A HIGHLY ADAPTIVE ACOUSTIC MODEL FOR ACCURATE MULTI-DIALECT SPEECH RECOGNITION

Sanghyun Yoo[1] *Inchul Song*[1] *Yoshua Bengio*[2*]

[1]Samsung Advanced Institute of Technology, Republic of Korea
[2]Mila, Université de Montréal, Canada
{sam.yoo, inchul2.song}@samsung.com yoshua.bengio@mila.quebec

ABSTRACT

Despite the success of deep learning in speech recognition, multi-dialect speech recognition remains a difficult problem. Although dialect-specific acoustic models are known to perform well in general, they are not easy to maintain when dialect-specific data is scarce and the number of dialects for each language is large. Therefore, a single unified acoustic model (AM) that generalizes well for many dialects has been in demand. In this paper, we propose a novel acoustic modeling technique for accurate multi-dialect speech recognition with a single AM. Our proposed AM is dynamically adapted based on both dialect information and its internal representation, which results in a highly adaptive AM for handling multiple dialects simultaneously. We also propose a simple but effective training method to deal with unseen dialects. The experimental results on large scale speech datasets show that the proposed AM outperforms all the previous ones, reducing word error rates (WERs) by 8.11% relative compared to a single all-dialects AM and by 7.31% relative compared to dialect-specific AMs.

Index Terms— Acoustic modeling, multi-dialect speech recognition, adaptation

1. INTRODUCTION

Every language has variations in terms of pronunciation. For example, English spoken by a native speaker is different from the one spoken by a non-native speaker, e.g., an Asian. Therefore, an acoustic model (AM) trained with speech data from only native speakers often fails to recognize speech from non-native speakers. An effective approach to dealing with multi-dialect speech recognition is to train a dialect-specific AM for each dialect [1]. Although it performs well in general, one disadvantage is that a separate AM needs to be maintained for each dialect, which increases operational cost. Thus, there have been demands for a single AM that can be used to recognize many dialects accurately.

A straightforward way to build a single AM for multiple dialects is to train the model on mixed data from many dialects. Although such an AM usually underperforms dialect-specific AMs, it can be used as a good starting point and hence is commonly considered as the baseline model in the literature [1, 2, 3, 4]. Recently, there have been many attempts to improve the single AM approach for multi-dialect speech recognition. One such approach is based on multi-task learning, where an AM is trained to not only predict phonemes but also identify dialects [1, 2, 5, 6]. There are other approaches that provide the AM with an auxiliary input such as i-vectors [7] or dialect information to make it adaptive to different dialects [2, 3, 6].

Feature-wise Linear Modulation (FiLM) is a recently introduced adaptive neural network modeling technique [8, 9, 10, 11]. By applying feature-wise transformations based on an auxiliary input, it enables neural networks to effectively adapt to multiple sources of information in a large set of problems [11]. Although FiLM was originally proposed for a visual QA problem [8, 9, 10], it is a general approach and hence can also be applied to acoustic modeling.

In this paper, we propose a novel approach to building a single AM for multi-dialect speech recognition. Motivated by FiLM, we apply feature-wise transformations to the AM based on dialect information. However, unlike the original FiLM architecture that adapts neural networks based on only external input, we also use internal representation extracted within the network as additional conditioning information. Using the combination of both external and internal information in feature-wise transformations makes our AM more adaptive and able to deal with multiple dialects more effectively. We also propose a simple but effective training method to handle unseen dialects during training. Through experimental evaluation on large scale speech datasets, we show that our proposed AM outperforms all the previous ones in terms of WERs.

The rest of this paper is organized as follows. First, Section 2 discusses past research related to this work. Section 3 describes our proposed AM that is adapted by both dialect information and internal representation. Section 4 shows the experimental results, and Section 5 concludes the paper.

2. RELATED WORK

The traditional approach to dealing with multi-dialect speech is to build a different AM per dialect as if each dialect is a different language [1]. However, when dialect-specific data is scarce, such a dialect-specific AM underperforms even a single AM trained with data from all dialects [5]. This is because a dialect-specific AM is not able to learn similarities between different dialects. To deal with such a resource-scarcity problem, there are some approaches that jointly train dialect-independent parts of an AM with data from all dialects and separately train dialect-dependent parts with dialect-specific data [3, 5, 12]. In other approaches, dialect-specific AMs are obtained simply by training a single AM using all available data first and then fine-tuning it on dialect-specific data [2]. However, the main drawback of these methods is that they all need to maintain several models for a given language, which incurs high maintenance cost especially when speech recognition service is provided for many languages and dialects [1].

A single AM for all dialects has merit in terms of maintenance cost, and hence there have been many attempts to improve its performance. Some of them are based on multi-task learning [1, 2, 5, 6]

[*] CIFAR Senior Fellow & AI Chair.

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

where an AM is trained to predict phonemes as well as classify dialects. However, it is not easy to train such a model because deciding the weights given to two different tasks is not straightforward [5]. Other approaches use an auxiliary input such as i-vectors or dialect information [2, 3, 6]. [2] proposes to provide dialect information as an additional input of an end-to-end speech recognition model, and [6] proposes to use utterance-level dialect embedding extracted from a dialect classifier. It has been shown that they outperform dialect-specific models [2].

Conditioning is an effective method for adaptive neural network modeling. [13] shows that conditional instance normalization, which dynamically generates the scaling and shifting parameters in instance normalization, can transfer styles of images succinctly. Motivated by this work, [4] proposes to dynamically generate the parameters in layer normalization [14] for speaker adaptation. Later, a general-purpose conditioning approach called Feature-wise Linear Modulation (FiLM) is proposed [8, 9, 10, 11]. In this approach, special layers called FiLM layers are inserted into the network, and each such layer applies feature-wise affine transformations to its input as follows:

$$\hat{\mathbf{x}} = \gamma \odot \mathbf{x} + \beta$$

where $\odot$ indicates pointwise product of vectors, x is a FiLM layer's input vector, and γ and β are the scaling and shifting vectors that are dynamically generated based on an auxiliary input. Although FiLM is proposed to solve a visual QA problem, such feature-wise transformations can also be applied to solve a diverse set of problems [11]. However, to the best of our knowledge, there has been no work that uses FiLM for multi-dialect speech recognition. In this paper, we examine how FiLM can be applied to acoustic modeling for accurate multi-dialect speech recognition and propose an improved network architecture.

3. MULTI-DIALECT ACOUSTIC MODELING

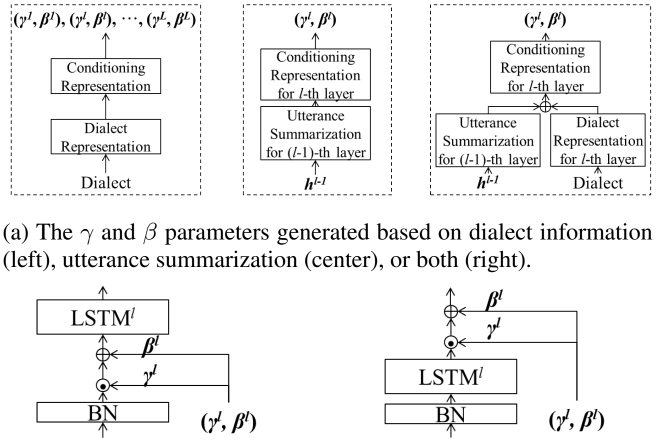
3.1. Baseline Acoustic Models

All AMs considered in this paper use a uni-directional recurrent neural network with LSTM layers followed by a softmax layer. We apply batch normalization to the input of each LSTM layer as in [15] and also add a lookahead convolutional layer [16] after each LSTM layer to provide the model with some future context. We call this basic AM that does not take any dialect information a *dialect-unaware AM*. Then, we obtain *dialect-specific AM*s by fine-tuning the dialect-unaware AM with data from the corresponding dialect. Our third baseline AM is to take dialect information as an additional input [2, 6, 17]. As in [2], dialect information is represented as a one-hot vector and concatenated with the other input features[1]. We call this model a *dialect-aware AM*.

3.2. Conditioning with External Information

The FiLM architecture can be applied to acoustic modeling for multiple dialects by using dialect information as a conditioning input. Based on this conditioning information, the scaling and shifting parameters for feature-wise affine transformations, γ and β, are dynamically generated. If dialect information d is represented as a D-dimensional one-hot vector where D is the number of dialects, the γ and β for all layers can be generated at once as follows:

$$a_c = tanh(W_c(tanh(W_d d + b_d)) + b_c)$$



(γ1, β1), (γ2, β2), ⋯, (γL, βL)
Conditioning Representation
Dialect Representation
Dialect
(γl, βl)
Conditioning Representation for l-th layer
Utterance Summarization for (l-1)-th layer
hl-1
(γl, βl)
Conditioning Representation for l-th layer
Utterance Summarization for (l-1)-th layer
Dialect Representation for l-th layer
hl-1
Dialect
LSTMl
βl
γl
BN
(γl, βl)
βl
γl
LSTMl
BN
(γl, βl)

(a) The γ and β parameters generated based on dialect information (left), utterance summarization (center), or both (right).

(b) Conditioning applied to each layer's input (left) or output (right).

Fig. 1: Illustration of (a) three kinds of conditioning information and (b) two different conditioning positions.

$$(\gamma^1, ...\gamma^L) = tanh(W_\gamma a_c + b_\gamma), \quad (\beta^1, ...\beta^L) = tanh(W_\beta a_c + b_\beta),$$

where L is total number of LSTM layers, Ws are weight matrices, and b's are bias terms. This is shown on the left picture in Fig. 1a.

The generated γ and β parameters are applied to either the input of each LSTM layer after batch-normalization (as shown on the left picture in Fig. 1b) or the output of each layer after activation (as shown on the right picture in Fig. 1b).

3.3. Conditioning with Internal Representation

[4] shows that an AM dynamically adapted based on the hidden representation of each layer performs well in speaker adaptation. Their method can also be applied to multi-dialect speech recognition. Similar to [4], we attach a separate utterance-level feature extractor network to each layer and jointly train it with the main AM. The feature vector a_s^{l-1}, called utterance summarization of $(l-1)$-th layer, is extracted as follows:

$$a_s^{l-1} = \frac{1}{T}\sum_{t=1}^{T} tanh(W_s^{l-1} h_t^{l-1} + b_s^{l-1}),$$

where T is the number of total frames in the utterance, and h_t^{l-1} is the output of the $(l-1)$-th hidden layer at time step t (h_t^0 is the input features x_t). W_s^{l-1} and b_s^{l-1} are the weight matrix and bias term, respectively. We apply one more non-linear transformation before parameter generation. Therefore, the γ^l and β^l at layer l are generated as follows:

$$a_c^l = tanh(W_c^l a_s^{l-1} + b_c^l)$$

$$\gamma^l = W_\gamma^l a_c^l + b_\gamma^l, \qquad \beta^l = W_\beta^l a_c^l + b_\beta^l,$$

where W^ls are weight matrices, and b^ls are bias terms. This method is depicted on the center picture in Fig. 1a.

3.4. Conditioning with External and Internal Information

Finally, we propose to use not only dialect information and but also utterance summarization in parameter generation. Conditioned by both external input and internal representation at the same time, an

[1] Although the dialect information can be added to the input of every layer, there is not much difference in performance in our experiments.

Table 1: The statistics of various datasets in terms of the number of utterances. (The numbers in parentheses refer to hours.)

Speaker	Training set	Speaker	Training set
Native	281,241 (960)	Chinese	73,413 (131)
German	55,836 (74)	Spanish	75,762 (102)
French	47,881 (59)	Italian	43,863 (57)
Portuguese	35,933 (46)	Korean	N/A (N/A)

AM can be more adaptive in dealing with multiple dialects simultaneously than models described above.

First, a one-hot vector indicating the dialect is transformed into a hidden representation a_d^l and concatenated with utterance summarization of the previous layer. Then, it is fed into another non-linear layer to make the final conditioning representation a_c^l as follows:

$$a_d^l = tanh(W_d^l d + b_d^l)$$

$$a_c^l = tanh(W_c^l(a_d^l \oplus a_s^{l-1}) + b_c^l),$$

where d is a one-hot vector representing the dialect, a_s^{l-1} is the utterance summarization of $(l-1)$-th layer, which is introduced in Section 3.3. The symbol $\oplus$ indicates the concatenation operation. The γ^l and β^l at l-th layer are generated based on this conditioning representation a_c^l as shown on the right picture in Fig. 1a:

$$\gamma^l = W_\gamma^l a_c^l + b_\gamma^l, \qquad \beta^l = W_\beta^l a_c^l + b_\beta^l.$$

As before, Ws and b's indicate the weight matrices and bias terms.

3.5. Unseen Dialect Handling

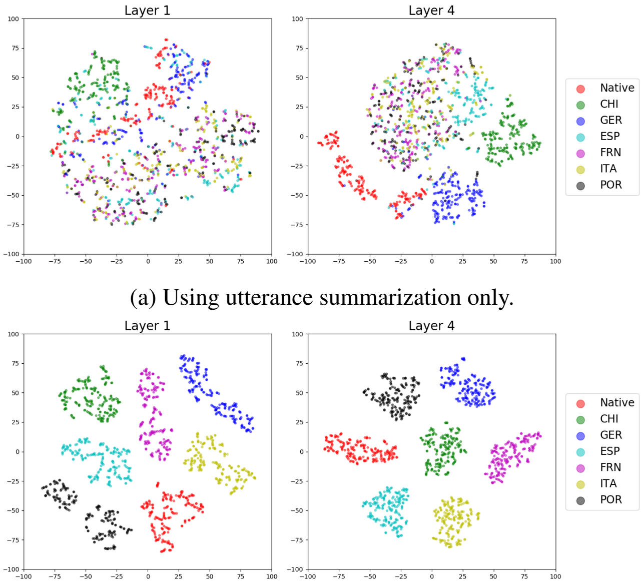
One limitation of our model is that users always need to provide dialect information. Therefore, even for users speaking unseen dialects during training, one dialect (usually native) must be chosen, which may result in performance degradation. Using a dialect classifier might be one solution [5, 6], but it makes the overall performance largely dependent on the classification accuracy. In this paper, we handle the unseen dialect problem by simply adding one more dialect, called $unknown$. During training, utterances randomly sampled with a certain probability (e.g., 0.1) are treated as if they were from the unknown dialect. At test times, users speaking unseen dialects are considered to use the unknown dialect and conditioning representation is generated accordingly.

4. EXPERIMENTAL RESULTS

4.1. Datasets

We evaluate the models described above on large vocabulary speech recognition datasets, which are English speeches spoken by users of 8 different dialects including native, Chinese, German, Spanish, French, Italian, Portuguese, and Korean. For the native speech data, we use 960 hours of speeches from the LibriSpeech corpus [18]. For the nonnative speech data, we use the commercial datasets from Speech Ocean [19]. Among these, the Korean English dataset is not used during training for the ablation test. The statistics of each dataset are shown in Table 1. Note that nonnative datasets are much smaller than the native one, which is a common setting in multi-dialect speech recognition.

In all experiments, we use 80-dimensional log-mel features, computed with a 25ms window and shifted every 10ms. We follow the standard Kaldi recipe s5 [20] for preparing data.



(a) Using utterance summarization only.

(b) Using both dialect information and utterance summarization.

Fig. 2: γ and β parameters for the input of the first layer (left) and the last layer (right) generated based on the corresponding information, plotted using t-SNE in 2D-space. Each color represents each dialect.

4.2. Network Architectures

We evaluate 10 different architectures in our experiments as shown in Table 2. All AMs considered in the experiments have 4 unidirectional LSTM layers, each of which has 640 LSTM units. During training, the mini-batch size is set to 320, and the Adam optimizer [21] is used with the initial learning rate 0.1 and learning rate scheduling is applied.

M1 - M3 are the baseline models. M1 is the dialect-unaware model, which is trained with data from all dialects except Korean English. M2 indicates the dialect-specific models, i.e., seven models trained by fine-tuning M1 on data from corresponding dialects. M3 is the dialect-aware model that takes dialect information as an additional input.

M4 - M10 are AMs conditioned on dialect information (D-Info), utterance summarization (Utt-Sum), or both (Both). As indicated by the second column of Table 2, M4 - M6 apply feature-wise transformations to the input of each layer, and M7 - M10 apply them to the output. Although these two groups take the same information, M7 - M10 generate a smaller number of γ and β parameters than M4 - M6[2], and hence the model sizes are reduced as shown on the third column in the table.

M4 and M7 are conditioned on only dialect information and use 64 hidden units for dialect representation. Similarly, M5 and M8 are conditioned on only utterance summarization and also use 64 hidden units for dialect representation. M6, M9, and M10, conditioned on both information, use 32 hidden units for each representation and then concatenate them together. In all cases, another fully-connected layer with 64 hidden units is located on top of the representation to make the final conditioning representation (the upper boxes in Fig. 1a). M10 is the model using our unseen dialect handling technique

[2] The number of parameters for the input is four times larger than that for the output because an LSTM unit has three gates and one cell, and each one has an input.

Table 2: The performance comparison among 10 different AMs in terms of WERs (%) for Native, Chinese (CHI), German (GER), Spanish (ESP), French (FRN), Italian (ITA), Portuguese (POR), and Korean (KOR) Ensligh. Since KOR is treated as the unseen dialect during training, there are two different overall WERs, -KOR and +KOR, which indicate that KOR is removed from or included in test cases, respectively.

Models			WER									
Description	Cond. Pos.	Size (M)	Nat	CHI	GER	ESP	FRN	ITA	POR	KOR	Overall	
											-KOR	+KOR
[M1] Dialect-unaware	N/A	15.44	5.93	29.40	19.36	21.54	17.96	15.47	15.96	30.30	15.87	17.39
[M2] Dialect-specific		15.44	**5.79**	28.32	18.00	20.36	16.98	14.86	15.16	33.20	15.12	17.24
[M3] Dialect-aware		15.46	5.95	28.51	18.51	20.55	17.33	14.96	15.17	37.70	15.33	17.69
[M4] Cond. on D-Info	Input	16.76	5.87	26.68	17.95	20.06	**16.54**	14.60	14.91	38.14	14.80	17.26
[M5] Cond. on Utt-Sum		16.90	5.96	27.69	18.06	20.46	17.87	15.10	15.38	29.04	15.28	16.73
[M6] Cond. on Both		16.84	5.91	26.31	17.39	19.63	16.63	14.55	**14.27**	36.79	14.58	16.92
[M7] Cond. on D-Info	Output	15.78	6.00	27.78	18.61	20.54	16.93	15.00	15.24	37.01	15.24	17.54
[M8] Cond. on Utt-Sum		15.92	5.92	27.27	17.89	20.13	17.46	14.68	15.30	29.22	15.05	16.54
[M9] Cond. on Both		15.86	5.87	**26.14**	17.12	**19.37**	16.65	**14.07**	14.37	37.96	**14.44**	16.92
[M10] Cond. on Both + Unk-D		15.86	5.80	26.61	**16.76**	19.48	16.92	14.31	14.36	**28.51**	14.51	**15.98**

(Unk-D) described in Section 3.5[3]. With a probability of 0.1, this model treats an utterance as the one coming from the unknown dialect.

4.3. Results

The Word Error Rates (WERs) of different AMs over various dialects and overall WERs are shown on the corresponding columns in Table 2. As we do not use Korean English at training time for the ablation test, the overall WERs are differently calculated according to whether Korean English is removed from or included in test cases.

M1 is our baseline and its WERs for other dialects are much higher than that for the native one, so we can see that the characteristics of the native speech are different from those of the other dialects. The performance of M2 is improved for all dialects except for Korean English, which is handled by the native English model that is more tuned to the native English than M1. Feeding dialect information as input vectors seems effective, as also shown in [2], because M3 outperforms M1 and is competitive with M2.

M4 is the model that applies feature-wise transformations to the input of each layer based on only dialect information. This model outperforms even M2 with a 6.74% relative WER reduction, which indicates that feature-wise transformations are effective.

M5 is the AM adapted by only utterance summarization, which is similar to the model proposed in [4]. Although it underperforms M4, its performance is competitive with M2 or M3. Moreover, its performance for the unseen dialect case is better than that of M1. This means that internal representation plays an important role to make an AM adaptive to multiple dialects. To show how different the γ and β parameters are depending on the dialect, we visualize them by using t-SNE [22] in Fig. 2a. Although no dialect information is given to the model, the parameters tend to be clustered according to the dialect, especially in the higher layers. This shows that conditioning information based on utterance summarization is able to catch dialect-specific aspects well.

M6 is the AM conditioned on both dialect information and utterance summarization. The overall WER is reduced by 3.57% relative compared to M2 when Korean English test cases are not considered. When we compare M6 with the single baseline AM, M3, the overall WER is reduced by 4.89% relative. Fig. 2b shows the visualization of γ and β parameters generated in M6. We can see that they are more strongly clustered by dialects, compared to those from M5. Similar to M5, the parameters in the higher layers seem to be more related to dialect-specific information.

M7 - M9 are equivalent to M4 - M6, respectively, except that these models apply feature-wise transformations to the output of each layer. Both of M7 and M8 show competitive performance with M2 or M3. Unlike M4, using only dialect information does not obtain better performance than the baseline models. However, M9, which uses the combination of dialect information and utterance summarization is superior to all the other AMs in terms of the overall WER. Compared to M2, the WER is reduced from 15.12% to 14.44% (4.50% relative). Moreover, it even outperforms M6 despite its smaller model size, which shows that the output of the layer is a better place to apply conditional transformations.

Finally, M10 is M9 with the unseen dialect handling technique applied. Although M10 performs similarly as M9 for all known dialects, it performs the best for Korean English, which is unseen during training. If Korean English is considered, the overall WER is reduced from 17.24% to 15.98% (7.31% relative), compared to M2. As a single AM for all dialects, the overall WER is reduced by 8.11% relative and 9.67% relative, compared to M1 and M3, respectively.

5. CONCLUSION

In this paper, we propose a highly adaptive AM for accurate multi-dialect speech recognition. Unlike previous work, our AM is conditioned on both dialect information and internal representation. We also propose a simple but effective method for handling unseen dialects. We show through experimental evaluation that our proposed AM is a single model with low maintenance cost that outperforms all the previous AMs for multi-dialect speech recognition.

As future work, we plan to apply our proposed highly adaptive acoustic modeling technique to end-to-end speech recognition. Since end-to-end speech recognition models contain both the acoustic and language models, we expect that each part can benefit from our proposed technique.

6. ACKNOWLEDGEMENTS

YB would like to thank NSERC and CIFAR AI Chairs for funding.

[3] We assume that all models except M10 treat Korean English as the native one due to the reason described in Section 3.5.

7. REFERENCES

[1] Mohamed Elfeky, Meysam Bastani, Xavier Velez, Pedro Moreno, and Austin Waters, "Towards acoustic model unification across dialects," in *Proc. SLT*. IEEE, 2016.

[2] 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 *Proc. ICASSP*. IEEE, 2018.

[3] Mingming Chen, Zhanlei Yang, Jizhong Liang, Yanpeng Li, and Wenju Li, "Improving deep neural networks based multi-accent mandarin speech recognition using i-vectors and accent-specific top layer," in *Proc. INTERSPEECH*. ISCA, 2015.

[4] Taesup Kim, Inchul Song, and Yoshua Bengio, "Dynamic layer normalization for adaptive neural acoustic modeling in speech recognition," in *Proc. INTERSPEECH*. ISCA, 2017.

[5] Xuesong Yang, Kartik Audhkhasi, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, and Mark Hasegawa-Johnson, "Joint modeling of accents and acoustics for multi-accent speech recognition," in *Proc. ICASSP*. IEEE, 2018.

[6] Abhinav Jain, Minali Upreti, and Preethi Jyothi, "Improved accented speech recognition using accent embeddings and multi-task learning," in *Proc. INTERSPEECH*. ISCA, 2018.

[7] Najim Dehak, Patrick J. Kenny, Reda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, August 2011.

[8] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI*. IEEE, 2017.

[9] Harm de Vries, Florian Strub, Jeremie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron Courville, "Modulating early visual processing by language," in *Proc. NIPS*, 2017.

[10] Florian Strub, Mathieu Seurin, Ethan Perez, Harm de Vries, Jérémie Mary, Philippe Preux, Aaron Courville, and Olivier Pietquin, "Visual reasoning with multi-hop feature modulation," in *arXiv preprint arXiv:1808.04446*, 2018.

[11] Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio, "Feature-wise transformations," *Distill*, 2018, https://distill.pub/2018/feature-wise-transformations.

[12] Yan Huang, Dong Yu, Chaojun Lie, and Yifan Gong, "Multi-accent deep neural network acoustic model with accent-specific top layer using the kld-regularized model adaptation," in *Proc. INTERSPEECH*. ISCA, 2014.

[13] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur, "A learned representation for artistic style," in *Proc. ICLR*, 2015.

[14] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton, "Layer normalization," in *CoRR, vol. abs/1607.06450*, 2016.

[15] César Laurent, Gabriel Pereyra, Philémon Brakel, Ying Zhang, and Yoshua Bengio, "Batch normalized recurrent neural networks," in *Proc. ICASSP*. IEEE, 2016.

[16] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Vaino Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, and Zhenyao Zh, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. ICML*, 2016.

[17] Ossama Abdel-Hamid and Hui Jiang, "Fast speaker adaptation of hybrid nn/hmm model for speech recognition based on discriminative learning of speaker code," in *Proc. ICASSP*. IEEE, 2013.

[18] Vassil Panayotov, Cuoguo Chen, Danial Povey, and Sanjeev Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*. IEEE, 2015.

[19] KingLine Data Center, "ASR corpus," in *http://kingline.speechocean.com*.

[20] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2011.

[21] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *arXiv preprint arXiv:1412.6980*, 2014.

[22] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.