INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Implicit Transfer of Privileged Acoustic Information in a Generalized Knowledge Distillation Framework

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

IBM Research AI
[1]Chuo-ku Hakozaki, Tokyo, 103-8510, JAPAN
[2]Yorktown Heights, NY, USA
fukuda1@jp.ibm.com, sthomas@us.ibm.com

Abstract

This paper proposes a novel generalized knowledge distillation framework, with an implicit transfer of privileged information. In our proposed framework, teacher networks are trained with two input branches on pairs of time-synchronous lossless and lossy acoustic features. While one branch of the teacher network processes a privileged view of the data using lossless features, the second branch models a student view, by processing lossy features corresponding to the same data. During the training step, weights of this teacher network are updated using a composite two-part cross entropy loss. The first part of this loss is computed between the predicted output labels of the lossless data and the actual ground truth. The second part of the loss is computed between the predicted output labels of the lossy data and lossless data. In the next step of generating soft labels, only the student view branch of the teacher is used with lossy data. The benefit of this proposed technique is shown on speech signals with long-term time-frequency bandwidth loss due to recording devices and network conditions. Compared to conventional generalized knowledge distillation with privileged information, the proposed method has a relative improvement of 9.5% on both lossless and lossy test sets.

Index Terms: speech recognition, acoustic modeling, knowledge distillation, privileged information, bandwidth loss.

1. Introduction

As automatic speech recognition (ASR) systems become ubiquitous, these systems have to process user inputs from various acoustic settings. Typical speech processed by these systems include inputs from environments with stationary and non-stationary noises like restaurants or exhibition halls, meeting recordings in very reverberant settings, emotional speech, and distorted signals with time-frequency bandwidth loss due to faulty devices or transmission network conditions [1–3]. To effectively handle these noisy inputs, dedicated acoustic models for specific target domains are usually developed. ASR system are often made robust to additive and convolutive distortions known to be present in these various environments, using data augmentation techniques that incorporate these noises [4, 5]. In contrast, lossy spectra due to distortions introduced by recording devices or communication networks, are hard to predict in advance and are hence difficult to process effectively [6, 7]. To cope with such novel acoustic degradations to the input signals, it is hence important to improve the robustness of acoustic models to various lossy spectra. This paper proposes a method to construct robust acoustic models that can process speech signals with such lossy spectra.

As described earlier, one straightforward approach to handle this problem is to include speech data with artificially distorted lossy spectra to the training data set. While this method can increase the robustness of ASR models to lossy spectra, it has been observed that the performance of such models on normal (lossless) speech often degrades, given that the training data distribution has changed. This paper investigates a novel method based on knowledge distillation [8] to alleviate this problem, while also being robust to lossy spectra. Knowledge distillation is a technique to mimic complicated teacher networks with a simple student model for test time deployment [9–15]. As an extension to the standard knowledge distillation framework, generalized knowledge distillation techniques have recently attracted attention to supplement missing information or compensate for inferior quality of acoustic features, with privileged information [16], which is available only during training time. Several methods related to this generalized knowledge distillation framework with privileged information have been previously investigated for ASR [17–19].

In our proposed method, the teacher network has two branches similar to a Siamese network [20] with shared weights. While one of the two branches processes privileged information available during training (privileged view branch), the other part is trained on a data view available to the student network that eventually will be trained and deployed (student view branch). In terms of input features, this framework allows the privileged information branch to receive lossless features available only during training. The student view branch, on the other hand, processes paired time synchronous lossy features. By incorporating lossless features also into training, we hypothesize that the network will be able to incorporate this privileged knowledge into soft labels needed to train a student network. This proposed teacher network is trained with a composite two-part cross entropy loss. Our proposed technique is in particular more useful than the simple data augmentation approach when target signal distortions are relatively unusual such as lossy spectra because the degradation of acoustic quality on input features can be recovered from the privileged view branch. In experiments with the Switchboard corpus and an artificially created lossy version of this corpus, we show that our proposed technique significantly improves performance on lossy feature inputs without any degradations to lossless feature inputs. The technique provides a significant relative improvement of 9.5% over conventional generalized knowledge distillation techniques.

2. Modeling Components

Our proposed technique builds on two key components to train robust acoustic models: knowledge distillation to train efficient student networks and privileged knowledge to enhance the soft

http://dx.doi.org/10.21437/Interspeech.2020-1575

labels used to train the student networks.

2.1. Knowledge Distillation

Knowledge distillation is a technique 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 [9–15]. 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 [21], and ResNet [22] models are initially trained using hard targets. Compact acoustic models or 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. This technique has been shown to be very successful in various settings - fully supervised [11], semi-supervised [9], multilingual [23], sequence training [24], CTC models [25, 26] to train student networks that perform better than training similar models from scratch using just hard targets.

Instead of using the ground truth labels, the knowledge distillation training approach defines the loss function with an index of context dependent phones i as

$$\mathcal{L}(\theta) = -\sum_{i} q(i|\boldsymbol{x}) \log p(i|\boldsymbol{x}), \tag{1}$$

where $q(i|\boldsymbol{x})$ 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. The KL-divergence criterion used for training the student model equivalently also minimizes the cross entropy of the soft target labels. Usually, the same acoustic feature inputs are used to generate posteriors $q(i|\boldsymbol{x})$ and $p(i|\boldsymbol{x})$. Within this knowledge distillation framework, several techniques have been proposed to create better student networks using multiple teacher networks [18, 19, 27]. The proposed algorithm is detailed in Section 3.

2.2. Privileged Information

Knowledge distillation techniques addressed in Section 2.1 have recently been extended to the generalized distillation framework, where in addition to distillation of information from teacher networks, privileged information available only during training is also factored in [17–19]. The generalized knowledge distillation training with privileged information is expressed with soft label $q(i|\hat{\boldsymbol{x}})$ generated with better quality feature $\hat{\boldsymbol{x}}$ which is time-aligned to the degraded quality feature $\boldsymbol{x}$ as

$$\mathcal{L}(\theta) = -\sum_{i} q(i|\hat{\boldsymbol{x}}) \log p(i|\boldsymbol{x}). \tag{2}$$

where the teacher network is trained with acoustically better quality features $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}$ (or both $\boldsymbol{x} \in \mathcal{X}$ and $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}$). Instead of using $\boldsymbol{x}$, the corresponding better feature $\hat{\boldsymbol{x}}$ is used to generate better soft labels $q(i|\hat{\boldsymbol{x}})$ while degraded features are used to estimate posteriors $p(i|\boldsymbol{x})$ for the student network. By training acoustic models with this scheme, the student network tries to incorporate special knowledge from $\hat{\boldsymbol{x}}$, which results in creating robust acoustic models against adverse acoustic conditions.

3. Modeling Algorithm

Figure 1 gives an overview of our proposed method. Similar to standard knowledge distillation training techniques, the proposed method has a teacher network training stage, followed by a subsequent training of a student network as an acoustic model for test deployment. A key difference of this approach compared to standard knowledge distillation techniques, is that the teacher network has two input branches that process a pair of lossless and lossy features. While training the teacher network, a cross entropy loss is computed not only from hard label targets but also uses the loss from the privileged information view branch as described later. On the other hands, the student network has only a single branch that is used for test time decoding. Soft labels used for updating weights in the student are created from the teacher network after the privileged view branch is removed out. Although this paper focuses on acoustic models that are robust to lossy spectra, the proposed method can be applied to any other different combination of acoustic factors, for example clean and distorted speech with additive noise. Each training stage in our framework is next described in detail below.

3.1. Teacher Network Training

As illustrated in Figure 1, the teacher network has two branches similar to a Siamese network: one branch receives lossless inputs as a privileged information branch while the other receives lossy (or lossless) inputs as a student network view branch. The teacher network processes a pair of lossy features $\boldsymbol{x} \in \mathcal{X}$ and lossless features $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}$ which are time synchronous. In this network, several layers at the input are considered as feature processing layers. Weights of these layers are hence tied between the two feature input heads. The remaining classifier layers in the teacher network are shared. To update weights of this teacher network, the training loss is computed in parts from two separate passes of the data. The cross entropy loss $\mathcal{T}_{prv}(\theta)$ corresponding to the privileged knowledge branch is first computed between predicted output for lossless data $p_{prv}(i|\hat{\boldsymbol{x}})$ and grand truth labels t_i by passing lossless data as

$$\mathcal{T}_{prv}(\theta) = -\sum_{i} t_i \log p_{prv}(i|\hat{\boldsymbol{x}}). \tag{3}$$

In the next training step, lossy data is passed through the network and a second loss $\mathcal{T}_{st}(\theta)$ corresponding to the student view branch is computed between the predicted output for lossy data $p_{st}(i|\boldsymbol{x})$ and the predicted output for lossless data $p_{prv}(i|\hat{\boldsymbol{x}})$, estimated in the first pass. $\mathcal{T}_{st}(\theta)$ is expressed as:

$$\mathcal{T}_{st}(\theta) = -\sum_{i} p_{prv}(i|\hat{\boldsymbol{x}}) \log p_{st}(i|\boldsymbol{x}). \tag{4}$$

The teacher network's parameters are updated to minimize the combined cross entropy loss, given as:

$$\mathcal{T}(\theta) = (1-\lambda)\mathcal{T}_{prv}(\theta) + \lambda\mathcal{T}_{st}(\theta), \tag{5}$$

where λ is the loss weight. For our experiments, we fix λ at 0.5 which showed the best performance. The proposed two part loss used to train the teacher network can be considered as a *pseudo* knowledge distillation step on the teacher side. By being given access to two input features, we hypothesize that the network is able to learn differences between the features. The proposed loss function also encourages the network to incorporate privileged information unique to the lossless data into the learning process. In our experiments, we train teacher networks on a mix of both lossy and lossless features. While training

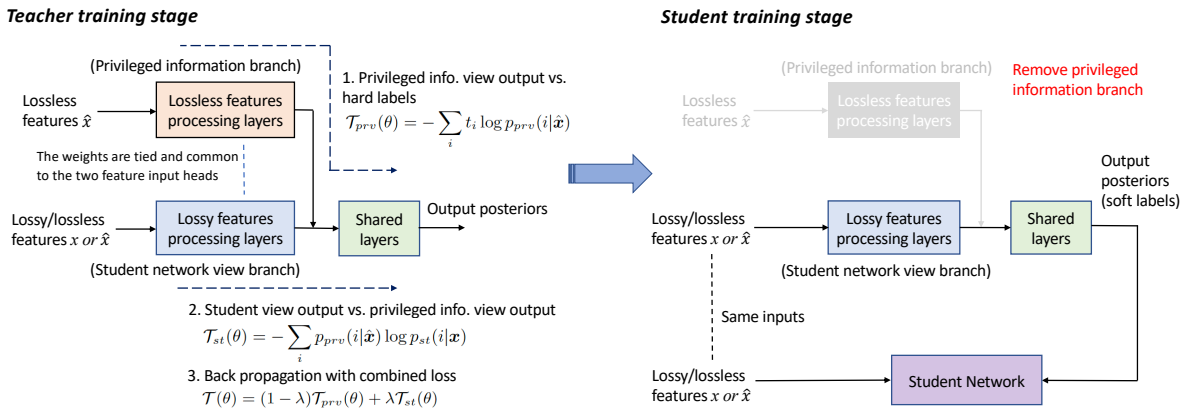

Figure 1: *Schematic of the proposed multi-view style student-teacher training framework (best viewed in color).*

models on this data set, the student view branch receives either lossless or lossy features. In both these cases, the input at the privileged information branch are always lossless features. When the student branch receives lossless features, the teacher model is updated equal to a single cross entropy loss.

3.2. Student Network Training

Similar to the teacher network, the student networks are also trained on a mix of lossy and lossless features. Soft targets corresponding to these features are generated using trained teacher networks described above. The teacher network can however be used in two modes to generate the soft targets. In the first mode (Teacher mode-1) both the privileged information branch and the student view branch are used to generate soft targets, similar to how the teacher network is trained. In a second mode (Teacher mode-2), the privileged view branch in the teacher is discarded and instead the student view branch is only used. As described earlier, for both these modes, the privileged view branch receives only lossless features while the student view can process either lossless or lossy features. Once soft targets have been extracted, the student network parameters are updated using the criterion outlined in Equation (1).

Student training with (Teacher mode-2) is a novel addition to the general knowledge distillation training framework. When the student network is updated with Teacher mode-2, corresponding soft labels $q(i|\boldsymbol{x})$ are generated from the proposed teacher network only using the student view branch with the same feature pair as the student network. Although these soft labels are created without using privileged features corresponding to lossy inputs, we hypothesize that privileged knowledge is implicitly transferred by the teacher network. This implicit information is obtained by the teacher network from lossless features during its training procedure.

4. Experiments and Results

4.1. Data

The efficacy of our proposed knowledge distillation framework is measured on a series of experiments using the Switchboard English conversational telephone corpus and an in-house telephony speech collection, both sampled at 8kHz. A training set is first constructed by randomly selecting 25-hours of telephony data from the Switchboard corpus and an in-house data collection. This data is further artificially corrupted with different ambient noises: more than 100 types of noises such as babble and office room noises at 5-25 dB SNR range are used to create a modified 100-hour training set that models a realistic diversity of acoustic conditions. Because the 100-hour training set has no bandwidth loss in the signal, it is used as a lossless training set although the data set has only been corrupted with additive noises. A lossy training set is further created by corrupting the lossless training set with bandwidth distortions. These distortions are introduced in 1-8 contiguous frequency bins at different frequency positions by zeroing out information in those frequency bins. The lossless and lossy training data sets are then used to train various models using the proposed framework. The artificial lossy spectra distortion we currently use is similar to frequency masking in SpecAugment [28]. The time masking in SpecAugment is another possible distortion that can be used.

We use the standard Hub5-2000 Switchboard (SWB) and CallHome (CH) test set as lossless test sets to measure the efficacy of our proposed method. Similar to the lossy training data creation, bandwidth distortions at 1-8 contiguous frequency bins are also added at various frequency positions to the lossless SWB and CH test data to create lossy versions of the test sets (SWB-LS and CH-LS).

4.2. Teacher and Student Network Topology

In our experiments, neural network based acoustic models are trained on the data set described above. CNN based acoustic models are constructed as student networks on a mixed set of the lossless and lossy training data sets 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 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.

With the same training data set, residual networks (ResNet)

Table 1: *Performances (WER%) of baseline CNNs, ResNet teachers, and student CNNs on lossless and lossy test sets*

Models	SWB	CH	SWB-LS	CH-LS	AVG
#1 Baseline CNN (no T/S, hard labels only)	14.3	24.9	18.9	30.1	22.1
#2 Baseline ResNet Teacher	12.7	21.9	13.5	23.5	17.9
#3 ResNet–Student CNN with privileged information	14.1	24.0	17.6	28.6	21.0
#4 ResNet–Student CNN without privileged information	13.9	24.3	14.9	25.8	19.7
#5 Proposed Multiview ResNet Teacher	12.5	21.5	13.3	23.2	17.6
#6 Proposed Multiview ResNet–Student CNN (Teacher mode-1)	13.8	23.7	17.2	28.1	20.7
#7 Proposed Multiview ResNet–Student CNN (Teacher mode-2)	13.3	23.4	14.3	25.1	19.0

are trained as the teacher networks. These networks have 12 convolutional layers, with a shortcut connection inserted every 3 convolutional layers to compose residual blocks, followed by 4 fully connected layers. The convolutional layers in each residual block have 64, 128, 256, and 512 nodes with 3×3 filters from the bottom of the network. Batch normalization is also applied to every layer. We trained two ResNet teachers for comparison: one teacher is trained with a single mixed set of the lossless and lossy training sets while the other is trained on a pair of lossless and lossy sets as we discussed in Section 3.

4.3. Experimental Results

We evaluate the effectiveness of our proposed method on both the lossless and lossy SWB and CH test sets. Given that student networks are trained on both lossy and lossless data, it is challenging for these networks to perform consistently well on both test sets, compared to domain specific models trained on just one kind of data, because of the data balance. We hypothesize that consistent improvements on both test sets are hence from knowledge incorporated by the student networks through knowledge distillation. Experimental results with various student and teacher models are shown in Table 1.

The baseline CNN (#1) corresponds to a CNN trained with hard labels. The topology of this network is the same as the other CNN student networks in our experiments, except that it is not trained using the knowledge distillation framework. The baseline ResNet teacher (#2) is a teacher network trained also on hard labels only. These models, trained on the mixed training set of lossless and lossy data, will be used to compare between standard knowledge distillation training and our proposed algorithm. As seen Table 1, the ResNet teacher shows significantly better performance than the baseline CNN although the decoding speed is very slow.

We train two student networks using this teacher model. The difference between the two student models is based on how the soft targets from the teacher network are created. Since the teacher network is trained on a mixed training set of lossless and lossy data, it can be used to produce soft targets using either lossless or lossy data. In our first experiment, student network (#3) is trained using soft labels created with only lossless inputs to the teacher. In contrast, for the student network in (#4), soft labels to train the student network are generated by the teacher network (#2) using corresponding lossless or lossy input features. Both networks in (#3) and (#4) are trained with the standard knowledge distillation technique as expressed in Equations (1) and (2) and show improved performance over the baseline CNN system. In most cases student (#4) performs better than (#3), suggesting that while extracting soft labels for training the student network, generating the labels with corresponding matched lossy or lossless inputs is more advantageous in this experimental setting. However by choosing this option, potential benefits of using lossless inputs for soft label generation are missed. It is possible that there is a significant acoustic mismatch between the lossy and lossless data and this prevents the student models from learning from just the lossless data.

The proposed knowledge distillation framework attempts to circumvent this issue with a multi-view style training with both lossless and lossy data. As discussed earlier, with the proposed two part loss, a *pseudo* knowledge distillation step is performed during the teacher training. This encourages the network to incorporate privileged information unique to the lossless data into the learning process. During the soft label generation stage, lossy data can then be used to effectively generate soft targets. A multi-view style teacher network trained using the proposed framework (#5) has results similar to the ResNet teacher (#2) on average, but contains special acoustic knowledge obtained from the differences between lossy and lossless data. We now use this teacher network to generate soft targets for student network training. We first generate soft targets using a pair of lossless and lossy features with (Teacher mode-1) of the teacher network as described in Section 3 (#6). In this setting both branches of the teacher network are used to generate soft targets for the student network. The student network is better than (#3) but lags behind (#4). This result shows the benefit of the proposed network while still highlighting that the gains might be lower because of the use of mismatched lossless data for soft target generation against the actual student inputs.

In the next experiment we use only the single features in (Teacher mode-2) of the teacher network to generate soft labels as described in earlier in Section 3. In this setting, only the student view branch of the teacher network is used. The teacher network does not explicitly use any privileged lossless data to generate soft targets for lossy data. The proposed student network (#7) improves across all test sets and provides 9.5% and 3.6% relative improvements over the baseline student systems (#3) and (#4). Unlike earlier experiments the soft targets to train the student network are estimated just from the same feature inputs. The model is however able to perform well on both lossy and lossless test sets. These results indicate that the the teacher network can successfully integrate privileged knowledge derived from the corresponding lossless features to the student network using the proposed training process.

5. Conclusions

In this paper we have proposed a novel knowledge distillation training strategy to construct neural network based acoustic model that are robust to signals with time frequency bandwidth loss. Our proposed method transfers privileged acoustic knowledge to student networks using a multi-view style knowledge distillation framework. Experiments on signals with lossy spectra show that our proposed technique can effectively leverage information from lossless data available only training. The technique provides a significant relative improvement of 9.5% over conventional generalized knowledge distillation techniques.

6. References

[1] R. Hsiao, J. Ma, W. Hartmann, M. Karafiát, F. Grezl, L. Burget, I. Szöke, J. Černocký, S. Watanabe, Z. Chen, S. H. Mallidi, H. Hermansky, S. Tsakalidis, and R. Schwartz, "Robust speech recognition in unknown reverberant and noisy conditions," *Proc. IEEE ASRU*, 2015.

[2] Z. Zhang, J. Geiger, J. Pohjalainen, A. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," *arXiv:1705.10874*, 2017.

[3] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 2263–2276, 2016.

[4] A. Ragni, K. M. Knill, S. P. Rath, and M. J. F. Gales, "Data augmentation for low resource languages," *Proc. Interspeech*, 2014.

[5] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," *Proc. IEEE ICASSP*, 2017.

[6] S. Thomas, G. Saon, M. V. Segbroeck, and S. S. Narayanan, "Improvements to the IBM speech activity detection system for the DARPA RATS program," *Proc. IEEE ICASSP*, 2015.

[7] G. Saon, S. Thomas, H. Soltau, S. Ganapathy, and B. Kingsbury, "The IBM speech activity detection system for the DARPA RATS program," *Proc. Interspeech*, 2013.

[8] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *arXiv:1503.02531v1*, 2015.

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

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

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

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

[13] 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.

[14] 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," *Proc. IEEE ICASSP*, pp. 6475–6479, 2019.

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

[16] V. Vapnik and R. Izmailov, "Learning using privileged information: Similarity control and knowledge transfer," *Machine Learning Research*, vol. 16, pp. 2023–2049, 2015.

[17] K. Markov and T. Matsui, "Robust speech recognition using generalized distillation framework," *Proc. Interspeech*, 2016.

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

[19] T. Fukuda and S. Thomas, "Mixed bandwidth acoustic modeling leveraging knowledge distillation," *Proc. IEEE ASRU*, 2019.

[20] G. Kock, R.Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," *ICML Deep learning workshop*, 2015.

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

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

[23] J. Cui, B. Kingsbury, B. Ramabhadran, G. Saon, T. Sercu, K. Audhkhasi, A. Sethy, M. Nussbaum-Thom, and A. Rosenberg, "Knowledge distillation across ensembles of multilingual models for low-resource languages," *Proc. IEEE ICASSP*, pp. 4825–4829, 2017.

[24] J. H. M. Wong and M. J. F. Gales, "Sequence student-teacher training of deep neural networks," *Proc. Interspeech*, pp. 2761–2765, 2016.

[25] G. Kurata and K. Audhkhasi, "Guiding CTC posterior spike timings for improved posterior fusion and knowledge distillation," *Proc. Interspeech*, 2019.

[26] R. Takashima, S. Li, and H. Kawai, "Investigation of sequence-level knowledge distillation methods for CTC acoustic models," *Proc. IEEE ICASSP*, 2019.

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

[28] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019.