Knowledge Distillation for Throat Microphone Speech Recognition

Conference Paper · September 2019

DOI: 10.21437/Interspeech.2019-1597

5 authors, including:

Takahito Suzuki
Shizuoka University
3 PUBLICATIONS **1** CITATION
SEE PROFILE

Takashi Tsunakawa
Shizuoka University
13 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Masafumi Nishida
Shizuoka University
77 PUBLICATIONS **469** CITATIONS
SEE PROFILE

INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Knowledge Distillation for Throat Microphone Speech Recognition

Takahito Suzuki[1], *Jun Ogata*[2], *Takashi Tsunakawa*[1], *Masafumi Nishida*[1], *Masafumi Nishimura*[1]

[1] Graduate School of Integrated Science and Technology, Shizuoka University, Shizuoka, Japan
[2] National Institute of Advanced Industrial Science and Technology, Japan

suzuki.takahito.14@shizuoka.ac.jp

Abstract

Throat microphones are robust against external noise because they receive vibrations directly from the skin, however, their available speech data is limited. This work aims to improve the speech recognition accuracy of throat microphones, and we propose a knowledge distillation method of hybrid DNN-HMM acoustic model. This method distills the knowledge from acoustic model trained with a large amount of close-talk microphone speech data (teacher model) to acoustic model for throat microphones (student model) using a small amount of parallel data of throat and close-talk microphones. The front-end network of the student model contains a feature mapping network from throat microphone acoustic features to close-talk microphone bottleneck features, and the back-end network is a phonetic discrimination network from close-talk microphone bottleneck features. We attempted to improve recognition accuracy further by initializing student model parameters using pretrained front-end and back-end networks. Experimental results using Japanese read speech data showed that the proposed approach achieved 9.8% relative improvement of character error rate (14.3% → 12.9%) compared to the hybrid acoustic model trained only with throat microphone speech data. Furthermore, under noise environments of approximately 70 dBA or higher, the throat microphone system with our approach outperformed the close-talk microphone system.

Index Terms: speech recognition, throat microphone, deep learning, knowledge distillation

1. Introduction

Deep learning has been recently applied to speech recognition, and reported results indicate that the recognition accuracy was greatly improved and reached human parity in easy tasks [1, 2]. However, the recognition accuracy degrades under reverberant and noise environments, which has been subject of additional studies, as reported in [3]. Throat microphones, which are placed around the neck and receive vibrations directly from the skin, are more robust against external noise than close-talk microphones. Accordingly, researchers have conducted studies using them for speech recognition [4-10], voice activity detection [11], and speaker recognition under noise environments [12, 13].

Throat microphone speech signals are narrowband and less clear than those of close-talk microphones. Due to large acoustic mismatches between throat and close-talk microphones, low recognition accuracy results if throat microphone speech signals are simply input to a general speech recognition system. Furthermore, it is difficult to train a large-scale acoustic model from throat microphone speech data due to limitations in their available speech data.

We previously proposed the approach using general Gaussian mixture model-hidden Markov model (GMM-HMM) trained with a large amount of close-talk microphone speech data and conducting feature mapping from throat microphones to close-talk microphones as a pre-processing step to suppress acoustic mismatches [9]. However, it is difficult to train mapping perfectly owing to limitations in available parallel data for training. In a hybrid deep neural network (DNN)-HMM system (hereinafter, referred to as hybrid model), input features could be mapped to space for phonetic discrimination more effectively by multistage non-linear transformation of DNN, and it is expected that the hybrid model would perform better than the GMM-HMM system. However, it is extremely difficult to train the hybrid model with only a small amount of throat microphone speech data; therefore, we propose a training approach for a hybrid model of a throat microphone utilizing a large amount of close-talk microphone speech data and a small amount of parallel throat and close-talk microphone data. Specifically, we propose to use parallel data to distill knowledge from a DNN of the hybrid model for close-talk microphones (teacher model) to a DNN of the hybrid model for throat microphones (student model). The front-end network of the student model contains a feature mapping network from acoustic features of throat microphones to bottleneck features (BNF) of close-talk microphones, and the back-end network is a phonetic discrimination network from BNF of close-talk microphones. We attempted to improve recognition accuracy further by initializing student model parameters using pretrained front-end and back-end networks. We conducted recognition experiments using reading speech data from Japanese newspaper articles and compared the proposed system with conventional systems, evaluated the initialization and training approach of the student model, and compared the performance of close-talk and throat microphones under noise environments.

The rest of this paper is organized as follows. Section 2 reviews previous studies of throat microphone speech recognition and knowledge distillation, and Section 3 describes the proposed training approach for a hybrid model of throat microphones. Section 4 elucidates the conditions and results of recognition experiments, and Section 5 discusses conclusions and future research prospects.

2. Related Work

2.1. Throat microphone speech recognition

Previous studies of speech recognition using throat microphones can be classified into two main approaches: 1) those using a combination of close-talk and throat microphones for recognition [4-7], and 2) those that used only a throat microphone for recognition [8, 9].

http://dx.doi.org/10.21437/Interspeech.2019-1597

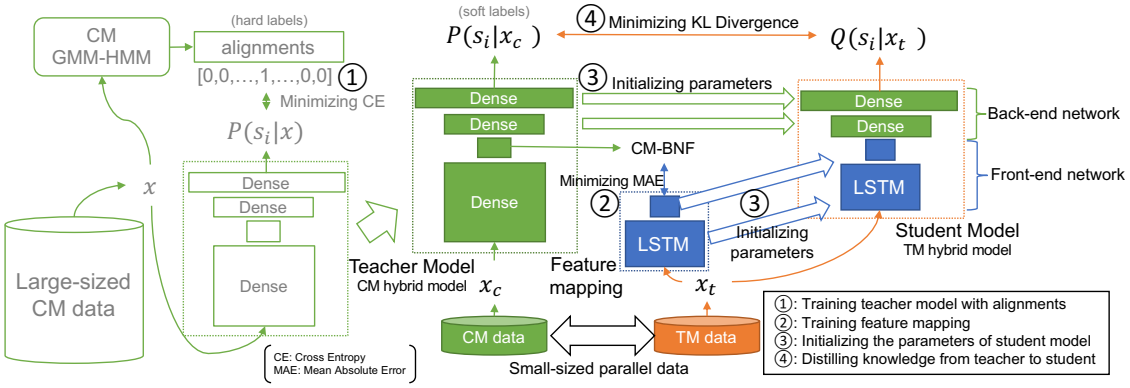

Figure 1: *Framework of knowledge distillation using parallel data of throat microphones (TM) and close-talk microphones (CM)*

The combination approaches used each acoustic score provided by HMM [4] or posterior probability provided by the hybrid artificial neural networks (ANN) [5] of two microphones by linear interpolation. However, these methods require the generation of reliable score or posterior probability from the acoustic feature of throat microphones. Methods of transformation from the noisy feature of two microphones to clean features of air-conducted microphones based on probable optimum filter (POF) [6] or DNN [7] have also been proposed; however, such approaches require dynamic weighting for two input features according to external noise levels.

For approaches using only a throat microphone for recognition, feature mapping techniques to suppress acoustic mismatches between throat and close-talk microphones have been proposed as a pre-processing step of a general speech recognition system [8, 9]. The feature mapping is trained with the parallel data, and several techniques have been applied as the feature mapper such as GMM [14], a feed forward neural network (FFNN) [9, 15], a combination of GMM and NN [16], and long short-term memory (LSTM) [8]. Although feature mapping can suppress acoustic mismatches, it is difficult to train mapping perfectly.

2.2. Knowledge distillation

As a method of model compression, an approach known as knowledge distillation has been proposed that entails training a small-scale model utilizing knowledge of a large-scale model with high prediction accuracy [17]. A large-scale model is often referred to as "teacher model" and a small-scale model is called "student model." In knowledge distillation, the student model is trained by minimizing the loss between the output distributions of the student and teacher models. The loss is calculated using L2 loss [18], softmax with temperature [17] or Kullback-Leibler (KL) divergence [18]. Hinton *et al.* reported that the model trained with hard labels was over-fitted when using small training data, whereas the model trained by knowledge distillation was not [17]. Hence, the latter is the preferred regularization method.

Some studies have applied knowledge distillation to speech recognition [19, 20]. Yi *et al.* proposed knowledge distillation for far-field speech recognition using parallel data from single distant microphones (SDM) and individual headset microphones (IHM), such that the teacher model is a DNN of the hybrid model for IHM trained with the alignments (hard labels) provided from the GMM-HMM [19]. The DNN of the hybrid model for SDM (student model) was trained by minimizing KL divergence between output distributions provided from student and teacher models using the parallel data.

3. Knowledge Distillation for Throat Microphone Speech Recognition

It is expected that the hybrid model would perform better than the GMM-HMM system; however, it is extremely difficult to train a hybrid model with only a small amount of throat microphone speech data. Inspired by Yi's work [19], we propose a training approach for a hybrid model of throat microphones by distilling knowledge from a DNN of the hybrid model trained from large-sized close-talk microphone data. In this work, the front-end network of student models performs mapping from the acoustic features of throat microphones to BNFs of close-talk microphone, whereas the back-end network produces posterior probabilities over tied-state triphone HMM states from the BNFs of close-talk microphones, and these networks are trained respectively in advance. Lin *et al.* reported that LSTM-based feature mapping had better performance than FFNN-based [8]; therefore, the front-end network was implemented by LSTM. We attempted to improve recognition accuracy further by initializing the student model parameters with these networks.

Yoshioka *et al*. reported that a hybrid model trained using log mel-filter bank (FBANK) applied feature-space maximum likelihood linear regression (fMLLR) outperformed that trained using mel-frequency cepstral coefficient (MFCC) applied fMLLR [22]. Therefore, we used MFCC applied fMLLR for training the GMM-HMM, and FBANK applied fMLLR for input feature of the hybrid model and LSTM for feature mapping. The fMLLR features are extracted as follows: 13-dimensional MFCCs or 40-dimensional FBANK features were spliced in a context size of nine frames ($\pm$4 frames), and then linear discriminant analysis (LDA) [29] was applied for decorrelation and dimensional reduction to compress into 40 dimensions. Subsequently, maximum likelihood linear transform (MLLT) [30] was applied for further decorrelation. Finally, fMLLR transforms were applied for speaker normalization.

Figure 1 illustrates the knowledge distillation framework for throat microphone speech recognition. The proposed training approach was divided into the following four stages:

1. Training teacher model with alignments;
2. Training feature mapping;
3. Initializing the student model parameters; and
4. Distilling knowledge from the teacher to the student.

First, the alignments of the close-talk microphone data were obtained by the GMM-HMM and the DNN of the hybrid model for the close-talk microphone (teacher model) was trained using close-talk microphone acoustic features as input signals and corresponding alignments (hard labels) as supervised signals. The teacher model has a bottleneck layer with fewer units than the other hidden layers. Subsequently, the LSTM for feature mapping was trained by minimizing mean absolute error between throat microphone acoustic features and the corresponding close-talk microphone BNFs obtained from the bottleneck layer of the teacher model. Next, the front-end network parameters of the student model were initialized with the LSTM for feature mapping, and the back-end network parameters were initialized using the network from the bottleneck layer to the output layer of the teacher model. Finally, the student model was fine-tuned using knowledge distillation. Specifically, the close-talk microphone feature x_c of parallel data was input to the teacher model to obtain the posterior probability distribution $P(s_i|x_c)$ (soft labels). The acoustic states are denoted by s_i. Similarly, the corresponding throat microphone feature x_t was input to the student model to obtain $Q(s_i|x_t)$. The student model was trained to minimize the KL divergence of the following equation:

$$\begin{aligned} D_{KL}(P||Q) &= \sum_i P(s_i|x_c) \log \frac{P(s_i|x_c)}{Q(s_i|x_t)} \\ &= \sum_i P(s_i|x_c) \log P(s_i|x_c) \\ &\quad - \sum_i P(s_i|x_c) \log Q(s_i|x_t) \end{aligned} \quad (1)$$

We can ignore the first term of Equation (1) because it is not related to the optimization of the student model. Therefore, only the second term was used in the minimization. The second term is same as the cross entropy (CE).

During recognition, the student model computed $Q(s_i|x_t)$ from x_t, and output probability distribution $Q(x_t|s_i)$ for HMM-based decoding was computed from $Q(s_i|x_t)$ and priori probability distribution $Q(s_i)$. $Q(s_i)$ was computed in advance from the alignments of a large-sized close-talk microphone data. The HMM was identical to that used in generating hard labels for training the teacher model.

4. Experiments

4.1. Datasets

Figure 2 shows the throat microphone used in this study. A small condenser microphone unit was attached to the tip of the neckband of this throat microphone. Parallel data were recorded simultaneously with a close-talk microphone (SHURE WH20XLR) and a throat microphone using a multitrack recorder (ZOOM R24). We recorded 11 male speakers reading Japanese phonetically balanced sentences for approximately 6 hours to serve as parallel data for training feature mapping and knowledge distillation. For test data, we recorded 6 male speakers reading Japanese newspaper articles for approximately 30 minutes. Approximately 240 hours of lecture speeches from the Corpus of Spontaneous Japanese (CSJ) [28] were used for training the teacher model.

4.2. Experimental setup

The Kaldi toolkit [21] was used for feature extraction, training GMM-HMM, and recognition experiments. The frame length was 25 ms and the frame shift was 10 ms. 440-dimensional features obtained by splicing ± 5 frames were used for the teacher model's input. The (7×40)-dimensions features obtained by combining 6 past frames were used for inputting LSTM for feature mapping and the student model.

Figure 2: *Throat microphone*

Table 1: *Character error rates (CER) of conventional models and the proposed model*

Model	CER
CM-GMM-HMM	82.3 %
TM-GMM-HMM	18.2 %
FM-GMM-HMM	15.8 %
TM-DNN-HMM	14.3 %
KD-LDNN-HMM	**12.9 %**

The GMM-HMM used for generating the hard labels had 3216 tied-state triphone HMM states. The teacher model had seven dense layers with a [1024, 1024, 1024, 1024, 42, 1024, 3216] architecture. The activation of the output layer was done using softmax, and sigmoid was used for the other layers. Kaldi+PDNN [23] was used for pre-training by a stacked denoising autoencoder [24] and fine-tuning of the teacher model. The mini-batch size at pre-training was 128, whereas that of fine-tuning was 256. Stochastic gradient descent was used for optimization in training the teacher model. The LSTM for feature mapping had 512 cells, and the output layer had 42 sigmoid units. The front-end network of the student model had the same architecture as LSTM for feature mapping, and the back-end network had the same architecture as the bottleneck layer of the teacher model. Keras [25] was used for training feature mapping and knowledge distillation, for which the mini-batch size was 256, and Adam [26] was used for optimization. The 3-gram language model was generated from CSJ transcripts.

4.3. Comparison with conventional systems

To compare the proposed system with conventional models, we evaluated the performances of a GMM-HMM trained with approximately 240 hours of close-talk speech data (CM-GMM-HMM), a GMM-HMM and a hybrid model trained with approximately 6 hours of throat microphone data (TM-GMM-HMM, TM-DNN-HMM), and the proposed system (KD-LDNN-HMM). The GMM-HMM of TM-GMM-HMM had 3288 tied-state triphone HMM states, and the DNN of TM-DNN-HMM had seven dense layers with a [1024, 1024, 1024, 1024, 42, 1024, 3288] architecture. We also evaluated the performance of a system using BNF-mediated feature mapping (FM-GMM-HMM), which we previously proposed in [9]. In this study, feature mapping from the FBANK features of throat microphone to the BNFs of close-talk microphone was implemented by LSTM, which had the same architecture as that of the front-end of the student model. The GMM-HMM of the FM-GMM-HMM was trained with approximately 240 hours of close-talk microphone BNFs obtained from the teacher model.

Table 2: *Character error rates (CER) of the student model fine-tuned by knowledge distillation (KD) or using alignments in each initialization method. "Random": random initial-ization; "FM-LSTM": initializing using the LSTM parameter for feature mapping; "Teacher-DNN": initializing using the network parameter from bottleneck layer of the teacher model*

Initializing method		**Fine-tuning method**	
The front-end of student model	The back-end of student model	KD (soft labels)	Alignments (hard labels)
Random	Random	13.5 %	16.2 %
Random	Teacher-DNN	**12.9 %**	17.2 %
FM-LSTM	Random	**12.8 %**	15.5 %
FM-LSTM	Teacher-DNN	**12.9 %**	15.6 %

Table 1 shows the character error rates (CER) of conventional systems and the proposed system. CM-GMM-HMM had high CER due to a large acoustic mismatch. Meanwhile, the proposed system had a 29.1% lower CER relative to the TM-GMM-HMM, and 9.8% lower CER relative to the TM-DNN-HMM. Each of those systems was trained with only throat microphone data. However, the proposed system achieved a 18.4 % CER reduction relative to the FM-GMM-HMM, although like the proposed system, the latter system was also trained utilizing a large-sized close-talk microphone data and small-sized parallel data.

4.4. Comparison of student model training approaches

To verify the effectiveness of the proposed training approach, we evaluated the performance of the student model fine-tuned with soft (knowledge distillation) and hard labels. Additionally, we compared the proposed initializing approach with approaches initializing only the front-end or the back-end of the student model using the pretrained DNN parameter and an approach that initialized all parameters randomly. For the hard labels fine-tuning, the alignments for supervised signals are generated from the GMM-HMM trained with close-talk microphone data. Alignments were generated using parallel data from a close-talk microphone because acoustic mismatch increases the difficulty of estimating reliable alignments using throat microphone data. We used Glorot uniform distribution as a random initialization method [27].

Table 2 summarizes the experimental results. In all the initialization approaches, the model fine-tuned by knowledge distillation had a lower CER than that fine-tuned using hard labels, thus, confirming the better training performance of the former. Moreover, the proposed initialization approach achieved further improvement of recognition accuracy compared with the model that randomly initialized all parameters. However, there was no significant difference in recognition accuracy compared with approaches that initialized only the front-end or back-end.

4.5. Evaluation under noise environment

We evaluated the recognition accuracy of close-talk and throat microphones in various noise environments. To artificially add noise to the test data, white noise generated at different levels (ranging from 40 dB to 90 dB) and several speeches were recorded simultaneously with a close-talk microphone and a throat microphone, and we measured the signal-to-noise ratio of each microphone at each noise level. We also recorded non-stationary noise in restaurants with both microphone types. The restaurant noise whose level was digitally adjusted according to the measurement results shown in Table 3 was added to the close-talk and throat microphone test data, respectively.

The acoustic model of the close-talk microphone was the teacher model, whereas that of throat microphone was the student model. Both were hybrid models. Table 4 shows the experimental results. The recognition accuracy of close-talk microphones was greatly degraded as the noise level increased, whereas the deterioration of the recognition accuracy of the throat microphones was small. The proposed system outperformed the hybrid model of close-talk microphones in noise environments at and greater than 70 dB.

Table 3: *Signal-to-noise ratio (SNR) of close-talk microphone (CM) and throat microphone (TM) per noise level*

Noise level	**CM**	**TM**
40 dB	44.4 dB	40.1 dB
50 dB	39.1 dB	39.2 dB
60 dB	26.7 dB	39.2 dB
70 dB	17.7 dB	34.6 dB
80 dB	13.9 dB	30.3 dB
90 dB	4.7 dB	18.9 dB

Table 4: *Character error rates (CER) of a close-talk microphone (CM) and a throat microphone (TM) for each noise level using test data artificially added with non-stationary noise (restaurants)*

Noise level	**CM**	**TM**
clean	9.2 %	12.9 %
40 dB	9.3 %	14.0 %
50 dB	9.5 %	13.9 %
60 dB	10.7 %	14.5 %
70 dB	17.6 %	**15.4 %**
80 dB	27.2 %	**17.3 %**
90 dB	78.5 %	**34.1 %**

5. Conclusions

In this work, we proposed a training approach for a hybrid model of throat microphones (student model) by distilling knowledge from a DNN trained with large-sized close-talk microphone data (teacher model). The student model had a higher performance than that trained with hard labels, thus, confirming the effectiveness of knowledge distillation in training a hybrid DNN with small-sized parallel data. Moreover, the recognition accuracy of the student model was better using the proposed initializing approach compared with the random initialization, and the proposed system outperformed conventional systems. Furthermore, when evaluating the performance of throat and close-talk microphones using test data with artificially added non-stationary noise, the results indicate that the proposed system had higher recognition accuracy than the hybrid model of close-talk microphone under noise environments at or greater than 70 dB. In future work, we plan to also introduce noise suppression methods and distill knowledge from multiple large-scale teacher models.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers (16H01817) and (16K01543).

7. References

[1] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, G. Zweig, "Achieving human parity in conversational speech recognition," *arXiv:1610.05256*, 2016.

[2] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L. Lim, B. Roomi, P. Hall, "English Conversational Telephone Speech Recognition by Humans and Machines," *Proc. Interspeech 2017*, pp.132-136, 2017.

[3] Z. Zhang, J. Geiger, J. Pohjalainen, A.E. Mousa, W. Jin, B. Schuller, "Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments," *ACM Transaction on Intelligent Systems and Technology*, Vol.9, Issue 5, Article No.49, 2018.

[4] P. Heracleous, J. Even, C.T. Ishi, T. Miyashita, N. Hagita, "Fusion of Standard and Alternative Acoustic Ssnsors for Robust Acoustic Speech Recognition," *ICASPP,* pp.4837-4840, 2012.

[5] S. Dupont, C. Ris, D. Bachelart, "Combined Use of Close-Talk and Throat Microphones for Improved Speech Recognition Under Non-Stationary Background Noise," *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*, pp. 1-4, 2004.

[6] M. Graciarena, H. Franco, G. Myers, C. Cowan, F. Cesari, V. Abrash, "Combination of Standard and Throat Microphones for Robust Speech Recognition in Highly Noisy Environments," *Interspeech*, pp. 1-4, 2004.

[7] S. Lin, T. Tsunakawa, M. Nishida, M. Nishimura, "Conversational Speech Recognition Using Multiple Wearable Microphones," *Proc. of NCSP 2018*, pp.363-366, 2018.

[8] S. Lin, T. Tsunakawa, M. Nishida, M. Nishimura, "DNN-based Feature Transformation for Speech Recognition Using Throat Microphone," *APSIPA ASC 2017*, pp. 596–599, 2017.

[9] T. Suzuki, J. Ogata, T. Tsunakawa, M. Nishida, M. Nishimura, "Bottleneck Feature-Mediated DNN-Based Feature Mapping for Throat Microphone Speech Recognition," *Proc. of APSIPA 2018,* WE-A1-P.10, pp.1738-1741, 2018.

[10] T. Winkler, S. Pronkine, R. Bardeli, J. Köhler, "A study of throat microphone performance in automatic speech recognition on motorcycles," *NAG/DAGA 2009*, pp.1659-1662, 2009.

[11] T. Dekens, W. Verhelst, F. Capman, F. Beaugendre, "Improved speech Recognition in Noisy Environments by Using a Throat Microphone for Accurate Voicing Detection," *Signal Processing Conference*, pp. 1978-1982, 2010.

[12] M. Sahidullah, R. Hautamaki, D. Thomsen, T. Kinnunen, Z. Tan, V. Hautamaki, R. Parts, M. Pitkanen, "Robust Speaker Recognition with Combined Use of Acoustic and Throat Microphone Speech," *Proc. Interspeech 2016*, pp.1720-1724

[13] A. Shahina, B. Yegnanarayana, and M.R. Kesheorey, "Throat microphone signal for speaker recognition," in Proceedings of *ICSLP,* 2004.

[14] M.A.T. Turan, E. Erzin, "Source and Filter Estimation for Throat-Microphone Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 24, no. 2, pp. 265-275, 2016.

[15] A. Shahina, B. Yegnanarayana, "Mapping Speech Spectra from Throat Microphone to Close-Speaking Microphone: A Neural Network Approach," *EURASIP Journal on Advances in Signal Processin*, vol. 2007, no. 2, pp. 1-10, 2007.

[16] K. Vijayan, K.S.R. Murty, "Comparative Study of Spectral Mapping Techniques for Enhancement of Throat Microphone Speech," *Twentieth National Conference on Communications*, pp. 1-5, 2014.

[17] G. Hinton, O. Vinyals, J. Dean, "Distilling the knowledge in a neural network," i*n Deep Learning and Representation Learning Workshop at NIPS 2014*, *arXiv preprint aeXiv:1503:02531*, 2014.

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

[19] J. Yi, J. Tao, Z. Wen, B. Liu, "Distilling Knowledge Using Parallel Data for Far-field Speech Recognition," *arXiv:1802.06941*, 2018

[20] J. Cui, B. Kingsbury, B. Ramabhadran, G. Saon, T. Sercu, K. Audhkhasi, A. Sethy, M. Nussbaum-Thom, A. Rosenberg, "Knowledge Distillation Across Ensemble of Multilingual Models for Low-Resource Languages," *ICASSP 2017,* pp.4825-4829, 2017.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[22] T. Yoshioka, A. Ragni, M.J.F. Gales, "Investigation of unsupervised adaptation of DNN acoustic models with filter bank input," *Proc. ICASSP*, pp.6344-6348, 2014.

[23] Y. Miao, "Kaldi+PDNN: building DNN-based ASR systems with Kaldi and PDNN," *arXiv preprint arXiv:1401.6984*, 2014.

[24] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371-3408, 2010.

[25] F. Chollet, Keras, https://github.com/fchollet/keras, 2015.

[26] D.P. Kingma, J.L. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] X. Glorot, Y. Bengio, "Understanding the Difficulty of Training Deep Feedforward Neural Networks," *in Proc. AIS-TATS*, 2014.

[28] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," *in International Conference on Language Resources and Evaluation (LREC)*, vol. 2, pp.947-952, 2000.

[29] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," *Proc. ICASSP 92*, pp.13-16, 1992.

[30] R. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," *Proc. ICASSP 98*, pp.661-664, 1998.