INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Age-Invariant Training for End-to-End Child Speech Recognition using Adversarial Multi-Task Learning

Lars Rumberg[1], *Hanna Ehlert*[2], *Ulrike Lüdtke*[2], *Jörn Ostermann*[1]

[1]Institut für Informationsverarbeitung, Leibniz University Hannover, Germany
[2]Institut für Sonderpädagogik, Leibniz University Hannover, Germany
rumberg@tnt.uni-hannover.de

Abstract

Automatic speech recognition for children's speech is a challenging task mainly due to scarcity of publicly available child speech corpora and wide inter- and intra-speaker variability in terms of acoustic and linguistic characteristics of children's speech. We propose a framework for age-invariant training of the acoustic model of end-to-end speech recognition systems based on adversarial multi-task learning. We use age information additionally to just differentiating between the child and adult domains and thus force the acoustic model to learn age invariant features. Our results on publicly available data sets show that this leads to better leveraging of existing data during training. We further show that usage of adversarial multi-task learning should not necessarily be regarded as a substitute for traditional feature space adaptation methods, but that both should be used together for best performance.

Index Terms: speech recognition, child speech, domain adaptation

1. Introduction

Automatic speech recognition (ASR) for adult speech has become highly accurate in recent years. However, its performance for child speech is significantly worse. Since there are many important application areas of ASR of child speech, research to improve upon its current performance is essential. Besides application for human-computer-interaction [1, 2], accurate child ASR is especially desirable for therapeutic applications. Exemplary use cases are automatic classification of disordered speech [3], pronunciation assessment [4, 5, 6] and linguistic analysis. It would greatly facilitate the work of speech language therapists and could therefore give more children access to early diagnosis and intervention.

Two main reasons for ASR still performing significantly worse for child speech than for adult speech can be identified. First, the high inter- and intra-speaker variability in terms of acoustic and linguistic characteristics of children's speech [7, 8] makes it inherently difficult for models to generalize well. Second, much less child speech data to train an ASR system, compared to adult speech data exits [9].

Due to this data scarcity, researchers can not rely on child speech alone. Instead they often additionally use adult speech. The existence of major differences in the characteristics of adult and child speech poses difficulties for the knowledge transfer. To maximise the benefits of using additional adult data, an important research area is the development of techniques to bridge the domain gap between adult and child speech [6, 8, 10, 11, 12, 4]. In addition to the domain gap between adult and child speech a large gap between children of different ages exists [13]. Training with adult and child speech should therefore not be regarded as a two domain problem. Additional performance gains are expected when the differences between speech of children of different ages are considered when designing the system.

The domain gap can be addressed both at the acoustic model as well as at the language model of the ASR pipeline. In this work we focus solely on the acoustic model and do not use a language model.

We propose a framework for age-invariant training of the acoustic model, based on adversarial multi-task learning [14]. While simultaneously training on child and adult speech, we train a discriminator model to estimate the speaker's age given the features of the last layer of the acoustic model. We then define an adversarial loss, which forces the feature extraction of the acoustic model to only extract age-invariant features. Our results show that this is advantageous to domain invariant training, where only child and adult domains are discriminated.

We compare our framework to the more traditional feature space adaptation method from [15] and show that the best results are achieved when adversarial multi-task learning and feature space adaption are combined.

In recent years, hybrid DNN/HMM speech recognition systems have been increasingly replaced by end-to-end systems. In hybrid systems a frame-level forced alignment from a GMM-HMM system is used to train the DNN to estimate likelihoods, used as the HMM state observation likelihoods [16]. An end-to-end system directly transforms the input sequence of acoustic features to an output sequence of tokens [17]. Most end-to-end systems are either based on connectionist temporal classification (CTC) [18], attention-based encoder-decoder networks [19], RNN transducer models [20], or a combination of them. In this work we use a simple CTC-based time delay neural network (TDNN) with letters (graphemes) as target tokens.

2. Related Work

2.1. Child speech recognition

[2] presents a large vocabulary ASR for child speech. Using a large proprietary child speech corpus, they achieve a high recognition accuracy. Since most researchers do not have access to such resources, a large part of recent work into child ASR focuses on how to incorporate adult data into the training. [6, 8] and [10] investigate how to best fine tune models trained on adult speech recognition with child data. Multi-task learning, where child and adult speech recognition is treated as two complementary tasks is explored in [11] and [12].

[21] proposes a data augmentation scheme, simulating vowel prolongation that is typically associated with speech produced by children. [13] shows that child ASR is especially difficult for children in kindergarten age and younger. They also

http://dx.doi.org/10.21437/Interspeech.2021-1241

show that even a few years age difference between children in training and testing data drastically reduces performance.

2.2. Speech domain adaptation

Feature space adaptation methods like vocal tract length normalization (VTLN) [22] and feature space maximum likelihood linear regression (fMLLR) [23] are commonly used for speaker adaptation in hybrid systems and have also been shown to increase performance for children's speech [8]. While these adaptation methods have been used for training of end-to-end models [24], they still require a hybrid model to obtain the adapted features. [15] show that the linear relationship between formants and f_0 can be used for an effective frequency normalization for children's speech. Their results show similar, or in some cases even better performance than VTLN for child ASR. Since their technique only relies on f_0 computation its usage in end-to-end systems is straight forward.

Adversarial multi-task learning [14] has been used for speaker [25] and domain [26] invariant training of hybrid ASR systems. In [27] the speaker invariant training is adapted for end-to-end models.

[4] first applies adversarial learning to domain adaptation for acoustic modeling of children's speech using a hybrid model. Using a child-adult domain discriminator, they learn a front-end feature adaptation network, mapping child speech to adult speech.

In this work we use adversarial multi-task learning for an end-to-end acoustic model for children's speech. Instead of only discriminating between the child and adult domains, we additionally use the age of the speaker. [25, 26, 27] and [4] use a gradient reversal layer (GRL) for adversarial multi-task learning. When using age information, the discriminator task becomes non-binary. Due to this, we investigate whether substituting the GRL by a loss function like the domain confusion loss from [28] is beneficial. We further compare our framework with the f_0-Normalization from [15], and investigate whether usage of adversarial multi-task learning can be a be regarded as a substitute for traditional feature space adaptation methods, or whether both should be used together.

3. Age invariant training

An overview of the proposed age invariant training is given in Figure 1. Given a mini-batch of utterances $X = \{x_1, ..., x_N\}$, where each x_i is the Mel spectrogram of a whole utterance, with corresponding transcripts $Y = \{y_1, ..., y_N\}$ and age labels $A = \{a_1, ..., a_N\}$, we train the encoder network with parameters θ_{enc} and the ASR-decoder network with parameters θ_{asr} using the CTC criterion $\mathcal{L}_{\text{ctc}}$.

To enforce an age invariant output of the encoder network, we use a discriminator network with parameters θ_{age}, which maps the encoder output to a value between 0 and 1. The discriminator gets trained using the cross-entropy loss $\mathcal{L}_{\text{age}}$ between its output and the age label of the utterance. The age label a_i is a value between 0 and 1 and gets computed from the age of the speaker. It can be interpreted as a soft domain label, representing the likeness to adult speech. We set it to zero for the youngest children, 0.8 for the oldest children in the corpus and linear in between. For all adults it is set to 1.

Instead of using a GRL [14] to optimize the encoder network, such that it maximises the discriminator loss $\mathcal{L}_{\text{age}}$, we define an adversarial loss $\mathcal{L}_{\text{adv}}$. This loss is similar to the domain confusion loss in [28], and is the cross entropy between the output $p_{\text{a},i}$ of the age discriminator network and an adversarial target a_{adv}

$$\mathcal{L}_{\text{adv}} = -\frac{1}{N}\sum_{n=1}^{N} a_{\text{adv}} \log(p_{\text{a},i}) + a_{\text{adv}} \log(1 - p_{\text{a},i}). \quad (1)$$

By choosing

$$a_{\text{adv}} = 0.5, \quad (2)$$

the loss is minimized, when the confusion of the discriminator is maximal. We discuss our reasons for choosing this adversarial loss (AL) over usage of a GRL in Section 4.3.2.

Given these losses, the search for the optimal parameters $\hat{\theta}_{enc}$, $\hat{\theta}_{asr}$, and $\hat{\theta}_{age}$ is formulated by

$$\hat{\theta}_{enc}, \hat{\theta}_{asr} = \underset{\theta_{\text{enc}},\theta_{\text{asr}}}{\operatorname{argmin}} \mathcal{L}_{\text{task}}\left(\theta_{\text{enc}}, \theta_{\text{asr}}, \hat{\theta}_{\text{age}}\right), \quad (3)$$

$$\hat{\theta}_{age} = \underset{\theta_{\text{age}}}{\operatorname{argmin}} \mathcal{L}_{\text{age}}\left(\hat{\theta}_{\text{enc}}, \theta_{\text{age}}\right), \quad (4)$$

where the task loss $\mathcal{L}_{\text{task}}$ is the sum of the CTC loss and the adversarial loss, weighted by a hyper-parameter λ

$$\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{ctc}}\left(\theta_{\text{asr}}, \theta_{\text{enc}}\right) + \lambda\mathcal{L}_{\text{adv}}\left(\theta_{\text{age}}, \theta_{\text{enc}}\right). \quad (5)$$

4. Experiments

4.1. Data sets

4.1.1. Adult speech

As adult speech we use the "train-clean-100" subset of the LibriSpeech [29] dataset. LibriSpeech is an English read speech dataset commonly used for large vocabulary continuous speech recognition.

4.1.2. Child speech

The OGI kids' speech corpus [30] contains speech from approximately 1100 unique speakers from kindergarten age to 10th grade. We use only the scripted part of the corpus and utterances shorter than 10 seconds. This results in a total of 47532 utterances. Since the corpus uses only 321 unique prompts, the variety in speakers is much bigger than the variety in words. Different strategies for splitting the corpus into training, development and test sets have been used in prior work. [31] randomly chooses utterances for each set and [13, 15] enforce unique speakers in each set. Instead, we enforce unique prompts and show in Section 4.3.1 that this is a more difficult splitting strategy for this corpus. This decision resulted in a training set of 38010 utterances, and development and test sets of 4697 and 4825 utterances respectively.

4.2. Implementation details

From the audio, we extract Mel Spectrogram features with 64 filter banks. The features are then normalized using the mean and variance of each individual feature channel and utterance. During training we augment the data using SpecAugment [32] frequency and time masking. For each utterance we apply two frequency and time masks, with size of 6 feature channels, respectively time steps.

For f_0-based frequency normalization [15] we extract kaldi-pitch-features [33]. In addition to the pitch of each frame, these provide a probability for whether the frame contains voice or not. We weight the pitch of each frame by this probability of voicing to compute the mean f_0 for each utterance. We warp

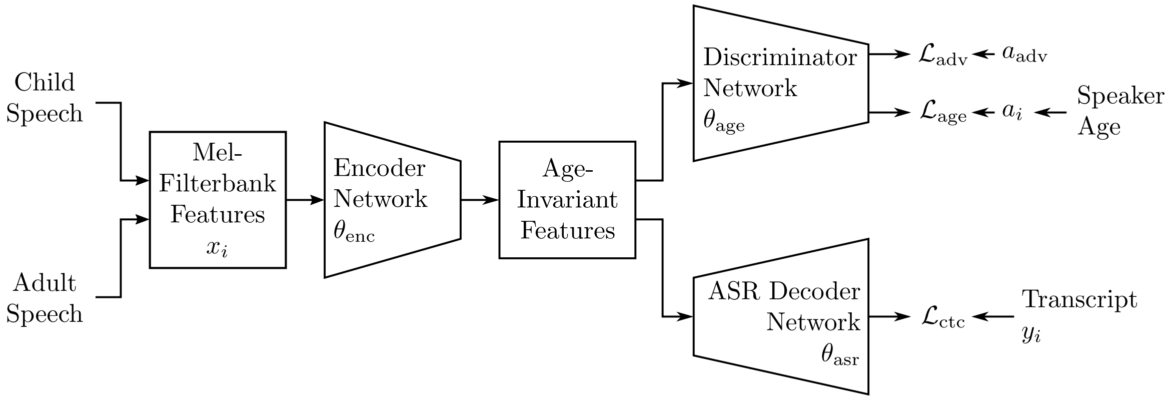
Child Speech
Adult Speech
Mel-Filterbank Features x_i
Encoder Network θ_{enc}
Age-Invariant Features
Discriminator Network θ_{age}
$\mathcal{L}_{\text{adv}}$
a_{adv}
$\mathcal{L}_{\text{age}}$
a_i
Speaker Age
ASR Decoder Network θ_{asr}
$\mathcal{L}_{\text{ctc}}$
Transcript y_i

Figure 1: *Overview of the proposed model for age invariant training. A discriminator network trained to estimate the age of the speaker is used to compute an adversarial loss, which forces the encoder network to extract age invariant features.*

the speech spectrum like [15] before computing Mel-filterbank features.

For the encoder network we use a 10-layer TDNN. Each Layer has a kernel size of 11, a dilation of one and 512 channels. The disciminator network consists of one 1-D convolution with kernel size 11, dilation of one and stride of 3, followed by an average pooling over the time dimension and two fully connected layers with 64 neurons. The output gets mapped to a value between 0 and 1 by a sigmoid. The ASR-decoder network is just one 1-D convolution layer with kernel size 1 followed by softmax. Since we do not use a language model its output gets decoded using greedy CTC decoding. Each layer in all networks, except for the output layers, is followed by a batch normalization [34] and a ReLU non-linearity. We implement all models using PyTorch [35].

As a baseline we train only the encoder and ASR-decoder network without the discriminator network using only the CTC criterion. To avoid overestimating the gains of the adaptation techniques, all hyper-parameters, except the weight of the adversarial loss, are tuned to maximise validation accuracy of the baseline model and are kept the same for all other models.

All models are trained on adult and child data simultaneously with a batch size of 64, with half of the batch child data and the other half adult data, for 50 epochs. Since the adult corpus contains more utterances than the child corpus, not the whole adult corpus is seen each epoch. For the evaluation of the splitting strategies of the OGI kid's corpus (see Section 4.3.1) we train only on the child data for only 25 epochs, to avoid overfitting due to the reduced amount of data. Training is done using the Adam optimizer [36], a 1cycle learning rate scheduler [37] with a maximum learning rate of $5 * 10^{-4}$ and gradient clipping.

The weight of the adversarial loss λ (see (5)) is set to zero at the beginning of the training. At this point, the discriminator has not learned to differentiate between speaker ages and thus, the adversarial loss can not give any meaningful signal. We linearly increase λ from zero at epoch 10 to $\lambda_{\max} = 0.5$ at epoch 40.

4.3. Results and discussion

In the the following section the experimental results are discussed. We report the character error rate (CER) on the test subset of the OGI kids' speech corpus for each model as mean and standard deviation over five training runs with different random seeds for model initialization.

4.3.1. Splitting OGI kids' corpus by prompt

As a preliminary study, we train the baseline model only on child speech. Due to the relatively low variety in words, compared to the variety in speakers of the OGI kid's speech corpus (see Section 4.1.2), we hypothesize that the model's ability to generalize over speakers will be much higher than its ability to generalize to unseen words. This is confirmed by our results, shown in Table 1, when comparing different strategies for splitting the data set into train, development and test subsets. When splitting the data set, such that no speaker occurs in multiple subsets, a low test CER of 7.81 % is achieved. When instead splitting, such that no prompts occur in multiple subsets, the CER increases to a much worse 55.18 %, which confirms our hypothesis. We therefore chose the later splitting strategy for all further experiments.

Table 1: *Test CER (%) on OGI kids' corpus, when training without adult data and using different splitting strategies. Mean and standard deviation are estimated over five training runs with different random seeds.*

Split strategy	**CER**	
	mean	std
Split by speaker	7.81	0.10
Split by prompt	55.18	0.48

4.3.2. Using age information and replacing GRL with adversarial loss

When training the baseline with adult and child speech simultaneously (baseline in Table 2), the CER gets already reduced by relative 10 % compared to training only on child data. We then add the domain discriminator network. We assess the effect of using age information via the age labels instead of hard domain labels for both AL and GRL. When using hard domain labels we set the age label a_i to zero for all children and to 1 for all adults. The model using a GRL and hard labels increases performance by relative 4 % over the baseline, but the GRL model does not significantly benefit from soft age labels. When instead using the AL, performance gets increased by relative 6.5 % over

Table 2: *Test CER $(\%)$ on OGI kids' corpus when using age information or only hard adult-child labels, both when training with the AL or using a GRL. Mean and standard deviation are estimated over five training runs with different random seeds. The significance of the improvement over the baseline, estimated using one-sided Welch's t-test, is shown with* * $p < 0.05$ *and* ** $p < 0.01$

Model	CER		CERR
	mean	std	
Baseline	49.41	1.43	-
GRL Hard Label	47.37	1.01	4.13*
GRL Age	47.14	0.87	4.60**
AL Hard Label	46.17	0.78	6.56**
AL Age	44.34	1.06	10.25**

the baseline. The model trained with the AL also significantly benefits from soft age labels, resulting in a relative improvement of 10 % over the baseline.

We believe the reason for the GRL model not being able to leverage the age information is as follows: Adversarial multi-task learning with a GRL is not designed for continuous domain labels. Given hard domain labels, the sign of the gradient of the discriminator loss is fixed for all samples from one domain. The reversed gradient will always push the encoder network to extract features more similar to the other domain. When instead continuous domain labels exist, the sign of the gradient depends on the discriminator output. When the discriminator underestimates the age of a young child, the reversed gradient will push the encoder to extract features that are more similar to even younger children. The desired behaviour would be to push the encoder to extract features more similar to older children and adults.

The AL has another potential advantage over the usage of a GRL. Since the AL does only depend on the output of the discriminator network and not on a label, the speaker age does not need to be known for all utterances. While this is not explored in this work, it may simplify the usage of additional data in the future.

4.3.3. f_0-Normalization

We compare the adversarial age invariant training with f_0-Normalization from [15]. We chose f_0-Normalization as a comparison over more widespread feature adaptation methods like VTLN, because it has shown similar or better performance for adapting differences between different age groups of the OGI kids' corpus, while being much easier to integrate into an end-to-end system. The results are shown in Table 3.

f_0-Normalization reduces the CER 4 % relative to the baseline, while the age invariant training (AL Age) results in 10 % decreased CER. When using f_0-Normalization together with the age invariant training, the best results are achieved with more than 13 % relative improvement over the baseline. The age variant training appears to benefit from f_0-Normalization to a similar extend as the baseline model. This suggests, that the age invariant training as implemented in this work, adapts other differences between speech from different aged children and adults than f_0-Normalization does.

Table 3: *Test CER $(\%)$ on OGI kids' corpus and relative improvement over the Baseline (CERR) $(\%)$ when training with f_0-Normalization from [15], adversarial loss (AL), and when combining both. Mean and standard deviation are estimated over five training runs with different random seeds. The significance of the improvement over the baseline, estimated using one-sided Welch's t-test, is shown with* * $p < 0.05$ *and* ** $p < 0.01$

Model	CER		CERR
	mean	std	
Baseline	49.41	1.43	-
Baseline + f_0	47.29	2.20	4.28
AL Age	44.34	1.06	10.25**
AL Age + f_0	42.82	0.28	13.34**

5. Conclusions

In this paper we proposed a framework for age invariant training of the acoustic model of end-to-end automatic speech recognition systems to improve joint training on child and adult speech, using adversarial multi-task learning. Instead of treating child and adult speech as only two domains, additional performance gains were achieved, when utilizing information about the speaker's age. Applied to a simple TDNN acoustic model and the OGI kids' corpus, the proposed framework shows promising results of more than 10 % relative improvement over the baseline.

Our results further show, using the f_0-normalization from [15], that usage of adversarial multi-task learning should not necessarily be regarded as a substitute for traditional feature space adaptation methods, but that both should be used together for best performance.

In this work a proof of concept using a simple model and a small data set was given. In further work we want to apply the proposed framework to more sophisticated models and use additional data. Even though in this work the fully supervised scenario was investigated, additional unlabeled data can be used in a straightforward manner.

6. References

[1] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme, "Child speech recognition in human-robot interaction: evaluations and recommendations," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 82–90.

[2] H. Liao, G. Pundak, O. Siohan, M. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," in *Proceedings INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*, 2015.

[3] M. Shahin, U. Zafar, and B. Ahmed, "The automatic detection of speech disorders in children: Challenges, opportunities, and preliminary results," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 400–412, 2020.

[4] R. Duan and N. F. Chen, "Unsupervised feature adaptation using adversarial multi-task training for automatic evaluation of children's speech," in *Proceedings INTERSPEECH 2020 – 21th Annual Conference of the International Speech Communication Association*, 2020, pp. 3037–3041.

[5] S. Dudy, S. Bedrick, M. Asgari, and A. Kain, "Automatic analysis of pronunciations for children with speech sound disorders," *Computer speech & language*, vol. 50, pp. 62–84, 2018.

[6] D. Smith, A. Sneddon, L. Ward, A. Duenser, J. Freyne, D. Silvera-Tawil, and A. Morgan, "Improving child speech disorder assessment by incorporating out-of-domain adult speech," *Proceedings INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association*, pp. 2690–2694, 2017.

[7] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[8] P. Gurunath Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," *Computer Speech & Language*, vol. 63, p. 101077, 2020.

[9] P. B. Ramteke, S. Supanekar, P. Hegde, H. Nelson, V. Aithal, and S. G. Koolagudi, "NITK kids' speech corpus," in *Proceedings INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, 2019, pp. 331–335.

[10] R. Gale, L. Chen, J. Dolata, J. van Santen, and M. Asgari, "Improving ASR systems for children with autism and language impairment using domain-focused DNN transfer techniques," *Proceedings INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, pp. 11–15, 2019.

[11] R. Tong, L. Wang, and B. Ma, "Transfer learning for children's speech recognition," in *2017 International Conference on Asian Language Processing (IALP)*. IEEE, 2017, pp. 36–39.

[12] J. Wang, S. I. Ng, D. Tao, W. Y. Ng, and T. Lee, "A study on acoustic modeling for child speech based on multi-task learning," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 389–393.

[13] G. Yeung and A. Alwan, "On the difficulties of automatic speech recognition for kindergarten-aged children," in *Proceedings INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association*, 2018, pp. 1661–1665.

[14] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[15] G. Yeung and A. Alwan, "A frequency normalization technique for kindergarten speech recognition inspired by the role of fo in vowel perception," in *Proceedings INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, 2019, pp. 6–10.

[16] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *2013 IEEE workshop on automatic speech recognition and understanding*. IEEE, 2013, pp. 273–278.

[17] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*. PMLR, 2014, pp. 1764–1772.

[18] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[19] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, 2015, pp. 577–585.

[20] A. Graves, "Sequence transduction with recurrent neural networks," in *ICML Representation Learning Workshop*, 2012.

[21] T. Nagano, T. Fukuda, M. Suzuki, and G. Kurata, "Data augmentation based on vowel stretch for improving children's speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 502–508.

[22] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1, 1996, pp. 346–348 vol. 1.

[23] M. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75–98, 1998.

[24] N. Tomashenko and Y. Estève, "Evaluation of feature-space speaker adaptation for end-to-end acoustic models," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[25] Z. Meng, J. Li, Z. Chen, Y. Zhao, V. Mazalov, Y. Gong, and B. Juang, "Speaker-invariant training via adversarial learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5969–5973.

[26] P. Denisov, N. T. Vu, and M. F. Font, "Unsupervised domain adaptation by adversarial learning for robust speech recognition," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.

[27] Z. Meng, Y. Gaur, J. Li, and Y. Gong, "Speaker adaptation for attention-based end-to-end speech recognition," in *Proceedings INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, 2019, pp. 241–245.

[28] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[30] K. Shobaki, J.-P. Hosom, and R. A. Cole, "The ogi kids' speech corpus and recognizers," in *Sixth International Conference on Spoken Language Processing*, 2000.

[31] F. Wu, L. P. García-Perera, D. Povey, and S. Khudanpur, "Advances in automatic speech recognition for child speech using factored time delay neural network," in *Proceedings INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, 2019, pp. 1–5.

[32] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proceedings INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, 2019, pp. 2613–2617.

[33] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 2494–2498.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 448–456.

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[36] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic gradient descent," in *3rd International Conference for Learning Representations*, 2015, pp. 1–15.

[37] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006. International Society for Optics and Photonics, 2019, p. 1100612.