Hybrid Unsupervised and Supervised Multitask Learning For Speech Recognition in Low Resource Languages

Srinivasa Raghavan [1,2], *Kumar Shubham*[1]

[1]International Institute of Information Technology, Bangalore, India
[2]Navana Tech India Private Limited, Bangalore, India

srinivasaraghavan.km@iiitb.org, kumar.shubham@iiitb.org

Abstract

With the recent developments of highly parametric deep learning architectures, many recent works have demonstrated their effectiveness with End-to-End approaches in achieving state-of-the-art performance in automatic speech recognition (ASR) tasks. But training such models requires a large amount of supervised labeled data to achieve good performance. This makes training ASR using such architectures in low resources languages challenging. Recent works have focused on rich representation learning using unsupervised data and fine-tuning them on task-specific limited supervised data. But it requires additional efforts of learning the rich unsupervised representations separately and then fine-tuning on task-specific data which is both time-consuming and resource exhaustive. Additionally, it does not impose any prior about downstream task over the unsupervised representation learning. In our work, we propose a hybrid multitask learning approach where both unsupervised and supervised datasets are used within the same architecture to train acoustic models under low-resource settings. Our hybrid learning approach provides required constraints over unsupervised representation learning by using a joint loss comprising CTC, Attention and reconstruction losses. We show that this leads to better utilization of the available datasets from both unsupervised and supervised paradigms and is readily applicable to both mono-lingual and cross-lingual scenarios.

Index Terms: speech recognition, unsupervised learning, low-resource setting, multitask learning

1. Introduction

Rapid advancement in End-to-End automatic speech recognition (E2E ASR) paradigm particularly with highly parametric Transformer based models [1, 2, 3], have demonstrated the effectiveness of End-to-End ASR system in real-world scenarios [4, 5]. Training heavily parameterized models are computationally expensive and require a large amount of clean labeled data. Collecting a large amount of labeled data is a time-consuming, expensive and labor-intensive process. This poses a serious limitation to train effective E2E ASR models for low resource languages.

Recently, two-stage training approaches like [6, 7, 8, 9] have been explored to circumvent the need for the large supervised dataset, where an ASR model is first trained in an unsupervised manner on large unlabeled data and then the model is further fine-tuned on a small labeled dataset. Such two-stage training approaches require additional training efforts to learn model parameters in an unsupervised setting and do not necessarily incorporate prior information about the downstream task.

Mathieu et al., [10] proposes a generalization approach for disentangling latent factors and demonstrates that matching latent space distribution of variational autoencoders with the desired target prior structure help provide interpretability of the learnt rich unsupervised representations. Without incorporating information on prior about downstream tasks leads to under-utilization of existing resources in unsupervised training for downstream task-specific use-cases in low-resource scenarios. The fine-tuning of learnt unsupervised weights can even lead to forgetting of essential language-specific features and might result in overfitting on the limited supervised data.

Recent works in the area of computer vision and natural language processing have explored various multitask learning approaches, where different subtasks of a higher-level task can be jointly trained to achieve better performance in-depth estimation [11], scene parsing [12] and natural learning understanding tasks [13]. For Robust ASR, [14] introduce enhancement as multitask learning loss and [15] use a problem agnostic speech encoder with multiple decoders also called workers for different multitask learning tasks for self-supervised learning scenarios.

Inspired by multitask learning approaches, we propose a hybrid learning paradigm using a joint loss, where supervised ASR learning and unsupervised representation learning are considered as two different tasks for a given network architecture which is trained using both supervised and unsupervised datasets in a single training process. Under the given setting, our ASR model jointly learns to use both supervised and unsupervised datasets, where the supervised dataset provides an additional prior about the task of interest for unsupervised representation learning and at the same time the larger unsupervised dataset acts as a regularizer for the ASR model training over the smaller supervised training set.

Nadig et al., [16], in their work have shown that different layers of encoders learn a different hierarchy of speech abstraction. Encouraged from such works, in our hybrid learning approach the unsupervised task taps encoder representations from different encoder layers to compute reconstruction loss and the supervised task uses the last layer of encoded representation from the encoder. This provides additional freedom for our model to preserve and use the low-level abstraction of the speech signal. Our results empirically show the benefit of using different layers of the encoder for unsupervised reconstruction.

We also show the effectiveness of our proposed architecture in a cross-lingual setting for ASR tasks, where we leverage labeled data from the target language as well as unlabelled data from other languages and demonstrate improvement in ASR performance in resource-constrained scenarios.

2. Proposed Framework

In machine learning, we generally train a model for a specific task using the objective of our interest, where the performance and generalization are often limited by the amount of dataset for a given task. Recently proposed multitask learning approaches

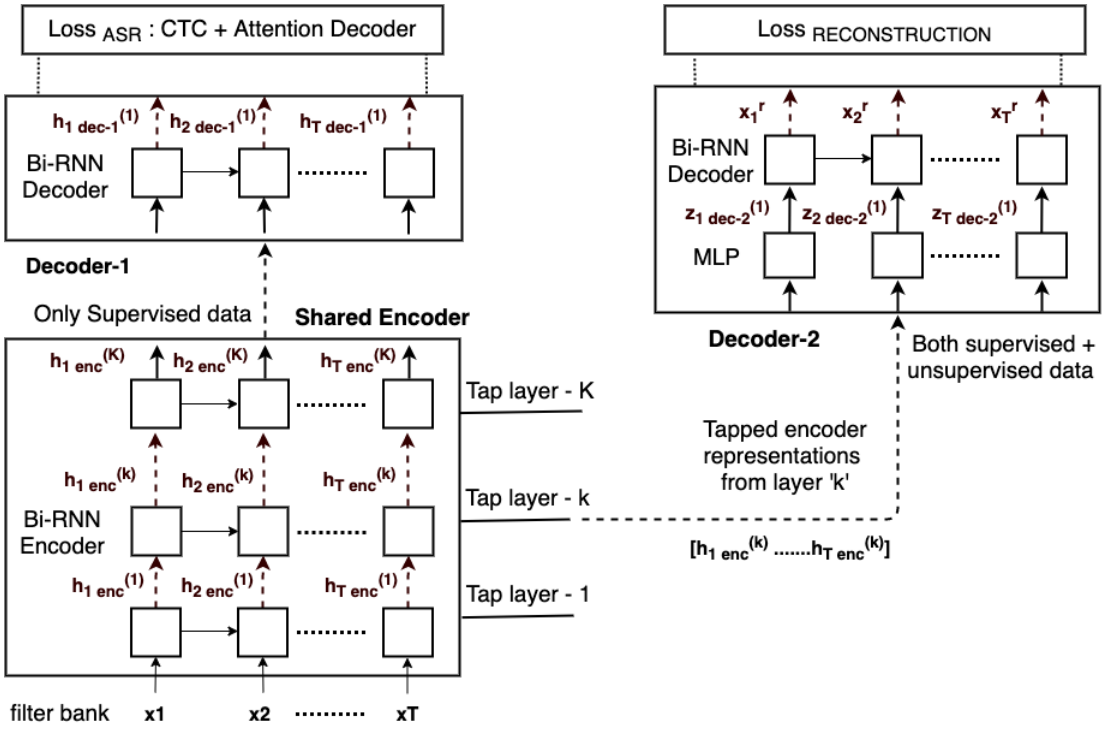

Figure 1: *Architecture Diagram*

attempt to overcome this limitation by learning representations that are robust to the variabilities across several related tasks. This enforces essential prior information about the relationship between related tasks into the model training, where generating a task-specific large dataset is challenging.

In [17], the prior information about alignment is introduced into the Joint CTC-Attention system using multitask learning approach to address errors in alignment and transcription. The advantages of such multitask learning become even more important in resource-constrained scenarios which often suffer from a lack of a large amount of labeled dataset.

In our work, we take inspiration from multitask learning paradigm and propose a hybrid learning architecture that can utilize both supervised and unsupervised datasets jointly in one training process and show its effectiveness in low resource scenarios. We hypothesize that such an approach allows supervised learning to provide necessary prior about the task of interest for the unsupervised representation learning, while unsupervised learning provides the required regularization to ensure that the model doesn't overfit on the limited supervised dataset.

Our End-to-End architecture consists of a shared-encoder and a multi-decoder network as shown in Figure-1 where each decoder tries to minimize a given task-specific loss. Decoder-1 uses labeled speech data to train for ASR, while Decoder-2 uses shared encoder representation to reconstruct the input raw filter-bank features for better parameter learning of shared encoder.

In our work, we use a tapping scheme where the latent representation from a different layer of the encoder is passed directly to the decoder for unsupervised representation learning and hence provide a regularizer using different levels of abstraction of the speech signal in overall ASR learning. Our experiments reveal that using encoded representation directly from the first layer of the shared encoder provided the best performance. These results align with the hypothesis of [16], that within a multitask learning framework, different layers of shared encoder learns different level of abstraction. The supervised ASR task might need a condensed higher-level representation of speech signal whereas the same level of abstraction might not be suitable for unsupervised representation learning.

To perform ASR, we compute Joint CTC and Attention loss [18] ($Loss_{ASR}$) to train Decoder-1 on available supervised subset of data using Equation-1.

$$Loss_{ASR} = \alpha * Loss_{CTC} + (1 - \alpha) * Loss_{Att} \quad (1)$$

where $0 \leq \alpha \leq 1$

To perform unsupervised representation learning, we compute reconstruction loss ($Loss_{Recons}$) to train Decoder-2 on both supervised and unsupervised data to reconstruct input raw filter-bank features.

For the given input filter-bank feature sequence, $\mathbf{X}_{input} = [\mathbf{x}_1, \mathbf{x}_2 ... \mathbf{x}_T]$, we use a selective tapping scheme to choose which corresponding encoder layer's latent representation to be passed to Decoder-2 for reconstruction. Using this encoder representation, the Decoder-2 generates $\mathbf{X}_{recons} = [\mathbf{x}_1^r, \mathbf{x}_2^r ... \mathbf{x}_T^r]$ and the reconstruction loss $Loss_{Recons}$ is computed as mean squared error as in Equation-2.

$$Loss_{Recons} = |\mathbf{X}_{input} - \mathbf{X}_{recons}|^2 \quad (2)$$

With this architecture, the Decoder-1 and Decoder-2 can be trained on a given limited supervised dataset. The Decoder-2 has the flexibility to utilize available large unsupervised datasets for better representation learning in a low resource setting. In each training batch, with mixed samples from both unsupervised and supervised datasets, we compute both ASR and reconstruction loss for the supervised data, and only consider re-

construction loss with an unsupervised set to obtain joint loss $Loss_{Hybrid}$ as a multitask learning loss using Equation-3

$$Loss_{Hybrid} = \alpha_1 * Loss_{CTC} + \alpha_2 * Loss_{Att} + \alpha_3 * Loss_{Recons} \quad (3)$$

where, $\alpha_1 + \alpha_2 + \alpha_3 = 1$ and α_1 , α_2 and α_3 are tuned hyperparameters.

3. Experiments

3.1. Data

To evaluate the effectiveness of our proposed hybrid unsupervised and supervised multitask learning approach, at first we experiment with only supervised dataset TIMIT (standard 3696 Train and 192 Test utterances) exploring a different set of α's considering only last layer encoder representations to compute joint loss $Loss_{Hybrid}$. We further assess the effectiveness of including reconstruction loss, by tapping hidden representations from different encoder layers for the same set of α's.

In the next set of experiments, to leverage the advantage of unsupervised data with our proposed multitask learning approach in a low resource setting, we carry out experiments on subsets of Tamil and Telugu datasets which is the publicly available data provided by SpeechOcean.com and Microsoft for the Indic languages Tamil, Telugu and Gujarathi released for the low-resource ASR challenge [19]. We consider the Tamil 5000 and 500 utterances subset, from the provided Train set as our supervised Train set Ta-Sup5k and Development set Ta-500 respectively. We also consider other non-overlapping sets of 10000 utterances subsets from each of Tamil and Telugu Train set namely Ta-Unsup10k and Te-Unsup10k respectively as unsupervised data (ignoring the transcriptions corresponding to these utterances) for our experiments. We report ASR performance on the standard Tamil Test set Ta-3026 (3026 utterances).

3.2. Training and decoding

We have used Encoder-Decoder modules of ESPNET Toolkit [20] to develop our proposed hybrid unsupervised and supervised multitask learning architecture as shown in Figure-1. The shared encoder consists of 5 layered Bi-directional GRU with projections (BGRUP) encoder, 320 encoder units with projections. The Decoder-1 consists of a single-layer decoder, 300 decoder units supporting CTC and Coverage Attention mechanism. The Decoder-2 consists of a 5 layer BGRUP decoder with 320 decoder units with projection and a linear layer. We use an Adadelta optimizer and train for 20 epochs with training batch size 30 without subsampling.

For the shared encoder representations to jointly support the supervised ASR task and unsupervised reconstruction task, we train Decoder-1 for ASR loss to predict phonemes instead of characters or words. So we consider Phoneme error rate (PER) as an evaluation metric for our study.

3.3. Results and Discussion

3.3.1. Adding reconstruction loss to the Joint CTC-Attention loss for supervised multitask learning with TIMIT

As shown in Table-1, when compared to the setting where only CTC loss and Attention based loss of Equation-3 is used as total loss $Loss_{Hybrid}$ estimated on the supervised dataset with α_3 = 0, α_1 = α_2 = 0.5, the inclusion of reconstruction based loss (i.e., for α_3 = 0 to 0.33) with only supervised TIMIT Train set leads to improvement in the PER. Here the Decoder-2 is provided with hidden representations from the last encoder layer during the forward pass for calculating reconstruction loss.

With our experiments on equal weights α's (α_1=α_2=α_3) for CTC, Attention and reconstruction loss we observe PER of 20%. With higher weight $\alpha_3 = 0.6$ for $Loss_{CTC}$, and comparatively lesser weights α_2 and α_3, for $Loss_{Att}$ and $Loss_{Recons}$ respectively we observe improvements in PER.

Table 1: *PER% observed for Supervised training with TIMIT, with different α's considering representations from the last encoder layer for training Decoder-2*

α_1	α_2	α_3	PER
0.5	0.5	**0.0**	21.5
0.33	0.33	**0.33**	20
0.2	0.6	0.2	20
0.6	0.2	0.2	19.7

We notice that incorporating reconstruction loss in model training helps in achieving better performance as shown in Table-1. On the TIMIT dataset, without considering reconstruction loss ($\alpha_3 = 0$) the PER is 21.5%, but with the reconstruction loss, the PER drops by 1.5%. This performance improvement can be attributed to the regularization effect of reconstruction loss as an additional task, which prevents the model from overfitting on a given supervised dataset.

3.3.2. Adding reconstruction loss to the Joint CTC-Attention loss for supervised multitask learning with TIMIT, considering tapping encoder representations from different encoder layers

With equal weights for the α's, we observe consistently lower PERs when the hidden representations are tapped from different layers of the shared encoder to compute reconstruction loss. Next, we observe that with this set of best α's (α_1=α_2=α_3), we get lower PERs when we tap representations from the first layer of the shared encoder.

Table 2: *PER% observed for Supervised training with TIMIT, with different α's, considering tapping representations from different encoder layers (Enc. Tap) for training Decoder-2*

Enc. Tap	α_1	α_2	α_3	PER
	0.33	**0.33**	**0.33**	**19.4**
1	0.2	0.6	0.2	19.8
	0.6	0.2	0.2	20.2
	0.33	**0.33**	**0.33**	**19.4**
2	0.2	0.6	0.2	20.3
	0.6	0.2	0.2	19.9
	0.33	**0.33**	**0.33**	**20.1**
4	0.2	0.6	0.2	20.8
	0.6	0.2	0.2	21
	0.33	**0.33**	**0.33**	**20**
5	0.2	0.6	0.2	20
	0.6	0.2	0.2	19.7

Compared to the representations from the last layers of the

encoder network, the lower layers of encoder networks learn lower-level characteristics of the speech signal that is more useful for reconstruction loss decoder (Decoder-2). As the shared layer is subject to both ASR and reconstruction losses, a reconstruction loss using the initial layers of encoded representation helps the hybrid model architecture to utilize different levels of abstraction of speech signal representations for the reconstruction task, which in turn influences a more effective representation for the ASR task from the last layer of the encoder.

3.3.3. Hybrid Supervised and Unsupervised multitask learning in a low resource cross-lingual setting with Indian languages Tamil and Telugu

As shown in Table-3 with the model trained on supervised Tamil subset Train Ta-Sup5k, when reconstruction loss is not included ($\alpha_3 = 0$), we observe a PER of 16.1%. When we include the reconstruction loss the overall PER drops by 2.1%. This result is consistent with our observations on Supervised TIMIT data (Table-1). All the results reported in this section are based on PER evaluated on the Tamil Test set Ta-3026.

Table 3: *%PER for hybrid multitask learning approach on low resource Indian languages in both monolingual and cross-lingual settings*

Unsup. Train	Sup. Train	α_1	α_2	α_3	PER
–	Ta-Sup5k	0.5	0.5	0.0	16.1
Ta-Sup5k	Ta-Sup5k	0.33	0.33	0.33	14
Ta-Unsup10k	Ta-Sup5k	0.33	0.33	0.33	12.8
Te-Unsup10k	Ta-Sup5k	0.33	0.33	0.33	12.6

We intend to utilize existing unsupervised data to improve ASR performance with our proposed hybrid unsupervised and supervised multitask learning architecture in low-resource scenarios. For this, we consider both supervised set Ta-Sup5k and unsupervised set Ta-Unsup10k. In each training batch, with mixed samples from both unsupervised and supervised sets, we compute both ASR and reconstruction loss for the supervised data and only consider reconstruction loss with an unsupervised set.

By this way of utilizing unsupervised data in a monolingual setting, we observe 12.8% PER i.e., an absolute 1.2% PER improvement when compared to 14% PER with joint loss training that does not use any additional unsupervised dataset.

We further demonstrate leveraging unsupervised data from phonetically similar languages (languages that have a substantial commonality in their phone-sets) in a cross-lingual setting using our proposed hybrid unsupervised and supervised multitask learning framework. To do this, we consider two languages Telugu and Tamil which belong to the Dravidian family of Indian languages. We consider unsupervised Telugu subset Te-Unsup10k and supervised Tamil subset Ta-Sup5k to train our hybrid unsupervised and supervised architecture. We observe 12.6% PER i.e., an absolute 1.4% PER improvement with additional unsupervised Telugu data when compared to 14% PER with joint loss training that does not use any additional unsupervised dataset.

4. Conclusions

Low-resource languages often suffer from a lack of availability of large labeled corpus for training highly parametrized ASR models. We present a hybrid ASR system that can leverage data from both unsupervised and supervised settings to better utilize the available resources. Our joint learning approach uses a joint loss comprising CTC, Attention and reconstruction losses. We use a multi-tapping scheme where representation from different layers of encoder network can be used for reconstruction loss and helps the model to utilize different levels of abstraction of speech signal representations for the reconstruction task, which in turn influences a more effective representation for the ASR task from the last layer of the encoder. Results show that the inclusion of reconstruction loss and tapping from initial layers of encoded representation helps in achieving better PER performance.

We benchmark the performance of our model in a low-resource setting by experimenting with low-resource Indic languages. We observe an improvement in ASR performance with the inclusion of an additional unsupervised dataset in a monolingual setting. We further demonstrate that we can leverage unsupervised data from phonetically similar language in a cross-lingual setting to improve the performance of ASR.

As future work, we will further explore ways of incorporating data from different languages in a multi-lingual setting within the multitask framework. We believe that multitask learning approaches that integrate unsupervised representation learning will provide a practical way to develop speech recognition for low-resource languages.

5. Acknowledgements

We would like to thank Shreekantha Nadig from IIIT Bangalore for his support and for sharing his insights from his thesis work on Joint Multitask learning End-to-End ASR framework.

6. References

[1] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[2] Y. Shi, Y. Wang, C. Wu, C. Fuegen, F. Zhang, D. Le, C.-F. Yeh, and M. L. Seltzer, "Weak-attention suppression for transformer based speech recognition," *arXiv preprint arXiv:2005.09137*, 2020.

[3] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7829–7833.

[4] N. Moritz, T. Hori, and J. Le, "Streaming automatic speech recognition with the transformer model," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6074–6078.

[5] P. Baquero-Arnal, J. Jorge, A. Giménez, J. A. Silvestre-Cerda, A. S. Iranzo-Sánchez, J. Civera, and A. Juan, "Improved hybrid streaming asr with transformer language models," *Proc. Interspeech 2020*, pp. 2127–2131, 2020.

[6] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[7] A. Baevski, M. Auli, and A. Mohamed, "Effectiveness of self-supervised pre-training for speech recognition," *arXiv preprint arXiv:1911.03912*, 2019.

[8] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

[9] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[10] E. Mathieu, T. Rainforth, N. Siddharth, and Y. W. Teh, "Disentangling disentanglement in variational autoencoders," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4402–4412.

[11] A. Atapour-Abarghouei and T. P. Breckon, "To complete or to estimate, that is the question: A multi-task approach to depth completion and monocular depth estimation," in *2019 International Conference on 3D Vision (3DV)*. IEEE, 2019, pp. 183–193.

[12] D. Xu, W. Ouyang, X. Wang, and N. Sebe, "Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 675–684.

[13] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," *arXiv preprint arXiv:1901.11504*, 2019.

[14] S. Parveen and P. Green, "Multitask learning in connectionist robust asr using recurrent neural networks," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[15] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-task self-supervised learning for robust speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6989–6993.

[16] S. Nadig, V. Ramasubramanian, and S. Rao, "Multi-target hybrid ctc-attentional decoder for joint phoneme-grapheme recognition," in *2020 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2020, pp. 1–5.

[17] S. Nadig, S. Chakraborty, A. Shah, C. Sharma, V. Ramasubramanian, and S. Rao, "Jointly learning to align and transcribe using attention-based alignment and uncertainty-to-weigh losses," in *2020 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2020, pp. 1–5.

[18] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[19] B. M. L. Srivastava, S. Sitaram, R. K. Mehta, K. D. Mohan, P. Matani, S. Satpal, K. Bali, R. Srikanth, and N. Nayak, "Interspeech 2018 low resource automatic speech recognition challenge for indian languages." in *6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages, SLTU 2018, 29-31 August 2018, Gurugram, India*. ISCA, 2018, pp. 11–14.

[20] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.