Preprint:
Farooq, M.U. orcid.org/0000-0002-6610-0923, Ahmad, R. and Hain, T. orcid.org/0000-0003-0939-3464 (Submitted: 2023) MUST: a multilingual student-teacher learning approach for low-resource speech recognition. [Preprint - arXiv] (Submitted)

https://doi.org/10.48550/arXiv.2310.18865

MUST: A MULTILINGUAL STUDENT-TEACHER LEARNING APPROACH FOR LOW-RESOURCE SPEECH RECOGNITION

Muhammad Umar Farooq, Rehan Ahmad, Thomas Hain

Speech and Hearing Research Group, University of Sheffield, UK

ABSTRACT

Student-teacher learning or knowledge distillation (KD) has been previously used to address data scarcity issue for training of speech recognition (ASR) systems. However, a limitation of KD training is that the student model classes must be a proper or improper subset of the teacher model classes. It prevents distillation from even acoustically similar languages if the character sets are not same. In this work, the aforementioned limitation is addressed by proposing a MUltilingual Student-Teacher (MUST) learning which exploits a posteriors mapping approach. A pre-trained mapping model is used to map posteriors from a teacher language to the student language ASR. These mapped posteriors are used as soft labels for KD learning. Various teacher ensemble schemes are experimented to train an ASR model for low-resource languages. A model trained with MUST learning reduces relative character error rate (CER) up to 9.5% in comparison with a baseline monolingual ASR.

Index Terms— multilingual, knowledge distillation, automatic speech recognition, low-resource languages

1. INTRODUCTION

State-of-the-art automatic speech recognition models nowadays require huge amounts of data for training. However, only 23 out of 7000 language are spoken by more than half of the world's population [1]. Thus a large number of languages lack enough data resources to train a modern ASR system. Multilingual and cross-lingual systems have got a lot of attention in recent years to exploit resources of other languages to overcome the data scarcity issue for training of speech technologies for low-resource languages [2, 3, 4, 5]. Although multilingual ASR systems are considered to perform better when compared with their monolingual counterparts of low-resource languages, the performance of these systems often degrades due to mixing of unrelated languages [6, 7, 8]. This has given rise to various studies with an aim to improve a monolingual ASR using multilingual or cross-lingual resources rather than training a unified model [9, 10, 11]. Recently, some efforts have been made towards multilingual knowledge distillation where multilingual models are used for knowledge distillation to train a language-specific student ASR model [12].

Student-teacher training or knowledge distillation (KD) [13] has widely been used to distil the knowledge from either a single or multiple teacher models [14] to train a student model. This technique of transferring a teacher's knowledge to a student model either at output layer [13] or at intermediate stages [15] has been used for many tasks such as model compression [14, 16] and domain generalisation [17, 18, 19]. The student model is trained with a combined objective of minimising the KL-divergence loss for prediction of the teacher's posteriors (soft labels) and a classification loss with the original training labels (hard labels).

Since KL-divergence loss is used as KD loss between a teacher's soft labels and the student model posteriors [13], the output classes of student model must be a subset of the teacher model. Studies on multilingual knowledge distillation have used teacher models where student model output classes are an improper subset of the teacher model classes [12]. Nevertheless, it still constrains a teacher model to cover all the student classes yet a lot of languages have diverse character sets and writing scripts. As it happens, a number of languages which are acoustically similar or belong to same language families are written in different scripts such as Turkish and Kazakh (Turkic), Urdu and Hindi (Indo-Aryan), and Greek and Armenian (Indo-European). It prevents various languages to distil their knowledge for training of a closer language ASR model. Though a lot of previous works have explored knowledge distillation for domains where student and teachers are from same language and have same output classes, to the best of authors knowledge no work has been done for either cross-lingual knowledge distillation or to overcome the aforementioned problem. This paper presents a step towards overcoming the obstacle for applying KD in cross-lingual settings.

To that end, a posteriors mapping technique is exploited here which has recently been proposed with an objective to analyse the cross-lingual acoustic-phonetic similarities [20]. A mapping model has been trained to map posteriors from a source language ASR to those of a target language ASR given a target language speech utterance. The work has been employed for multilingual and cross-lingual model fusion for speech recognition on phonemes level [21] and end-to-end

979-8-3503-0689-7/23/$31.00 ©2023 IEEE

ASR systems [22]. In this work, we make use of a source (teacher) language ASR model followed by a source-target (teacher-student) mapping model to act as a teacher for student model training. Source and target are synonymously used as teacher and student respectively for rest of the paper. For N languages, one mapping model is trained for each target language to map posteriors from other $N-1$ source languages ASR models to the posteriors of the target language ASR. Posteriors from ASR model of a source language followed by the target language mapping model are used as soft labels for knowledge distillation. Having multiple teachers from $N-1$ source languages, different existing weighting schemes along with a proposed self-adaptive weighting (SAW) are experimented for teachers ensemble to generate soft-labels. The key contribution of MUST learning is to overcome the limitation of multilingual KD and use teachers from diverse languages for multilingual knowledge distillation. ASR models trained with MUST learning for low resource languages yield a gain of up to 9.5% in terms of relative character error rate (CER).

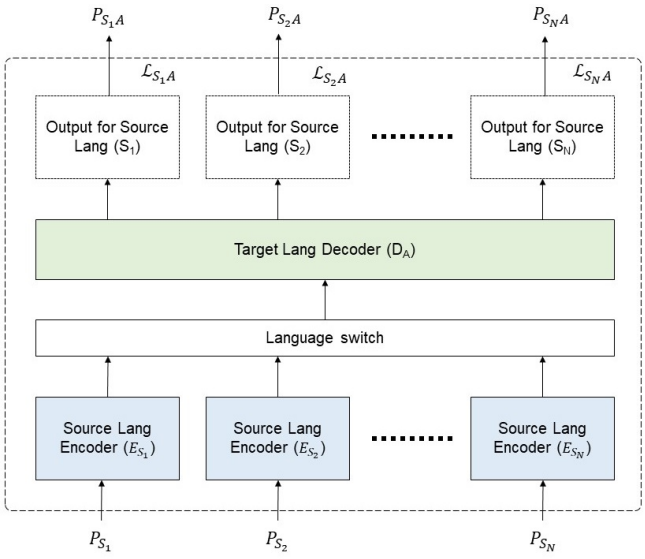

Fig. 1. Architecture of the MESD mapping model [22]

2. MAPPING MODELS

Let the monolingual acoustic models of the target language (M_A) and the i^{th} source language (M_{S_i}), a mapping model N_{S_iA} is trained to map posteriors from M_{S_i} (P^{S_i} of dimension d_{S_i}) to the posteriors of M_A (P^{S_iA} of dimension d_A). Given a set of target language observations $X = \{x_1, x_2, \ldots, x_T\}$, posterior distributions from the target and the i^{th} source acoustic models ($P^Z = \{p_1^Z, p_2^Z, \ldots, p_T^Z\}$ where $Z \in \{A, S_i\}$) are attained. A sequence-to-sequence mapping model is trained to map posteriors from i^{th} source acoustic model (P^{S_i}) to the target language posteriors (P^{S_iA}) using the KL-divergence loss as follows;

$$\mathcal{L}_{S_iA}(\theta) = \sum_{n=1}^{B} p_n^A \cdot (\log p_n^A - \log p_n^{S_iA}) \tag{1}$$

where B is the number of frames in one batch for training of a mapping model.

Mapping models are trained to learn the mappings between posterior distributions from a source language and the target language ASR given a target language utterance. An underlying assumption is that these mapping models are able to learn some language-related relationships between posterior distributions of a source and the target language acoustic models.

Multi Encoder Single Decoder (MESD) architecture, as proposed in [22] and shown in Figure 1, is used for all the mapping models. A single MESD model is trained for each target language which consists of multiple encoders (same as number of source languages) and a single decoder with a language switch in-between.

For training of the MESD mapping model, outputs from all the source acoustic models (P^{S_i}) for a given utterance u, are fed to source-language dependent encoders successively. Embeddings from the final layer of the encoders are then passed to a single target-language dependent decoder. Target posteriors (P^A) are generated by decoding utterance u through the target-language ASR. KL loss (Equation 1) is calculated between target posteriors (P^A) and the output of mapping model decoder (P^{S_iA}). Mapping model loss is calculated as the mean of the losses of all encoder-decoder pairs.

$$\mathcal{L}_A(\theta) = \sum_{K} w_k \cdot \mathcal{L}_{S_kA} \tag{2}$$

where K is the number of the source languages ($N-1$). In the case of mean average, w_k is given as $w_k = \frac{1}{K}$. $\mathcal{L}_{S_kA}$ is given in Equation 1 where each frame serves as a training example. It enables the mapping models to converge in low-resource setting as a small amount of data provides millions of training examples. Since the average loss of all encoder-decoder pairs for a mapping model causes unbalanced training across languages, rank sum dynamic weighting scheme [23] is applied to weight the losses for each encoder-decoder loss. In this scheme, the weights are assigned based on their normalised ranks. w in Equation 2 then becomes

$$w_r = \frac{2(K+1-r)}{K(K+1)} \tag{3}$$

where r is rank of the language when the source languages are sorted in descending order of their losses. It restricts model from biasing towards a specific language or a group of languages.

Though a mapping model contains multiple encoders, any encoder can be used with decoder during decoding and MESD does not require data stream from all the encoders for a given utterance. It implies that mappings can be obtained

having input even from only one source language at a time. Training of these mapping models allows to use any source language ASR for decoding the data of a target language followed by the source-target mapping model.

3. MULTILINGUAL STUDENT-TEACHER (MUST) LEARNING

As described in the Section 1, output classes of the student model are required to be a proper or improper subset of the teacher model classes for knowledge distillation. It prevents a teacher language to distil its knowledge to train a student model if writing scripts or character sets are not the same. In this work, mapping models are employed to overcome this issue and distil knowledge from diverse teacher languages to train a student model of a low-resource language.

For a given target language (L_{tgt}), an encoder-decoder sequence-to-sequence monolingual acoustic model is trained using hybrid CTC loss as given in Equation 4.

$$\mathcal{L}_{ASR}(\theta) = \alpha\mathcal{L}_{CTC} + (1-\alpha)\mathcal{L}_{seq} \tag{4}$$

where $\mathcal{L}_{CTC}$ is applied on top of the encoder after an affine projection layer. $\mathcal{L}_{seq}$ is cross-entropy loss which is applied on the decoder's output.

For MUST learning, soft-labels from a single teacher or an ensemble of multiple teacher models are used to distil knowledge for the training of a model for low-resource language. $\mathcal{L}_{seq}$ loss in Equation 4 is modified as

$$\mathcal{L}_{seq}(\theta) = \lambda\mathcal{L}_{KD} + (1-\lambda)\mathcal{L'}_{seq} \tag{5}$$

where $\mathcal{L'}_{seq}$ is still cross-entropy loss and $\mathcal{L}_{KD}$ is the knowledge distillation loss which is ensemble of multiple teachers and given as

$$\mathcal{L}_{KD} = \sum_{K} \mathcal{W}_k \mathcal{L}_{T_k} \tag{6}$$

$\mathcal{L}_{T_k}$ is KL-divergence loss between posteriors from k^{th} teacher model and the student model.

$$\mathcal{L}_{T_k} = \sum_{B} p^{T_k} \log \frac{p^s}{p^{T_k}} \tag{7}$$

where p^{T_k} and p^s are the posterior distributions from k^{th} teacher and the student model respectively. A teacher model is a source language ASR model followed by a target language mapping model N_A as shown in Figure 2. α and λ in Equations 4 and 5 are hyper-parameters and different teacher weighting strategies are experimented for $\mathcal{W}$ in Equation 6.

Given an utterance u of a target language, it is decoded through all the source language ASR systems (M_{S_i}) which generate posteriors for their output classes (P^{S_i}). Then a pre-trained target language mapping model (N_A) is used to map the output posteriors from source language ASR systems to the target language ASR (P^{S_iA}). Output posteriors from the mapping model are used as soft targets for student model training. So, ASR of each source language along with a target language mapping model act as a teacher model for MUST learning. For the target language student learning, experiments are conducted using an ensemble of multiple teachers (source languages) and a single teacher to generate soft labels for KD training.

3.1. Self-adaptive weighting

Performance of the ensemble teacher models depends on the choice of $\mathcal{W}$ in the Equation 6 for each teacher loss. A straightforward approach is the teacher-averaging (TA) where all the teachers are assigned equal weights. However, all the teachers have different relationships with the student task and thus impact differently. In case of multilingual systems, all teacher languages are not equally similar and assigning the equal weights does not prove to be an optimal way.

In this work, a self-adaptive weighting (SAW) scheme is proposed. Motivated by a recent work which makes use of posterior distributions [24], teacher models get relative weights based on their confidence in soft-labels. Furthermore, rather than assigning the same weights for a batch, teachers weights are calculated on-the-fly for each utterance. Given an utterance u of T frames, mean of $max(p_t)\forall t \in \{1, 2, \cdots, T\}$ is calculated where p_t is the posterior distribution at time t.

$$\mu_k = \frac{1}{T}\sum_{t} \max(p_t^{T_k})$$

Then the weight of each teacher is set to

$$\mathcal{W}_k = \frac{\tau^{\mu_k}}{\sum_K \tau^{\mu_k}} \tag{8}$$

where $\sum_K \mathcal{W} = 1$ and τ is a hyper-parameter for the sake of statistically significant weights distribution across the teachers. Increasing the value of τ increases the deviation of the teachers weights from the mean weight.

4. EXPERIMENTAL SETUP

4.1. Data set

In this work, all the experiments are conducted using the same data sets as in the previous work on mapping models [21].

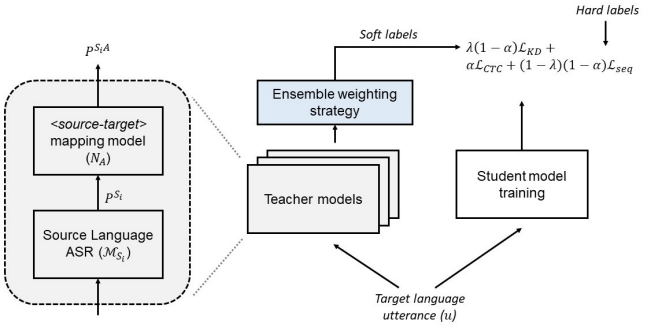

Fig. 2. Architecture of Multilingual Student-Teacher (MUST) learning

Table 1. Details of BABEL data sets used for the experimentation

Lang	Train		Eval	
	# hours	# spks	# hours	# spks
Tamil *(tam)*	59.11	372	7.8	61
Telugu *(tel)*	32.94	243	4.97	60
Cebuano *(ceb)*	37.44	239	6.59	60
Javanese *(jav)*	41.15	242	7.96	60

Four low-resource languages (Tamil (*tam*), Telugu (*tel*), Cebuano (*ceb*) and Javanese (*jav*)) from the IARPA BABEL speech corpus [25] with their Full Language Packs (FLP) are used for ASR training and evaluation. Most of the BABEL data sets consist of conversational telephone speech with real-time background noises and is quite challenging because of conversation styles, limited bandwidth, environment conditions and channel. All the utterances without any speech are discarded. The details of the data sets are given in Table 1.

For training of the mapping models, a subset of 30 hours is randomly selected from each BABEL language pack. This data is further split into 29 hours of train set and 1 hour of dev set.

4.2. Student and teacher models

As described earlier, Hybrid CTC/attention architecture [26] is used to train all speech recognition models which consists of three modules that are; a shared encoder, an attention decoder and a CTC module. The training process jointly optimises the weighted sum of CTC and attention model as given in Equation 4 but $\mathcal{L}_{seq}$ is a cross-entropy loss for the training of teacher models which implies that $\mathcal{L}_{seq}$ in Equation 4 is same as $\mathcal{L}'_{seq}$ of Equation 5.

The input to the model are 40 filterbanks and the output of the model is byte-pair encoded (BPE) tokens. All the models are trained for 100 BPE tokens for each language and SentencePiece library [27] is used for tokenisation. Both student and teacher models are of the same capacity ($\sim$ 170.9 million) throughout the experimentation.

During decoding, the final prediction is made based on a weighted sum of log probabilities from both the CTC and attention components. Given a speech input X, the final prediction $\hat{Y}$ is given by;

$$\hat{Y} = \arg\max_{Y \in \mathcal{Y}} \{\gamma \log P_{CTC}(Y|X) + (1-\gamma) \log P_{seq}(Y|X)\} \tag{9}$$

where γ is a hyper-parameter.

For speech recognition task, results are reported in terms of percent character error rate. The SpeechBrain toolkit [28] is used for training of all ASR systems.

4.3. Mapping models

A multi encoder single decoder model is trained for each target language. In an MESD model, there are three encoders and only one attention decoder. Each encoder and single decoder consists of one bidirectional RNN layer. Mapping models are also of the same capacity ($\sim$ 2.59 millions) for all the languages and trained on equal amounts of data.

4.3.1. Performance metric

Performance of mapping models is reported in terms of accuracy. Accuracy of a mapping model is measured as the ratio of correctly mapped frames (CMF) to the total number of frames (TF). Correctly mapped frames are defined as the frames where the most probable classes from mapping model and the target posteriors are the same (that is $\arg\max_k(p^A_{t,k}) == \arg\max_k(p^{S_iA}_{t,k})$ where k is the index of a class in the output vector p_t).

4.4. MUST learning

For the experimentation in this work, values of α and λ of Equation 4 and 5 are varied between the range of $[0, 1]$ and the numbers are reported with the best configuration. The values of α and γ are kept constant for all the experimentation while λ may vary for different languages. For teachers' ensemble, various weighting strategies are experimented to assign the weights ($\mathcal{W}$). Conventional teacher averaging (TA) is compared with proposed self-adaptive weighting (SAW). In teacher averaging, all the teachers get the equal weights and does not change during whole training. Frame-wise max (FWM) selects posteriors from a different teacher for each frame of a given utterance. For each frame, the teacher having a maximum value of posteriors among all the teachers is selected. Recently, an elitist sampling (ES) has been proposed and prove to outperform TA and FWM weighting strategy for speech recognition domain generalisation task [24]. ES takes mean of maximum posterior values of all the frames for a given utterance. Then the soft labels of the teacher having the highest value are used for that given utterance.

In previous work, posterior distributions from all the mapping models have been fused as an acoustic model which have outperformed the monolingual acoustic models [21]. However, the fused weights have been fine-tuned for test set. An experiment is also done here by assigning fine-tuned weights (FTW) for the test set. These weights are manually fine-tuned and might be a sub-optimal solution. Lastly, a comparison is shown with using only one teacher model for knowledge distillation rather than ensemble of all the teachers to reduce the computational complexity. The objective is to analyse the gap in performance by reducing the teacher models. In case of single teacher (ST) distillation, only the teacher from the closest language is selected. 'Closest' language is defined in terms of mapping models accuracy. For a target language, the source language with maximum mapping model accuracy is selected as the teacher model.

5. RESULTS AND DISCUSSION

5.1. Teacher models

As described in Section 3, a teacher model for MUST learning is a combination of a teacher language ASR and a student-

Table 2. Accuracy of the pre-trained mapping models

Source Lang.	*Target/Student languages*			
	tam	*tel*	*ceb*	*jav*
tam	-	48.88	60.53	62.24
tel	47.46	-	48.32	54.64
ceb	45.98	46.22	-	65.51
jav	46.97	47.40	65.04	-

teacher mapping model. Since most of the languages included in this study have different scripts and character sets, ASR of a language cannot be used for decoding the data of the other one. Pre-trained mapping models are used for each student-teacher pair in this work. Performance of the pre-trained mapping models is tabulated in Table 2 in terms of accuracy. For each target language, accuracy of the mapping model is shown for all the source-target mapping modules.

5.2. MUST learning

For multilingual student-teacher learning, various teacher ensemble strategies are explored. Before training a student model, the ensemble strategies are applied for teacher models fusion. For a given target language, outputs from all the teacher models are fused together in a weighted sum. CER is calculated by applying greedy search on fused teacher outputs. All the discussed ensemble strategies (in Section 4.4) including teacher averaging (TA), frame-wise max (FWM), elitist sampling (ES), self-adaptive weighting (SAW) and fine-tuned weights (FTW) are experimented and results are tabulated in Table 3.

Analysis shows that the trend of student model performance with different ensemble strategies is same as the trend for model fusion. So, the student models here are trained using only top three best performing teachers' ensemble techniques in Table 3 which are SAW, TA and FTW. %CER of student model are shown in Table 4. First row is the %CER from a baseline monolingual ASR using an explicit RNNLM trained on limited text of train set transcriptions. Although the performance of TA and SAW is almost same for model fusion (in Table 3), student models trained with SAW ensemble reduces average CER by 1.27% relative if compared with the models trained with TA weighting (Table 4). With the weights fine-tuned for test set, average CER is reduced to 42.13% from 42.93% of SAW trained models which is a relative improvement of 2.4% compared to monolingual models.

Table 3. MUST teachers performance in terms of %CER

MUST teachers	*Target/Student languages*				
	tam	*tel*	*ceb*	*jav*	*avg*
ES	57.24	83.23	72.09	75.93	72.12
FWM	57.39	82.54	62.45	70.14	68.13
SAW	57.34	84.31	61.99	67.32	67.74
TA	57.38	84.31	61.98	67.26	67.73
FTW	57.34	83.36	60.03	59.45	65.04

Table 4. Performance (%CER) of student model trained using MUST learning

MUST teachers	*Target/Student languages*				
	tam	*tel*	*ceb*	*jav*	*avg*
mono	44.28	56.18	31.26	40.90	43.16
TA	44.72	57.02	32.43	42.61	44.20
SAW	44.59	56.14	31.87	39.11	42.93
FTW	44.42	55.79	30.80	37.50	42.13
ST	**43.77**	**55.56**	**29.43**	**36.98**	**41.44**

In another experiment, knowledge from only a single teacher model is distilled for student model training (ST in Table 4). For each target language, the closest language is chosen as a teacher model. As described earlier, a closest source language for a target language is the one which has highest mapping model accuracy for the target language. Student models trained using the single teacher outperform all other students for all the languages by an average improvement of 4% in performance of monolingual model. For *jav* target language, a relative improvement of 9.5% is observed.

Both *ceb* and *jav* yield more gains in performance than *tam* and *tel* because the mapping models' accuracies are higher for these two languages. It is evident that the gain for each language depends directly on the performance of corresponding mapping model. Student training with ST does not have any test set information and performs even better than FTW which has fine-tuned weights for the test set. The results are inline with the performance of mapping models and the results reported for data augmentation using the mapping models in [22]. Since some source-target mapping models does not perform very well for some of the teacher languages, the teacher knowledge introduces noise in student training and makes it hard for student to learn. Knowledge distillation from only a single student not only improves ASR performance but also reduces the computational complexity.

6. CONCLUSION

This paper presents a multilingual student-teacher (MUST) approach to address a limitation of knowledge distillation systems to apply in a cross-lingual settings. In MUST learning, a teacher model is a combination of a source language ASR followed by a source-target mapping model. Pre-trained mapping models are used to map posteriors from a source language ASR to those of the target language ASR (Table 2). Various weighting strategies are explored for teachers ensemble (Table 3). Student models are trained for each language with top performing ensemble strategies. A student model trained with MUST learning proves to outperform baseline monolingual ASR by a relative gain of up to 9.5%.

7. ACKNOWLEDGEMENTS

This work was partly supported by LivePerson Inc. at the Liveperson Research Centre.

8. REFERENCES

[1] "Languages of the world," https://www.ethnologue.com/guides/how-many-languages, accessed: 2023-07-18.

[2] S. T. Abate, M. Y. Tachbelie, and T. Schultz, "Multilingual acoustic and language modeling for ethio-semitic languages," in *Proc. Interspeech 2020*, 2020, pp. 1047–1051.

[3] M. Y. Tachbelie, S. T. Abate, and T. Schultz, "Development of multilingual asr using globalphone for less-resourced languages: The case of ethiopian languages," in *Proc. Interspeech 2020*, 2020, pp. 1032–1036.

[4] M. Karafiát, M. K. Baskar, P. Matějka, K. Veselý, F. Grézl, and J. Černocky, "Multilingual blstm and speaker-specific vector adaptation in 2016 but babel system," in *IEEE SLT*, 2016, pp. 637–643.

[5] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

[6] V. Pratap, A. Sriram, P. Tomasello, A. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert, "Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters," in *Proc. Interspeech 2020*, 2020, pp. 4751–4755.

[7] W. Hou, Y. Dong, B. Zhuang, L. Yang, J. Shi, and T. Shinozaki, "Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning," in *Proc. Interspeech 2020*, 2020, pp. 1037–1041.

[8] N. Gaur, B. Farris, P. Haghani, I. Leal, P. J. Moreno, M. Prasad, B. Ramabhadran, and Y. Zhu, "Mixture of informed experts for multilingual speech recognition," in *ICASSP*, 2021, pp. 6234–6238.

[9] Q. Xu, A. Baevski, and M. Auli, "Simple and Effective Zero-shot Cross-lingual Phoneme Recognition," in *Proc. Interspeech 2022*, 2022, pp. 2113–2117.

[10] O. Klejch, E. Wallington, and P. Bell, "Deciphering Speech: a Zero-Resource Approach to Cross-Lingual Transfer in ASR," in *Proc. Interspeech 2022*, 2022, pp. 2288–2292.

[11] M. Morshed and M. Hasegawa-Johnson, "Cross-lingual articulatory feature information transfer for speech recognition using recurrent progressive neural networks," in *Proc. Interspeech 2022*, 2022, pp. 2298–2302.

[12] I. Leal, N. Gaur, P. Haghani, B. Farris, P. J. Moreno, M. Prasad, B. Ramabhadran, and Y. Zhu, "Self-Adaptive Distillation for Multilingual Speech Recognition: Leveraging Student Independence," in *Proc. Interspeech 2021*, 2021, pp. 2556–2560.

[13] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015. [Online]. Available: http://arxiv.org/abs/1503.02531

[14] K. P. Huang, T.-H. Feng, Y.-K. Fu, T.-Y. Hsu, P.-C. Yen, W.-C. Tseng, K.-W. Chang, and H.-Y. Lee, "Ensemble knowledge distillation of self-supervised speech models," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[15] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," 2015.

[16] H.-G. Kim, H. Na, H. Lee, J. Lee, T. G. Kang, M.-J. Lee, and Y. S. Choi, "Knowledge distillation using output errors for self-attention end-to-end models," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6181–6185.

[17] Y. Wang, H. Li, L.-p. Chau, and A. C. Kot, "Embracing the dark knowledge: Domain generalization using regularized knowledge distillation," in *Proceedings of the 29th ACM International Conference on Multimedia*, ser. MM '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 2595–2604.

[18] B. Kim, S. Yang, J. Kim, and S. Chang, "Domain generalization on efficient acoustic scene classification using residual normalization," 2021.

[19] G. Fang, Y. Bao, J. Song, X. Wang, D. Xie, C. Shen, and M. Song, "Mosaicking to distill: Knowledge distillation from out-of-domain data," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 11 920–11 932.

[20] M. U. Farooq and T. Hain, "Investigating the Impact of Crosslingual Acoustic-Phonetic Similarities on Multilingual Speech Recognition," in *Proc. Interspeech 2022*, 2022, pp. 3849–3853.

[21] M. U. Farooq, D. A. H. Narayana, and T. Hain, "Non-Linear Pairwise Language Mappings for Low-Resource Multilingual Acoustic Model Fusion," in *Proc. Interspeech 2022*, 2022, pp. 4850–4854.

[22] M. U. Farooq and T. Hain, "Learning Cross-lingual Mappings for Data Augmentation to Improve Low-Resource Speech Recognition," in *Proc. Interspeech 2023*, 2023.

[23] E. Roszkowska, "Rank ordering criteria weighting methods – a comparative overview," *Optimum. Studia Ekonomiczne*, no. 5(65), p. 14–33, 2013.

[24] R. Ahmad, M. A. Jalal, M. U. Farooq, A. Ollerenshaw, and T. Hain, "Towards domain generalisation in asr with elitist sampling and ensemble knowledge distillation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[25] M. J. F. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED," in *Proc. 4th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2014)*, 2014, pp. 16–23.

[26] 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)*, 2017, pp. 4835–4839.

[27] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," 2018.

[28] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," 2021.