INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Knowledge Distillation Based Training of Universal ASR Source Models for Cross-lingual Transfer

Takashi Fukuda[1], *Samuel Thomas*[2]

[1]IBM Research AI, Chuo-ku Hakozaki, Tokyo, 103-8510, Japan
[2]IBM Research AI, Yorktown Heights, NY, USA

fukuda1@jp.ibm.com, sthomas@us.ibm.com

Abstract

In this paper we introduce a novel knowledge distillation based framework for training universal source models. In our proposed approach for automatic speech recognition (ASR), multilingual source models are first trained using multiple language-dependent resources before being used to initialize language specific target models in low resource settings. For the proposed source models to be effective in cross-lingual transfer to novel target languages, the training framework encourages the models to perform accurate universal phone classification while ignoring any language-dependent characteristics present in the training data set. These two goals are achieved by applying knowledge distillation to improve the models' universal phone classification performance along with a shuffling mechanism that alleviates any language specific dependencies that might be learned. The benefits of this proposed technique are demonstrated in several practical settings, where either large amounts or only limited quantities of unbalanced multilingual data resources are available for source model creation. Compared to a conventional knowledge transfer learning method, the proposed approaches achieve a relative WER reduction of 8-10% in streaming ASR settings for various low resource target languages.

Index Terms: speech recognition, transfer learning, knowledge distillation, multi-lingual data, low-resource

1. Introduction

In ideal settings, a neural network based speech recognition model is trained on large amounts of data that covers a wide variety of acoustic conditions including pronunciation and dialect variations [1–3]. Unlike high resource languages like US English or Japanese that have been actively been studied [4–7], sufficient data resources for such training, does not exist in a majority of other world languages and dialects [8–10]. Transfer learning has recently attracted attention as a technique that can help improve the performance of various machine learning tasks in such low resource settings [11–13]. Several variants of these transfer learning techniques have been proposed for speech recognition and have shown to be useful for training ASR systems in low resource languages with insufficient amounts of data [14–17].

In its basic idea, with cross-lingual knowledge transfer, an ASR model trained on large amounts of language data in a source domain, is used to first initialize model parameters of a target model in new language. The target ASR model is then adapted or fine-tuned, often in low resource ASR settings, using just the target language data. When only limited amounts of training data are available in the target language, this kind of cross-lingual transfer learning often provides significant performance improvements while also helping cut the cost of data collection in the target language [14, 15].

For effective cross-lingual transfer learning, while training a source model we need to simultaneously achieve two competing goals: (a) increase the model's ability to discriminate between sound classes (phonemes) of each language available to train the source model, while at the same time, (b) discourage the model from learning any language specifics present in the source language data sets. In practice, language resources available to train such multilingual source models are also usually very unbalanced. From a data size point of view, the amount of data can range from dozens of hours to several thousands of hours across various languages. The data can also be acoustically very diverse. The success of cross-lingual transfer learning with such unbalanced multilingual resources is hence dependent on how well a universal source model can be created with respect to the above mentioned training objectives. In [15], Yi *et al.* propose a language-adversarial transfer learning approach to alleviate any language dependent information from being learned by a source model by adding language identification layers to the model training. In this method, the source model is adversarially trained to make it impossible to identify the input language. The shared layers of the source model in turn, are hence encouraged to learn more language invariant features. With this technique, [15] demonstrates an improvement on source model training using a small, balanced multilingual data set.

The goal of our method is to improve cross-lingual transfer learning in more practical situations, especially when much larger, unbalanced multilingual data sets are available. On the source model side, while we focus our approach on creating a better model within the framework of transfer learning, as our target model, we consider a computationally inexpensive and low-latency streaming ASR system. We first propose a framework to construct a multilingual source model that leverages *knowledge distillation* to improve the phone classification capabilities of the model in each language of the training set [18–23]. Within this framework, language-dependent teacher networks are constructed with corresponding language data. A single student network is then trained with soft labels derived from the multiple language-dependent teachers. This student model serves as the source model for subsequent transfer learning. To train the source model effectively, we next introduce two shuffling mechanisms which are applied on the input data of the teacher networks and the language dependent layers of the source model. These methods are used to discourage the source network from learning any language dependent information as we explicitly train the source model on *mismatched* language

http://dx.doi.org/10.21437/Interspeech.2021-796

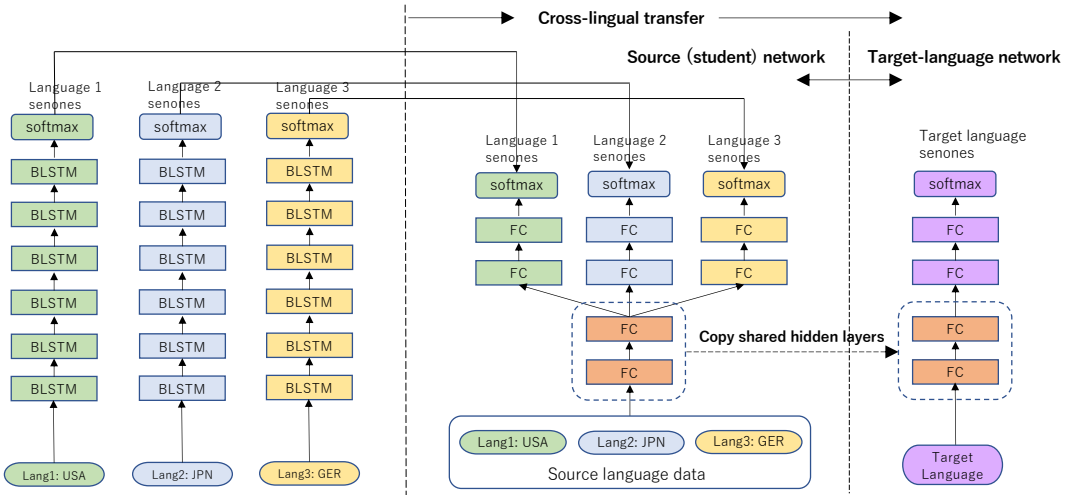


Figure 1: *Schematic of the proposed transfer learning framework exploiting knowledge distillation (best viewed in color).*

data. In the first of these methods, while training on a specific language, mismatched data is derived from teacher networks corresponding to other languages. In the second method, by shuffling language-dependent layers of the source network, subsequently language independent layers are presented with mismatched internal data representations during the training process. In experiments with multilingual language data resources containing ten languages (US English, UK English, AU English, Japanese, Korean, German, Spanish, Italian, French, and Brazilian Portuguese), we show that our proposed technique significantly improves performance especially under a condition where training data in the target language is very small. The technique provides a significant relative improvement of 8-10% over conventional techniques in practical experimental conditions with multilingual data.

2. Modeling Components

This section briefly reviews conventional transfer learning methods and knowledge distillation techniques related to our proposed framework.

2.1. Knowledge Transfer Learning

Conventional cross-lingual transfer learning [14, 15] corresponds to a part of our proposed method and is illustrated on the right side of Figure 1. In this case, a multilingual source model is composed of several shared hidden layers followed by multiple language dependent layers. Both these sets of layers are jointly optimized using a multilingual training set. Unlike our proposed method described in Section 3, the source model is trained on hard targets without utilizing any teacher networks. After training, the shared layers are treated as a universal feature transformation that works for other novel languages as well. The shared layers are expected to exhibit language invariance while being able to discriminate between various speech sounds. In addition to these shared layers in the source model, each language in the training data set has its own hidden layers and a softmax layer. The output labels of the softmax layer correspond to language specific speech sounds (tied quinphone states in our experiments).

Once a source model has been created, the shared language independent parameters of the model are transferred to a target model and set as its initial parameters. The target model is subsequently trained only on the target language data, which is often available only in limited amounts. With this initialization from the source model, the target model can hence be better constructed, than training from scratch. The final goal in this setting is hence to construct a better source model that effectively serves as an initial set of parameters for every target language, without the need for any prior adjustments except the direct copy of the parameters to the target model.

2.2. Knowledge Distillation

In our proposed framework, knowledge distillation is used to enhance phoneme classification abilities for each language in the source model. Knowledge distillation is a well-known technique used to mimic complicated teacher networks with a simple student network. Applying techniques based on knowledge distillation to ASR have recently received considerable attention in the community [18–26]. In the knowledge distillation framework, instead of training models which have reduced computational requirements and improved latency performances directly on hard targets in a single step, training is performed in two separate steps. In the first step, complex teacher neural network such as bidirectional LSTM, VGG [27] , and ResNet [28] models are initially trained using hard targets. Student networks are then trained on the soft outputs of teachers using a training criteria that minimizes the differences between the student and teacher distributions as

$$\mathcal{L}(\theta) = -\sum_{i} q_i(\boldsymbol{x}) \log p_i(\boldsymbol{x}), \tag{1}$$

where i is an index of context dependent phones, $q_i(\boldsymbol{x})$ is the so-called soft label from the teacher network for input feature $\boldsymbol{x} \in \hat{\mathcal{X}}$, which also works as a pseudo label. $p_i(\boldsymbol{x})$ is output probability of the class from the student network. With soft labels $q_i(\boldsymbol{x})$, competing classes will have small but non-zero posterior probabilities for each training example.

Within this knowledge distillation framework, several techniques have been proposed to create better student networks using multiple teacher networks [22, 29]. These prior works focus on how various teacher networks with different topologies, and trained with the same data, can be effectively exploited. In contrast, our proposed method in this paper, focuses on using multiple language-dependent teachers with the same topology but designed with different language data sets that are maximally

dissimilar from each other. This topology is selected in order to inject enhanced language-invariant classification abilities to the shared layers of the source model as it is being trained for cross-lingual transfer.

3. Proposed Framework

3.1. Training the source model as a student network

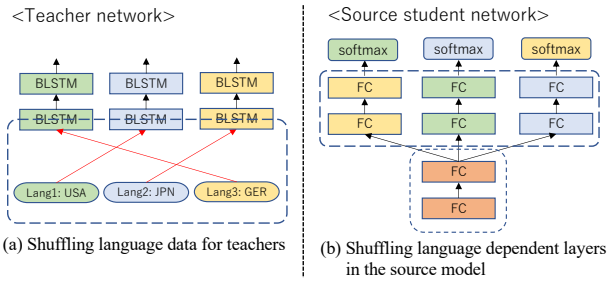
The proposed framework combined with knowledge distillation is shown in Fig. 1. Unlike conventional transfer learning, the source model is first trained as a student network, learning from teacher networks corresponding to each language data. Similar to standard knowledge distillation training techniques, the teacher models used in this step are independently trained on hard targets of corresponding language data prior to being used for training the student model. Given our final goal of creating a light-weight, streaming universal source model that can be used as initial parameters for any novel language, the teacher networks we use are complex, offline but highly accurate, bidirectional LSTM based models. In contrast the student model is a much simpler network with fully connected layers.

3.2. Shuffling mechanisms

We next introduce two novel shuffling steps to discourage the source network from learning any language specific constructs during its training. As mentioned earlier, for each language in the multilingual training set, we train language specific teacher networks. We denote the language a particular teacher network is trained on, as called the network's L1 language. Other languages of the multilingual training pool are referred to as L2 languages of that network.

In a knowledge distillation framework, teacher networks are traditionally only used to produce soft labels corresponding to their L1 languages. These teacher networks can however also process and produce cross-lingual soft labels for L2 languages in terms of its L1 language output set. In the first of our shuffling methods, the input data processed by various teacher networks is reordered in terms of languages, to produce such cross-lingual labels. As illustrated in Fig.2-A, US English data is input to Japanese teacher, as if the English data was Japanese. Given that English is an L2 language for this network, cross-lingual soft labels generated by the model are then used to update the Japanese branch of the student source model. To do this, the same English data is fed to the student source model and used to update the Japanese branch of the network with the corresponding soft labels generated from the Japanese teacher. Given that the Japanese phonetic space does not fully cover the English phonetic space, this kind of cross-lingual training for the source model is difficult to achieve with conventional hard targets, but possible with our proposed framework.

In the second shuffling mechanism, instead of reordering data, the language dependent layers in the source student model are replaced by copying weights across languages as shown in Fig.2-B. Unlike the preceding shared layers, these language dependent layers are likely to learn language specifics during training. By reordering these layers, cross-lingual knowledge from other languages is injected into various language specific parts of the network. We hypothesize that these operations will make the source model more robust to various language data inputs and alleviate language dependencies of the source model. This kind of shuffling can also be thought of as some kind of localized self-transfer learning, as sub-parts of the network are initialized with various other parts of the same network.



(a) Shuffling language data for teachers

(b) Shuffling language dependent layers in the source model

Figure 2: *Shuffling methods (best viewed in color).*

Table 1: *Data size for source model training*

Source languate	Size (hours)
US-English, UK-English, AU-English	5000, 630, 275
Japanese, Korean	1000, 800
French, German, Italian, Spanish	1850, 1150, 910, 790
Brazilian Portuguese	330

4. Experiments

4.1. Data

The efficacy of our proposed knowledge distillation based training is measured on a series of experiments using a collection of in-house multilingual speech data sets, each sampled at 8kHz. The various languages and the amount of data in each language we use to train various models are listed in Table 1. Each data set consist of speech from different kinds of realistic and scripted settings, including telephone conversations, face-to-face dialogues, lectures, command and control utterances, and read speech. Parts of the speech corpora contain segments with additive ambient noises such as babble and vehicle noise, some contain room reverberation distortions, while others have cellular noise from mobile phone recordings. The noise range in all these recordings is roughly between 5-30dB. In general, this is a realistic, large-scale data set spanning several languages and domains compared to other publicly available multilingual corpora. In our experiments given the significant differences we observe across each dialect, we consider various dialects of English as separate languages.

4.2. Teacher and Source Student Network Topology

In our experiments, CNN based acoustic models are constructed as student source networks with 40 dimensional log Mel-frequency spectra augmented with Δ and $\Delta\Delta$s as inputs. The log Mel-frequency spectra are extracted by first applying mel scale integrators on power spectral estimates in short analysis windows (25 ms) of the signal followed by the log transform. Each frame of speech is also appended with a context of 11 frames after applying a speaker independent global mean and variance normalization. The CNN systems use two convolutional layers with 128 and 256 hidden nodes each in addition to four fully connected layers with 2048 per layer to estimate posterior probabilities of 9300 output targets. All of the 128 nodes in the first feature extracting layer are attached with

Table 2: *Comparisons between proposed modeling components.*

Modeling components	CC Set-1	CC Set-2
Conventional source model training	35.1	28.3
+ Knowledge distillation	34.0	27.9
+ Shuffling: Input lang. for teacher	33.2	26.2
+ Shuffling: Lang. dependent layers	32.9	25.5

Table 3: *Comparisons between source models for cross-lingual transfer (%WERs). Target language is AU English.*

Source Models	CC Set-1	CC Set-2	CC Set-3	CC Set-4	Reading-Noisy	Reading-Clean
No initialization	29.7	22.5	15.7	26.7	24.2	20.9
US English model	29.4	22.2	14.4	25.2	22.3	19.3
Conventional multilingual model	29.2	22.4	14.6	25.4	22.2	19.5
Proposed multilingual model	28.2	22.1	14.1	24.8	21.0	19.2

9×9 filters that are two dimensionally convolved with the input log Mel-filterbank representations. The second feature extracting layer with 256 nodes has a similar set of 3×4 filters that processes the non-linear activations after max pooling from the preceding layer. The non-linear outputs from the second feature extracting layer are then passed onto the subsequent fully connected layers. All the layers use the ReLU non-linearity. In this topology, two CNN layers followed by two fully connected layers are considered as shared layers in the source model, and the remaining layers are language-dependent layers.

With the same training data set, LSTM models are trained as language-dependent teacher networks. Each teacher network consists of 6 bidirectional LSTM layers with 512 units per direction and a linear bottleneck layer with 256 units. The teacher models are sequence trained after being constructed with a cross entropy criterion. Each teacher model is trained on its corresponding language data set as shown in Table 1. After these multiple teacher models are constructed, the source student model is trained on the soft labels generated from corresponding teacher models. For the shuffling method in Fig.2-A, cross-lingual soft labels equal to an amount of 5% in each L1 language is derived from L2 languages.

4.3. Experimental Results

In our first set of experiments, we evaluate the effectiveness of various component of our proposed method in a *low resource* multilingual training setting. In this experiment, only a subset of each language in the training data (roughly one-tenth) is used to train the multilingual source model. This training data set is still however unbalanced in terms of the amount of data available for each language. We report results on a wide variety of call center (CC) test sets in Australian English, a low resource language in our data set. Table 2 shows the results we obtain on this setup. In this experiment, we train the target model on only close to 30 hours of Australian English as well. As shown in Table 2, significant improvements are obtained using the various proposed methods: training the source model via knowledge distillation, shuffling input languages across teachers and reordering of language specific layers of the source network. Gains from each method are observed to be additive and result in a combined relative improvement of up to 10% reduction in WER with all the different proposed shuffling mechanisms in place. Similar trends are observed on other language test sets as well (not reported in this paper).

In our next set of experiments, we evaluate our proposed method on a *high resource* multilingual source model trained on the full set of multilingual resources listed in Table 1. Experimental results on various experiments in this setting are reported in Tables 3 and 4. During training, when a target model is constructed for a specific language, the corresponding language data is not included in its source model training. This ensures that the gains we observe from the improved source model training are only from the L2 languages present in the multilingual data set corresponding to that task. In experiments reported in Table 3, all 275 hours of English AU data were used to train the target AU model. It is interesting to observe that even in this *high resource* setting, significant gains are observed. Similar results are observed for other target languages. In experiments on Table 4, we vary the amount of another target language, German, from 50 to 1150 hours. The results in the table are averaged WERs calculated over a call center test suite. The proposed methods perform similarly in this case as well. It is interesting to observe that a model trained on just 50 hours of data performs quite as well as a model trained on significantly larger amounts of data (1150 hours of data). By being able to create accurate models using limited amounts of data, these results point to savings in terms of data collection costs, an equally important benefit of cross-lingual transfer.

Table 4: *Comparisons between training data size used for source model creation (%WERs). Target language is German.*

Source Models	D=50h	D=200h	D=1150h
No initialization	22.9	21.6	20.6
Conv. multilingual model	21.9	21.0	20.5
Proposed multilingual model	21.0	20.5	20.2

In both tables, the 'no initialization' experiment uses a target CNN model trained from random initialization and does not use any source model initialization. In the 'US English model' experiment, a source model is constructed using only US English, the language with the most amount of data in our multilingual data set. As another baseline, a 'conventional multilingual model' is also trained on the multilingual data using the method without using knowledge distillation. Both these baselines are compared with the 'proposed multilingual model' developed using techniques introduced in this paper. In Table 3, the US English model is regarded as a very accurate monolingual source model. In comparison, the 'conventional multilingual model' is trained in much larger amounts of multilingual data. Both these methods perform better than the 'no initialization' baseline. However, due to the significant data imbalance between languages in the multilingual data set, we hypothesize that the 'conventional multilingual model' model cannot effectively be trained to leverage the benefits of the large multilingual corpus. This model is hence not able to perform better than the monolingual 'US English model' baseline. In contrast, our proposed method maximizes the benefits of using the multilingual data set even when there is an imbalanced among the various languages and demonstrates up to 8.3% relative WER reduction in the *high resource* setting for the source model creation.

5. Conclusions

In this paper we have proposed a novel training strategy based on knowledge distillation to construct better source ASR models for cross-lingual transfer. The knowledge distillation based training improves the model's universal phone classification performance and an associated shuffling mechanism alleviates any language specific dependencies present in the data. Experiments on unbalanced multilingual resources show that our proposed technique can effectively leverage information from multilingual resources. The technique provides a significant relative improvement of up to 8.3% over the baseline system in various multilingual settings.

6. References

[1] L. Lu, C. Liu, J. Li, and Y. Gong, "Exploring transformers for large-scale speech recognition," in *Proc. Interspeech*, 2020, pp. 5041–5045.

[2] Y. Long, Y. Li, S. Wei, Q. Zhang, and C. Yang, "Large-scale semi-supervised training in deep learning acoustic model for ASR," in *IEEE Access*, 2019, pp. 133 615–133 627.

[3] M. Harper, "The automatic speech recognition in reverberant environments (ASpIRE) challenge," in *Proc. IEEE ASRU*, 2015, pp. 547–554.

[4] G. Saon, Z. Tüske, D. Bolanos, and B. Kingsbury, "Advancing RNN transducer technology for speech recognition," in *arXiv:2103.09935v1*, 2021.

[5] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription," in *IEEE ASRU*, 2013.

[6] S. Thomas, M. Suzuki, Y. Huang, G. Kurata, Z. Tuske, G. Saon, B. Kingsbury, M. Picheny, T. Dibert, A. Kaiser-Schatzlein, and B. Samko, "English broadcast news speech recognition by humans and machines," in *Proc. IEEE ICASSP*, 2019.

[7] T. Nagano, T. Fukuda, M. Suzuki, and G. Kurata, "Data augmentation based on vowel stretch for improving children's speech recognition," in *IEEE ASRU*, 2019.

[8] S. Thomas, K. Audhkhasi, and B. Kingsbury, "Transliteration based data augmentation for training multilingual ASR acoustic models in low resource settings," in *Proc. Interspeech*, 2020.

[9] M. Karafiatetal, "Multilingual BLSTM acoustic model with i-vector based adaptation," in *Proc. Interspeech*, 2017, pp. 719–723.

[10] T. Sercu, G. Saon, J. Cui, X. Cui, B. Ramabhadran, B. Kingsbury, and A. Sethy, "Network architectures for multilingual speech representation learning," in *Proc. IEEE ICASSP*, 2017, pp. 5295–5299.

[11] X. Li, Y. Grandvalet, F. Davoine, J. Cheng, Y. Cui, H. Zhang, S. Belongie, Y. Tsai, and M. Yan, "Transfer learning in computer vision tasks: Remember where you come from," in *Image and Vision Computing*, vol. 93, 2020.

[12] J. Zhang, W. Li, P. Ogunbona, and D. Xu, "Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective," in *ACM Computing Surveys*, 2019.

[13] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in *NAACL*, 2019.

[14] J. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. IEEE ICASSP*, 2013.

[15] J. Yi, J. Tao, Z. Wen, and Y. Bai, "Language-adversarial transfer learning for low-resource speech recognition," in *IEEE/ACM Transactions on Audio, Speech, And Language Processing*, vol. 27, no. 3, 2019, pp. 621–630.

[16] J. Ma, F. Keith, T. Ng, M. H. Siu, and O. Kimball, "Improving deliverable speech-to-text systems with multilingual knowledge transfer," in *Proc. Interspeech*, 2017, pp. 127–131.

[17] P. Swietojanski, A. Ghoshal, and S.Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Proc. IEEE SLT*, 2013, pp. 246–251.

[18] J. Li, R. Zhao, J. Huang, and Y. Gong, "Learning small-size DNN with output-distribution-based criteria," in *Proc. Interspeech*, September 2014, pp. 1910–1914.

[19] W. Chan, N. R. Ke, and I. Lane, "Transferring knowledge from a RNN to a DNN," in *Proc. Interspeech*, 2015, pp. 3264–3268.

[20] K. J. Geras, A. Mohamed, R. Caruana, G. Urban, S. Wang, O. Aslan, M. Philipose, M. Richardson, and C. Sutton, "Blending LSTMs into CNNs," in *ICLR Workshop*, 2016.

[21] Z. Tang, D. Wang, and Z. Zhang, "Recurrent neural network training with dark knowledge transfer," *Proc. IEEE ICASSP*, pp. 5900–5904, 2016.

[22] T. Fukuda, M. Suzuki, G. Kurata, T. Samuel, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers," in *Proc. Interspeech*, 2017, pp. 3697–3701.

[23] T. Fukuda and S. Thomas, "Implicit transfer of privileged acoustic information in a generalized knowledge distillation framework," in *Proc. Interspeech*, 2020, pp. 41–45.

[24] V. Manohar, P. Ghahremani, D. Povey, and S. Khudanpur, "A teacher-student learning approach for unsupervised domain adaptation of sequence-trained ASR models," in *Proc. IEEE SLT*, 2018.

[25] L. Mošner, M. Wu, A. Raju, S. H. Krishnan Parthasarathi, K. Kumatani, S. Sundaram, R. Maas, and B. Hoffmeister, "Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning," in *Proc. IEEE ICASSP*, 2019, pp. 6475–6479.

[26] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc, 2014, pp. 2654–2662.

[27] T. Sercu, C. Puhrsch, B. Kingsbury, and Y. LeCun, "Very deep multilingual convolutional neural networks for LVCSR," in *ICASSP*, 2016.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.

[29] Y. Chebotar and A. Waters, "Distilling knowledge from ensembles of neural networks for speech recognition," in *Proc. Interspeech*, 2016, pp. 3439–3443.