as its transferability score. However, this TIH-based SWD score **cannot** be used directly for evaluating SSL tasks as the unsupervised pre-trained data and target tasks come from the same domain (e.g., Librispeech [34]). In the next section, we investigate a more general framework for estimating the transferability of PSMs.

2.3. Transferability Estimation by Likelihood

We illustrate a connection between the aforementioned TIH and likelihood-based estimation in LogME [18] by computing the probability of the target labels conditioned on the source features. LogME measures the suitability of the encoded features for predicting labels via a probability density, which is estimated by mapping F to $\boldsymbol{y}$ using a linear transformation parameterized by w. The goal is to find the optimal weight w^*:

$$p(\boldsymbol{y}|F) \rightarrow p(\boldsymbol{y}|F, w^*) \rightarrow \int p(w)p(\boldsymbol{y}|F,w)dw, \quad (2)$$

where Eq. (2) denotes a proxy for feature suitability.

Next, the probability densities of w and $p(\boldsymbol{y}|F, w)$ are modeled using positive hyper-parameters $\alpha \in \mathbb{R}^+$ and $\beta \in \mathbb{R}^+$. The prior distribution of the weights is modeled as an isotropic multivariate Gaussian $w \sim \mathcal{N}(0, \alpha^{-1}I)$, and each observation's distribution is modeled as a one-dimensional normal distribution $p(\boldsymbol{y}_i|f_i, w, \beta) \sim \mathcal{N}(\boldsymbol{y}_i|w^T f_i, \beta^{-1})$, which is computed by:

$$p(\boldsymbol{y}|F, \alpha, \beta) = \int p(w|\alpha)p(\boldsymbol{y}|F, w, \beta)dw \quad (3)$$

The logarithm maximum evidence $\mathcal{L}(\alpha^*, \beta^*) = \log p(\boldsymbol{y}|F, \alpha^*, \beta^*)$ is used to evaluate the compatibility between features and labels. $\mathcal{L}(\alpha^*, \beta^*)/n$ is used as our LogME evaluation metric.

Based on Equations (1) and (3), we propose a new formulation of LogME for speech processing tasks. The model is parameterized by hyper-parameters α and β and is given by the following equation:

$$p(\boldsymbol{y}|F, \alpha, \beta) \rightarrow \sum_{A \in A_F} \prod_{t=1}^{T} \int p_t(w_t|\alpha)p(a_t|F, w_t, \beta)dw_t.$$
$$(4)$$

We compute the log-likelihood of the labels in the aligned sequence and sum them to obtain the LogME score for the speech model. It is important to note that this revised LogME measurement can be used to estimate both supervised speech model training with labeled data and SSL models, without a need for source data.

3. Experiments

In our speech task experiments, we conduct layer-wise exploration to estimate the transferability of pre-trained models.

3.1. Experimental setup

Continuous ASR: Our experimental setup follows the same training procedure as reported in [35] using the official test data. We conduct ASR adaptation experiments using a pre-trained English-only Conformer model [36]. We evaluate the average word error rate (WER) on the Multilingual Librispeech (MLS) benchmark, which includes seven languages [35, 1]. The supervised Conformer-ASR model was trained with a batch size of 1024 on 64 TPUs following the same setup as in [35]. **Phoneme and Isolated Word Recognition**: We evaluated the performance of our models on the LibriSpeech dataset using the train-clean-100, dev-clean, and test-clean subsets. The evaluation metric was Phone Error Rate (PER) in Table 2. For isolated word recognition tasks, we used the Speech Commands dataset v1.0 [37], which includes ten keyword classes, silence, and an unknown class. The evaluation metric for this task was accuracy (ACC) in Tables 3 and 4. In the SSL setup, the raw waveform was used as input, and a linear head was added to the upstream SSL model for phoneme recognition and speech command tasks. The SSL models were trained on a single RTX3090 GPU using the experimental settings from the SUPERB benchmark [14]. **A tSNE-based Baseline Method:** The tSNE-based clustering [38] has been commonly used for model interpretability. We also calculate the median tSNE clustering points of the unpaired source ($\boldsymbol{x}_t^{\text{src}}$) and target input ($\boldsymbol{x}_t^{\text{trg}}$) distributions. As an extra baseline for transferability estimation, this tSNE score is similar to the SWD score, where a large distance indicates greater difficulty in aligning the two different representations.

3.2. Study 1: Layer-Wise Exploration

Table 1: *Fine-Tuning pre-trained English-only Conformer for Multilingual Librispeech (MLS) [39] for cross-lingual adaptation, which WER is reported as averaged of seven MLS languages as the setup in [35].*

RNN-T Layer	WER ($\downarrow$)	Rank$_{FT}$	Rank$_{tSNE}$	Rank$_{LogME}$	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 ($\uparrow$)		0.69	0.87	0.81
p-value ($\downarrow$)		1×10^{-3}	6×10^{-6}	7×10^{-5}

Table 2: *HuBERT-based SSL PSM layer-wise fine-tuning results*

HuBERT Layer	PER ($\downarrow$)	Rank$_{FT}$	Rank$_{LogME}$
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 ($\uparrow$)			0.94
p-value ($\downarrow$)			3×10^{-6}

Transferability estimation can aid in determining the optimal layer for fine-tuning pre-trained models. In Study 1, we performed a layer-wise analysis to investigate the fine-tuning of a pre-trained English-only Conformer for MLS cross-lingual adaptation [39], as detailed in Table 1. Our results suggest that fine-tuning the top layers achieves better performance. Using LogME, we evaluated the performance of fine-tuning each layer, obtaining a correlation of 0.87 with a p-value of 6×10^{-6}. Additionally, SWD accurately estimated the ranking with a correlation of 0.81 and a p-value of 7×10^{-5}. However, direct computation of distance with tSNE was found to be less precise. Our experimental results demonstrate that layer-wise analysis using the LogME score is a more accurate method for estimating transferability in speech processing tasks, which provide insights into selecting the optimal layer for fine-tuning in cross-lingual ASR transfer.

In SSL models, tSNE and SWD cannot be used to estimate transferability as they rely on source data, which is not available in this task. To address this issue, we performed a layer-wise analysis of a HuBERT base model [8] for phoneme recognition. Our results, presented in Table 2, indicate that the LogME score is a precise method for evaluating transferability in speech processing tasks. Our findings also suggest that fine-tuning the

top layers does not necessarily result in improved performance, which differs from the results of the RNN-T model presented in Table 1. Therefore, estimating transferability for individual layers is essential, as it can save time and computational costs and eliminate the need for expert knowledge in selecting the optimal layers for fine-tuning.

3.3. Study 2: Model-Wise Exploration

3.3.1. Pre-trained speech model tuned on classification

For model-wise exploration, we conducted experiments to explore the transferability of pretrained models tuned on the classification tasks. To this end, we evaluated six different pre-trained models and computed their transferability scores using various evaluation metrics. As presented in Table 3, all of the evaluation metrics produced accurate estimates of the transferability scores. Our experimental results thus demonstrate that the LogME score can effectively capture the transferability of a pre-trained model in speech classification tasks.

3.3.2. Different pre-train data on same neural architecture

We evaluated transferability metrics on the AST model [40] using three pre-trained datasets: frozen +Linear Probing, ImageNet [41], and AudioSet [42], to examine the effectiveness of transferability metrics evaluation on different pre-training data. Only the last MLP layer was fine-tuned during the experiments. The results in Table 4 indicate that the LogME score effectively estimates the transferability of pre-trained models in speech classification tasks. The model pre-trained on AudioSet has the best transferability, whereas the model with random initialization showed the worst transferability. These findings underscore the significance of selecting appropriate pre-training data to enhance model transferability.

Table 3: *Evaluation on GSC-v1 [37] on different PSMs.*

Models	Para.	Acc. ($\uparrow$)	Rank$_{FT}$	Rank$_{LogME}$	Rank$_{tSNE}$	Rank$_{SWD}$
[†]HuBERT [8]	95M	95.94	2	2	1	2
[†]Wav2Vec2 [7]	95M	92.27	5	5	5	5
[†]DeCoAR2.0 [32]	90M	92.63	4	4	4	4
[†]Vggish [43]	72M	96.78	1	1	2	1
Yamnet [44]	4M	94.32	3	3	3	3
fsFCNN [45]	20M	91.34	6	6	6	6
Time[†]			~ 0.61Day	24.09s	28.01s	10.86s

Table 4: *Same AST [40] as PSM but different pre-trained setups.*

PSM Setup	FT Acc. ($\uparrow$)	LogME$_{TIH}$
ImageNet + Audioset [42]	92.37	0.6483
ImageNet [41]	89.72	0.5668
Frozen + Linear-probing	32.56	-0.1247
Time	214s	1.75s

4. Conclusion

In this study, we explored the transferability of pre-trained upstream models in speech recognition tasks. Our results showed we are able to estimate the transferability of speech processing tasks. Moreover, our model-wise analysis showed that different pre-trained upstream models had varying adaptability to different downstream tasks. We believe our findings can provide insights into improving the efficiency and effectiveness of transfer learning in speech recognition.

5. References

[1] B. Li, D. Hwang *et al.*, "Efficient domain adaptation for speech foundation models," *arXiv:2302.01496*, 2023.

[2] Y. Zhang, D. S. Park *et al.*, "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1519–1532, 2022.

[3] A. Radford, J. W. Kim *et al.*, "Robust speech recognition via large-scale weak supervision," *arXiv:2212.04356*, 2022.

[4] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv:2108.07258*, 2021.

[5] C. Raffel, N. Shazeer *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv:1810.04805*, 2018.

[7] A. Baevski, Y. Zhou, A. Mohamed *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[8] W.-N. Hsu, B. Bolte *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[9] A. Xiao, W. Zheng *et al.*, "Scaling asr improves zero and few shot learning," *arXiv:2111.05948*, 2021.

[10] S. Kessler, B. Thomas, and S. Karout, "Continual-wav2vec2: an application of continual learning for self-supervised automatic speech recognition," *arXiv:2107.13530*, 2021.

[11] T. Brown, B. Mann, N. Ryder *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[12] Z. Ye, P. Liu, J. Fu *et al.*, "Towards more fine-grained and reliable nlp performance prediction," *arXiv:2102.05486*, 2021.

[13] D. Yu and L. Deng, *Automatic speech recognition*. Springer, 2016, vol. 1.

[14] S.-w. Yang, P.-H. Chi *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv:2105.01051*, 2021.

[15] Z.-C. Chen, C.-L. Fu *et al.*, "Exploring efficient-tuning methods in self-supervised speech models," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1120–1127.

[16] Z.-C. Chen, Y.-S. Sung, and H.-y. Lee, "Chapter: Exploiting convolutional neural network adapters for self-supervised speech models," *arXiv:2212.01282*, 2022.

[17] C.-H. H. Yang, Y.-Y. Tsai *et al.*, "Voice2series: Reprogramming acoustic models for time series classification," in *Proc. of ICML*. PMLR, 2021, pp. 11 808–11 819.

[18] K. You, Y. Liu *et al.*, "Logme: Practical assessment of pre-trained models for transfer learning," in *Proc. of ICML*. PMLR, 2021.

[19] A. Graves, S. Fernández, F. Gomez *et al.*, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. of ICML*, 2006, pp. 369–376.

[20] A. T. Tran, C. V. Nguyen, and T. Hassner, "Transferability and hardness of supervised classification tasks," in *Proc. of ICCV*, 2019, pp. 1395–1405.

[21] C. Nguyen, T. Hassner *et al.*, "Leep: A new measure to evaluate transferability of learned representations," in *Proc. of ICML*. PMLR, 2020.

[22] E. Bassignana, M. Müller-Eberstein, M. Zhang, and B. Plank, "Evidence > intuition: Transferability estimation for encoder selection," in *EMNLP*. ACL, 2022.

[23] A. Agostinelli, J. Uijlings *et al.*, "Transferability metrics for selecting source model ensembles," in *Proc. of CVPR*, 2022, pp. 7936–7946.

[24] A. Agostinelli, M. Pándy, J. Uijlings, T. Mensink, and V. Ferrari, "How stable are transferability metrics evaluations?" in *Proc. of ECCV*. Springer, 2022, pp. 303–321.

[25] R. Zhao, J. Xue *et al.*, "On addressing practical challenges for rnn-transducer," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 526–533.

[26] A. Graves, "Sequence transduction with recurrent neural networks," *Proc. of ICML Representation Learning Workshop*, 2013.

[27] A. Pasad, J.-C. Chou *et al.*, "Layer-wise analysis of a self-supervised speech representation model," in *Proc. of ASRU*. IEEE, 2021, pp. 914–921.

[28] B. Thomas, S. Kessler, and S. Karout, "Efficient adapter transfer of self-supervised speech models for automatic speech recognition," in *Proc. of ICASSP*. IEEE, 2022, pp. 7102–7106.

[29] J. Zhao and W.-Q. Zhang, "Improving automatic speech recognition performance for low-resource languages with self-supervised models," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[30] D. Alvarez-Melis and N. Fusi, "Geometric dataset distances via optimal transport," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 428–21 439, 2020.

[31] M. Ravanelli, J. a. Zhong *et al.*, "Multi-task self-supervised learning for robust speech recognition," in *ICASSP 2020*. IEEE, 2020, pp. 6989–6993.

[32] S. Ling and Y. Liu, "Decoar 2.0: Deep contextualized acoustic representations with vector quantization," *arXiv:2012.06659*, 2020.

[33] S. Kolouri, K. Nadjahi *et al.*, "Generalized sliced wasserstein distances," *Advances in neural information processing systems*, vol. 32, 2019.

[34] V. Panayotov, G. Chen *et al.*, "Librispeech: an asr corpus based on public domain audio books," in *Proc. of ICASSP*. IEEE, 2015, pp. 5206–5210.

[35] C.-H. H. Yang, B. Li *et al.*, "From english to more languages: Parameter-efficient model reprogramming for cross-lingual speech recognition," *arXiv:2301.07851*, 2023.

[36] A. Gulati, J. Qin *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv:2005.08100*, 2020.

[37] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv:1804.03209*, 2018.

[38] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[39] V. Pratap, Q. Xu *et al.*, "Mls: A large-scale multilingual dataset for speech research," *Proc. of Interspeech*, pp. 2757–2761, 2020.

[40] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv:2104.01778*, 2021.

[41] J. Deng, W. Dong *et al.*, "Imagenet: A large-scale hierarchical image database," in *Proc. of CVPR*. IEEE, 2009, pp. 248–255.

[42] J. F. Gemmeke, D. P. Ellis *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. of ICASSP*. IEEE, 2017, pp. 776–780.

[43] S. Hershey, S. Chaudhuri *et al.*, "Cnn architectures for large-scale audio classification," in *Proc. of ICASSP*. IEEE, 2017, pp. 131–135.

[44] E. Fonseca, M. Plakal *et al.*, "Audio tagging with noisy labels and minimal supervision," in *DCASE*, 2019, p. 69.

[45] H. Hu, C.-H. H. Yang *et al.*, "Device-robust acoustic scene classification based on two-stage categorization and data augmentation," *DCASE*, 2020.