Losses Can Be Blessings: Routing Self-Supervised Speech Representations Towards Efficient Multilingual and Multitask Speech Processing

Yonggan Fu[1], Yang Zhang[2], Kaizhi Qian[2], Zhifan Ye[1], Zhongzhi Yu[1]
Cheng-I Lai[3], Yingyan Lin[1]
[1]Rice University, [2]MIT-IBM Watson AI Lab, [3]MIT CSAIL
{yf22,zy50,zy42,yingyan.lin}@rice.edu,
{yang.zhang2,kqian}@ibm.com {clai24}@mit.edu

Abstract

Self-supervised learning (SSL) for rich speech representations has achieved empirical success in low-resource Automatic Speech Recognition (ASR) and other speech processing tasks, which can mitigate the necessity of a large amount of transcribed speech and thus has driven a growing demand for on-device ASR and other speech processing. However, advanced speech SSL models have become increasingly large, which contradicts the limited on-device resources. This gap could be more severe in multilingual/multitask scenarios requiring simultaneously recognizing multiple languages or executing multiple speech processing tasks. Additionally, strongly overparameterized speech SSL models tend to suffer from overfitting when being finetuned on low-resource speech corpus. This work aims to enhance the practical usage of speech SSL models towards a win-win in both enhanced efficiency and alleviated overfitting via our proposed S^3-Router framework, which for the first time discovers that simply discarding no more than 10% of model weights via only finetuning model *connections* of speech SSL models can achieve better accuracy over standard weight finetuning on downstream speech processing tasks. More importantly, S^3-Router can serve as an all-in-one technique to enable (1) a new finetuning scheme, (2) an efficient multilingual/multitask solution, (3) a state-of-the-art pruning technique, and (4) a new tool to quantitatively analyze the learned speech representation.

1 Introduction

Deep neural network (DNN) breakthroughs have tremendously advanced the field of Automatic Speech Recognition (ASR). However, one major driving force for powerful DNNs, i.e., the availability of a large amount of training data, is not always possible for ASR. This is because collecting large-scale transcriptions is costly, especially for low-resource spoken languages around the world, limiting the wide application of deep ASR models. Fortunately, recent advances in self-supervised learning (SSL) for rich speech representations [1, 2, 3, 4, 5, 6, 7, 8, 9] have achieved empirical success in low-resource ASR, where SSL models pretrained on raw audio data without transcriptions can be finetuned on low-resource transcribed speech to match the accuracy of their supervised counterparts.

However, there exists a dilemma between the trends of speech SSL models and the growing demand for speech processing applications on the edge. While advanced speech SSL models become increasingly larger to learn more generalizable features, it is highly desired to process the captured speech signals in real time on edge devices, which have limited resources and conflict with the prohibitive complexity of existing speech SSL models. Such an efficiency concern would be more severe in

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

multilingual/multitask scenarios where simultaneously recognizing multiple languages or executing multiple speech processing tasks is required: if one separate model is finetuned for each target language/task, the storage and computational cost will be significantly increased, thus prohibiting the practical deployment of existing speech SSL models. Additionally, strongly overparameterized speech SSL models tend to suffer from overfitting when being finetuned on a low-resource speech corpus [10, 11, 12, 13, 14], limiting the achievable accuracy improvement brought about by more parameters and thus the achievable performance-efficiency trade-off.

In this work, we aim to facilitate the practical usage of speech SSL models towards a win-win in enhanced efficiency and alleviated overfitting for boosting task accuracy under low-resource settings. Excitingly, we develop a framework, dubbed **S**elf-**S**upervised **S**peech Representation **Router** (S^3-Router), that can serve as an all-in-one technique to tackle the aforementioned challenges and largely enhance the practical usage of speech SSL models, contributing (1) a new finetuning scheme for downstream speech processing, i.e., finetuning the connections of the model structure via learning a binary mask on top of pretrained model weights, which notably alleviates model overfitting and thus improves the achievable accuracy over standard weight finetuning methods under a low-resource setting; (2) a multilingual/multitask technique via learning language-/task-specific binary masks on top of shared model weights inherited from SSL pretraining; (3) a competitive pruning technique to trim down the complexity of speech SSL models while maintaining task accuracy; and (4) a new tool to quantitatively analyze what is encoded in speech SSL models thanks to the learned masks' binary nature on top of shared model weights. We summarize our contributions below:

- We propose a framework dubbed S^3-Router, which offers an alternative to the mainstream finetuning of model *weights* that finetunes the *structure* of speech SSL models, by learning a binary mask on top of the pretrained model weights and integrating it with a novel mask initialization strategy customized for the pretrain-finetune paradigm;
- We are the first to discover that discarding no more than 10% of weights *without* finetuning pretrained model weights can achieve better task performances as compared to the mainstream method of weight finetuning on downstream speech processing tasks. Notably, our method can scale well to even larger models;
- We extend our S^3-Router framework for enhanced deployment efficiency, contributing (1) a novel multilingual/multitask solution, and (2) a competitive pruning technique achieving better performance-efficiency trade-offs than state-of-the-art (SOTA) ASR pruning techniques;
- We demonstrate the capability of S^3-Router as a tool to quantitatively understand what is encoded in speech SSL models thanks to the binary nature of the learned masks.

S^3-Router has opened up a new perspective for empowering efficient multilingual/multitask speech processing and enhancing our understanding about what is encoded in speech SSL models.

2 Related Work

Automatic speech recognition. Early ASR systems [15, 16, 17, 18, 19, 20] were mainly based on the combinations of hidden Markov models (HMM) with Gaussian mixture models or DNNs, and often contain multiple modules (e.g., an acoustic model, a language model, and a lexicon model) trained separately. Recent works process raw audio sequences end-to-end, including CTC [21]-based models [22, 23, 24, 25, 26], recurrent neural network(RNN)-transducers [27, 28, 29, 30], sequence-to-sequence models [31, 32, 33, 34, 35, 36], and transformer-based models [37, 38, 39]. Specifically, transformer-based models have been widely adopted as speech SSL models [6, 9, 7].

Self-supervised learning for speech representation. Considering the high cost of collecting large-scale transcriptions, learning rich speech representations via SSL has become crucial and promising for empowering low-resource ASR. Early works [40, 41, 42, 43, 44, 45, 46, 47] build generative models for speech with latent variables. Recently, prediction-based SSL methods have become increasingly popular. We refer the readers to a recent survey [48] for more details. Among prior arts, [49] is a pioneering work relevant to our method, and adopts masking as an alternative to weight finetuning on pretrained language models for natural language processing (NLP). Nevertheless, our S^3-Router is non-trivially different from [49] in that (1) we target SSL models in the speech domain and the learned speech representation is required to be generalizable across different spoken languages under a cross-lingual transfer setting, where the speech SSL models suffer from a higher risk of overfitting on downstream low-resource tasks as compared to NLP, making our findings non-trivial

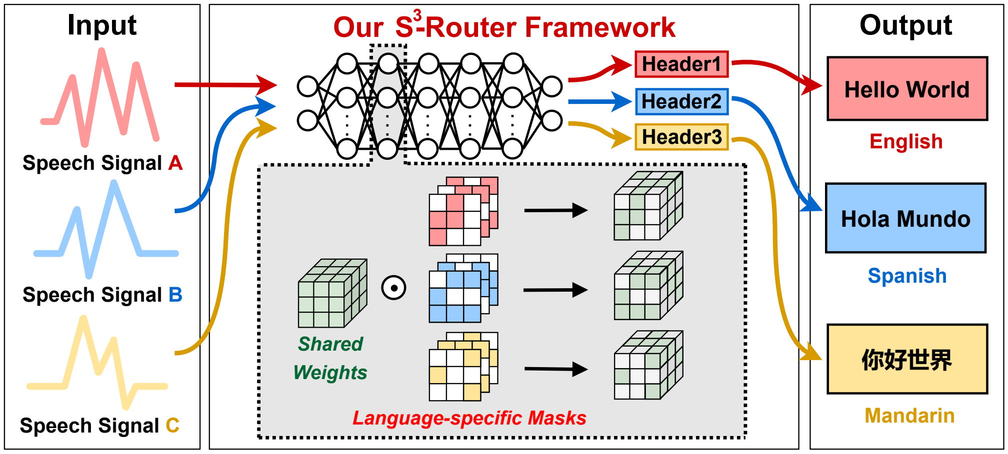

Figure 1: An overview of our S^3-Router framework, which receives multilingual speech signals denoted as A, B, and C here and then outputs the corresponding text transcript of predication, based on one *shared weight* model together with language-/task-specific *binary* masks.

contributions; (2) S^3-Router features a win-win in both efficiency and accuracy thanks to our proposed mask initialization strategy, which plays a crucial role in maintaining task accuracy under a high sparsity and enables the extension of S^3-Router towards a competitive pruning technique with SOTA accuracy; and (3) we further develop S^3-Router for multilingual/multitask speech processing and as a simple yet effective tool for analyzing language-wise similarities.

Multitask learning for speech processing. There exists a growing demand for DNNs to simultaneously process multiple tasks [50, 51]. In the speech domain, previous works attempt to train a single model to solve multiple tasks [52, 53, 54, 55, 56, 57] or use an auxiliary task to enhance the accuracy of a primary task [58, 59, 60, 61, 62, 63]. Motivated by the success of SSL in low-resource downstream tasks, multitask/multilingual learning has been adopted in both SSL pretraining [64, 65, 66, 67, 7] to enrich learned speech representations, and finetuning scenarios [68]. For example, a recent relevant work [67] adopts language-adaptive pretraining on top of [7], where different sparse sub-networks are activated for different languages during multilingual pretraining and the gradients are accumulated on the shared super-network to learn better multilingual representations. In contrast, S^3-Router is fundamentally different, as it targets the finetuning stage of a given pretrained speech SSL model and only tunes the connections of the model structure *without* tuning the SSL pretrained weights.

ASR pruning. As ASR models become increasingly overparameterized for abstracting generalizable representations, pruning large-scale ASR models has drawn a growing attention. Early works trim down either the decoding search spaces [69, 70, 71, 72, 73, 74] or HMM state space [75]. Modern ASR paradigm has gradually shifted its focus to pruning end-to-end ASR models [76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87]. Recently, [88, 89, 90] prune speech SSL models towards more efficient low-resource ASR, e.g., PARP [88] proposes a finetuning pipeline to identify the existence of lottery tickets within speech SSL models. In addition to the boosted pruning efficiency over PARP [88], S^3-Router emphasizes the role of sparsity in encoding language-/task-specific information which can empower multitask/multilingual speech processing *without* tuning the pretrained weights.

3 The Proposed S^3-Router Framework

3.1 Drawn Inspirations from Previous Work

Recent works [91, 92, 93] find that sub-networks featuring a decent inborn accuracy and adversarial robustness are hidden within randomly initialized networks *without* any weight training. Specifically, [91, 92] show that sub-networks with a decent accuracy, even matching that of their dense networks, can be identified from randomly initialized networks, and [93] shows an even stronger evidence: merely updating the sparsity patterns of model connections *without* modifying the randomly initiated model weights can produce both accurate and adversarially robust models. These pioneering works imply that tuning the connections of the model structure, which can be characterized by the learned connection sparsity patterns, can be as effective as training the model weights. We hypothesize that **model sparsity not only can favor model efficiency, but also can serve as a similar role, i.e., another optimization knob, as model weights to encode language-/task-specific information**.

3.2 Formulation and Optimization of S^3-Router

Inspired by the aforementioned intriguing hypothesis, we propose the S^3-Router framework to empower efficient multilingual and multitask speech processing on top of SSL speech representations.

Overview. As shown in Fig. 1, given the raw audios of different spoken languages (or different tasks), S^3-Router finetunes the connection patterns of the model structure for *each* target spoken language/task via optimizing language-/task-specific *binary masks* on top of the *shared weights* of a given speech SSL model, instead of finetuning the model weights as adopted in the common pretrain-finetune paradigm. Specifically, the learned binary masks of each language/task are multiplied with the shared model weights to mask out some connections within the given speech SSL model, where the remaining connections (or the induced connection sparsity) encode language-/task-specific information for different spoken languages and downstream tasks. Note that for each language or task, only one set of binary masks and one lightweight header, e.g., the classification head for ASR which is naturally non-shareable across languages due to diverse dictionary sizes, need to be independently trained, incurring negligible overhead (e.g.,$\leq$6.3% storage of the whole backbone model).

Formulation. Formally, our S^3-Router framework can be formulated as:

$$\arg\min_{m_t} \sum_{(x_t,y_t)\in D_t} \ell_t(f(m_t \odot \theta_{SSL}, x_t), y_t) \quad s.t. \ ||m_t||_0 \leqslant k_t \tag{1}$$

where (x_t, y_t) are the input audio and corresponding transcriptions/labels of a spoken language/task t in a downstream dataset D_t, and θ_{SSL} is the SSL pretrained weights of the given speech SSL model f. Specifically, the mask set m_t is applied on top of the model weights θ_{SSL}, and optimized to minimize the loss function l_t, e.g., a CTC loss [21] for ASR, subject to an L_0 sparsity constraint, where the number of non-zero elements in m_t is limited to k_t, which serves as a hyperparameter for controlling and balancing (1) the amount of language-/task-specific information encoded in the connection sparsity and (2) model efficiency. Intuitively, if the sparsity is 0% or 100%, no new information is introduced during finetuning on the corresponding new/downstream language/task.

Optimization. To differentiably optimize m_t in Eq. (1), we binarize m_t and activate only its top k_t elements during forward, while all the elements in m_t are updated via straight-through estimation [94] during backward. More specifically, during forward, we binarize m_t to $\hat{m}_t$ via enforcing its top k_t elements to 1 and other elements to 0, thus the forward function becomes $f(\hat{m}_t \odot \theta_{SSL}, x_t)$, which is the same for the inference process; during backward, we directly propagate the gradients from the binary mask $\hat{m}$ to m, i.e., $\frac{\partial l}{\partial m_t} \approx \frac{\partial l}{\partial \hat{m}_t}$, thus m_t can be learned in a gradient-based manner.

3.3 How to Initialize the Masks in S^3-Router?

Importance of mask initialization. A low learning rate is commonly adopted during finetuning to ensure effective inheritance of the SSL speech representations, resulting in merely smaller changes in each set of the masks m_t in Eq. (1) as compared to training from scratch. Therefore, the mask initialization strategy plays an important role for the quality of the finally optimized masks in S^3-Router. Next, we discuss the pros and cons of two intuitive mask initialization strategies:

① **Random initialization (RI).** We empirically find that adopting commonly used random initialization [95] in S^3-Router can achieve a decent accuracy under a low sparsity. However, since no prior knowledge of the speech SSL model is utilized, the accuracy can largely drop with a high sparsity under random mask initialization, limiting extending S^3-Router to be a practical pruning technique.

② **Weight magnitude based initialization (WMI).** As weights magnitude can quantify the importance of weights as commonly used in pruning [96, 97, 98, 99, 100], taking the magnitudes of the given SSL pretrained weights $||\theta_{SSL}||$ in Eq. 1 as the initial values of masks might help utilize the knowledge learned during SSL pretraining. However, we empirically find that doing so causes worse trainability, i.e., the ranking of mask values is then more stable during finetuning than that under random initialization, and learned binary masks seldom change and mostly stick to their initial values. This may inhibit the optimization process and lead to sub-optimal learned masks.

③ **Proposed Order-Preserving Random Initialization (ORI).** To marry the best of both above initialization strategies, we propose a new one, which can be viewed as a weight rank based initialization. In ORI, we first acquire mask values of the same dimension as the shared weights via random initialization like [95], and then perform magnitude-based sorting to assign a larger mask value to weight elements with a larger magnitude. Thus, the ranking order between the mask values of the weight elements is the same as that of their magnitudes. In this way, the mask trainability is maintained while the learned speech SSL model knowledge can be exploited (see Sec 4).

3.4 S^3-Router is Useful in Various Application Scenarios

A new finetuning scheme. S^3-Router offers a new and equally effective finetuning scheme as it finetunes model connections given a speech SSL model. As large-scale speech SSL models tend to overfit when finetuning their weights under a low-resource setting, we hypothesize that the binary optimization of the mask patterns in Eq. (1) can serve as regularization and thus alleviate overfitting.

An efficient multilingual and multitask method. S^3-Router can naturally enable multilingual and multitask speech processing by merely switching among its learned language-/task-specific binary masks. One advantage is that since the gradients of different spoken languages or tasks can now be independently accumulated on their corresponding masks *without* interfering each other, the commonly observed gradient conflict issue [101] can be alleviated.

A new pruning technique. Since S^3-Router encodes language/task-specific information via binary masks on top of shared model weights, it naturally introduces sparsity into the given model, and thus can naturally serve as a pruning technique. To further improve the achievable accuracy-efficiency trade-off and more fairly benchmark with SOTA ASR pruning methods, we propose a variant of S^3-Router dubbed S^3-Router-P, which first finetunes the model weights on the downstream audios and then prunes the model connections based on the learned binary masks in Eq. (1).

An effective and simple tool to analyze what is encoded across language/task-specific speech SSL models. Another exciting advantage of S^3-Router is that it can provide a quantitative metric about how the pretained SSL speech presentation is utilized by different spoken languages/tasks via masking out languages/tasks-specific weights.

4 S^3-Router: Discarding $\leq$10% Weights is All You Need

4.1 Experiment Setup

Here we evaluate S^3-Router as a new finetuning scheme for downstream speech processing. Models: We adopt wav2vec 2.0 base (wav2vec2-base) and large (wav2vec2-large) [6] (with 95M/317M parameters, respectively) pretrained on LibriSpeech 960 hours [102] and xlsr [7] (with 317M parameters) pretrained on 53 languages sampled from CommonVoice [103] as our speech SSL models. Datasets: We consider 10 speech processing tasks, including low-resource English ASR on LibriSpeech [7] with only 10min/1h/10h labeled data, following the dataset split in [6, 88], and low-resource phoneme recognition on CommonVoice [103] with 1h labeled data per language, following the dataset split in [104, 7], as well as 8 speech processing tasks from SUPERB [105]. Finetuning settings: Our code is built on top of Fairseq [106] and we follow the standard finetuning settings for each task, i.e., the default configurations in Fairseq [106] for ASR/phoneme recognition and those in SUPERB [105] for other tasks. In particular, all our experiments on ASR/phoneme recognition adopt an Adam optimizer with an initial learning rate of 5e-5 plus a tri-stage schedule [6] and we finetune wav2vec2-base/large for 12k/15k/20k steps on the 10m/1h/10h splits, respectively, and xlsr is finetuned for 12k steps for each spoken language. It takes about 10/24/24 GPU hours to finetune wav2vec2-base/large/xlsr for 12k steps with our S^3-Router. We do not freeze all the layers except the final linear layer for the first 10k steps [6], following [88]. S^3-Router settings: If not specifically stated, S^3-Router only tunes the connections of (i.e., applies learnable masks on) the feed-forward networks (FFNs) of transformer structures [107] and fixes all other weights, which is empirically found to be optimal based on the ablation studies in Sec. 6. In addition, if not stated, we adopt our ORI mask initialization by default and do not apply language models for a fair benchmark of ASR performances, following [88].

4.2 Benchmark on Low-resource English ASR

Finetuning on wav2vec2-base. We apply our S^3-Router on wav2vec2-base under different low-resources settings as shown in Fig. 2 (a)$\sim$(c). We can observe that our S^3-Router can consistently outperform the standard weight finetuning in terms of the achievable WER, i.e., the lowest WER at the corresponding optimal sparsity. In particular, our S^3-Router achieves a 1.72%/2.34% reduction in word error rate (WER) under a sparsity ratio of 8% on the test-clean/test-other set of LibriSpeech,

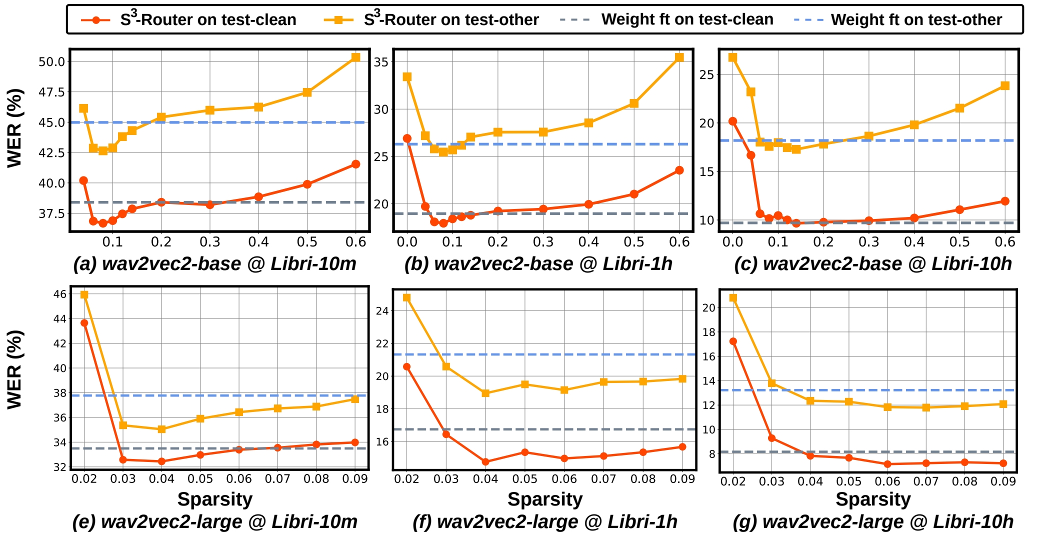

Figure 2: Benchmark our S^3-Router and standard weight finetuning on the test-clean/test-other sets of LibriSpeech on top of wav2vec2-base/large under different low-resource settings.

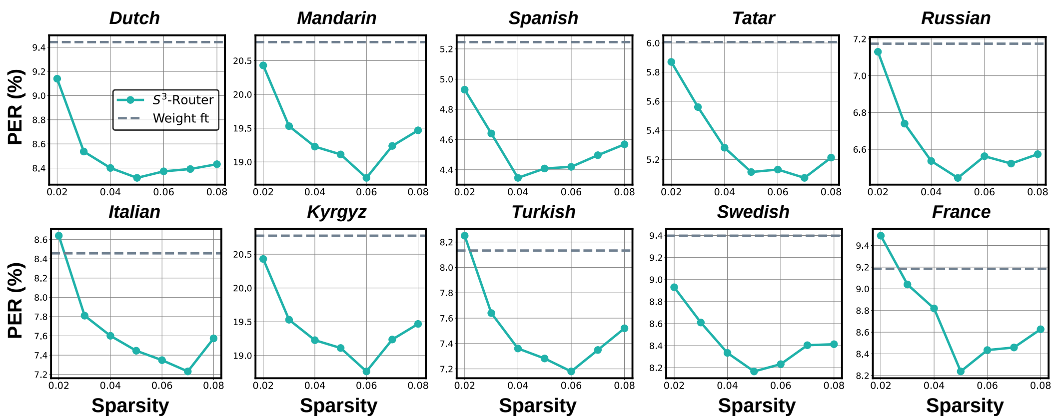

Figure 3: Benchmark our S^3-Router and weight finetuning on xlsr across 10 spoken languages.

respectively, when being finetuned on 10min labeled data. This indicates that finetuning connections via our S^3-Router can be a competitive alternative with reduced WER for finetuning weights under low-resource settings, which provides a new paradigm for finetuning speech SSL models.

Ablation studies of mask initialization. We equip our S^3-Router with different mask initialization schemes in Sec. 3.3 and show their achievable WER on low-resource English ASR in Tab. 1. We can observe that (1) the proposed ORI mask initialization consistently outperforms the other two schemes in terms of the achievable WER, and (2) even random mask initialization can match or surpass the performances of standard weight finetuning. We also provide their complete sparsity-WER trade-offs in the appendix and find that weight magnitude based initialization favors larger sparsity ratios, while it is harder to overturn the ranking between masks via gradients, resulting in inferior achievable WER.

Scalability to larger speech SSL models. We apply S^3-Router to a larger model wav2vec-large. As shown in Fig. 2 (e)$\sim$(g), we can see that the achievable WER of S^3-Router still consistently outperforms the standard weight finetuning across all downstream datasets, e.g., a 1.99%/2.37% WER reduction under a 4% sparsity on the test-clean/test-other set, respectively, when being finetuned on the 1h labeled data. Consistent results on xlsr are provided in the appendix.

Table 1: Benchmark different mask initialization schemes of S^3-Router with weight finetuning on LibriSpeech test-clean/test-other sets.

Method	Libri-10m	Libri-1h	Libri-10h
Weight ft	38.40/44.98	18.963/26.298	9.7/18.19
RI	36.98/43.42	18.19/26.31	9.78/18.17
WMI	40.09/46.63	19.99/27.16	11.17/18.97
ORI (Ours)	**36.68/42.64**	**17.95/25.47**	**9.66/17.27**

Insights. This set of experiments indicates that ❶ our S^3-Router features a good scalability to larger speech SSL models as well as a good generality for different pretraining schemes, and ❷ our method

Table 4: Benchmark S^3-Router with standard weight finetuning on 8 tasks from SUPERB [105].

Catagory	Content	Speaker			Paralinguistics	Semantics		
Task	Keyword Spotting	Speaker Identification	Speaker Verification	Speaker Diarization	Emotion Recognition	Intent Classification	Slot Filling	Speech Translation
Metric	Acc ↑	Acc ↑	EER ↓	DER ↓	Acc ↑	Acc ↑	F1 ↑	BLEU ↑
Weight ft	95.5	66.8	7.24	7.41	61.5	91.48	88.1	18.85
S^3-Router	**95.7**	**71.07**	**6.79**	**7.18**	**62.13**	**92.6**	**88.76**	**19.01**
Opt Spar.	0.1	0.06	0.1	0.08	0.1	0.1	0.1	0.08

effectively reduces the overfitting on more overparameterized speech SSL models according to the larger performance gains, thanks to the regularization effect of the binary optimization process in Eq. (1), making our method an appealing solution for more advanced speech SSL models.

Add language models (LMs). We further apply a 4-gram LM [108] (details in the appendix) as the decoder for both our S^3-Router at the optimal sparsity ratio in Fig. 2 and the weight finetuning baselines. As shown in Tab. 2, our S^3-Router still consistently outperforms weight finetuning with more notable reductions on wav2vec2-large, e.g., a 1.66%/3.22% WER reduction on the test-clean/test-other set of LibriSpeech when being finetuned on 10min labeled data.

Table 2: Benchmark our S^3-Router with weight finetuning when being equipped with a 4-gram LM [108].

Model	Method	Libri-10m	Libri-1h	Libri-10h
wav2vec2-base	Weight ft	26.55/32.95	11.54/18.45	6.65/13.97
	S^3-Router	**25.28/31.07**	**11.28/18.33**	**6.42/13.14**
wav2vec2-large	Weight ft	24.83/29.05	11.18/15.80	5.52/10.31
	S^3-Router	**23.17/25.83**	**9.74/13.77**	**4.98/9.27**

4.3 Benchmark on Low-resource Cross-lingual Transfer

We further evaluate our S^3-Router under two cross-lingual transfer settings, including (1) finetuning the multilingual pretrained xlsr on different low-resource languages in CommonVoice [103], and (2) a high-to-low resource transfer setting where the monolingual (English) pretrained wav2vec2-base is finetuned on multiple low-resource languages in CommonVoice.

Cross-lingual transfer on xlsr. As shown in Fig. 3, our S^3-Router consistently outperforms standard weight finetuning in terms of the achievable phoneme error rate (PER) across all the 10 spoken languages, e.g., a 1.12%/2.01% PER reduction on Dutch/Mandarin, respectively.

High-to-low resource transfer on wav2vec2-base. As shown in Tab. 3, on top of the English pretrained wav2vec2-base, finetuning the connections with S^3-Router still wins the lowest achievable PER over the baseline. Consistent results on wav2vec2-large are in the appendix.

Table 3: Benchmark our S^3-Router and weight finetuning on wav2vec2-base and CommonVoice.

Language	Dutch	Mandarin	Spanish	Tatar	Russian
Weight ft	19.821	26.674	13.864	11.143	17.052
S^3-Router	**18.511**	**26.098**	**13.369**	**10.936**	**16.329**
Language	**Italian**	**Kyrgyz**	**Turkish**	**Swedish**	**France**
Weight ft	19.265	13.409	15.699	20.807	19.349
S^3-Router	**18.292**	**12.303**	**14.818**	**19.641**	**17.937**

Insights. This set of experiments implies that for each downstream spoken language, there exist decent subnetworks for processing its language-specific information even in monolingual pretrained speech SSL models. In another words, properly learned sparsity can encode language-specific information with competitive performances.

4.4 Benchmark on More Downstream Speech Processing Tasks

We further evaluate our S^3-Router via finetuning wav2vec2-base on 8 speech processing tasks from SUPERB [105], covering different aspects of speech (content/speaker/semantics/paralinguistics). We show the achievable task performances as well as the corresponding optimal sparsity ratios in Tab. 4 and find that our method surpasses the task performances over standard weight finetuning across all 8 tasks via discarding $\leq$10% weights, indicating that properly learned sparsity can also effectively encode task-specific information. We provide the sparsity-performance trade-offs in the appendix.

4.5 Empowering Multilingual and Multitask Speech Processing

Since all the aforementioned results achieved on the same speech SSL model via S^3-Router share pretrained weights, it can simultaneously enable multilingual and multitask speech processing, thanks to the independent accumulation of task-specific gradients that avoids gradient conflicts [101], of superior scalability with the number of tasks. For example, as compared to independent weight finetuning for each language/task, S^3-Router can simultaneously support 11 languages in Sec. 4.2/ 4.3 and 8 tasks in Sec. 4.4 using one wav2vec2-base while achieving a win-win in both accuracy (as validated in Sec. 4.2/ 4.3/ 4.4) and efficiency, i.e., more than 88.5% reductions in model parameters.

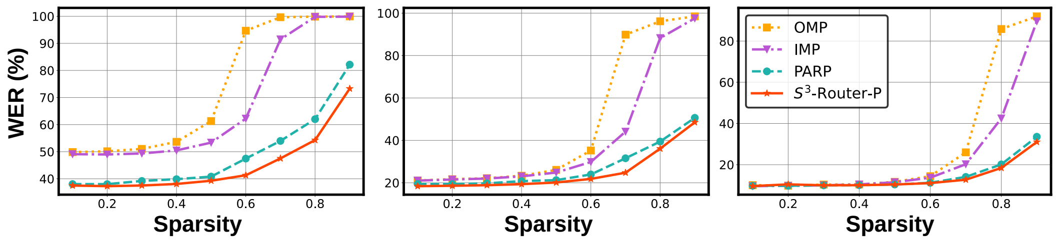

Figure 4: Benchmark our S^3-Router-P against OMP, IMP, and PARP [88] for pruning wav2vec2-base on LibriSpeech. The WER on the test-clean set is reported.

5 S^3-Router-P: Pruning ASR Models for Enhancing Efficiency

Setup. For evaluating S^3-Router-P (see Sec. 3.4), we conduct two modifications: (1) both FFN and self-attention (SA) modules are pruned for a fair comparison, and (2) we adopt weight magnitude based mask initialization for two reasons: firstly, it features the best scalability to high sparsity among the three initialization schemes; secondly, since the weights have been finetuned, their magnitudes can serve to indicate their importance on the downstream speech. Ablation studies for pruning under different mask initialization schemes are in the appendix.

For our baselines, we benchmark with three ASR pruning methods, i.e., one-shot/iterative magnitude pruning (OMP/IMP), and the SOTA method PARP [88] under the best setting (dubbed PARP-P), and directly adopt their reported performances in [88].

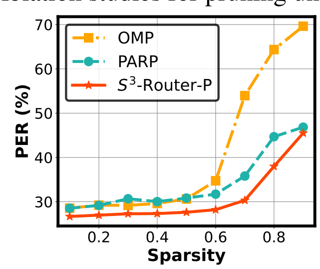

Figure 5: Benchmark our S^3-Router-P against OMP and PARP for pruning on Mandarin.

Pruning wav2vec2-base on LibriSpeech. As shown in Fig. 4, we can observe that (1) our S^3-Router-P consistently achieves the most competitive WER-sparsity trade-offs across the two models and three datasets, e.g., a 6.46% lower WER over PARP under a sparsity ratio of 0.7 with 10min labeled data, and (2) our method features a better scalability to more stringent low-resource scenarios according to the large performance gains on LibriSpeech-10m.

Pruning wav2vec2-base on Mandarin@CommonVoice. As shown in Fig. 5, consistent observations can be drawn that our method still wins the WER-sparsity trade-offs. More pruning results are provided in the appendix.

6 S^3-Router: What is Encoded in Speech SSL Models?

6.1 What is The Roles of Different Modules in Speech SSL Models?

We tune the connections of a subset of the modules in the speech SSL model via our S^3-Router to study their contributions with other modules fixed in terms of both weights and connections.

FFN vs. SA. We tune the connections of FFN, SA, or both on top of wav2vec2-base and LibriSpeech-1h as shown in Tab. 5. We can observe that tuning the connections of FFN only wins the lowest achievable WER, even outperforming tuning both FFN and SA. This indicates that the SSL pretrained SA modules are generalizable enough to capture temporal relationship between tokens for downstream speech thus only FFN needs to be tuned for encoding new information about downstream tasks. We provide more analysis about the roles of different blocks in the appendix.

Table 5: The achievable WER on LibriSpeech test-clean/test-other when finetuning different modules.

Sparsity	SA only	FFN only	Both
0.1	24.388/30.791	**18.607/26.124**	19.264/27.362
0.2	23.064/30.207	**19.541/27.07**	18.999/27.383
0.3	22.782/30.21	**19.585/27.576**	19.645/28.399
0.4	23.22/31.28	20.021/28.97	**19.783/27.975**

6.2 How Much New Information is Learned by Finetuning?

We visualize the optimal sparsity ratio for achieving the lowest WER on LibriSpeech-10m/1h/10h, which serves as an indicator about the amount of new information learned by finetuning, across wav2vec2-base/large/xlsr in Tab. 6. We can observe that (1) larger speech SSL models reach their optimal performances under lower sparsity, i.e., relatively less amount of tuning could lead to their decent downstream performances thanks to their overparameterization, and (2) finetuning on more resources leads to higher optimal sparsity, indicating that speech SSL models have more confidence to update the SSL pretrained representations given more resources.

Table 6: The optimal sparsity ratios for the lowest WER across different models and resources.

Model	Libri-10m	Libri-1h	Libri-10h
wav2vec2-base	0.08	0.10	0.20
wav2vec2-large	0.04	0.04	0.06
xlsr	0.04	0.05	0.07

6.3 How is The Learned Masks Correlated to Phonetics?

Given the decent performances achieved by the learned masks, we are interested their correlations with human expertise in phonetics. In particular, we wonder whether the similarity between the learned masks of two languages aligns with that between their phoneme inventories.

Setup. Given the 11 languages adopted in Sec. 4.2 and 4.3, we calculate layer-wise cosine similarities of their learned masks, under a sparsity ratio of 0.1 with near-optimal performances, between each pair of them. We pick the mask similarity of the first layer as our metric without losing generality. We further measure the cosine similarity between their corresponding phoneme inventories acquired from Phoible [109], following [110], and more detailed settings are in the appendix.

Results and analysis. We visualize the two similarity metrics of each language pair in Fig. 6 and find that the Pearson Correlation Coefficient [111] and the Spearman correlation coefficient [112] between the two similarity metrics are 0.527 and 0.548, respectively. This indicates that the similarities of our learned masks are non-trivially correlated with that of phonetics while the former also provides new insights in the eyes of speech SSL models, which could shed light on future advances in zero-shot cross-lingual transfer [110]. We provide more insightful correlation analysis as well as the visualization of layer-wise mask similarities between different languages in the appendix.

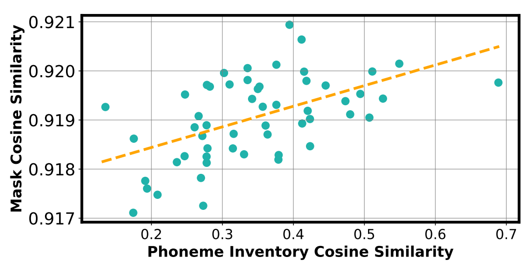

Figure 6: Visualizing the correlation between the similarity of the learned masks and that of the corresponding phoneme inventories of language pairs.

7 Conclusion

Motivating by the empirical success of SSL speech representations in low-resource speech processing, we propose S^3-Router to facilitate the practical usage of speech SSL models via finetuning their connections instead of weights and encoding language-/task-specific information via sparsity. Extensive experiments validate that S^3-Router not only serves as a stronger alternative with alleviated overfitting and enhanced accuracy for standard weight finetuning, but also empowers efficient multilingual and multitask speech processing. Our insights could enhance our understandings about what is encoded in speech SSL models and thus shed light on future advances in SSL speech representations.

Acknowledgements

The work is supported by the National Science Foundation (NSF) through the CCRI program (Award number: 2016727) and an IBM faculty award received by Dr. Yingyan Lin.

References

[1] Po-Han Chi, Pei-Hung Chung, Tsung-Han Wu, Chun-Cheng Hsieh, Shang-Wen Li, and Hung-yi Lee. Audio albert: A lite bert for self-supervised learning of audio representation. *arXiv*

preprint arXiv:2005.08575, 2020.

[2] Shaoshi Ling and Yuzong Liu. Decoar 2.0: Deep contextualized acoustic representations with vector quantization. *arXiv preprint arXiv:2012.06659*, 2020.

[3] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[4] Eugene Kharitonov, Morgane Rivière, Gabriel Synnaeve, Lior Wolf, Pierre-Emmanuel Mazaré, Matthijs Douze, and Emmanuel Dupoux. Data augmenting contrastive learning of speech representations in the time domain. *arXiv preprint arXiv:2007.00991*, 2020.

[5] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[6] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[7] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

[8] Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. Effectiveness of self-supervised pre-training for speech recognition. *arXiv preprint arXiv:1911.03912*, 2019.

[9] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[10] Wenxin Hou, Han Zhu, Yidong Wang, Jindong Wang, Tao Qin, Renjun Xu, and Takahiro Shinozaki. Exploiting adapters for cross-lingual low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:317–329, 2021.

[11] Meng Cai, Yongzhe Shi, Jian Kang, Jia Liu, and Tengrong Su. Convolutional maxout neural networks for low-resource speech recognition. In *The 9th International Symposium on Chinese Spoken Language Processing*, pages 133–137. IEEE, 2014.

[12] Meng Cai, Yongzhe Shi, and Jia Liu. Stochastic pooling maxout networks for low-resource speech recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3266–3270. IEEE, 2014.

[13] William Chan and Ian Lane. Deep convolutional neural networks for acoustic modeling in low resource languages. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2056–2060. IEEE, 2015.

[14] Yajie Miao and Florian Metze. Improving low-resource cd-dnn-hmm using dropout and multilingual dnn training. In *Interspeech*, volume 13, pages 2237–2241, 2013.

[15] Fei Sha and Lawrence Saul. Large margin hidden markov models for automatic speech recognition. *Advances in neural information processing systems*, 19, 2006.

[16] Shuiyang Mao, Dehua Tao, Guangyan Zhang, PC Ching, and Tan Lee. Revisiting hidden markov models for speech emotion recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6715–6719. IEEE, 2019.

[17] Stephen Adams and Peter A Beling. A survey of feature selection methods for gaussian mixture models and hidden markov models. *Artificial Intelligence Review*, 52(3):1739–1779, 2019.

[18] Xian Tang. Hybrid hidden markov model and artificial neural network for automatic speech recognition. In *2009 Pacific-Asia Conference on Circuits, Communications and Systems*, pages 682–685. IEEE, 2009.

[19] John S Bridle. Alpha-nets: A recurrent 'neural'network architecture with a hidden markov model interpretation. *Speech Communication*, 9(1):83–92, 1990.

[20] Navdeep Jaitly, Patrick Nguyen, Andrew Senior, and Vincent Vanhoucke. Application of pretrained deep neural networks to large vocabulary speech recognition. 2012.

[21] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

[22] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[23] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016.

[24] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772. PMLR, 2014.

[25] Yajie Miao, Mohammad Gowayyed, and Florian Metze. Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 167–174. IEEE, 2015.

[26] Florian Eyben, Martin Wöllmer, Björn Schuller, and Alex Graves. From speech to letters-using a novel neural network architecture for grapheme based asr. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 376–380. IEEE, 2009.

[27] Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

[28] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013.

[29] Kanishka Rao, Haşim Sak, and Rohit Prabhavalkar. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 193–199. IEEE, 2017.

[30] Linhao Dong, Shiyu Zhou, Wei Chen, and Bo Xu. Extending recurrent neural aligner for streaming end-to-end speech recognition in mandarin. *arXiv preprint arXiv:1806.06342*, 2018.

[31] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. *Advances in neural information processing systems*, 28, 2015.

[32] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4945–4949. IEEE, 2016.

[33] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4960–4964. IEEE, 2016.

[34] Yu Zhang, William Chan, and Navdeep Jaitly. Very deep convolutional networks for end-to-end speech recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4845–4849. IEEE, 2017.

[35] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778. IEEE, 2018.

[36] Rohit Prabhavalkar, Tara N Sainath, Yonghui Wu, Patrick Nguyen, Zhifeng Chen, Chung-Cheng Chiu, and Anjuli Kannan. Minimum word error rate training for attention-based sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4839–4843. IEEE, 2018.

[37] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[38] Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888. IEEE, 2018.

[39] Yongqiang Wang, Abdelrahman Mohamed, Due Le, Chunxi Liu, Alex Xiao, Jay Mahadeokar, Hongzhao Huang, Andros Tjandra, Xiaohui Zhang, Frank Zhang, et al. Transformer-based acoustic modeling for hybrid speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6874–6878. IEEE, 2020.

[40] Wei-Ning Hsu, Yu Zhang, and James Glass. Learning latent representations for speech generation and transformation. In *INTERSPEECH*, 2017.

[41] Jan Chorowski, Ron J Weiss, Samy Bengio, and Aäron van den Oord. Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM transactions on audio, speech, and language processing*, 27(12):2041–2053, 2019.

[42] Aaron van den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017.

[43] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In *NeurIPS*, 2017.

[44] Janek Ebbers, Jahn Heymann, Lukas Drude, Thomas Glarner, Reinhold Haeb-Umbach, and Bhiksha Raj. Hidden markov model variational autoencoder for acoustic unit discovery. In *INTERSPEECH*, 2017.

[45] Thomas Glarner, Patrick Hanebrink, Janek Ebbers, and Reinhold Haeb-Umbach. Full bayesian hidden markov model variational autoencoder for acoustic unit discovery. In *INTERSPEECH*, 2018.

[46] Sameer Khurana, Shafiq Rayhan Joty, Ahmed Ali, and James Glass. A factorial deep markov model for unsupervised disentangled representation learning from speech. In *ICASSP*, 2019.

[47] Sameer Khurana, Antoine Laurent, Wei-Ning Hsu, Jan Chorowski, Adrian Lancucki, Ricard Marxer, and James Glass. A convolutional deep markov model for unsupervised speech representation learning. *arXiv preprint arXiv:2006.02547*, 2020.

[48] Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabeleiro, Kun Qian, Xin Jing, Alexander Kathan, Bin Hu, and Bjoern W Schuller. Audio self-supervised learning: A survey. *arXiv preprint arXiv:2203.01205*, 2022.

[49] Mengjie Zhao, Tao Lin, Fei Mi, Martin Jaggi, and Hinrich Schütze. Masking as an efficient alternative to finetuning for pretrained language models. *arXiv preprint arXiv:2004.12406*, 2020.

[50] Rich Caruana. Multitask learning. *Machine Learning*, 1997.

[51] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv:1706.05098*, 2017.

[52] Yi-Chen Chen, Po-Han Chi, Shu-wen Yang, Kai-Wei Chang, Jheng-hao Lin, Sung-Feng Huang, Da-Rong Liu, Chi-Liang Liu, Cheng-Kuang Lee, and Hung-yi Lee. Speechnet: A universal modularized model for speech processing tasks. *arXiv preprint arXiv:2105.03070*, 2021.

[53] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Listening while speaking: Speech chain by deep learning. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 301–308. IEEE, 2017.

[54] Yi Ren, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Almost unsupervised text to speech and automatic speech recognition. In *International Conference on Machine Learning*, pages 5410–5419. PMLR, 2019.

[55] Zhuo Chen, Shinji Watanabe, Hakan Erdogan, and John R Hershey. Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[56] Tomi Kinnunen, Lauri Juvela, Paavo Alku, and Junichi Yamagishi. Non-parallel voice conversion using i-vector plda: Towards unifying speaker verification and transformation. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5535–5539. IEEE, 2017.

[57] Mingyang Zhang, Xin Wang, Fuming Fang, Haizhou Li, and Junichi Yamagishi. Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet. *arXiv preprint arXiv:1903.12389*, 2019.

[58] Abhinav Jain, Minali Upreti, and Preethi Jyothi. Improved accented speech recognition using accent embeddings and multi-task learning. In *Interspeech*, pages 2454–2458, 2018.

[59] Yi-Chen Chen, Zhaojun Yang, Ching-Feng Yeh, Mahaveer Jain, and Michael L Seltzer. Aipnet: Generative adversarial pre-training of accent-invariant networks for end-to-end speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6979–6983. IEEE, 2020.

[60] Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Yu-An Chung, Yuxuan Wang, Yonghui Wu, and James Glass. Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5901–5905. IEEE, 2019.

[61] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31, 2018.

[62] Jing-Xuan Zhang, Zhen-Hua Ling, and Li-Rong Dai. Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:540–552, 2019.

[63] Jing-Xuan Zhang, Zhen-Hua Ling, Yuan Jiang, Li-Juan Liu, Chen Liang, and Li-Rong Dai. Improving sequence-to-sequence voice conversion by adding text-supervision. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6785–6789. IEEE, 2019.

[64] Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993. IEEE, 2020.

[65] Salah Zaiem, Titouan Parcollet, and Slim Essid. Pretext tasks selection for multitask self-supervised speech representation learning. *arXiv preprint arXiv:2107.00594*, 2021.

[66] Santiago Pascual, Mirco Ravanelli, Joan Serrà, Antonio Bonafonte, and Yoshua Bengio. Learning problem-agnostic speech representations from multiple self-supervised tasks. In *INTERSPEECH*, 2019.

[67] Yizhou Lu, Mingkun Huang, Xinghua Qu, Pengfei Wei, and Zejun Ma. Language adaptive cross-lingual speech representation learning with sparse sharing sub-networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6882–6886. IEEE, 2022.

[68] Yi-Chen Chen, Shu-wen Yang, Cheng-Kuang Lee, Simon See, and Hung-yi Lee. Speech representation learning through self-supervised pretraining and multi-task finetuning. *arXiv preprint arXiv:2110.09930*, 2021.

[69] Sherif Abdou and Michael S Scordilis. Beam search pruning in speech recognition using a posterior probability-based confidence measure. *Speech Communication*, 42(3-4):409–428, 2004.

[70] Janne Pylkkönen. New pruning criteria for efficient decoding. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[71] Vesa Siivola, Teemu Hirsimaki, and Sami Virpioja. On growing and pruning kneser–ney smoothed n-gram models. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1617–1624, 2007.

[72] Tianxing He, Yuchen Fan, Yanmin Qian, Tian Tan, and Kai Yu. Reshaping deep neural network for fast decoding by node-pruning. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 245–249. IEEE, 2014.

[73] Hainan Xu, Tongfei Chen, Dongji Gao, Yiming Wang, Ke Li, Nagendra Goel, Yishay Carmiel, Daniel Povey, and Sanjeev Khudanpur. A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5929–5933. IEEE, 2018.

[74] Yuekai Zhang, Sining Sun, and Long Ma. Tiny transducer: A highly-efficient speech recognition model on edge devices. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6024–6028. IEEE, 2021.

[75] Hugo Van Hamme and Filip Van Aelten. An adaptive-beam pruning technique for continuous speech recognition. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 4, pages 2083–2086. IEEE, 1996.

[76] Dong Yu, Frank Seide, Gang Li, and Li Deng. Exploiting sparseness in deep neural networks for large vocabulary speech recognition. In *2012 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 4409–4412. IEEE, 2012.

[77] Sankaran Panchapagesan, Daniel S Park, Chung-Cheng Chiu, Yuan Shangguan, Qiao Liang, and Alexander Gruenstein. Efficient knowledge distillation for rnn-transducer models. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5639–5643. IEEE, 2021.

[78] Yuan Shangguan, Jian Li, Qiao Liang, Raziel Alvarez, and Ian McGraw. Optimizing speech recognition for the edge. *arXiv preprint arXiv:1909.12408*, 2019.

[79] Zhaofeng Wu, Ding Zhao, Qiao Liang, Jiahui Yu, Anmol Gulati, and Ruoming Pang. Dynamic sparsity neural networks for automatic speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6014–6018. IEEE, 2021.

[80] Ganesh Venkatesh, Alagappan Valliappan, Jay Mahadeokar, Yuan Shangguan, Christian Fuegen, Michael L Seltzer, and Vikas Chandra. Memory-efficient speech recognition on smart devices. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8368–8372. IEEE, 2021.

[81] Yangyang Shi, Yongqiang Wang, Chunyang Wu, Ching-Feng Yeh, Julian Chan, Frank Zhang, Duc Le, and Mike Seltzer. Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6783–6787. IEEE, 2021.

[82] Shengqiang Li, Menglong Xu, and Xiao-Lei Zhang. Efficient conformer-based speech recognition with linear attention. *arXiv preprint arXiv:2104.06865*, 2021.

[83] Stefan Braun and Shih-Chii Liu. Parameter uncertainty for end-to-end speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5636–5640. IEEE, 2019.

[84] Sharan Narang, Erich Elsen, Gregory Diamos, and Shubho Sengupta. Exploring sparsity in recurrent neural networks. *arXiv preprint arXiv:1704.05119*, 2017.

[85] Dawei Gao, Xiaoxi He, Zimu Zhou, Yongxin Tong, Ke Xu, and Lothar Thiele. Rethinking pruning for accelerating deep inference at the edge. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 155–164, 2020.

[86] Jian Xue, Jinyu Li, and Yifan Gong. Restructuring of deep neural network acoustic models with singular value decomposition. In *Interspeech*, pages 2365–2369, 2013.

[87] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur. Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Interspeech*, pages 3743–3747, 2018.

[88] Cheng-I Jeff Lai, Yang Zhang, Alexander H Liu, Shiyu Chang, Yi-Lun Liao, Yung-Sung Chuang, Kaizhi Qian, Sameer Khurana, David Cox, and Jim Glass. Parp: Prune, adjust and re-prune for self-supervised speech recognition. *Advances in Neural Information Processing Systems*, 34, 2021.

[89] Lodagala VSV Prasad, Sreyan Ghosh, and S Umesh. Pada: Pruning assisted domain adaptation for self-supervised speech representations. *arXiv preprint arXiv:2203.16965*, 2022.

[90] Yingzhu Zhao, Chongjia Ni, Cheung-Chi Leung, Shafiq Joty, Eng Siong Chng, and Bin Ma. A unified speaker adaptation approach for asr. *arXiv preprint arXiv:2110.08545*, 2021.

[91] Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. What's hidden in a randomly weighted neural network? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11893–11902, 2020.

[92] Eran Malach, Gilad Yehudai, Shai Shalev-Schwartz, and Ohad Shamir. Proving the lottery ticket hypothesis: Pruning is all you need. In *International Conference on Machine Learning*, pages 6682–6691. PMLR, 2020.

[93] Yonggan Fu, Qixuan Yu, Yang Zhang, Shang Wu, Xu Ouyang, David Cox, and Yingyan Lin. Drawing robust scratch tickets: Subnetworks with inborn robustness are found within randomly initialized networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[94] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[95] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[96] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[97] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

[98] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.

[99] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. *arXiv preprint arXiv:1810.05270*, 2018.

[100] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[101] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836, 2020.

[102] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[103] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

[104] Morgane Riviere, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. Unsupervised pretraining transfers well across languages. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7414–7418. IEEE, 2020.

[105] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021.

[106] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*, 2019.

[107] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[108] Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, 2013.

[109] Steven Moran and Daniel McCloy, editors. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena, 2019.

[110] Heting Gao, Junrui Ni, Yang Zhang, Kaizhi Qian, Shiyu Chang, and Mark Hasegawa-Johnson. Zero-shot cross-lingual phonetic recognition with external language embedding. In *Proc. Interspeech*, pages 1304–1308, 2021.

[111] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[112] Philip Sedgwick. Spearman's rank correlation coefficient. *Bmj*, 349, 2014.