Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7314–7320
Marseille, 20-25 June 2022

Multilingual Transfer Learning for Children Automatic Speech Recognition

Thomas Rolland[1,2] **Alberto Abad**[1,2] **Catia Cucchiarini**[3] **Helmer Strik**[3]
[1]INESC-ID, Lisbon, Portugal
[2]Instituto Superior Técnico, Universidade de Lisboa, Portugal
[3]Centre for Language and Speech Technology (CLST), Radboud University Nijmegen, The Netherlands

Abstract

Despite recent advances in automatic speech recognition (ASR), the recognition of children's speech still remains a significant challenge. This is mainly due to the high acoustic variability and the limited amount of available training data. The latter problem is particularly evident in languages other than English, which are usually less-resourced. In the current paper, we address children ASR in a number of less-resourced languages by combining several small-sized children speech corpora from these languages. In particular, we address the following research question: Does a novel two-step training strategy in which multilingual learning is followed by language-specific transfer learning outperform conventional single language/task training for children speech, as well as multilingual and transfer learning alone? Based on previous experimental results with English, we hypothesize that multilingual learning provides a better generalization of the underlying characteristics of children's speech. Our results provide a positive answer to our research question, by showing that using transfer learning on top of a multilingual model for an unseen language outperforms conventional single language-specific learning.

Keywords: children speech, children speech recognition, ASR, multilingual training, transfer learning

1. Introduction

Recently, significant improvements have been achieved by Automatic Speech Recognition (ASR) systems through the application of deep learning approaches. However, children ASR still represents a significant challenge, as testified by the performance drop of the current state-of-the-art systems compared to adult speech.

This degradation can be partially attributed to the high acoustic variability in children speech caused by developmental changes of the speech production apparatus (Gerosa et al., 2009; Wilpon and Jacobsen, 1996). Such physical changes lead to different formant and fundamental frequency locations (Lee et al., 1999). Moreover, limited linguistic and phonetic knowledge of children can also contribute to performance degradation (Wilpon and Jacobsen, 1996). Finally, the performance gap between children and adult ASR can be explained by a data scarcity problem. Indeed, current speech recognition systems are based on deep learning, for which the amount of data used is essential. Despite recent efforts to collect larger datasets of children speech (Ward et al., 2013), most of the publicly available children corpora contain less than fifty hours of speech, while adult speech corpora containing hundreds (or even thousands) of hours can be easily found (Panayotov et al., 2015). The problem of data scarcity is particularly acute for languages other than English, for which, in general, fewer resources are available. This can be seen in (Liao et al., 2015), in which the authors used a large amount of children speech –comparable to an adult speech corpus– to train a convolutional long-short-term-memory deep neural network. This system achieved state-of-the-art performances (9.4% WER) competitive with adult speech recognition systems. Thus, this work demonstrates that neural networks can learn from complex and variable children's speech data as long as there is enough data for training.

To tackle the different challenges of children ASR, several strategies have been proposed over the years. Vocal tract length normalisation (VTLN) has been commonly used to wrap spectral features onto a canonical space (Serizel and Giuliani, 2014). Acoustic model adaptation and speaker adaptive training have been also found to be effective to improve the performance of children ASR (Shivakumar et al., 2014; Gray et al., 2014). Improved acoustic model architectures, such as factorized time-delay neural network (TDNN-F) based models, have also been proposed (Wu et al., 2019). Due to data constraints, the vast majority of recent children ASR literature focuses on the hybrid Hidden-Markov-Model Deep-Neural-Networks (HMM-DNN) paradigm. In fact, some recent studies have reported the limitations of end-to-end approaches for this task (Gelin et al., 2021).

In this work, we study whether the performance of children ASR for less-resourced languages can be improved by using a novel approach in which we combine resources from different languages. We propose to address the aforementioned large acoustic variability and data scarcity challenges by exploiting several small-sized corpora of children from these different languages. To leverage information from heterogeneous data, the present study extends conventional multilingual training and transfer learning for hybrid HMM-DNN ASR combining them in a meaningful way in a new context. First, a multilingual model trained with a multi-task learning objective attempts to optimize the network parameters to the particular characteristics of children speech on multiple languages/tasks in parallel.

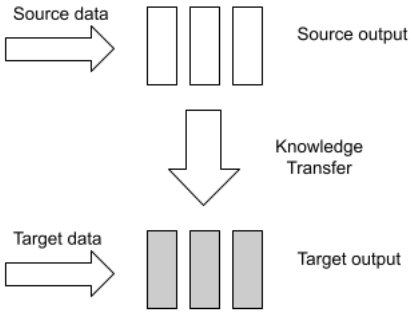

Figure 1: Transfer learning procedure. The white blocks are randomly initialised layers while the grey blocks are layers initialised with the pre-trained source parameters.

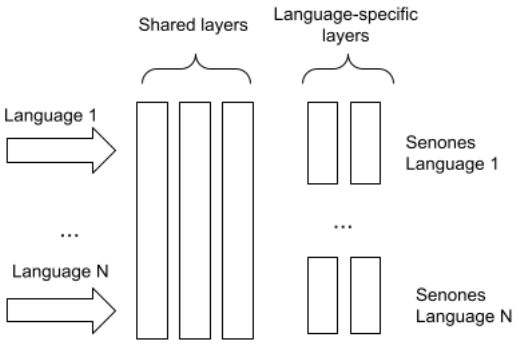

Figure 2: Multilingual approach using each language as a task in a multi-task learning context.

Subsequently, this multilingual model is used to improve ASR for a target language –potentially different from those used in the multilingual training stage– by using transfer learning. We address the following research question: Does this two-step training strategy outperform conventional single language/task training for children speech, as well as multilingual and transfer learning alone?

The rest of this paper is organized a follows. Section 2 reviews transfer learning and multi-task. Section 3 introduces our combined multilingual and transfer learning approach for children ASR. The different corpora used for this work are described in section 4. Section 5-7 presents the experimental setup and results. Finally section 8 gives the conclusions and presents potential perspectives for future work.

2. Transfer learning and multi-task learning for children ASR

2.1. Transfer learning

Transfer learning (TL) is a training procedure in which model parameters are initialized using knowledge gained from a model trained on a source related task (see figure 1). The resulting model leverages various underlying characteristics that have been captured by the different layers of the neural network during the learning process. A generally accepted interpretation is that the bottom layers, close to the input, capture more signal specific characteristics. While higher layers, close to the output, are more task-specific (Bengio et al., 2013). Furthermore, because the target model relies on a pre-trained model, one advantage of TL is the reduced training (or adaptation) data requirements.

TL has been successfully used in a large variety of applications, including language understanding (Devlin et al., 2018) and dysarthric speech recognition (Takashima et al., 2020) among others. This success has motivated its use for children speech recognition. Thus, some works (Gurunath Shivakumar and Georgiou, 2020; Tong et al., 2017) have explored the use of transfer learning for children speech recognition, using adult speech as source data, reporting remarkable WER improvements. Finally, (Matassoni et al., 2018) proposed a variation of the traditional adult to children TL where the source task is a multi-view learning setting based on multiple languages of children speech by using a shared lexicon across all languages. In a similar way, (Tachbelie et al., 2020) has used this approach for low-resource Ethiopian languages.

2.2. Multi-task learning

Multi-task learning (MTL) aims to learn shared representations between related tasks by jointly training all tasks in parallel. This procedure can enable the model to better generalize. In general, a typical model consists of two distinct parts, the first part, a sub-network shared by all tasks, and the second part, a task-specific sub-network. In the context of a multilingual system trained with multi-task objective, the outputs are the number of senones for the language of the corresponding subnetwork (see figure 2).

MTL has also been successfully applied to many areas, including natural language processing (Collobert and Weston, 2008) and automatic speech recognition (Madikeri et al., 2020). In the context of speech recognition, MTL has found its direct application in the field of low-resource ASR (Abad et al., 2020; Madikeri et al., 2021). (Tong et al., 2017) and (Wei et al., 2019) successfully applied MTL using children speaking Mandarin and English, leading to a relative improvement of 16.96% WER in English children case.

3. Proposed approach

Motivated by the reported success of MTL and TL for children ASR, we propose to combine them together for improved acoustic modelling of hybrid HMM-DNN ASR. The main motivation for focusing on hybrid ASR –in contrast to more recent end-to-end (E2E) paradigms– is the limited current success of E2E ap-

proaches in low-resource tasks (Gelin et al., 2021). The proposed approach consists of a two stage procedure combining MTL and TL that extends the existing techniques, since these are usually applied separately. First, a multilingual model trained with a multi-task learning objective attempts to optimize the network parameters to the particular characteristics of children speech on multiple languages in parallel. In this work, the model is considered multilingual because all the tasks trained during multitask learning are a corpus of children from different languages. Secondly, we adapt this model for a specific children corpus with TL.The motivation for using TL as a second stage is to take advantage of the robust pre-trained model trained during the MTL phase. Indeed, this pre-trained model has potentially learned cross-linguistic information of children speech, but has also seen more children data than a model trained in a single language. For this purpose, the acoustic model is divided in two parts: the layers close to the input are shared across all languages and the top layers are language-specific. That is, there are as many output layers as there are languages, i.e. children corpora. Notice that one can incorporate a new language/task in this second stage adding a new language-specific output, even if this new language/task has not been seen during MTL training (figure 3). Our hypothesis is that the more data we use, the better the shared layers can capture the underlying characteristics of children speech during the first stage of the procedure. These characteristics can be used effectively, later, by the language-specific layers and during the second step of the procedure (figure 3).

Although the approaches adopted in this work have been used previously in other studies, for instance (Tong et al., 2017) and (Wei et al., 2019) where they successfully applied MTL using children speaking Mandarin and English, obtaining a relative improvement of 16.96% WER in the English children case, it is clear that successful performance of a methodological approach in the case of English cannot be expected to generalize to other contexts and languages. As we all know, English is a large-size, resource-rich pluricentric language which should be seen more as an exceptional case, rather than an average representative. Against this background, it is important to emphasize that there is a need for research that investigates whether methods that have already been tested for English also work in new contexts such as those of mid-sized languages with fewer resources than English, like Dutch, Portuguese, Swedish and German.

4. Corpora

All experiments were conducted using five children corpora, each from a different language. This section briefly presents each corpus and how it was used in the present study. In addition, more information about the duration and language can be found in Table 1. Notice that in this work we have only used small datasets to better reflect the average size of the available children's speech corpora.

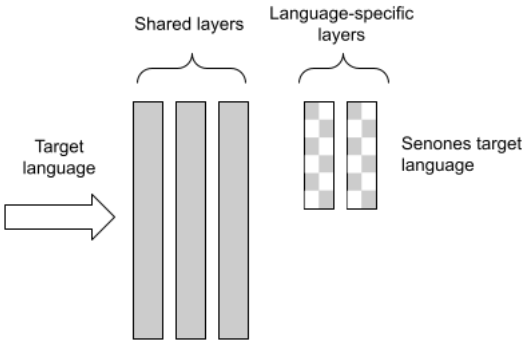

Figure 3: Multilingual transfer learning approach. Language-specific layers can be randomly initialized for a language not present during the MTL phase or use the corresponding pre-trained layers in case the target language was present during the MTL phase. Grey blocks are pre-trained during MTL phase.

Corpus name	Language	Train	Test
PFSTAR_SWE	Swedish	6030 utt 04h00	2879 utt 01h48
ETLTDE	L2 German	1445 utt 04h41	339 utt 01h06
CMU	English	3637 utt 06h26	1543 utt 02h45
LETSREAD	Portuguese	3590 utt 12h00	1039 utt 02h30
CHOREC	Dutch	2490 utt 20h12	575 utt 04h42

Table 1: Statistics on the different corpora of children's speech.

4.1. PFSTAR_SWEDISH

The PFStar children's speech corpus (Batliner et al., 2005) was collected as part of the EU FP5 PFSTAR project. It contains more than 60 hours of speech. This corpus is divided in two parts: native-language speech and non-native language part. The native-language speech part contains recordings of British English, German and Swedish children, from 4 to 14 years old. The non-native language part consists of speech by Italian, German and Swedish children speaking English. In this work, we only used the native language Swedish part, consisting of speech by 198 native Swedish children, between 4 and 8 years old recorded in the Stockholm area, imitating an adult who read the text from a screen.

4.2. ETLTDE

Extended Trentino Language Testing (ETLT) corpus (Gretter et al., 2020) has been collected in northern Italy for assessing English and German proficiency

of Italian children between 9 and 16 years old, by asking them to answer questions. The data collection was carried out in schools. On average the signal quality is good, but some background noise is often present (doors, steps, keyboard typing, background voices, street noises if the windows are open, etc). In addition, many answers are whispered and difficult to understand. In our experiments, we only used the German-transcribed subset (named ETLTDE), around 6h divided into training and test partitions.

4.3. CMU_KIDS

The CMU kids corpus (Eskenazi et al.,) contains English sentences read aloud by children, 24 males and 52 females, from 6 to 11 years old. In total, 5,180 utterances were recorded with one sentence per utterance. This database was created to train the SPHINX II (Huang et al., 1993) automatic speech recognition system within the LISTEN project at Carnegie Mellon University (CMU).

4.4. LETSREAD

LetsRead (Proença et al., 2016) is a corpus of European Portuguese read speech of children from 6 to 10 years old. In total, 284 children from private and public Portuguese schools were asked to carry out two tasks: reading sentences and a list of pseudo-words. The difficulty of the tasks varies depending on the school year of the child. In our experiments, we excluded all utterances from the pseudo-word reading task because we do not include pseudo-words in the language model and lexicon.

4.5. CHOREC

The Chorec corpus (Cleuren et al., 2008) consists of 400 Dutch-speaking elementary school children, between 6 and 12 years old, reading words, pseudo-words and stories. The difficulty of the reading task was adapted to children with 9 different levels. Recordings were made in schools, leading to some environmental noises (school bells, children entering the playground etc.). For our experiments, similar to the LETSREAD dataset, we discarded pseudo-word utterances.

5. Experimental setup

All experiments were carried out using the Kaldi open-source toolkit (Povey et al., 2011). First, for each language, an independent HMM-GMM acoustic model was trained to produce the necessary alignment to the HMM-DNN model. Then, HMM-DNN acoustic models were trained using 40-dim filter-banks (*fbanks*) in addition with a 40-dim Spectral Subband Centroid (SSC) features (Paliwal, 1998). These features are known to have similar properties to formant frequencies. Thus, we expect them to help vowel recognition and lead to better recognition of children's speech. The resulting 80-dim input features are then augmented by 100-dim i-vector. Concatenating speaker embeddings to the input features helps to improve model speaker robustness (Senior and Lopez-Moreno, 2014). For our experiments, we use an i-vector extractor trained on a set of pooled children data from different languages.
Data augmentation was applied to all training corpora by perturbing the speaking rate of each training utterance by 0.9 and 1.1 factor; as well as volume perturbation. This helps the network to be more robust to rate and volume variability on the test sets. To further improve the robustness of the model, Specaugment (Park et al., 2019) was applied on top of the fbanks and SSC features by randomly masking time and frequency bands.
For all experiments, we kept the same HMM-DNN acoustic model architecture using lattice-free maximum mutual information (LF-MMI) objective with a learning rate of 2.0E-4. The architecture is divided in two parts: i) six convolutional neural network layers and seven TDNN-F layers of dimension 1024 shared across all languages, and ii) two TDNN layers of dimension 450 and a fully-connected layer for the languages-specific part where the output dimension correspond to the number of senones for the branch's language. Each corpus, i.e. each language, uses an independent language model and lexicon, fixed in all experiments, in order to evaluate only the contribution of the acoustic model.

6. Multilingual-transfer learning experiment

Table 2 presents the WER results of the multilingual transfer learning (MLTL) approach compared to three different methods: baseline, trained on each corpus individually for 4 epochs; Multi-task Training (MTL) alone, trained jointly using all corpora for 4 epochs ; Transfer Learning (TL) alone, adapted for the target language using in turn one of the other 4 baseline models as a source, leading to 4 results per target language. In addition, for clarity, we summarise the transfer learning scores with the average of the 4 scores and the best of the 4 for each target.
Firstly, it is important to emphasise that the baseline scores correctly reflect the different tasks the children were asked to perform and the corresponding amount of data available for each corpus. The best WER score, 21.26% for CMU, can be explained by the reading-aloud-sentences task nature of this corpus. Thus, the language model can more easily compensate the acoustic model errors. In addition, Chorec and LetsRead, as the largest corpora in our experiment, also yield relatively good results for children speech recognition. On the other hand, ETLTDE and PFSTAR_SWE show the worse WER results with 44.69% and 54.36% WER, respectively. This can be explained by the amount of data available and by the language model which does not compensate as much as the CMU model. Especially for ETLTDE, since it is the only corpus that does not contain scripted text, but extemporaneous responses. In

	PFSTAR_SWE	ETLTDE	CMU	LETSREAD	CHOREC
Single language	54.36%	44.69%	21.26%	26.88%	25.15%
MTL	54.95%	42.46%	23.01%	27.45%	25.10%
TL from PFSTAR_SWE	-	42.23%	20.62%	26.47%	24.65%
TL from ETLTDE	53.60%	-	20.90%	26.61%	25.42%
TL from CMU	52.83%	41.54%	-	26.49%	24.58%
TL from LETSREAD	52.50%	41.77%	20.41%	-	24.60%
TL from CHOREC	52.20%	40.28%	19.77%	26.05%	-
TL Average	52.78%	41.46%	20.43%	26.41%	24.81%
TL Best	52.20%	40.28%	19.77%	26.05%	24.58%
MLTL	**51.67%**	**38.04%**	**19.33%**	**25.75%**	**23.78%**
MLTL-olo	**51.58%**	40.05%	**19.67%**	26.20%	**24.57%**

Table 2: WER results of multilingual-transfer learning and cross-lingual experiments. MTL: Multi-Task Learning, TL: Transfer Learning, MLTL: Multilingual Transfer Learning, MLTL: Multilingual Transfer Learning one-language-out

addition, the age range of PFSTAR_SWE children also plays a critical role in performance, since younger children generally yield worse performance scores (Gurunath Shivakumar and Georgiou, 2020).

Turning to multi-task learning, among all the approaches presented, only MTL fails to improve the baseline performance for almost all languages, which is in contradiction with(Tong et al., 2017). However, it can be explained by the differences in terms of the size of the child speech corpora used. The smaller the size of the corpora used, the more difficult it is to model the acoustic variation in the children speech.

Concerning TL, all performance scores outperform their corresponding baseline, confirming that TL is an adequate method for children ASR since it allows the system to be confronted with more children, thus with more variation. Precisely, table 2 shows that the best pre-trained model for knowledge transfer is Chorec. This makes sense since Chorec is the largest corpus, representing about 40% of the total data used in our experiments.

Finally, MLTL shows an average relative improvement in WER of 7.73% compared to the baseline, slightly higher than the average (TL Avg) and the best (TL Best) transfer learning performance, with an average relative improvement of 4.50% and 2.66%, respectively.

The strength of MLTL is that it can benefit both from MTL and TL, minimizing some of their associated weaknesses. Attending to our results, MTL does not improve single language training. We believe that the unbalanced amount of data, the significant differences among data sets and the use of segmental optimization (lattice-free MMI) can partially explain these results. Nevertheless, we hypothesize that the multi-task objective leans the network towards a better optimization of the lower layers, rather than optimizing the upper language-specific layers, that can still be beneficial for TL. Regarding TL, one can observe considerable performance variations depending on the pre-trained model used as the source model, probably due to a poorer initialisation of lower layers that is less efficient for TL. The MLTL experiments show that we can overcome these drawbacks combining both MTL and TL, thus, validating the effectiveness of this approach for robust speech recognition of children.

7. Cross-lingual validation

In the previous section, we saw that the MLTL approach yields better results than separate multi-task and transfer-learning frameworks.

To further validate the hypothesis that the shared lower layers are able to learn meaningful information of children's speech characteristics, regardless of the language, we perform a cross-language experiment following a leave one-language-out cross-validation setting. In this experiment, we keep one language out of the multi-task training and use it only during the TL phase to adapt the acoustic model parameters.

We repeated this procedure for each corpus in our experiment.As in the previous experiment, we used 4 epochs for each learning phase. Last row of Table 2 presents the results of the cross-language experiment.

For all corpora, the MLTL one-language-out (MLTL-olo) approach outperforms the baseline WER score with an average relative improvement of 5.56%. Improvements are more important for the small corpora ETLTDE and CMU, with a relative improvement of 14.88% and 9.07%, respectively. PFSTAR_SWE does not benefit as much, with only 5.05% relative improvement. This is mainly due to the age differences with the children in the other corpora used in the MTL phase. Indeed, the children in PFSTAR_SWE are much younger (see section 4 for more details). Therefore, we conclude that the shared layers have learned the underlying multilingual features of children.

It is also interesting to compare MLTL-olo with the results of transfer learning alone. In both cases, the pre-trained models used have never seen the target language data. We observe that the results between the

MLTL-olo and TL Best are extremely close, with small improvement with the MLTL-olo, only the best transfer learning model on LetsRead is slightly better than MLTL. This means that during multilingual training the system learned, at least, the best representation of the available children's characteristics. This is consistent with our hypothesis of the important role of the multilingual training phase in our two-step procedure.

8. Conclusions

In this work, we addressed the following research question: Does the two-step training strategy we propose in the current paper outperform conventional single language/task training for children speech, as well as multilingual and transfer learning alone. Our results provide a positive answer to this question, by showing that the limitations of MTL and TL can be overcome by the multilingual transfer learning approach, even in a low-resource scenario, leading to an average relative improvement of 7.73%. Multilingual pre-training is also beneficial for transfer learning with an unseen language, with an average relative improvement of 5.56%. Multilingual transfer learning thus seems to be an appropriate method to address children speech recognition in a challenging context. In future work, it would be interesting to investigate the effect of a larger children corpus or an adult speech corpus in the multilingual learning phase, as this would allow the model to be more acoustically robust. In addition, it would be interesting to explore the effect of non-European languages, as previous works has shown an improvement by combining Madarin and English. Furthermore, a more detailed comparison between age groups on the systems' performance would be an interesting next step. Finally, assessing the importance of the nature of the task within the multi-task phase and transfer learning phase would also be a possible avenue for future research.

9. Acknowledgements

This work was partially supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020 and European Union funds through Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 766287.

10. Bibliographical References

Abad, A., Bell, P., Carmantini, A., and Renais, S. (2020). Cross lingual transfer learning for zero-resource domain adaptation. In *ICASSP*, pages 6909–6913.

Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M., Steidl, S., and Wong, M. (2005). The pf_star children's speech corpus. pages 2761–2764, 01.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Cleuren, L., Duchateau, J., Ghesquiere, P., et al. (2008). Children's oral reading corpus (chorec): description and assessment of annotator agreement. *LREC 2008 Proceedings*, pages 998–1005.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Eskenazi, M., Mostow, J., and Graff, D.). The cmu kids speech corpus.

Gelin, L., Daniel, M., Pinquier, J., and Pellegrini, T. (2021). End-to-end acoustic modelling for phone recognition of young readers. *Speech Communication*, 134:71–84.

Gerosa, M., Giuliani, D., Narayanan, S., and Potamianos, A. (2009). A review of asr technologies for children's speech. New York, NY, USA. Association for Computing Machinery.

Gray, S. S., Willett, D., Lu, J., Pinto, J., Maergner, P., and Bodenstab, N. (2014). Child automatic speech recognition for us english: child interaction with living-room-electronic-devices. In *WOCCI*, pages 21–26.

Gretter, R., Matassoni, M., Bannò, S., and Falavigna, D. (2020). Tlt-school: a corpus of non native children speech.

Gurunath Shivakumar, P. and Georgiou, P. (2020). Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. *Computer Speech Language*, 63:101077.

Huang, X., Alleva, F., Hwang, M.-Y., and Rosenfeld, R. (1993). An overview of the sphinx-ii speech recognition system. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, page 81–86, USA. Association for Computational Linguistics.

Lee, S., Potamianos, A., and Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468.

Liao, H., Pundak, G., Siohan, O., Carroll, M. K., Coccaro, N., Jiang, Q.-M., Sainath, T. N., Senior, A., Beaufays, F., and Bacchiani, M. (2015). Large vocabulary automatic speech recognition for children. In *Proc. Interspeech 2015*, pages 1611–1615.

Madikeri, S. R., Khonglah, B. K., Tong, S., Motlicek, P., Bourlard, H., and Povey, D. (2020). Lattice-free maximum mutual information training of multilingual speech recognition systems. In *INTERSPEECH*, pages 4746–4750.

Madikeri, S., Motlicek, P., and Bourlard, H. (2021). Multitask adaptation with lattice-free mmi for multi-genre speech recognition of low resource languages. *Proc. Interspeech 2021*, pages 4329–4333.

Matassoni, M., Gretter, R., Falavigna, D., and Giuliani,

D. (2018). Non-native children speech recognition through transfer learning. In *ICASSP*, pages 6229–6233.

Paliwal, K. (1998). Spectral subband centroid features for speech recognition. In *ICASSP*, volume 2, pages 617–620 vol.2.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. In *ICASSP*, pages 5206–5210.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*, pages 2613–2617.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., and Vesel, K. (2011). The kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 01.

Proença, J., Celorico, D., Candeias, S., Lopes, C., and Perdigão, F. (2016). The letsread corpus of portuguese children reading aloud for performance evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 781–785.

Senior, A. and Lopez-Moreno, I. (2014). Improving dnn speaker independence with i-vector inputs. In *ICASSP*, pages 225–229.

Serizel, R. and Giuliani, D. (2014). Vocal tract length normalisation approaches to dnn-based children's and adults' speech recognition. In *SLT Workshop*, pages 135–140.

Shivakumar, P. G., Potamianos, A., Lee, S., and Narayanan, S. S. (2014). Improving speech recognition for children using acoustic adaptation and pronunciation modeling. In *WOCCI*, pages 15–19.

Tachbelie, M. Y., Abate, S. T., and Schultz, T. (2020). Development of multilingual asr using globalphone for less-resourced languages: The case of ethiopian languages. In *Interspeech 2020*, pages 1032–1036.

Takashima, R., Takiguchi, T., and Ariki, Y. (2020). Two-step acoustic model adaptation for dysarthric speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6104–6108. IEEE.

Tong, R., Wang, L., and Ma, B. (2017). Transfer learning for children's speech recognition. *IALP*, pages 36–39.

Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., and Weston, T. B. (2013). My science tutor: A conversational multimedia virtual tutor. *Journal of Educational Psychology*, 105:1115–1125.

Wei, L., Dong, W., Lin, B., and Zhang, J. (2019). Multi-task based mispronunciation detection of children speech using multi-lingual information. In *APSIPA ASC*, pages 1791–1794. IEEE.

Wilpon, J. and Jacobsen, C. (1996). A study of speech recognition for children and the elderly. In *ICASSP*, volume 1, pages 349–352 vol. 1.

Wu, F., García-Perera, L. P., Povey, D., and Khudanpur, S. (2019). Advances in automatic speech recognition for child speech using factored time delay neural network. In *Interspeech 2019*, pages 1–5.