INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Unsupervised Feature Adaptation using Adversarial Multi-Task Training for Automatic Evaluation of Children's Speech

Richeng Duan, Nancy F. Chen

Institute for Infocomm Research, A*STAR, Singapore
duan_richeng@i2r.a-star.edu.sg, nfychen@i2r.a-star.edu.sg

Abstract

Processing children's speech is challenging due to high speaker variability arising from vocal tract size and scarce amounts of publicly available linguistic resources. In this work, we tackle such challenges by proposing an unsupervised feature adaptation approach based on adversarial multi-task training in a neural framework. A front-end feature transformation module is positioned prior to an acoustic model trained on adult speech (1) to leverage on the readily available linguistic resources on adult speech or existing models, and (2) to reduce the acoustic mismatch between child and adult speech. Experimental results demonstrate that our proposed approach consistently outperforms established baselines trained on adult speech across a variety of tasks ranging from speech recognition to pronunciation assessment and fluency score prediction.

Index Terms: computer assisted language learning, computer assisted pronunciation training

1. Introduction

Globalization has driven the demand for many of us (adult and child) to learn more than one language. In Finland, all children are required to learn at least two foreign languages, while there are four official languages in Singapore. On the other hand, there is dire need of human resources for performing spoken language assessment. Therefore, developing methods for speech evaluation by leveraging automatic speech recognition (ASR) technology has been an established research area [1, 2, 3, 4, 5, 6]. Fig.1 shows the framework of a standard automatic evaluation system [7], which includes an acoustic model to extract features from the learner's speech for training a classifier for predicting assessment scores. Acoustic modeling, therefore, plays a critical role in the entire pipeline system. Despite devoted research efforts in acoustic modeling for several decades, it is still challenging in the context of spoken language education.

Acoustic modeling challenges of children's speech stem from the physiological differences of their articulatory apparatus size and the high heterogeneity of vocal effort, speaking rate, and proficiency levels within children [8, 9, 10, 11, 12, 13]. While large-scale linguistic resources can enable high performance acoustic modeling [14], such privileged scenarios is often unavailable [15]. To overcome the paucity of transcribed linguistic resources for children's speech, we propose to learn an unsupervised feature adaptation model to transform child speech to the adult feature space. This adaptation module could then be placed prior to an acoustic model trained on large amounts of adult speech. This child-to-adult feature adaptation model is trained through adversarial multi-task training, where the objective function is to minimize the error rate of a fixed back-end adult acoustic model (i.e. senone classifier) while maximizing the error rate of a child-adult classifier. There has been work on learning domain invariant feature representations [16, 17, 18, 19], but few, if any, has done so at the front-end feature prepossessing stage. We show that the proposed feature adaptation model is effective in applications such as speech recognition and evaluation of children's speech.

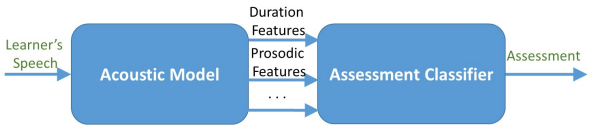

Figure 1: *Automatic spoken language assessment system.*

2. Related work

2.1. Automatic spoken language assessment

In the field of automatic spoken language assessment, past research efforts mainly focus on investigating better acoustic models [20, 21] or improving the assessment classifier [7, 22]. In this work, we investigate acoustic modeling approaches to generate more robust features using unsupervised adaptation for the subsequent assessment classifier.

2.2. Acoustic modeling of children's speech

Past approaches that adopted statistical machine learning based acoustic models (e.g. Gaussian mixture models) use vocal tract length normalization (VTLN) and maximum linear likelihood transform (MLLT) [13, 23] to account for speaker variability. Recent deep learning based acoustic models rely on large corpora (thousands of hours of child speech) to fuel its computational power [14]. With limited amounts of transcribed child speech, an alternative approach proposed in [15] is conducted by freezing lower layers of the pre-trained DNN while only the output layer is updated. However, these methods usually require at least some quantity of human transcribed child speech for supervised model training or fine-tuning, while in reality such annotated resources are publicly unavailable to most academic researchers. In addition, leveraging transcribed data from new domains to sequentially retrain pre-trained DNN models usually suffers from model forgetting previously learned information [24, 25]. Such data sparsity and catastrophic forgetting motivated us to investigate unsupervised multi-task training approaches in this work, where we propose a front-end neural feature adapter to learn a non-linear transformation to convert the feature space from child speech to adult speech without using transcribed child speech.

Copyright © 2020 ISCA

http://dx.doi.org/10.21437/Interspeech.2020-1657

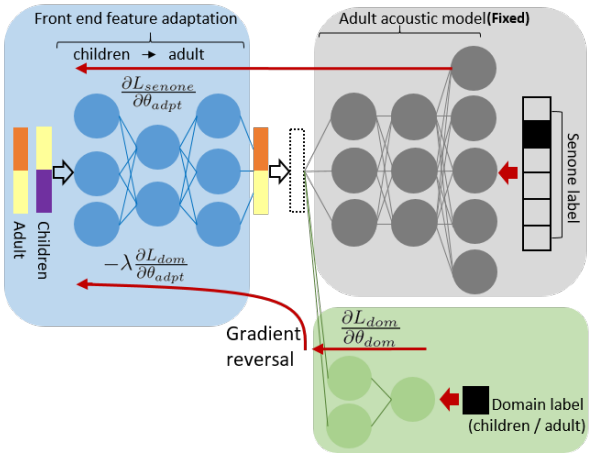

Figure 2: *Front-end child to adult speech adaptation with adversarial multi-task training. The parameters in the gray part is fixed during model training.*

2.3. Adversarial training

Adversarial training, especially adversarial multi-task training, has been proposed for domain invariant feature representations and applied in various application areas such as image classification [16], robust speech recognition [17], speaker adaptation [18], and spoken keyword spotting [19]. However, few have applied such approaches to front-end speech processing. Instead of learning a domain-invariant feature representation for different speaking populations, we explicitly learn a neural feature transformation that maps child speech to adult speech.

2.4. Front-end feature processing

Feature space adaptation is a widely-adopted approach used in front-end speech processing tasks such as speaker adaptation or speech enhancement, where the goal is to either map the features of target speaker(s) to the back-end acoustic model [26, 27] or learn a feature transformation that converts (noise) corrupted speech to clean speech [28, 29] based on transcribed training samples. Most of such approaches are supervised. By contrast, in this work, we investigate unsupervised approaches to learn a non-linear transformation to map untranscribed child speech to the adult feature space.

3. Unsupervised feature space adaptation using adversarial multi-task training

In this section, we delineate the proposed approach which trains an acoustic model without using transcribed children's speech. Fig. 2 illustrates the proposed neural architecture, which is an unsupervised model training framework based on adversarial multi-task training. There are three sub-networks in total: A standard DNN acoustic model is trained using transcribed adult speech and its corresponding senone labels. A front-end feature adaptation DNN is attached to the input layer of the aforementioned adult acoustic model. A discriminator network, which is used to map the front-end acoustic features to a binary label (adult or child), is connected to the output layer of the front-end feature adaptation DNN.

To ensure the output of the front-end feature adapter is transformed to the same feature space as the adult features used to train the original adult acoustic model, during training we minimize the errors from the senone classifier while maximizing the loss of the child-adult classifier. The latter is to ensure that after the transformation of the front-end feature adapter, the differences between adult and child speech features are as minimal as possible.

Assuming that there are N training samples in total and n samples of them belong to adult speech, the objective function is computed as:

$$E(\Theta_{adpt}, \Theta_{dom}) = \frac{1}{n}\sum_{i=1}^{n} L^i_{senone}(\Theta_{adpt}) - \frac{1}{N}\sum_{i=1}^{N} L^i_{dom}(\Theta_{adpt}, \Theta_{dom}) \quad (1)$$

where L_{senone} is the cross-entropy loss between the ground truth senone labels of adult speech and the predicted posteriors. L_{dom} is the child-adult domain classification loss, which is applied to both adult and child speech samples.

During model training, the parameters in the adult acoustic model are fixed while only the parameters in the front-end feature adapter (Θ_{adpt}) and domain classifier (Θ_{dom}) are optimized such that:

$$\Theta_{adpt} = \underset{\Theta_{adpt}}{\operatorname{argmin}}\, E(\Theta_{adpt}, \Theta_{dom}) \quad (2)$$

$$\Theta_{dom} = \underset{\Theta_{dom}}{\operatorname{argmax}}\, E(\Theta_{adpt}, \Theta_{dom}) \quad (3)$$

Following the implementation in [16], a non-parametric gradient reversal layer is inserted between the feature adapter and the domain classifier, which is used to change the sign of the gradient from the subsequent layer during backward propagation. The hyper-parameter λ is added to balance the two gradients during the model training process. All parameters can thus be updated as follows via stochastic gradient descent:

$$\theta_{adpt} \leftarrow \theta_{adpt} - \frac{\partial L^i_{senone}}{\partial \theta_{adpt}} + \lambda \frac{\partial L^i_{dom}}{\partial \theta_{adpt}} \quad (4)$$

$$\theta_{dom} \leftarrow \theta_{dom} - \frac{\partial L^i_{dom}}{\partial \theta_{dom}} \quad (5)$$

Via joint optimization, the adapted features generated from child speech are not only mapped to become closer to the adult speech features, but they are also optimized to ensure minimal senone classification error rate performance on the adult acoustic model.

4. Experimental setup

4.1. Data sets

4.1.1. Speech corpora for acoustic modeling

Adult speech: We used LibriSpeech [30], a widely adopted dataset for English large-vocabulary continuous speech recognition. We selected 100 hours from the "train-clean-100" subset (251 speakers) to train the adult acoustic model.

Child speech: We used *SingaKids-English*, consisting of 46 hours from 193 speakers (90 male, 103 female) ranging from ages 6 to 12. The training, developmental, and test sets were split into 40 hours (175 speakers), 2 hours (6 speakers), and 4 hours (12 speakers), respectively. We merged the six grades in primary school into three groups: G12 (grade 1

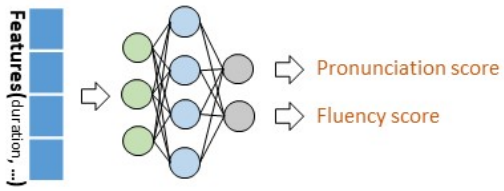

Figure 3: *Multi-task neural network for pronunciation and fluency scoring.*

and 2), G34 (grade 3 and grade 4), and G56 (grade 5 and 6). The age range for the developmental (2 speakers from each group) and test sets (4 speakers from each group) were equally distributed.

4.1.2. Score dataset for assessment classifier

Pronunciation and fluency scores were obtained for a subset of 1,547 utterances from an English school teacher certified by the Ministry of Education in Singapore. The scores come in 5 levels, where level 1 is the worst, and level 5 is the best. Detailed descriptions for each individual level are not shown due to space constraints.

4.2. Acoustic model

The acoustic feature consists of 40-dimensional log Mel-scale filter-bank outputs plus first and second temporal derivatives. The input to the network are 11 contiguous frames, 5 frames on each side of the current frame. The back-end neural network for acoustic modeling of adult speech has 6 hidden layers with 2048 nodes per layer. The front-end feature adaptation network consists of 6 layers with 2048, 2048, 512, 2048, 2048 and 1320 nodes in each layer. To prevent over-fitting, batch normalization [31] and dropout [32] were used. Hyper-parameters were optimized on the developmental data sets of "Dev-clean" of Librispeech and the developmental set SingaKids-Enlgish. The DNN model was implemented using PyTorch [33]. For decoding, we used Kaldi's WFST decoder [34].

4.3. Assessment model

For pronunciation evaluation, DNN-based goodness of pronunciation (GOP) score [20] is adopted and computed as:

$$GOP(p) \approx log \frac{P(p|\mathbf{o}; t_s, t_e)}{max_{\{q \in Q\}} P(q|\mathbf{o}; t_s, t_e)} \quad (6)$$

where t_s and t_e are the start and end frame indices of acoustic features $\mathbf{o}$. Q represents the entire phone set. It is approximated by the log posterior ratio between the target canonical phone p and the competing hypothesis phone q, which has the highest posterior probability. The log posterior of p given the observation o is computed as:

$$logP(p|\mathbf{o}; t_s, t_e) \approx \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} log \sum_{s \in p} P(s|o_t), \quad (7)$$

where o_t is acoustic observations of frame t, s is the senone label belonging to the phone p.

The duration of phones and pauses, which are derived from the transcription that is time aligned to its corresponding audio, are used as features to characterize the speaking fluency.

Table 1: *WER (%) on LibriSpeech*

	Kaldi DNN[1]	Baseline
Dev-clean	9.19	9.32
Test-clean	9.66	9.49

Table 2: *PER (%) on Child test data set. Numbers in the parentheses are relative improvements (%).*

	Baseline	Baseline + adaptation	
		FMLLR	Proposed
G12	85.11	83.73	75.89 (10.83)
G34	73.79	73.24	67.26 (8.85)
G56	69.08	65.78	62.31 (9.80)
Overall	74.43	72.55	67.19 (9.73)

Though pronunciation and fluency are different aspects of spoken language capabilities, a language learner's pronunciation and fluency performance are often correlated in our pilot studies. This observation motivates us to implement a classifier using multi-task neural network in Fig. 3 for both pronunciation and fluency scoring. The input to the classifier network consists of GOP based pronunciation features as well as duration based fluency features. The classifier consists of 3 hidden layers (128 nodes per layer) and 2 softmax output layers with 5 nodes in each branch.

5. Experimental results

5.1. Speech recognition

To validate our baseline model (DNN model trained on adult speech), we used the pruned version of the standard Wall Street Journal (WSJ) 5k trigram language model that is distributed with the corpus, and compared the results with Kaldi's benchmark results on Librispeech "test-clean" set. Results are in Table 1: our baseline model achieved a slightly lower word error rate than that of Kaldi's DNN model, suggesting the trained model is a competitive baseline.

We then compared the proposed approach with the baseline and the adaptation method of lightly supervised feature space maximum likelihood linear regression (FMLLR) on the child test set[2]. To obtain FMLLR transformations, time-stamped phonetic transcriptions were first obtained through passing child speech through the aforementioned baseline speech recognizer by deactivating the language model and using free phone loop decoding instead. The transformation matrix was then trained to maximize likelihood at the feature level given those generated phonetic labels. Note that this paper focuses on investigating unsupervised approaches without using annotated data. Dedicated approaches for data selection or post-processing generated labels using various confidence scores for FMLLR are therefore beyond the scope of this paper.

From the baseline column in Table 2, we first see the phone error rate (PER) reached 85.11% for the child speech of G12,

[1]https://github.com/kaldi-asr/kaldi/blob/master/egs/librispeech/s5/RESULTS

[2]FMLLR is also known as constrained maximum likelihood linear regression (CMLLR) [35].

though the WER is less than 10% on the adult speech using the Librispeech corpus. This observation aligns with our understanding that the acoustic variability of children's speech is high and their pronunciation is quite different from that of adult speech [8, 9, 10, 11, 13]. As the level increased from G12 to G56, the PER decreased significantly from 85.11% to 69.08%, which matches our intuition that speech of younger children presents a larger acoustic mismatch from that of adults. Two adaptation methods of lightly supervised FMLLR and the proposed unsupervised feature adaptation achieved better performance, where PER results are 72.55% and 67.19% respectively, which are lower than the baseline PER of 74.43%. Compared to lightly supervised FMLLR, our proposed method achieved lower PER in all three conditions of G12, G34 and G56, demonstrating that proposed approach more significantly reduces the mismatch between the trained and test conditions. Compared to the baseline, PER of the proposed method consistently decreased for at least 8.85% relative for all conditions, though the absolute PER is still more than 60%. Such results are consistent with recent literature (PER=56.11% reported in [36] when only adult speech are adopted for model training and WER is 53.2% on processing Italian children's speech [15]).

5.2. Speech evaluation

To generate more robust pronunciation features, we averaged phone-level GOP scores to obtain word-level GOP scores. The three highest word-level GOP scores, the three lowest word-level GOP scores, and the average word-level GOP score of a sentence were used to represent pronunciation quality. Similarly, for fluency features, the mean duration of words, the time duration of the three longest and the three shortest silence regions were adopted.

To complement classification accuracy, we also adopted mean squared error (MSE) between the predicted score level and the human rated score level, as there is a ranked relationship among the class categories. In terms of classification accuracy, there is no difference between misclassifying score level 1 and score level 5 vs. misclassifying score level 3 and score level 4, but MSE will take into account that the former is a much more serious error than the latter.

Table 3 and Table 4 show that the proposed approach improves the prediction accuracy on both pronunciation and fluency evaluation. The average relative improvement over the baseline is up to 9.6% in pronunciation prediction while it is 6.2% when predicting the fluency score, which indicates our approach generates more accurate features for the scoring classifier. The classification accuracy is around 50%, twice the probability at chance level, which means half of the machine scores are precisely equal to the human scores. For the proposed approach, when the system prediction is wrong, the prediction MSE between the system score and the teacher reference is fewer than 2 scales in all conditions. The overall MSE is reduced up to 10.7% relative over the baseline when predicting fluency scores. Therefore, our approach is capable of improving the prediction accuracy as well as the reliability of generating more acceptable scoring errors if the system makes a mistake.

Note that since fluency scoring has to be done at the utterance level (or longer), the available data for training the assessment classifier is much less (at least 500 times less) than that for acoustic modeling, which limits the classification performance. Approaches to relieve the need of scores from teachers (a labor-intensive and time consuming process) is a topic of future research.

Table 3: *Pronunciation prediction performance comparison. Numbers in the parentheses are relative improvements (%).*

	Baseline	Proposed
	Classification Accuracy(%)	
G12	42.1	47.3 (12.3)
G34	37.3	38.9 (4.2)
G56	47.3	52.3 (11.3)
Overall	43.3	47.5 (9.6)
	MSE	
G12	1.30	1.10 (15.2)
G34	1.14	1.14 (0.0)
G56	1.44	1.32 (8.2)
Overall	1.32	1.25 (5.1)

Table 4: *Fluency prediction performance comparison. Numbers in the parentheses are relative improvements (%).*

	Baseline	Proposed
	Classification Accuracy(%)	
G12	31.6	42.1 (33.3)
G34	40.3	38.8 (-3.7)
G56	50.9	53.6 (5.3)
Overall	44.2	47.0 (6.2)
	MSE	
G12	2.25	1.99 (11.6)
G34	1.96	1.80 (8.4)
G56	2.22	1.96 (11.7)
Overall	2.13	1.90 (10.7)

6. Conclusions

To train a spoken language assessment system for children, acoustic modeling of children's speech is important for extracting accurate model features. Acoustic and linguistic variability of children's speech is high in variance, where such variations and heterogeneity cast technical challenges on acoustic modeling. The linguistic resources suitable for children's speech assessment is also scarce. Considering such data sparsity challenges, in this paper, we proposed an unsupervised acoustic model adaptation framework to transform children's speech feature space to that of adults' through adversarial multi-task training. Experimental results show that the proposed approach is able to model the children's speech better and achieves lower speech recognition error rate. For pronunciation and fluency evaluation, we conducted sentence level assessments based on features derived from the acoustic models. Empirical results demonstrate that the proposed method not only improves proficiency prediction accuracy but also reduces the difference between the system generated scores and human ratings.

Feed-forward neural networks were adopted in this work to seed capabilities in developing unsupervised adversarial training for speech adaptation. Future endeavors using more sophisticated neural models with attention mechanisms are in the pipeline. Other languages such as Malay, Mandarin Chinese and Tamil are also part of our on-going efforts.

7. References

[1] C. Cucchiarini, H. Strik, and L. Boves, "Automatic evaluation of dutch pronunciation by using speech recognition technology," in *IEEE workshop on automatic speech recognition and understanding proceedings*, 1997, pp. 622–629.

[2] H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, J. Butzberger, R. Rossier, and F. Cesari, "The SRI EduSpeakTM system: Recognition and pronunciation scoring for language learning," *Proceedings of InSTILL*, pp. 123–128, 2000.

[3] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[4] A. Metallinou and J. Cheng, "Using deep neural networks to improve proficiency assessment for children english language learners," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[5] Y. Qian, X. Wang, K. Evanini, and D. Suendermann-Oeft, "Self-adaptive dnn for improving spoken language proficiency assessment." in *INTERSPEECH*, 2016, pp. 3122–3126.

[6] Y. Lu, M. Gales, K. Knill, P. Manakul, and Y. Wang, "Disfluency detection for spoken learner english," in *8th ISCA Workshop on Speech and Language Technology in Education*, 2019, pp. 74–78.

[7] Y. Wang, G. Mark, K. Kate, M. Andrey, R. van Dalen, and Rashid, "Towards automatic assessment of spontaneous spoken english," vol. 67, 2018, pp. 47–56.

[8] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, no. 10-11, pp. 847–860, 2007.

[9] M. Russell and S. D'Arcy, "Challenges for computer recognition of children's speech," in *Workshop on Speech and Language Technology in Education*, 2007.

[10] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on speech and audio processing*, vol. 11, no. 6, pp. 603–616, 2003.

[11] S. S. Gray, D. Willett, J. Lu, J. Pinto, P. Maergner, and N. Bodenstab, "Child automatic speech recognition for US English: child interaction with living-room-electronic-devices." in *WOCCI*, 2014, pp. 21–26.

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

[13] P. G. Shivakumar, A. Potamianos, S. Lee, and S. Narayanan, "Improving speech recognition for children using acoustic adaptation and pronunciation modeling." in *WOCCI*, 2014, pp. 15–19.

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

[15] M. Matassoni, R. Gretter, D. Falavigna, and D. Giuliani, "Non-native children speech recognition through transfer learning," in *ICASSP*. IEEE, 2018, pp. 6229–6233.

[16] 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.

[17] Y. Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition." in *INTERSPEECH*. San Francisco, CA, USA, 2016, pp. 2369–2372.

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

[19] J. Hou, P. Guo, S. Sun, F. K. Soong, W. Hu, and L. Xie, "Domain adversarial training for improving keyword spotting performance of esl speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8122–8126.

[20] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.

[21] Y. Qian, X. Wang, K. Evanini, and D. Suendermann-Oeft, "Self-adaptive dnn for improving spoken language proficiency assessment," in *INTERSPEECH*, 2016.

[22] Z. Yu, X. Wang, K. Zechner, L. Chen, J. Tao, A. Ivanou, and Y. Qian, "Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech," in *ASRU*. IEEE, 2015, pp. 338–345.

[23] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to dnn-based children's and adults' speech recognition," *IEEE SLT*, pp. 135–140, 2014.

[24] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[25] S.-W. Lee, J.-H. Kim, J. Jun, J.-W. Ha, and B.-T. Zhang, "Overcoming catastrophic forgetting by incremental moment matching," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, pp. 4652–4662.

[26] 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.

[27] Y. Miao, H. Zhang, and F. Metze, "Speaker adaptive training of deep neural network acoustic models using i-vectors," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 11, pp. 1938–1949, 2015.

[28] T. Gao, J. Du, and C.-H. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *ICASSP*. IEEE, 2015, pp. 4375–4379.

[29] M. Mimura, S. Sakai, and T. Kawahara, "Joint optimization of denoising autoencoder and dnn acoustic model based on multi-target learning for noisy speech recognition." in *INTERSPEECH*, 2016, pp. 3803–3807.

[30] 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)*. IEEE, 2015, pp. 5206–5210.

[31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[33] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[34] 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. CONF. IEEE Signal Processing Society, 2011.

[35] M. J. Gales and P. C. Woodland, "Mean and variance adaptation within the mllr framework," *Computer Speech & Language*, vol. 10, no. 4, pp. 249–264, 1996.

[36] R. Serizel and D. Giuliani, "Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children," *Natural Language Engineering*, vol. 23, no. 3, pp. 325–350, 2017.