INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Joint Learning Feature and Model Adaptation for Unsupervised Acoustic Modelling of Child Speech

Richeng Duan

Institute for Infocomm Research (I2R), A*STAR, Singapore

Duan_Richeng@i2r.a-star.edu.sg

Abstract

Due to the high acoustic variability of child speech and the lack of publicly available datasets, acoustic modeling for child speech is challenging. In this work, we address these challenges by leveraging the large amounts of resources for adult speech (well-trained acoustic models and transcribed speech dataset) and proposing a joint acoustic feature and model adaptation framework to minimize acoustic mismatch between adult and child speech. Empirical results on three tasks of speech recognition, pronunciation assessment, and fluency assessment show that our proposed approach consistently outperforms competitive baselines, achieving up to 31.18% phone error reduction on speech recognition and around 7% gains on speech evaluation tasks.

Index Terms: unsupervised acoustic modelling, feature and model adaptation, speech evaluation

1. Introduction

Thanks to the availability of vast amounts of linguistic resources and deep learning models, automatic speech recognition (ASR) technology has achieved great success in recent years. However, it is still challenging when processing child speech (e.g. Google's voice-activated application [1] and automatic speech evaluation tasks for child speech [2, 3, 4, 5, 6]). Physiological differences, proficiency differences, as well as different speaking habits all cast technical challenges on acoustic modelling for child speech [7, 8, 9, 10, 11, 12]. Though large-scale annotated resources enable to develop superior acoustic models, this privileged scenario is often unavailable when applied to child speech [13, 14].

To overcome such challenges, our previous work proposed to perform unsupervised feature adaptation that transforms child acoustic features to the adult acoustic feature space based on adversarial multi-task learning. We strategically incorporated phonemic information during the model training process to learn the fine-grained transformation adaptation. Although exploiting phonemic information can help anchor more targeted transformations, the demerit is that the information is represented by pseudo phonetic labels that are generated by an adult acoustic model. The acoustic feature alignment relies on the quality of pseudo labels and incorrect pseudo labels could result in wrong feature adaptation. In this work, we introduce model adaptation in our previous feature adaptation framework to alleviate the harm from noisy pseudo labels and further reduce the acoustic domain mismatch. We propose a novel bidirectional learning algorithm to jointly learn feature and model adaptation, in which two adaptation modules are trained alternatingly to promote each other through providing more accurate information to the other. We validate our method on three tasks and experiment results demonstrate that the proposed approach outperforms established baselines by a large margin.

2. Related work

2.1. Acoustic modeling for child speech

Training speaker dependent models with large amounts of labeled data [1] is the most straightforward way to reach good performance. Using a small amount of labeled data, it is also possible to train a good speaker dependent model by tailoring it to child. In [13], it freezes lower layers of the pre-trained model while only updating the output layer with a few transcribed child speech samples to better fit the characteristics of child speakers. Rather than training speaker dependent models, adaptation techniques such as feature space maximum likelihood linear regression (FMLLR [15]) can be applied. However, most of these methods require at least some amount of manually transcribed resources for supervised training, fine-tuning, or adaptation, which are often time consuming and in reality difficult to obtain because of privacy issues. Furthermore, sequentially retraining pre-trained DNN models with labeled data from new domains usually suffers from catastrophic forgetting [16, 17]. Such challenges motivate us to explore unsupervised acoustic adaptation approaches.

2.2. Adversarial learning for domain adaptation

Adversarial learning, which is inspired by generative adversarial networks (GAN)[18], has become popular in recent years. It has been investigated for learning domain-invariant models in areas such as image processing [19, 20], text processing [21, 22], and speech processing [23, 24]. Instead of training domain-invariant models from scratch, we applied adversarial training to explicitly learn global child-to-adult adaptation transformations at the feature level [25]. Although it achieved favorable performance, the quality of the model-generated soft labels could cast limitations in achieving robust feature alignment. To address this problem, we introduce model adaptation and bidirectional adaptation learning algorithm in this work, where both feature adaptation and model adaption can improve each other in a closed training loop.

3. Adversarial multi-task training based acoustic feature and model adaptation

Figure 1 shows our proposed model architecture for acoustic modeling of child speech. The output layer of the feature adapter (parametrized by Θ_{f_adpt}) is connected to the input layer of the adult acoustic model. The parameters of the adult acoustic model are copied from a well-trained adult model.

10.21437/Interspeech.2023-1302

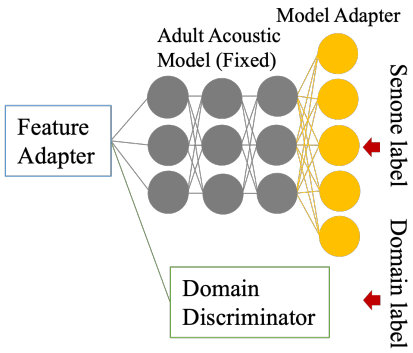

Figure 1: *Overview of our proposed framework.*

We replace its last layer with a trainable lightweight model adapter (parametrized by Θ_{m_adpt}) that performs model adaptation while the layers used for learning feature representation are frozen during model training. The domain discriminator (parametrized by Θ_{dom}) is attached to the output layer of the front-end feature adapter for mapping the transformed features to domain labels. During inference, the domain discriminator is removed and the senone labels are output from the model adapter.

3.1. Acoustic feature adaptation

The feature adapter reduces the acoustic space shift by maximizing the domain discriminator loss while minimizing the senone classification loss. Assuming that the model is trained with N speech samples, of which n samples are adult speakers, the loss function is defined as:

$$\begin{aligned} L(\Theta_{f\_adpt}, \Theta_{dom}) =& L_{senone} - L_{dom} \\ =& \frac{1}{n}\sum_{i=1}^{n} L^i_{senone}(\Theta_{f\_adpt}) \\ &- \frac{1}{N}\sum_{i=1}^{N} L^i_{dom}(\Theta_{f\_adpt}, \Theta_{dom}) \end{aligned} \tag{1}$$

where L_{senone} is the cross-entropy loss for the senone classification on adult speech samples only while the domain classification loss L_{dom} is calculated on all speech samples. For each sample i, it is computed as follows:

$$\begin{aligned} L^i_{dom} =& -(1 - I_{dom}) \sum_{k=1}^{K} \alpha^i_k \log P\left(dom = a, sen = k \mid x_i\right) \\ &- I_{dom} \sum_{k=1}^{K} \alpha^i_k \log P\left(dom = c, sen = k \mid x_i\right) \end{aligned} \tag{2}$$

where I_{dom} is the domain indicator function. It is equal to 0 for adult training samples and 1 for child training samples. K is the number of senone categories and α^i_k is the kth entry of senone posteriors for speech sample i, which is extracted from the adult acoustic model. Considering inaccurate domain labels could lead to learn wrong feature alignment, instead of using one hot hard labels, we employ soft domain labels for model training. The soft domain labels consist of two parts: broad binary (child/adult) domain information (hard label) and the fine-grained senone information (probability label).

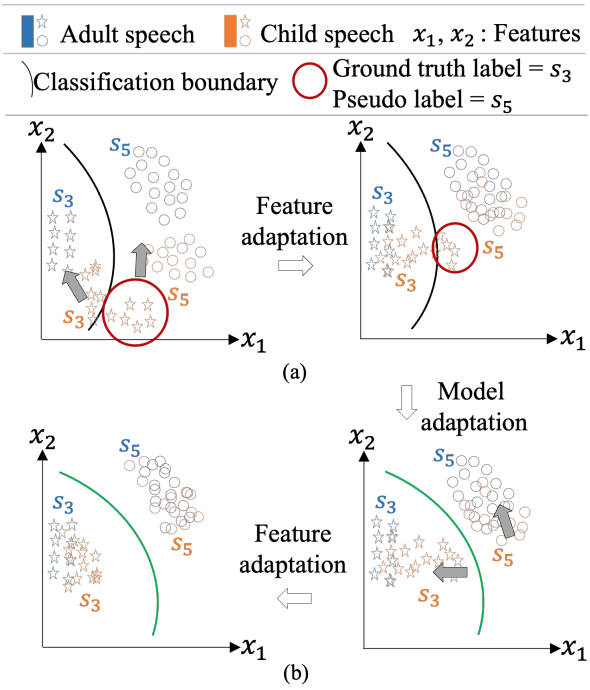

Figure 2: *(a) Acoustic feature adaptation (b) Joint learning feature and model adaptation. Adapted acoustic model can generate more accurate pseudo senone labels*

They are thus represented by a $2K$ dimension vector, [0;α] for child and [α;0] for adult. $P(dom = a, sen = k \mid x_i)$ and $P(dom = c, sen = k \mid x_i)$ are the kth entry of probability outputs from the domain discriminator for adult and child speech samples respectively.

To ensure the features generated by the feature adapter are able to perform senone classification and the adapted child acoustic features are closer to the adult speech features, Θ_{f_adpt} and Θ_{dom} are optimized such that:

$$\Theta_{f\_adpt} = \underset{\Theta_{f\_adpt}}{\operatorname{argmin}}\, L(\Theta_{f\_adpt}, \Theta_{dom}) \tag{3}$$

$$\Theta_{dom} = \underset{\Theta_{dom}}{\operatorname{argmax}}\, L(\Theta_{f\_adpt}, \Theta_{dom}) \tag{4}$$

Though we can accurately get the binary domain information according to the data sample itself, the senone information for child speech samples, however, can only be obtained by the adult acoustic model. Adopting inaccurate senone information may cause child speech samples with acoustic feature distributions that match the adult ones well but are somehow associated with wrong labels after the feature adaptation. Figure 2(a) shows an example of classifying two senones (s_3 and s_5). In the figure, we see that while the pre-trained adult acoustic model can predict some senone labels correctly, there are still many child samples (the stars in the red circle) falsely classified by the adult acoustic model, resulting in wrong feature alignment after the adaptation. Indeed, the pre-trained adult acoustic model does not necessarily predict correct labels for child speech. This limitation motivates us to incorporate acoustic model adaptation into feature adaptation training to produce an acoustic model more suited to generating labels for child speech, which we elaborate in the next section.

Algorithm 1 Bidirectional training for feature and model adaptation

Input: (X_{adult},Y_{adult}), X_{child}
For ep in 1 to N_{ep} **do**
  For $iter$ in 1 to N_{iter} **do**
    Generate Y_{child} with M^{ep}_{iter}
    Optimize θ_{m_adpt} with (X_{adult},Y_{adult}), (X_{child},Y_{child}) using equation 6
  End for
  Generate Y_{child} with $M^{ep}_{N_{iter}}$
  Optimize θ_{f_adpt} and θ_{dom} with (X_{adult},Y_{adult}), (X_{child},Y_{child}) using equation 3 and 4
End for
Output $M^{N_{ep}}_{N_{iter}}(\theta_{f\_adpt})$, $M^{N_{ep}}_{N_{iter}}(\theta_{m\_adpt})$

3.2. Acoustic model adaptation

As shown in Figure 1, the last layer of pre-trained adult acoustic model is replaced by a model adapter, the adapted acoustic model (adult acoustic model + model adapter), therefore, is expected to be more suitable for recognizing child speech. The loss function for training model adapter is computed as:

$$L(\Theta_{m\_adpt}) = \frac{1}{N}\sum_{i=1}^{N} L^{i}_{senone}(\Theta_{m\_adpt}) \quad (5)$$

where L_{senone} is the cross-entropy loss on all labeled adult and unlabelled child speech samples, which is different from the feature adaptation in equation 1 that only computes the loss on adult samples. Θ_{m_adpt} is optimized as:

$$\Theta_{m\_adpt} = \underset{\Theta_{m\_adpt}}{\operatorname{argmin}}\, L(\Theta_{m\_adpt}) \quad (6)$$

To generate more reliable labels for unlabelled child speech, we propose a bidirectional training algorithm, which incrementally optimizes feature adaptation and model adaptation layers. In Algorithm 1, each training cycle consists of multiple updates to the model adapter and one update to the feature adapter. The number of updates to model adapter, N_{iter}, is a hyperparameter and is tuned on the validation set. Given the labeled adult speech samples (X_{adult},Y_{adult}) and unlabeled child speech samples (X_{child}), the inner loop is mainly to learn the model adaptation with model predicted senone labels of high output probabilities (Y_{child}), and the adapted model, $M^{ep}_{N_{iter}}$ (green classification boundary in Figure 2(b)), is capable of generating more accurate pseudo labels than the fixed adult acoustic model (black classification boundary in Figure 2(a)). Utilizing soft labels that are closer to the ground truth can promote learning the feature adaptation. Similarly, better feature adaptation model would in return contribute to better model adapter. Such joint feature and model adaptation training allows two adaptation models to gradually refine and reinforce each other, ultimately resulting in a better acoustic model for processing child speech.

4. Experimental setup

4.1. Datasets

4.1.1. Adult speech corpus

The adult speech datasets used for acoustic feature and model adaptation are *LibriSpeech* [26] "train-clean-100" subset, "dev-clean" subset, and "test-clean" subset.

4.1.2. Child speech corpus

The child speech dataset is *SingaKids-English* corpus, which includes 46 hours (train-40h, dev-2h, test-4h) of phonetically transcribed children's speech data. There are 193 speakers in total, aged between 6 and 12 years old. 1547 utterances from the "test-4h" subset was scored for pronunciation and fluency using 5 proficiency levels by an English teacher certified by the Ministry of Education, Singapore.

4.2. Model implementation

4.2.1. Acoustic model

For easy deployment on low-power devices, the pre-trained adult acoustic model is a small-size DNN model (26.5906 millions parameters) that was well trained on large amounts of adult speech data. The acoustic feature dimension is 1320, which consists of 11 consecutive speech frames. Each speech frame is parameterized into 40-dimensional log Mel-scale filterbank features, along with their first and second difference coefficients. The acoustic model in Figure 1 (adult acoustic model + model adapter) is initialized with the pre-trained adult model while we freeze the feature representation layers and allow the last layer to be trainable. During model training, we adopt both batch normalization and dropout to prevent over-fitting. The *LibriSpeech* developmental set "dev-clean" and the *SingaKids-English* developmental set "dev-2h" were used to optimize the model hyperparameters.

4.2.2. Assessment model

A multi-task DNN was employed to conduct the pronunciation and fluency scoring. To train the assessment classifier with limited utterance-level scoring data, we used 3 dense layers with 128 nodes per layer for representation learning, and 2 softmax layers with 5 nodes to output the proficiency score. The input feature is a 30-dimensional vector consisting of a set of widely used speech evaluation features [2, 27, 28].

4.3. Evaluation metric

To better examine the acoustic modeling itself, instead of using language models, we employ the free phone decoding graph for the speech recognition task and use phone error rate (PER) to evaluate the performance. As for the speech assessment tasks, we adopt two widely used metrics [29, 30] of prediction accuracy and mean squared error (MSE) for evaluation. MSE is used to measure the difference between the model predicted score and the reference score rated by the teacher when the scoring classifier gives a wrong prediction. The performance metrics are computed by comparing the model's predicted scores with the scores rated by a teacher. The two-sided t-test is utilized to assess the statistical significance of performance metric differences between various methods.

5. Experimental results

5.1. Speech recognition

To validate the competitiveness of baseline model, we first conduct speech recognition on LibriSpeech "test-clean" set with the standard pruned version of the WSJ-5k tri-gram language model. It achieves a word error rate (WER) of 9.49%, which is better than Kaldi's benchmark result of 9.66%[1] adopts a simi-

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

Table 1: *Phone error rate (%) on SingaKids-English test set.*

Method		G12	G34	G56	Overall
Pre-trained adult acoustic model (Baseline)		85.11	73.24	69.08	74.43
+ Feature adaptation	FMLLR	83.73	73.24	65.78	72.55
	SAT	69.56	61.84	58.02	62.02
+ Model adaptation		54.67	48.12	46.30	48.86
+ Feature adaptation + model adaptation (Proposed)		46.26	43.44	41.42	43.25

larly sized model and same testing settings. This suggests that the pre-trained adult acoustic model can be used as a competitive baseline model. We also include comparisons to two feature adaptation approaches of FMLLR and SAT [29]. The FMLLR transformations are estimated using the labels predicted by baseline model. The six grades of primary school were merged into three groups of G12 (age 6-7), G34 (age 8-9), and G56 (age 10-12) to analyze the results.

As shown in Table 1, the proposed joint feature and model adaptation approach achieves the lowest PER in all conditions, which shows the effectiveness of our proposed approach. It reduces the overall PER by 31.18% over the pre-trained model, by 29.3% over the FMLLR adaptation, by 18.77% over the SAT adaptation, and by 5.61% over model adaptation. FMLLR adaptation being worse than the adversarial-based feature adaptation methods indicates that the one hot phonetic labels predicted by the pre-trained model are not reliable, and directly employing such hard labels will cause a tendency for learned transformations to overfit wrong phone classes.

We also observe that the PER is up to 85.11% for the child speech of G12 when directly employing the pre-trained adult acoustic model, even though its WER is less than 10% on the Librispeech "test-clean" subset. Meanwhile, as the school grade increases from G12 to G56, the PER decreases from 4.84% to 17.95% in different methods. Such performance differences match our understanding that there is a huge acoustic mismatch between adult speech and child speech. Moreover, the acoustic variability of younger child (G12) is higher than that of senior grades, and their pronunciation shifts across age groups and becomes closer to that of an adult's along their growth and development during childhood.

All models powered by speech adaptation achieved better performance than the pre-trained model, suggesting that speech adaptation is able to reduce the mismatch between the adult and child acoustic domains. When combining feature adaptation with model adaptation, the performance can be further improved, which means the feature adaptation and model adaptation are complementary to each other and helps achieve the best result in adapting child speech to adult speech. Although the PER is relatively high at around 43%, it is consistent with literature [13, 31, 32], illustrating that recognizing child speech is challenging.

5.2. Pronunciation and fluency assessment

When using speech technology to evaluate speaking skills such as pronunciation and fluency, acoustic modeling is a crucial aspect. The phone likelihood ratio and duration of phones and pauses, obtained from an acoustic model, are commonly used to evaluate pronunciation proficiency and speaking fluency. This section presents a performance comparison of these two speech assessment tasks.

Tables 2 shows the results on pronunciation and fluency assessment tasks. We first observe the proposed approach improves the overall prediction accuracy over the baseline by 6.9% on pronunciation evaluation task (Pron) and 6.7% on fluency evaluation task (Flu). The adapted acoustic model, therefore, can generate more precise forced alignments (containing pronunciations per word and time boundaries) of child speech for the scoring classifier. The fluency prediction accuracy on G12 is improved from 31.6% to 52.6% and the corresponding MSE is reduced from 2.25 to 1.38. It indicates our approach can better align the speech of younger students and provide more accurate duration features for the fluency scoring classifier. It is important to note that fluency scoring has to be performed at the utterance level or longer, resulting in significantly less labeled data being available for training the assessment classifier compared to acoustic modeling. This data scarcity limits the performance on utterance-level assessment task.

Table 2: *Pronunciation and fluency evaluation. Asterisks (*) indicate statistically significant differences at the 0.05 significance level between the baseline and the proposed method in the two-sided t-test.*

		Accuracy(%)		MSE	
		Baseline	Proposed	Baseline	Proposed
Pron	G12	42.1	50.0*	1.30	1.11*
	G34	37.3	47.7*	1.14	1.11
	G56	47.3	53.6*	1.44	1.17*
	Overall	43.3	50.2*	1.32	1.16*
Flu	G12	31.6	52.6*	2.25	1.38*
	G34	40.3	47.7*	1.96	1.45*
	G56	50.9	54.5*	2.22	1.31*
	Overall	44.2	50.9*	2.13	1.39*

6. Conclusions and future work

To tackle the challenges posed by the high acoustic variability and limited linguistic resources available for modeling of child speech, we proposed an unsupervised speech adaptation method based on adversarial learning to reduce the mismatch between child speech and adult speech. We designed a novel bidirectional learning algorithm that facilitates the joint learning of feature and model adaptation. The algorithm utilizes two adaptation modules that alternate in training to improve each other's performance by exchanging more accurate information. The results showed that the proposed joint feature and model adaptation method outperforms other comparable baselines across various speech tasks.

Our future efforts will involve leveraging more unlabelled child speech data (hundreds to thousands of hours) to improve performance and explore its ceiling. Developing strategies to reduce the requirement for teacher scores, which is a time-consuming and labor-intensive task, is another future endeavor.

7. References

[1] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, and Senior, "Large vocabulary automatic speech recognition for children," in *INTERSPEECH*, 2015, pp. 1611–1615.

[2] C. Cucchiarini, H. Strik, and L. Boves, "Automatic evaluation of dutch pronunciation by using speech recognition technology," in *ASRU*. IEEE, 1997, pp. 622–629.

[3] 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.

[4] J. Cheng and J. Shen, "Towards accurate recognition for children's oral reading fluency," in *2010 IEEE Spoken Language Technology Workshop*, 2010, pp. 103–108.

[5] A. Metallinou and J. Cheng, "Using deep neural networks to improve proficiency assessment for children english language learners," in *INTERSPEECH*, 2014.

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

[7] S. Ghai and R. Sinha, "Exploring the role of spectral smoothing in context of children's speech recognition," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[8] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *EUSPEECH*, 1997.

[9] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, pp. 847–860, 2007.

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

[11] 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.

[12] S. Lee, A. Potamianos, and S. Narayanan, "Developmental acoustic study of american english diphthongs," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 1880–1894, 2014.

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

[14] F. Claus, H. G. Rosales, R. Petrick, H.-U. Hain, and R. Hoffmann, "A survey about databases of children's speech." in *INTERSPEECH*, 2013, pp. 2410–2414.

[15] 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.

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

[17] 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.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[19] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[20] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. PMLR, 2018, pp. 1989–1998.

[21] J. X. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, "Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP," *arXiv preprint arXiv:2005.05909*, 2020.

[22] A. Ramponi and B. Plank, "Neural unsupervised domain adaptation in NLP—A survey," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 6838–6855.

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

[24] 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*. IEEE, 2019, pp. 8122–8126.

[25] R. Duan and N. F. Chen, "Unsupervised feature adaptation using adversarial multi-task training for automatic evaluation of children's speech," in *INTERSPEECH*, 2020, pp. 3037–3041.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.

[27] 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.

[28] K. Shi, K. M. Tan, R. Duan, S. U. M. Salleh, N. F. A. Suhaimi, R. Vellu, N. T. H. H. Thai, and N. F. Chen, "Computer-assisted language learning system: Automatic speech evaluation for children learning malay and tamil." in *INTERSPEECH*, 2020, pp. 1019–1020.

[29] R. Duan and N. F. Chen, "Senone-aware adversarial multi-task training for unsupervised child to adult speech adaptation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7758–7762.

[30] K. Kyriakopoulos, M. Gales, and K. Knill, "Automatic characterisation of the pronunciation of non-native english speakers using phone distance features," in *7th ISCA Workshop on Speech and Language Technology in Education*, 2018.

[31] 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.

[32] R. Gretter, M. Matassoni, D. Falavigna, A. Misra, C. Leong, K. Knill, and L. Wang, "ETLT 2021: Shared Task on Automatic Speech Recognition for Non-Native Children's Speech," in *Proc. Interspeech 2021*, 2021, pp. 3845–3849.