DOMAIN ADAPTATION OF END-TO-END SPEECH RECOGNITION IN LOW-RESOURCE SETTINGS

Lahiru Samarakoon[1]*, Brian Mak*[2]*, Albert Y.S. Lam*[1]

[1]Fano Labs, Hong Kong
[2]Hong Kong University of Science and Technology
lahiru@fano.ai, mak@cse.ust.hk, albert@fano.ai

ABSTRACT

End-to-end automatic speech recognition (ASR) has simplified the traditional ASR system building pipeline by eliminating the need to have multiple components and also the requirement for expert linguistic knowledge for creating pronunciation dictionaries. Therefore, end-to-end ASR fits well when building systems for new domains. However, one major drawback of end-to-end ASR is that, it is necessary to have a larger amount of labeled speech in comparison to traditional methods. Therefore, in this paper, we explore domain adaptation approaches for end-to-end ASR in low-resource settings. We show that joint domain identification and speech recognition by inserting a symbol for domain at the beginning of the label sequence, factorized hidden layer adaptation and a domain-specific gating mechanism improve the performance for a low-resource target domain. Furthermore, we also show the robustness of proposed adaptation methods to an unseen domain, when only 3 hours of untranscribed data is available with improvements reporting upto 8.7% relative.

Index Terms— ASR, end to end speech recognition, domain adaptation

1. INTRODUCTION

The recent success in machine learning relies on the availability of training data in large quantities. All machine learning techniques, including deep learning-based methods, are susceptible to performance degradation due to the training and testing mismatch. Therefore, when building automatic speech recognition (ASR) systems, it is necessary to find transcribed data from the real domain. However, it is expensive and time-consuming to transcribe large amount of speech training data for a new domain. To mitigate this issue, domain adaptation techniques have been proposed to bootstrap the training of a system for a new domain using a well-trained system [1, 2, 3, 4, 5].

Most of the domain adaptation techniques are evaluated In the context of hybrid deep neural network (DNN) / hidden Markov model (HMM)-based systems [1, 4, 5]. Another approach is unsupervised augmentation of data for the target domain as done using Variational autoencoder based method in [3]. In [2], domain adaptation is investigated for end-to-end acoustic models for distant speech recognition. However, their approach can only be used when a parallel corpus is available between source and target domains.

End-to-end ASR provides a simplified and elegant system building pipeline by optimizing a single model instead of the traditional way of building separate acoustic, language and pronunciation models. Thus, one of the most attractive advantages of end-to-end ASR is that these systems do not rely on expert linguistic knowledge for building pronunciation dictionaries. Therefore, end-to-end ASR fits well when building systems for new domains. However, one major drawback of end-to-end ASR is that, it is necessary to have a larger amount of labeled speech in comparison to hybrid DNN/HMM systems. This can be a restricting factor in terms of performance when deploying end-to-end ASR systems for new domains.

Therefore, in this paper, we explore domain adaptation approaches for end-to-end ASR systems in low-resource settings. There are two main stream for end-to-end ASR systems. Namely; connectionist temporal classification (CTC) [6, 7] and attention-based encoder-decoder mechanism [8, 9, 10, 11]. In this work we use the hybrid attention/CTC architecture [12]. This hybrid architecture achieved state-of-the-art performances for Chinese and Japanese ASR benchmarks [12]. We investigate joint domain identification and ASR, factorized hidden layer adaptation (FHL) [13], cluster adaptive training (CAT) [14] and a domain-specific gating mechanism [15] to improve the performance for a low-resource target domain. We evaluate the proposed techniques by building an end-to-end system for conversational Hong Kong Cantonese telephony speech when only 37 hours of training data is available. Furthermore, we also investigate the robustness in adapting to a new domain which is not seen during training, when only 3 hours of untranscribed data is available.

The rest of the paper is organized as follows. Section 2 describes the domain adaptation approaches to end-to-end speech recognition. In Section 3, we give the details of the experimental setup. Results are presented in Section 4. Finally, we summarize our findings in Section 5.

2. DOMAIN ADAPTATION

The simplest approach for domain adaptation is to use the validation data from the target domain to guide the training. In addition, this section discusses multiple ways of optimizing an end-to-end model for a low-resource target domain when the model is also trained with out-of-domain data. Moreover, we detail how to perform the unsupervised domain adaptation for a new target domain when the adaptation data is very limited.

2.1. Joint domain identification and speech recognition

The intuition behind this approach is to make the model explicitly identifies the domain for a given sequence. It is possible to achieve this through multi-task learning where the domain classification can be used as an auxiliary objective [16]. However, multi-task learning is not desirable as our end-to-end model uses multi-task learning with hybrid attention/CTC architecture. Therefore, in this work we propose to do domain identification by expanding the character dictionary using a set of special tags, each of which corresponds to the domain. This explicit identification by augmented output targets has been used in multi-lingual machine translation [17], multi-dialect end-to-end ASR [18], and also for language independent end-to-end speech recognition [19]. Furthermore, this approach has shown to be effective when there is code-switching [20].

In this work, we assume the domain information is always available. This is a realistic assumption because the target domain is known in advance and the out-of-domain data is only used during training. Moreover, there is no switching of domains in a sequence. So each utterance can be conditioned on the correct domain.

For multi-dialect ASR, it is shown that having dialect identification after speech recognition is more beneficial because that speech recognition is more fragile to the errors incurred in dialect identification. Therefore, it was found that having dialect identification at the end of the label sequence is more beneficial [18]. Therefore, in this work we explore domain identification at the start of the sequence as well as at the end. A Cantonese example for the target domain is given below (English translation : "Although I really don't understand"):

- original sequence :
  "雖然我真係唔係好明"
- domain identification before speech recognition:
  "[D]雖然我真係唔係好明"
- domain identification after speech recognition:
  "雖然我真係唔係好明[D]"

where [D] is the domain-specific tag.

2.2. Domain information as input vectors

To make the model aware of the domain, it is possible to pass the domain information as a feature to the model [21, 22, 23] to learn domain dependent biases. In this work, we concatenate input to each encoder layer with our domain-specific feature. We explore multiple ways of generating the domain specific feature. First, we use 1-hot vectors for each domain. Second, we investigate the passing of channel (whether it is telephony speech or smartphone recorded speech) or the style (conversational vs read) into the domain-specific feature. We also investigate linearly transforming the domain-specific features to a higher dimension before being concatenated with the output of the layer below.

2.3. Domain-specific gating

In [15], language-specific gating is used to modulate the hidden representations of a multi-lingual end-to-end model. In that work, outputs of each hidden layer are processed by a set of language-dependent gates before being inputted to the next layer. In this paper, we investigate the effectiveness of domain-specific gating to improve the performance of the target domain.

The domain specific gating is computed using the domain-specific vector ($\mathbf{v}_d$ for domain d) and the current (i-th) layer output ($\mathbf{h}_i$) as given below:

$$g(\mathbf{h}_i, \mathbf{v}_d) = \sigma(\mathbf{A}\mathbf{h}_i + \mathbf{B}\mathbf{v}_d + \mathbf{b}), \quad (1)$$

where $\mathbf{A}, \mathbf{B}$ and $\mathbf{b}$ are trainable parameters and σ is the sigmoid function. Then, domain-specific output ($\hat{\mathbf{h}}_i$) for the current layer is calculated as

$$\hat{\mathbf{h}}_i = g(\mathbf{h}_i, \mathbf{v}_d) \odot \mathbf{h}_i \quad (2)$$

Finally, $\hat{\mathbf{h}}_i$ and $\mathbf{v}_d$ is concatenated before being presented to the next layer.

2.4. Cluster adaptive training

Cluster adaptive training (CAT) [24] is used successfully for DNN adaptation [24, 14, 25]. Therefore, this paper also investigates the effectiveness of CAT for domain adaptation. Since we are dealing with domain adaptation in low-resource settings, we cannot estimate different bidirectional Long Short Term Memory with projections (BLSTMPs) for each domain and use 1-hot domain vectors to switch between clusters. Therefore, we propose to apply CAT to the projection ($\mathbf{W}$) of the BLSTMP layers as given below:

$$\mathbf{W}_d = \mathbf{W} + \sum_{i=1}^{|\mathbf{v}_d|} \mathbf{v}_d(i)\mathbf{B}(i) \quad (3)$$

where $\{\mathbf{B}(1), \mathbf{B}(2), .., \mathbf{B}(|\mathbf{v}_d|)\}$ is the set of basis matrices for clusters and $\mathbf{v}_d$ is the 1-hot domain vector.

2.5. Factorized hidden layer adaptation

Factorized hidden layer (FHL) adaptation has been introduced for acoustic model adaptation [13, 26, 27] and has shown to outperform CAT [28]. Therefore, we also investigate the effectiveness of FHL for domain adaptation. Since we are constrained from the availability of the data for some domains, we propose only to estimate the FHL basis matrices for the projections of the BLSTMP layers of the encoder.

The major difference between CAT and FHL is that FHL uses rank 1 basis matrices while CAT uses full-rank. Therefore, from (3) and by constraining basis matrices ($\mathbf{B}(i)$s) to rank 1:

$$\begin{aligned}\mathbf{W}_d &= \mathbf{W} + \sum_{i=1}^{|\mathbf{v}_d|} \mathbf{v}_d(i)\boldsymbol{\gamma}(i)\boldsymbol{\psi}^\top(i) \\ &= \mathbf{W} + \boldsymbol{\Gamma}\mathbf{V}_d\boldsymbol{\Psi}^\top, \end{aligned} \tag{4}$$

where $\mathbf{B}(i) = \boldsymbol{\gamma}(i)\boldsymbol{\psi}^\top(i)$ and $\mathbf{V}_d$ is a diagonal matrix ($\mathbf{V}_d = \mathrm{diag}(\mathbf{v}_d)$) and $\boldsymbol{\gamma}(i)$, $\boldsymbol{\psi}(i)$ are i-th column vectors for $\boldsymbol{\Gamma}$, $\boldsymbol{\Psi}$ respectively. Since, FHL uses rank 1 basis matrices, we first have to increase the dimensionality of the domain vector by estimating a non-linear projection during training as given below:

$$\hat{\mathbf{v}}_d = \sigma(\boldsymbol{\Lambda}\mathbf{v}_d) \tag{5}$$

where σ is the sigmoid function and $|\hat{\mathbf{v}}_d| > |\mathbf{v}_d|$. Then, this new domain vector can be used to estimate the bases for FHL basis matrices as given below:

$$\mathbf{W}_d = \mathbf{W} + \boldsymbol{\Gamma}\hat{\mathbf{V}}_d\boldsymbol{\Psi}^{l\top} \tag{6}$$

where $\hat{\mathbf{V}}_d = \mathrm{diag}(\hat{\mathbf{v}}_d)$.

2.6. Unsupervised adaptation

So far, we have detailed multiple approaches to train a model for a target domain that utilizes data from other domains. In this section, we briefly explain how these well-trained models can be adapted to a new domain when only a limited amount of untranscribed data is available. This falls into unsupervised adaptation where the data from the new target domain are transcribed using the trained model before the adaptation is conducted.

One issue with the joint domain identification and speech recognition approach is that it is not possible to introduce a new symbol for the new domain without retraining the entire model. Therefore, in this work we only perform the new domain adaptation for models trained with other approaches. When the model uses domain vectors as a bias, we propose to estimate a shift to the current target domain's vector using the new target domain data while keeping all other parameters of the model fixed. For the case of domain-specific gating, we propose to estimate parameters of the gating layers $\mathbf{A}, \mathbf{B}$ and $\mathbf{b}$ from (1) while keeping the other parameters fixed. For the case of CAT and FHL, it is also possible to estimate the basis matrices ($\{\mathbf{B}(1), \mathbf{B}(2), .., \mathbf{B}(|\mathbf{v}_d|)\}$) instead of updating the domain-specific interpolation vector ($\mathbf{v}_d$) when more than two hours of data are available.

Table 1. Details of the in-domain corpus (D) and the out-of-domain corpora (C1, C2) used in this work.

Corpus ID	Details
D	Hong Kong Cantonese mixed with English Conversational speech Telephone channel (8 kHz, 16bit) 46 hours
C1	Guangdong Cantonese mixed with English Read speech Smartphone channel (16 kHz, 16bit) 300 hours
C2	Guangdong Cantonese (non-mixed) Conversational speech Smartphone channel (16 kHz, 16bit) 700 hours

3. EXPERIMENTAL SETUP

Table 1 shows the corpora used in this work. The target domain (D) is Hong Kong (HK) Cantonese conversational telephone speech. HK conversational speech is also mixed with general English words. However, we only have 46 hours of target domain data which is considerably low for end-to-end speech recognition. We are using 80% of the data for training and 10% each for validation and testing. In addition, out-of-domain corpora C1 and C2 are used to improve the performance for the target domain (D). For all the corpora, simplified Chinese characters were converted to traditional Chinese symbols. The speech of the domain D is up-sampled from 8 kHz to 16 kHz. We extract 80 dimensional filterbank features with 3 dimensional pitch features [29] using Kaldi [30]. We also perform the speed perturbation for the training data of target domain D [31]. All our models use the hybrid attention/connectionist temporal classification architecture [32, 12]. All the models has a 4-layer BLSTMP encoder network. The final softmax layer has 5475 dimensions. These experiments are implemented with the Pytorch [33] backend on top of Espnet [32, 12] implementation. All results are reported on the test set of the target domain (D). Finally, for the unsupervised adaptation experiments, we use a new target domain which is not seen during training. This domain only has 3 hours of noisy telephony speech with considerable amount of overlapped speech.

Table 2. *Word error rates (WER %) for baselines.*

Model	WER	Corpus
Baseline (M_1)	46.0	D
+ speed peturb. (M_2)	42.6	D_{sp}

Table 3. *WER % for models trained using various combinations of out-of-domain data.*

Model	WER	Corpus
M_3	37.9	$D_{sp} + C1$
M_4	35.6	$D_{sp} + C2$
M_5	35.1	$D_{sp} + C1 + C2$

4. RESULTS

Table 2 shows the results for the baseline models that are trained only using the target domain data. The baseline which is trained only using the training set of the target domain reports 46.0%. During speed perturbation, two additional copies of the target domain training set is created by modifying the speed to 0.9 and 1.1 times of the original rate. This creates a 3-fold training set. As can be seen clearly, speed perturbation significantly improves the performance. For the rest of the experiments, this speed perturbed target domain training set is used.

Next, we investigate the effect of adding out-of-domain data as shown in Table 3. As mentioned before, in all these experiments, the validation set only contains the target domain data. The inclusion of out-of-domain data significantly improves the performance. The best performance is reported when both C1 and C2 are included. Therefore, in rest of our experiments, we use both out-of-domain training sets as well as the speed perturbed target domain training set.

In Table 4 we present the results for joint domain identification and speech recognition. As clearly seen, the best performance is reported when the domain identification is performed before speech recognition. This observation is contradictory to the findings of joint dialect identification and speech recognition results mentioned in [18]. In that work, authors found that if the dialect identification is performed before speech recognition, the errors in dialect identification degrade the final ASR performance. This can be due to multiple reasons. First, dialect identification is much harder than domain identification and more likely to negatively influence speech recognition. Second, our model is trained to optimize for the target domain whereas their model is to recognize multiple dialects. Third, they get dialect information based on the geographical locations of speakers which can be very noisy. Finally, we are use the hybrid attention/CTC architecture whereas they employ a pure attention-based model, which makes the model more sensitive to the previous predictions. For the rest of the experiments, domain identification is always performed before the speech recognition.

Table 4. *WER % for Joint domain identification and speech recognition.*

Model	WER	Domain Identification
M_6	34.0	Before SR
M_7	34.5	After SR

Table 5. *WER % for various models when the domain vector is used as a bias to the encoder.*

Model	WER	Domain Vector Dimension
M_8	35.1	3
M_9	34.4	7
M_{10}	34.0	100 (7)
M_{11}	34.2	400 (7)
M_{12}	34.2	500 (7)

Next, we investigate the effect of feeding domain information as an input feature. In the experiments listed in Table 5, we feed domain information as a bias to the encoder. We explore two different representations for domain vectors. First, we create a 1-hot domain indicator for each domain which gives us a 3-dimensional 1-hot vector. The result for that is shown in the first row. As can be clearly seen, the 3-dimensional 1-hot vector degrades the performance. Second, we concatenate 3-dimensional domain vector with a channel 1-hot vector and a speech style 1-hot vector. Since we have two channels (telephone and smartphone), and two speech styles (conversational and read), the final domain vector is 7-dimensional. For clarity, 7-dimensional domain vectors for each domain are listed below:

- D : [1, 0, 0, 1, 0, 1, 0]
- C1 : [0, 1, 0, 0, 1, 0, 1]
- C2 : [0, 0, 1, 0, 1, 0, 1].

As seen from the second row, the 7-dimensional domain vector outperforms the 3-dimensional domain vector. However, still degrading the performance compared to joint domain identification and speech recognition (M_7). Furthermore, we also investigate the effect of projecting the 7-dimensional domain vector to higher dimensions for improvement. This slightly improves the performance and 100 dimensional projections give us the best result. However, it seems that joint domain identification and domain vector-based bias is not complementary.

Table 6 shows the results for cluster adaptive training. As mentioned in Section 2.4, due to the limited availability of the target domain data, CAT is only performed on projections of BLSTMP layers. We use 7-dimensional domain vectors as interpolation vectors for clusters. As can be seen in the results, CAT degrades the performance significantly. We believe this degradation is mainly due to the limited availability

Table 6. *WER % for cluster adaptive training on the BLSTMP projections.*

Model	WER	Clustered Layer
M_{13}	37.5	All
M_{14}	34.2	1
M_{15}	34.6	2
M_{16}	34.2	3
M_{17}	39.6	4

Table 7. *WER % for FHL and domain-specific gating experiments.*

Model	WER	Details
M_{18}	33.7	FHL
M_{19}	33.7	Gating

of the target domain data as a full-rank basis matrix should be estimated for each cluster.

Next, we present the results for FHL and domain-specific gating experiments. For both cases, 7-dimensional domain vectors are used. We can see that both approaches report similar improvements. Therefore, for the new target domain adaptation experiments, we only consider FHL and domain-specific gating approaches.

Finally, we conduct experiments to evaluate these domain adaptation approaches for a new domain that is not seen during training. However, it is worth highlighting that this unseen domain is a bit similar to the original target domain D. Table 8 provides a comparison of these two different domains. In Table 9, we analyze the performances of various domain adaptation approaches to the new target domain. As can be clearly seen, the performance of the baseline (M_5) degrades significantly when domain is changed (from 35.1% to 58.6%) highlighting the differences between two target domains. In addition, without any adaptation, FHL enjoys the best performance for the new target domain (56.6%) . Next, we perform unsupervised adaptation only using 3 hours of new target domain data as illustrated in Section 2.6. Both FHL (53.5%) and domain-specific gating (53.7%) enjoys significant improvements from unsupervised adaptation.

Fig. 1 shows WERs for the number of adaptation iterations. This result indicates that 3 iterations over the adaptation data are sufficient. More importantly, it also highlights the robustness of both FHL and domain-specific gating since no overfitting is observed with more adaptation iterations.

5. CONCLUSION

In this paper, we detailed domain adaptation approaches for end-to-end speech recognition in low resource settings. We found that addition of out-of-domain data, while guiding the training with a validation set from the target domain is to be beneficial.In addition, joint domain identification and speech recognition reported improvements while performing the domain identification first gave us more gains. Moreover, factorized hidden layer adaptation (FHL) and domain-specific gating considerably improved the domain adaptation performances whereas providing domain information as a bias and cluster adaptive training (CAT) on the projections of encoder layers were not effective. Finally, we showed that both FHL and domain-specific gating are robust to unsupervised adaptation to a new domain when only 3 hours of untranscribed data is available.

Table 8. *Comparison of the new target domain with the original target domain (D).*

characteristics	original target domain	new target domain
language	HK Cantonese	HK Cantonese
channel	telephone	telephone
mixed with English	yes	yes
speaker overlaps	low	high
condition	clean	noisy
content	general	telecommunication
style	conversational	conversational

Table 9. *WER % for various approaches on the new target domain test data. The performance for the original target domain (D) is given in brackets.*

Model	First Pass	Second pass
$D_{sp} + C1 + C2$ (M_5)	58.6 (35.1)	-
Joint (M_7)	58.0 (34.0)	-
FHL (M_{18})	56.6 (33.7)	53.5
Gating (M_{19})	58.1 (33.7)	53.7

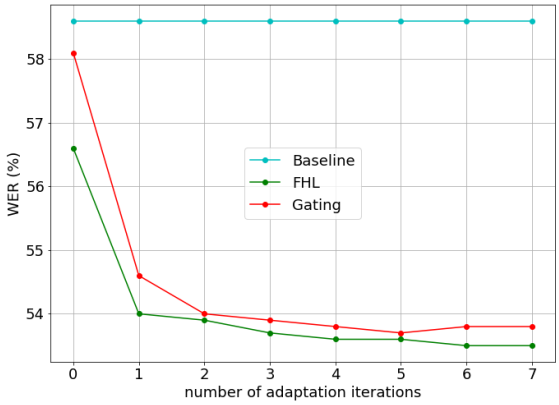

Fig. 1. *WER (%) for number of unsupervised adaptation iterations.*

6. REFERENCES

[1] Jinyu Li, Michael L Seltzer, Xi Wang, Rui Zhao, and Yifan Gong, "Large-scale domain adaptation via teacher-student learning," in *Interspeech*, 2017.

[2] Seyedmahdad Mirsamadi and John HL Hansen, "On multi-domain training and adaptation of end-to-end rnn acoustic models for distant speech recognition," in *Interspeech*, 2017.

[3] Wei-Ning Hsu, Yu Zhang, and James Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," *arXiv preprint arXiv:1707.06265*, 2017.

[4] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *ICASSP*, May 2013, pp. 7398–7402.

[5] Lahiru Samarakoon and Khe Chai Sim, "Multi-attribute factorized hidden layer adaptation for dnn acoustic models.," in *Interspeech*, 2016, pp. 3484–3488.

[6] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[7] Alex Graves and Navdeep Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014, pp. 1764–1772.

[8] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*. IEEE, 2016, pp. 4960–4964.

[9] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: first results," *arXiv preprint arXiv:1412.1602*, 2014.

[10] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, "End-to-end attention-based large vocabulary speech recognition," in *ICASSP*. IEEE, 2016, pp. 4945–4949.

[11] Liang Lu, Xingxing Zhang, and Steve Renais, "On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition," in *ICASSP*. IEEE, 2016, pp. 5060–5064.

[12] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, Dec 2017.

[13] Lahiru Samarakoon and Khe Chai Sim, "Factorized hidden layer adaptation for deep neural network based acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[14] T. Tian, Q. Yanmin, Y. Maofan, Z. Yimeng, and K. Yu, "Cluster adaptive training for deep neural network," in *ICASSP*. IEEE, 2015, pp. 4325–4329.

[15] Suyoun Kim and Michael L Seltzer, "Towards language-universal end-to-end speech recognition," in *ICASSP*. IEEE, 2018.

[16] Michael L Seltzer and Jasha Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *ICASSP*. IEEE, 2013, pp. 6965–6969.

[17] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al., "Google's multilingual neural machine translation system: enabling zero-shot translation," *arXiv preprint arXiv:1611.04558*, 2016.

[18] Bo Li, Tara N Sainath, Khe Chai Sim, Michiel Bacchiani, Eugene Weinstein, Patrick Nguyen, Zhifeng Chen, Yonghui Wu, and Kanishka Rao, "Multi-dialect speech recognition with a single sequence-to-sequence model," in *ICASSP 2018*.

[19] Shinji Watanabe, Takaaki Hori, and John R Hershey, "Language independent end-to-end architecture for joint language identification and speech recognition," in *ASRU*. IEEE, 2017, pp. 265–271.

[20] Hiroshi Seki, Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R Hershey, "An end-to-end language-tracking speech recognizer for mixed-language speech," in *ICASSP 2018*.

[21] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in *ICASSP*. IEEE, 2013, pp. 7942–7946.

[22] H. Huang and K.C. Sim, "An investigation of augmenting speaker representations to improve speaker normalization for DNN-based speech recognition," in *ICASSP*. IEEE, 2015, pp. 4610–4613.

[23] Lahiru Samarakoon and Khe Chai Sim, "On combining i-vectors and discriminative adaptation methods for unsupervised speaker normalization in DNN acoustic models," in *ICASSP*. IEEE, 2016.

[24] M.J.F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 417–428, 2000.

[25] C. Wu and M. J. F. Gales, "Multi-basis adaptive neural network for rapid adaptation in speech recognition," in *ICASSP*. IEEE, 2015, pp. 4315–4319.

[26] Lahiru Samarakoon and Khe Chai Sim, "Learning factorized transforms for speaker normalization," in *ASRU*. IEEE, 2015.

[27] Lahiru Samarakoon, Brian Mak, and Khe Chai Sim, "Learning effective factorized hidden layer bases using student-teacher training for lstm acoustic model adaptation," in *ICASSP*, 2018.

[28] Lahiru Samarakoon and Brian Mak, "Unsupervised adaptation of student dnns learned from teacher rnns for improved asr performance," in *ASRU*. IEEE, 2017, pp. 200–205.

[29] Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *ICASSP*. IEEE, 2014, pp. 2494–2498.

[30] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The Kaldi speech recognition toolkit," in *ASRU*. IEEE, 2011.

[31] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[32] Suyoun Kim, Takaaki Hori, and Shinji Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *ICASSP*. IEEE, 2017, pp. 4835–4839.

[33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in Pytorch," in *NIPS-W*, 2017.