STUDENT-TEACHER NETWORK LEARNING WITH ENHANCED FEATURES

Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R. Hershey

Mitsubishi Electric Research Laboratories (MERL), Cambridge MA, USA

ABSTRACT

Recent advances in distant-talking ASR research have confirmed that speech enhancement is an essential technique for improving the ASR performance, especially in the multichannel scenario. However, speech enhancement inevitably distorts speech signals, which can cause significant degradation when enhanced signals are used as training data. Thus, distant-talking ASR systems often resort to using the original noisy signals as training data and the enhanced signals only at test time, and give up on taking advantage of enhancement techniques in the training stage. This paper proposes to make use of enhanced features in the student-teacher learning paradigm. The enhanced features are used as input to a teacher network to obtain soft targets, while a student network tries to mimic the teacher network's outputs using the original noisy features as input, so that speech enhancement is implicitly performed within the student network. Compared with conventional student-teacher learning, which uses a better network as teacher, the proposed self-supervised method uses better (enhanced) inputs to a teacher. This setup matches the above scenario of making use of enhanced features in network training. Experiments with the CHiME-4 challenge real dataset show significant ASR improvements with an error reduction rate of 12% in the single-channel track and 15% in the 2-channel track, respectively, by using 6-channel beamformed features for the teacher model.

Index Terms— Distant-talking ASR, speech enhancement, student teacher learning, self-supervised learning, CHiME-4

1. INTRODUCTION

Speech enhancement and ASR techniques have been developing rapidly with the growing demand for distant-talking speech interfaces and recent challenge activities [1, 2, 3, 4]. In particular, when using multichannel speech enhancement techniques such as beamforming [5], significant improvements can be obtained compared to a single-channel ASR systems, and the state-of-the-art multichannel systems are approaching the performance of ASR systems in clean conditions [6, 7]. The enhancement component can greatly suppress the noise components, and is responsible for the largest share of the improvements in the above systems.

However, these studies also reveal an interesting phenomenon: the use of enhanced signals as training data tends to degrade the ASR performance despite the fact that this could mitigate a mismatch between training data (noisy speech) and test data (enhanced speech). This likely comes from the sporadic distortions that speech enhancement inevitably adds, which may introduce additional within-class variability that is difficult to model robustly without overfitting. Therefore, finding a way to take advantage of speech enhancement in an acoustic model's training phase is an interesting and potentially fruitful research direction.

In this paper, instead of simply using enhanced speech as training data in the network training, we take advantage of the more reliable accuracy displayed by acoustic models when enhanced speech is used as test data. That is, we use a student-teacher learning paradigm [8, 9] where we train the *student* network using noisy training data as input and the soft targets obtained from the *teacher* state posteriors computed on the enhanced speech data, instead of the conventional hard targets, as shown in Fig. 1. Conventional student-teacher learning is designed to train a so-called student model typically consisting of a small-complexity network, in order to obtain performance close to that of a so-called teacher model consisting of a large-complexity network (or sometimes even an ensemble thereof). The small student model tries to mimic the teacher performance by using the soft targets obtained with the teacher model, which corresponds to obtaining a compressed model with similar performance to the large teacher model [10]. Relatedly, [11] also points out the importance of using soft targets rather than hard targets for deep learning, referring to such soft assignment information as "dark knowledge." Compared with these conventional student-teacher learning methods, the proposed method does not aim to compress the model, but aims for the student model to mimic the teacher model as if the student network also performed speech enhancement. Therefore, unlike the conventional methods, the student and teacher models use different data as input, namely the noisy data and the corresponding enhanced data, respectively. The proposed method becomes especially attractive when we use multichannel speech enhancement in the teacher model. Indeed, the student model then tries to mimic the effects of multichannel speech enhancement even though it only uses single channel inputs.

This method recalls the traditional noise robustness technique of *single-pass retraining* [12], in which the state alignment obtained from a close-talking microphone signal is used to train an acoustic model for a distant-talking microphone in a stereo recording. In [13], simulation was used to obtain the clean and noisy stereo data in a student-teacher retraining framework that used soft posteriors instead of hard alignments. Like [13], the proposed method uses soft posteriors, however it differs from both [12] and [13] in that it does not require the availability of noisy/clean parallel data, and instead uses multichannel enhancement instead of clean input for the teacher model, which can be easily obtained. It can thus apply to realistic situations where we cannot obtain the reference clean speech features.

The use of a teacher model based on more accurate sensor information is also known as *self-supervised learning* [14, 15]. For example, in [15], a multi-camera teacher model was used to infer soft labels of depth-based categories, in order to supervise a single-camera student model. This allowed learning of distant depth estimation without parallel data, despite gaps in the teacher's depth inference. Inspired by this work, in our proposed method, we expect that the soft posteriors from the multichannel teacher's acoustic model will be relatively insensitive to the distortion anomalies in the enhanced data, and provide reliable supervision for the single-channel student model.

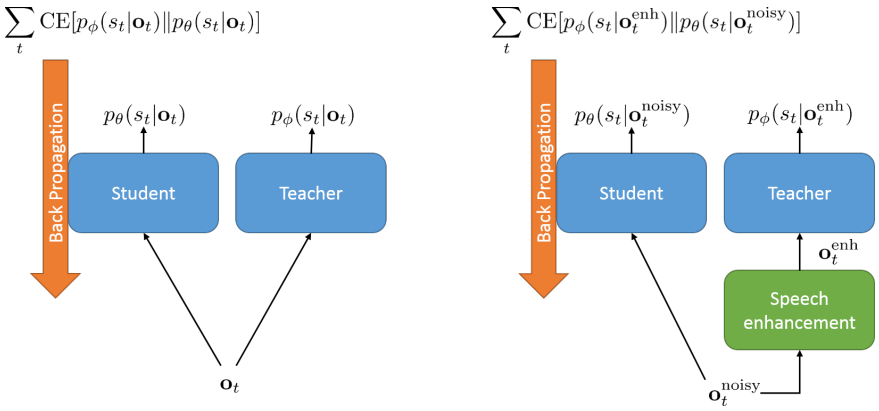

Fig. 1. Conventional student-teacher learning (left side) and the proposed learning with enhanced signals (right side). The proposed student-teacher learning uses enhanced feature $\mathbf{o}_t^{\mathrm{noisy}}$ to provide better soft targets $p_\phi(s_t|\mathbf{o}_t^{\mathrm{enh}})$.

This paper experimentally demonstrates the effectiveness of the method by using the CHiME-4 speech separation and recognition challenge dataset [16], in which 1-, 2-, and 6-channel tracks are prepared by setting the subset of microphones that can be used in the test stage, while all 6 channel data can be used in the training stage. We use here a delay-and-sum beamformed signal (BeamformIt [17]) from 5-channel signals as enhanced signal, and perform the proposed student-teacher learning. The obtained acoustic model improves the performance of all 1-, 2-, and 6-channel tracks, and especially obtains a large improvement on the 1-channel track.

2. FORMULATION

2.1. Conventional student-teacher learning

Let $O = \{\mathbf{o}_t \in \mathbb{R}^D | t = 1, \ldots, T\}$ be a set of D dimensional input feature vectors and $S^{\mathrm{ref}} = \{s_t^{\mathrm{ref}} \in \{1, \ldots, K\} | t = 1, \ldots, T\}$ be the corresponding target labels, where T is the number of samples and K is the number of distinct categories. The standard cross entropy criterion is defined using the reference label distribution $p_{\mathrm{ref}}(s_t)$ and an arbitrary posterior distribution $p_\theta(s_t|\mathbf{o}_t)$ with model parameter θ as follows:

$$\begin{aligned}\mathrm{CE}(\theta; O, S^{\mathrm{ref}}) &\triangleq \sum_t \mathrm{CE}[p_{\mathrm{ref}}(s_t)\|p_\theta(s_t|\mathbf{o}_t)] \\ &\triangleq \sum_t \sum_{s_t} -\delta(s_t, s_t^{\mathrm{ref}}) \log p_\theta(s_t|\mathbf{o}_t) \\ &= -\sum_t \log p_\theta(s_t^{\mathrm{ref}}|\mathbf{o}_t). \end{aligned} \quad (1)$$

Here the reference label distribution is represented by a Kronecker delta function, i.e., $p_{\mathrm{ref}}(s_t) = \delta(s_t, s_t^{\mathrm{ref}})$, which corresponds to the hard assignment to the labels obtained by the Viterbi algorithm in the acoustic model training case.

Instead of using the hard assignment labels, student-teacher learning uses the label posterior distribution $p_\phi(s_t|\mathbf{o}_t)$ obtained by using a well-trained teacher network with parameter ϕ as follows:

$$\begin{aligned}\mathrm{CE}(\theta; O, \phi) &\triangleq \sum_t \mathrm{CE}[p_\phi(s_t|\mathbf{o}_t)\|p_\theta(s_t|\mathbf{o}_t)] \\ &\triangleq -\sum_t \sum_{s_t} p_\phi(s_t|\mathbf{o}_t) \log p_\theta(s_t|\mathbf{o}_t). \end{aligned} \quad (2)$$

Contrary to the previous Kronecker delta case, which is a very sparse one-hot representation, there are many non-zero values in this dense representation, which may lead to more useful supervision. Hinton et al. [11] refer to this information as dark knowledge, and show that this training criterion can be used to efficiently obtain a compressed student network with comparable performance to the teacher model. Note that Eq. (2) does not depend on the supervision S^{ref}, although it is used when training the teacher model $p_\phi(s_t|\mathbf{o}_t)$.

2.2. Proposed student-teacher learning with enhanced signals

Although the main target of the conventional student-teacher learning is to use different network architectures between teacher and student models, our proposed method is to use different input features: a teacher model uses enhanced features $O^{\mathrm{enh}} = \{\mathbf{o}_t^{\mathrm{enh}} \in \mathbb{R}^D | t = 1, \ldots, T\}$ for predicting better posteriors, while a student model uses original noisy speech features $O^{\mathrm{noisy}} = \{\mathbf{o}_t^{\mathrm{noisy}} \in \mathbb{R}^D | t = 1, \ldots, T\}$.

The proposed teacher-student learning considers the following objective function:

$$\begin{aligned}&\mathrm{CE}(\theta; O^{\mathrm{enh}}, O^{\mathrm{noisy}}, \phi) \\ &\quad\triangleq \sum_t \mathrm{CE}[p_\phi(s_t|\mathbf{o}_t^{\mathrm{enh}})\|p_\theta(s_t|\mathbf{o}_t^{\mathrm{noisy}})] \\ &\quad\triangleq -\sum_t \sum_{s_t} p_\phi(s_t|\mathbf{o}_t^{\mathrm{enh}}) \log p_\theta(s_t|\mathbf{o}_t^{\mathrm{noisy}}) \end{aligned} \quad (3)$$

The proposed objective function is similar to Eq. (2) except that we use noisy/enhanced parallel data. We can alternately use speech features obtained by clean or close-talking microphone signals in lieu of O^{enh}, if available.

2.3. Multichannel extension

With multichannel signals, we can use powerful beamforming methods that exploit spatial information to enhance the speech signals. Our student-teacher learning with enhanced signals can be extended to this particular setup, considering each channel separately as a single-channel noisy signal used as input to the student network. When we have J-channel speech features $\mathcal{O}^{\mathrm{noisy}} = \{\mathbf{o}_{j,t}^{\mathrm{noisy}} \in \mathbb{R}^D | t = 1, \ldots, T, j = 1, \ldots, J\}$ and the corresponding enhanced

features O^{enh}, Eq. (3) is extended as:

$$\begin{aligned}
&\text{CE}(\theta; O^{\text{enh}}, \mathcal{O}^{\text{noisy}}, \phi) \\
&\quad \triangleq \sum_t \sum_j \text{CE}[p_\phi(s_t|\mathbf{o}_t^{\text{enh}}) \| p_\theta(s_t|\mathbf{o}_{j,t}^{\text{noisy}})] \\
&\quad \triangleq -\sum_t \sum_j \sum_{s_t} p_\phi(s_t|\mathbf{o}_t^{\text{enh}}) \log p_\theta(s_t|\mathbf{o}_{j,t}^{\text{noisy}}). \qquad (4)
\end{aligned}$$

Compared with the single channel enhancement case in Eq. (3), Eq. (4) has the same posterior target, but the student network now needs to learn that target with any of the channels as input.

In our experiments, we have combined the standard cross entropy in Eq. (2) and our proposed objective function in Eq. (4) with a weight factor γ, as follows:

$$(1-\gamma)\text{CE}(\theta; \mathcal{O}^{\text{noisy}}, S) + \gamma \text{CE}(\theta; O^{\text{enh}}, \mathcal{O}^{\text{noisy}}, \phi). \qquad (5)$$

With γ approaching 0, the objective function becomes the standard cross entropy, while with γ approaching 1, the training criterion only considers the proposed student-teacher learning objective function.

3. EXPERIMENTS

This section shows the efficacy of the proposed method by using the 1ch, 2ch, and 6ch tracks of the CHiME-4 speech separation and recognition challenge [16].

3.1. Experimental setup

CHiME-4 revisits the datasets originally recorded for CHiME-3 [3], i.e., Wall Street Journal (WSJ) corpus sentences spoken by talkers situated in challenging noisy environments recorded using a 6-channel tablet-based microphone array. The CHiME-4 data consist of real and simulation data for each of the training, development, and evaluation sets. There are three kinds of test data depending on the number of microphones used (1, 2, and 6), which form the 1ch, 2ch, and 6ch tracks in the challenge. Training data does not have such limitation, and we can use all 6 channel data to obtain acoustic models.

We use the same experimental conditions as the official CHiME-4 baseline [16] using Kaldi [18] with a 3-gram language model, except that the DNN acoustic model is here trained by using the noisy speech data for all 6 channel signals (i.e., 6 times more data than the official baseline). We use fMLLR features where the transformation matrices are obtained with the GMM acoustic model built on the same 6 channel signals. The DNN has 7 layers and each layer has 2048 neurons with sigmoid activation functions between the linear transformation layers. We also use the same hard targets for the standard CE training of both conventional and proposed methods. These were obtained by using the Viterbi alignments of the noisy speech data with the above GMM.

For the DNN training including the standard cross entropy and proposed student-teacher learning, we used the Chainer deep network toolkit [19]. DNNs were optimized by using the stochastic gradient descent algorithm with a mini-batch size of 512, with the learning rate initialized to 1.0 and halved when we observed a degradation of the validation score. As multichannel speech enhancement algorithm, we used a delay-and-sum beamformer based on BeamformIt [17]. At training time, the proposed student-teacher learning used the enhanced training data obtained using 5 microphone array signals (the 2nd channel was excluded following the baseline [16]), and the obtained model was used for all 1ch, 2ch, and 6ch track evaluation. At test time, in the 2ch and 6ch conditions, we apply the delay-and-sum beamformer to the noisy signal, and use the enhanced signal as input to the network. In the 1ch condition, the noisy signal is used as is. In the experiments, we used the same network architecture (i.e., 7-layer DNN) for both student and teacher models so that we simply evaluate the effectiveness of the proposed learning with enhanced features.

Table 1. WERs of the training data (closed condition) by using Noisy, Clean*, and Enhanced features. The clean data are obtained by using the original WSJ data for the simulation part and the close-talking microphone data for the real part.

	WER (%)
Noisy	29.54
Clean*	28.87
Enhanced	**28.05**

3.2. Quality of soft targets

Before moving to the student-teacher learning experiments, we analyzed the qualities of the soft targets obtained by using the noisy, clean (original WSJ data for the simulation part and close-talking microphone data for the real part), and enhanced features. Table 1 shows the WERs of the training sets with these three features. The multi-condition acoustic model was obtained by using the standard CE training with noisy speech features. This result shows that the soft targets obtained with enhanced features seem to act as better ground truths compared with those with noisy features. We also found that the result with clean features is not better than that with enhanced features, probably due to the large mismatch between the clean speech data and the multi-condition acoustic model trained with noisy speech features. Also, features obtained from close-talking microphones have their own specific distortions due to lip noises and channel distortions. Overall, it is very difficult to prepare ideal parallel data of clean and noisy speech features unless we fully use simulation data like [13], which is a less realistic scenario. This result motivated us to use enhanced features for soft targets rather than clean speech features.

3.3. Effect of soft targets

The second experiment investigates the performance of the soft targets without using enhanced features. We only report the results on the real development and evaluation sets, as these are more realistic tasks. We use the teacher state posterior obtained by using noisy speech features, and the student model is also trained by using the *same* noisy speech features. That is, we use the following objective function:

$$(1-\gamma)\text{CE}(\theta; \mathcal{O}^{\text{noisy}}, S) + \gamma \text{CE}(\theta; \mathcal{O}^{\text{noisy}}, \phi), \qquad (6)$$

where the second term $\text{CE}(\theta; \mathcal{O}^{\text{noisy}}, \phi)$ is obtained by substituting $\mathcal{O}^{\text{noisy}}$ into $\mathcal{O}$ in the conventional student-teacher learning equation (2). Figures 2 and 3 show the WERs of the 1, 2, and 6 channel tracks for the development and evaluation sets, respectively, depending on the weight factor γ in Eq. (6), where $\gamma = 0$ corresponds to the conventional CE. For all experiments, simply using the soft target improved the performance. It is expected that the Viterbi alignments for noisy speech data are prone to including misalignments, which is mitigated by the soft target training.

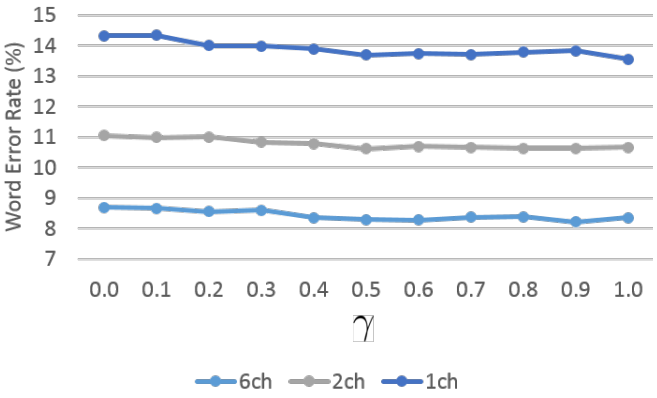

Fig. 2. Student-teacher learning with *noisy* speech for CHiME-4 *development* sets (1, 2, and 6 channel tracks).

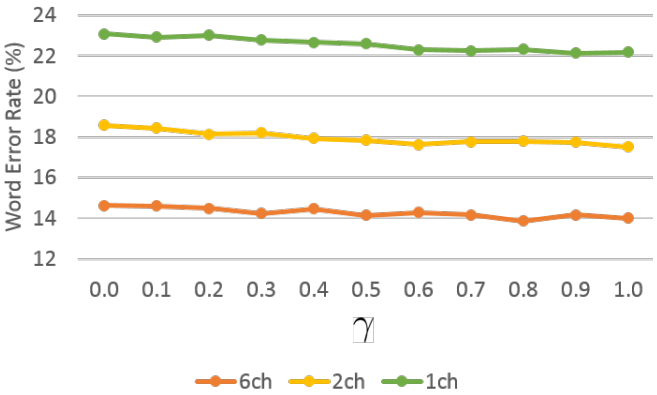

Fig. 3. Student-teacher learning with *noisy* speech for CHiME-4 *evaluation* set (1, 2, and 6 channel tracks).

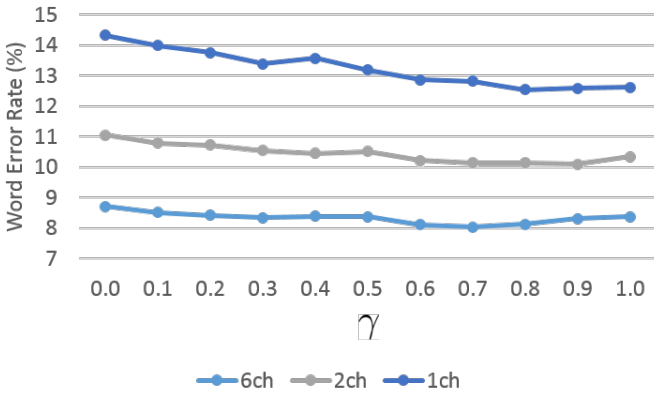

Fig. 4. Proposed student-teacher learning with *enhanced* speech for CHiME-4 *development* set (1, 2, and 6 channel tracks).

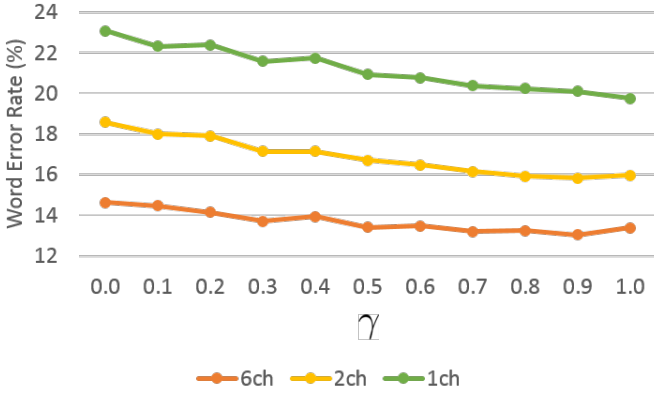

Fig. 5. Proposed student-teacher learning with *enhanced* speech for CHiME-4 *evaluation* set (1, 2, and 6 channel tracks).

Table 2. Summary of the WERs for standard CE with hard targets (Baseline), student-teacher learning with noisy speech features (Noisy), and proposed student-teacher learning with enhanced speech features (Proposed).

	6ch		2ch		1ch	
Method	Dev	Eval	Dev	Eval	Dev	Eval
Baseline	8.70	14.63	11.05	18.57	14.33	23.09
Noisy	8.21	14.17	10.64	17.75	13.55	22.19
Proposed	**8.03**	**13.19**	**10.09**	**15.83**	**12.54**	**20.25**

3.4. Student-teacher learning with enhanced features

The third experiment evaluates the proposed student-teacher learning with enhanced features. As shown in Fig. 4 for the development set and Fig. 5 for the evaluation set, the proposed method significantly improves the performance upon the standard CE training ($\gamma = 0$), and also upon the previous results simply using soft targets for training (Figs. 2 and 3). The improvements are especially large in the 1ch condition, where the signal is noisy, and the 2ch condition, where the enhanced signal did not benefit as much from beamforming as in the 6ch case. The student network is able to deal better with the remaining noise in the input than the networks that are trained without taking advantage of the soft targets obtained from the enhanced features. We also found that the weight factor γ was not so sensitive when we used a value around 1.0.

Table 2 summarizes the WERs for the standard CE with hard targets (Baseline), student-teacher learning with noisy speech features (Noisy) from Figures 2 and 3, and proposed student-teacher learning with enhanced speech features (Proposed) from Figures 4 and 5. The proposed method improved the performance in all cases, and achieved around 10% error reduction rates for all cases, which shows its effectiveness.

4. SUMMARY

This paper proposes a new student-teacher learning scheme for noise robust ASR, where the teacher model uses enhanced features while the student model uses noisy features during training. This encourages the student model to try to perform speech enhancement within the network. The experiments on the CHiME-4 speech separation and recognition challenge show significant improvement when using the proposed method, especially in the single-channel track.

Future work will consider extending the proposed method to handle sequence discriminative training and state sequence posteriors, as well as using state-of-the-art beamforming with mask estimation [20, 21, 22].

5. REFERENCES

[1] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, 2013.

[2] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, A. Sehr, W. Kellermann, S. Gannot, R. Maas, R. Haeb-Umbach, V. Leutnant, and B. Raj, "The REVERB challenge: a common evaluation framework for dereverberation and recognition of reverberant speech," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2013.

[3] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015.

[4] M. Harper, "The automatic speech recognition in reverberant environments (ASpIRE) challenge," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015.

[5] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[6] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2015.

[7] T. Hori, Z. Chen, H. Erdogan, J. R. Hershey, J. Le Roux, V. Mitra, and S. Watanabe, "The MERL/SRI system for the 3rd CHiME challenge using beamforming, robust feature extraction, and advanced speech recognition," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2015.

[8] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size DNN with output-distribution-based criteria." in *Proceedings of Interspeech*, 2014.

[9] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Proceedings of Neural Information Processing Systems*, 2014.

[10] C. Bucilu, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006.

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

[12] P. C. Woodland, M. J. F. Gales, and D. Pye, "Improving environmental robustness in large vocabulary speech recognition," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1. IEEE, 1996.

[13] K. Markov and T. Matsui, "Robust speech recognition using generalized distillation framework," in *Proceedings of Interspeech*, 2016.

[14] D. Lieb, A. Lookingbill, and S. Thrun, "Adaptive road following using self-supervised learning and reverse optical flow." in *Robotics: Science and Systems*, 2005.

[15] R. Hadsell, P. Sermanet, J. Ben, A. Erkan, M. Scoffier, K. Kavukcuoglu, U. Muller, and Y. LeCun, "Learning long-range vision for autonomous off-road driving," *Journal of Field Robotics*, vol. 26, no. 2, 2009.

[16] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, 2016, (to appear).

[17] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, 2007.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[19] S. Tokui, K. Oono, S. Hido, and J. Clayton, "Chainer: a next-generation open source framework for deep learning," in *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*, 2015.

[20] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, 2016.

[21] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, 2016.

[22] H. Erdogan, J. R. Hershey, S. Watanabe, M. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," *Proceedings of Interspeech*, 2016.