HHS Public Access

Author manuscript

Comput Speech Lang. Author manuscript; available in PMC 2021 September 01.

Published in final edited form as:

Comput Speech Lang. 2020 September ; 63: . doi:10.1016/j.csl.2020.101077.

Transfer Learning from Adult to Children for Speech Recognition: Evaluation, Analysis and Recommendations

Prashanth Gurunath Shivakumar, **Panayiotis Georgiou***

Signal Processing for Communication Understanding & Behavior Analysis (SCUBA) Lab, University of Southern California, Los Angeles, California, USA

Abstract

Children speech recognition is challenging mainly due to the inherent high variability in children's physical and articulatory characteristics and expressions. This variability manifests in both acoustic constructs and linguistic usage due to the rapidly changing developmental stage in children's life. Part of the challenge is due to the lack of large amounts of available children speech data for efficient modeling. This work attempts to address the key challenges using transfer learning from adult's models to children's models in a Deep Neural Network (DNN) framework for children's Automatic Speech Recognition (ASR) task evaluating on multiple children's speech corpora with a large vocabulary. The paper presents a systematic and an extensive analysis of the proposed transfer learning technique considering the key factors affecting children's speech recognition from prior literature. *Evaluations* are presented on (i) comparisons of earlier GMM-HMM and the newer DNN Models, (ii) effectiveness of standard adaptation techniques versus transfer learning, (iii) various adaptation configurations in tackling the variabilities present in children speech, in terms of (a) acoustic spectral variability, and (b) pronunciation variability and linguistic constraints. Our *Analysis* spans over (i) number of DNN model parameters (for adaptation), (ii) amount of adaptation data, (iii) ages of children, (iv) age dependent-independent adaptation. Finally, we provide *Recommendations* on (i) the favorable strategies over various aforementioned - analyzed parameters, and (ii) potential future research directions and relevant challenges/problems persisting in DNN based ASR for children's speech.

Keywords

Analysis of Children's Speech; Children Speech Recognition; Automatic Speech Recognition; Deep Learning; Transfer Learning; Deep Neural Network

*Corresponding author: georgiou@sipi.usc.edu (Panayiotis Georgiou).

The authors do not have any conflicts of interest

1. Introduction

Speech recognition is set to soon become an ubiquitous part of our life in the foreseeable future. A range of applications, such as human-machine interaction, communication, education, pronunciation and communication tutoring, entertainment and interactive gaming depend on such functionality. This has become partly possible due to high accuracies achieved by state-of-the-art speech recognition systems. An important user population for many such technologies are children. However, Automatic Speech Recognition (ASR) for children is still significantly less accurate than that of adults [1]. With the recent increased deployment of speech based technologies it becomes ever more important to be inclusive towards children. Thus there is a need to robustly address the challenges brought by the variability in kids speech.

Researchers have studied how the speech patterns of children differ to that of the adults. Prior studies have looked into the factors affecting, and degrading, the performance of ASR. Children speech was found to exhibit high level of variability. The research suggests that the variability exists in two levels. Firstly, the variability is embedded in the acoustic signals in the form of spectral and temporal variability, due to the physiological and developmental differences of children. Secondly, there is variability in kids pronunciation patterns, due to differing and partial linguistic knowledge.

Acoustic variability can be attributed to three main factors (i) shifted overall spectral content and formant frequencies for children [1], (ii) high within-subject variability in the spectral content, that affects formant locations [2], (iii) high inter-speaker variability observed across age groups, due to developmental changes, especially vocal tract [3]. Lee et al. [2] conducted a detailed study analyzing the temporal and spectral parameters of children speech. The study found that the within-subject variability decreased with increase in age from 5 years to 12 years, reaching adult levels at an age of 15.

The *word error rates* (WER) for children's ASR were found to be 2 to 5 times worse than adults [1]. Due to the unique acoustic characteristics of children's speech, training children-specific ASR models was found to be highly advantageous. Age dependent ASR models were also studied giving promising improvements, thereby confirming high inter-age dependent acoustic variability in children [4]. Li and Russell [5] studied the effect of speech bandwidth on recognition accuracy. The study found that the recognition performance degraded more rapidly for children when the bandwidth was reduced from 4kHz to 1.5kHz. Investigation of the possible causes showed that the average formant frequencies F1, F2 and F3 for children exceeded those of adults by more than 60% [6].

Several techniques to tackle the acoustic variability were proposed in recent times. Different front-end robust features such as *Mel-Frequency Cepstral Coefficients* (MFCC), *Perceptual Linear Prediction* (PLP) cepstral coefficients, and spectrum based filter bank features have been tried [7]. Several minor alterations of front-end features have also been investigated [7, 8, 6, 9, 10]. However, MFCC features have dominated due to their robustness and compatibility with adult ASR systems.

Potamianos and Narayanan [1] proposed several front-end frequency warping techniques and speaker normalization techniques with evaluations over different age groups. Particularly, *Vocal Tract Length Normalization* (VTLN) technique to suppress acoustic variability introduced by the developing vocal tracts in children has become a standard in children ASR systems [7, 11, 12], effectively reducing inter-speaker and inter-age-group acoustic variability. Adapting acoustic models with *Maximum Likelihood Linear Regression* (MLLR) and *Maximum A-Posteriori* (MAP) was found to be effective [13, 7, 14]. Further modest gains were achieved using *Speaker Adaptive Training* (SAT) based on *Constrained MLLR* (CMLLR) for children ASR [14, 7].

Some research efforts have also concentrated on dealing with the increased pronunciation variability and mispronunciations present in kids due to limited and developing linguistic knowledge. Performance gap between spontaneous speech recognition and read speech is particularly large for children [15]. Gerosa et al. [3] showed that spontaneous speech annotations are extremely useful. They showed that language usage efficiency increases with age for children reaching adult levels at 13 years of age i.e., disfluencies decrease with age. Potamianos and Narayanan [16] performed an in-depth analysis of linguistic variability in the context of spoken dialogue systems for children. Inter-speaker linguistic variability was found to be twice the intra-speaker variability. Mispronunciations in children were found to be twice as high for children of 8-10 years compared to that of 11-14 years, while the trend was reversed for filler pauses. Age dependencies were also found for the frequency of false-starts, duration, utterance length and breathing.

Das et al. [17] showed that language models trained on children speech were advantageous to using adult models suggesting children use different grammatical constructs. In [14], language model adaptation from adult to children showed improvements.

Children tend to also mispronounce, thus customized dictionaries for children can provide performance benefits [5]. Pronunciation variations among children vary with age. Data-driven pronunciation variation modeling is shown to be useful across children of all ages [7]. However, part of the variations are attributed towards the phonological processes and hence the customization of dictionaries have their limitations [18].

There are also significant efforts in speech applications for kids towards learning. For example [19, 20] focused on read speech assessment. Further, Tong et al. [21] focused on pronunciation assessment in Mandarin. Hagen et al. [22] proposed subword unit based speech recognition for children enabling assessment of children speech at finer details and detection of speech events such as partial words and mispronunciations.

More modern methods, specifically related to deep learning, have been extremely successful in improving ASR performance. The successes of Deep Neural Networks (DNN) have been attributed to DNN's ability to use vast amount of training data and to better approximate the non-linear functions needed to model speech, thus surpassing GMM based ASR systems. However, relatively less work has investigated DNNs for children's speech probably due to lack of large amounts of children's training data. [23, 24] conducted ASR experiments using a hybrid DNN-HMM based ASR system. They trained on approximately 10 hours of Italian

children's speech giving small improvements over traditional GMM based systems. Serizel and Giuliani [25] used a DNN to predict the frequency warping factors for VTLN which was later used to train a hybrid DNN-HMM system. [26] employed convolutional long short-term memory recurrent neural networks to train children ASR for use with Youtube Kids. They further employed data augmentation through artificially adding noise for more robustness. Combining adults' speech with children's speech for training improved results for both adults and children [27, 26, 28, 29]. Particularly, combining female adult speech in the training was shown to be more advantageous [27]. Multi-task learning frameworks for adapting adults' speech to children's speech were presented in [30, 21]. In [31, 32], a technique similar to [30] was adopted to overcome limited training data for DNN. Most recently, multi-lingual data adaptation in a transfer learning and multi-task learning framework was found to be useful for the task of ASR for children speaking in non-native language [33].

However, most of the prior works pertaining to analysis of children's speech in context of speech recognition has been on gaussian mixture based hidden markov models (GMM-HMM). Although there has been a wide consensus in the community about the advantages of DNN acoustic modeling for children's speech [27, 26, 28, 29, 30, 21, 31, 25, 23, 24], there has been no work to the best of our knowledge, which attempts to evaluate and analyze where the strengths of the DNNs lie in context to children's ASR. More importantly there is a need for an analysis of the shortcomings of the DNN based ASRs, i.e., problems and challenges persisting in children speech recognition using state-of-the-art speech recognition systems. Our study attempts to contribute to this gap and provide insights towards future developments.

In this work, we conduct ***Evaluations*** on large vocabulary continuous speech recognition (LVCSR) for children, to:

1. Compare older GMM-HMM models and newer DNN models.
2. Investigate different transfer learning adaptation techniques. Particularly we look at two factors degrading children ASR: acoustic variability and pronunciation variability in a DNN setup.
3. Assess effectiveness of different speaker normalization and adaptation techniques like VTLN, fMLLR, i-vector based adaptation versus the employed transfer learning technique.

Further, we conduct ***Analysis*** over the following parameters in context of transfer learning:

1. DNN model parameters.
2. Amount of adaptation data.
3. Effect of children's ages.
4. Age dependent transformations obtained from transfer learning and their validity, portability over the children's age span.

Recommendations are provided from the insights gained from conducting the aforementioned evaluations and analysis for:

1. Favorable transfer learning adaptation strategies for low data and high data scenarios.
2. Suggested transfer learning adaptation techniques for children of different ages.
3. Amount of adaptation data required for efficient performance over children's ages.
4. Potential future research directions and relevant challenges and problems persisting in children speech recognition.

The rest of the paper is formatted as follows: Section 2 motivates and describes the proposed transfer learning technique. Section 3 describes the databases used for recognition experiments. The experimental setup and baseline systems for both adult and children ASR models are described in Section 4. Section 5 presents experiment results and discussion. Section 6 analyzes the amount of adaptation data and its effect on the performance. We carry out analysis of transfer learning technique on children's age in Section 7. Section 8 discusses the study of age dependent transfer learning transformations and Section 9 provides comparisons between the age dependent and age independent transfer learning transformations. Finally, Section 10 discusses potential future work and concludes.

2. Proposed Transfer Learning Technique

Transfer learning is a method of seeding models of a new task by using the knowledge gained from a related task. The method has been used successfully, for cross-lingual knowledge transfer in DNN-based speech recognition [34, 35] and character recognition tasks [36]. Transfer learning often exploits the various level of information that are captured by the different neural network layers. Often layers closer to the signal capture signal specific characteristics, *e.g.* edge characteristics, basic shapes, or spectral content. Higher layers capture information more related to the task at hand, *e.g.* phoneme classes, object types [37].

Prior literature (see Section 1) establishes that childrens speech is significantly different to that of adults. ASR performance suffers acutely for cross-domain tasks (children vs. Adults). In this study, we consider children ASR and adult ASR as two different tasks. We attack the mismatch problem as a transfer learning between the two tasks children ASR and adult ASR. Children, as described above, differ (i) in acoustics and (ii) pronunciation from adults. This motivates us to investigate the transfer learning between adult and children ASR systems in two ways: (i) acoustic variability, as those relate to layers near input, and (ii) pronunciation variability as it relates to layers near output.

2.1. Accounting for Acoustic Variability

We assume that acoustic variability affects the lower-level network structures only and hence these layers need to be adapted to better represent the children's feature subspace. This could be thought of as retaining the knowledge of higher level abstract functions (mappings) from an adult's ASR, while accounting for the spectral variabilities. This parallels alternate approaches such as feature space transforms like VTLN, fMLLR. One important difference is the degrees of freedom and hence parameters that this technique allows, likely resulting in

better transformations but also much larger demands on adaptation data. Hence, to account for the acoustic variability we retain all the hidden layers from adult models except the bottom-most layer as shown in Figure 1. The figure comprises of two input layers, one corresponding to original task (Adult's ASR), and the second, a new estimate adapted to the target task (children's ASR). The DNN is retrained with children speech until convergence to estimate the optimal parameters of the lowest layer. Note, most of the transfer learning techniques adapt the output layers [30, 21, 31, 32] while for this task we adapt the input layer(s).

Moreover, we also augment the MFCC features with i-vector information. The i-vector subspace has been shown to capture speaker specific information efficiently [38]. It has also been successfully used for capturing speaker age characteristics [39]. Further speaker specific information is useful for speaker adaptation of DNN acoustic models [40]. The augmentation of i-vectors enables for better adaptation of the bottom layers during transfer learning by estimating speaker and age specific spectral transformations which are highly relevant for modeling children speech.

2.2. Accounting for Pronunciation Variability

We assume that phonemic variability affects the higher-level network structures only and hence these layers need to be adapted to better represent the children's pronunciation variance. Hence we propose to adapt higher layers towards modeling pronunciations as illustrated in Figure 2. The figure comprises of two output layers, one corresponding to original task (Adult's ASR), and the second, a new estimate adapted to the target task (children's ASR). This parallels work in adapting acoustic models across languages [34, 35] or for non-native speakers [33]. In this case we are only tackling the pronunciation variability and as such the lower-order layers will remain unchanged.

2.3. Accounting for Acoustic & Pronunciation Variability

Finally, to account for both the acoustic and pronunciation variability, we would like to update both the top-most and bottom-most layers and keep the rest of the layers fixed. This is attempted in two ways: (i) keeping weights of the middle hidden layers fixed and allow the top-most and bottom-most layer(s) to update simultaneously, (ii) dis-jointly and alternately training the various layers (top & bottom) until convergence. The motivation behind the disjoint training is to constrain the updatable parameters at any time, to limit the adaptation, and to regulate the amount of knowledge retained from adult acoustic models.

3. Databases

In this work we employ 5 different children speech databases and 1 adult speech corpora. All the data are processed at 16kHz.

The following children speech databases were used:

1. CU Kid's Prompted and Read Speech Corpus [41]
2. CU Kid's Read and Summarized Story Corpus [42]

3. OGI Kid's Speech Corpus [43]
4. ChIMP Corpus [16]
5. CID Children's Speech Corpus [2]

Using multiple children's speech corpora makes the problem more challenging and more relevant to real world scenarios. The CID Children's Speech Corpus is used for testing and the rest for training. The summary of breakup of databases and their split for training and testing is provided in table 1. The distribution of data over the age is illustrated in Figure 3.

The adults corpus employed in this work is the TED-LIUM ASR corpus [44]. It consists a total of 206 hours of speech data of 774 speakers giving TED talks.

4. Experimental Setup & Baseline System

4. 1. Experimental Setup

The experimental setup is very similar to the one used in our previous work [7].

GMM-HMM System: We employ as a baseline a Gaussian Mixture Model based Hidden Markov Model ASR. For this system the features used are standard Mel-Frequency Cepstral Coefficients (MFCC) of dimension 13 with window size of 25ms and shift of 10 ms with their first order and second order derivatives. The HMMs were modeled using 3 states for non-silence phones and 5 for silence phones. The GMM-HMM system consists of 3976 GMMs built from 100,124 gaussians. The choice of parameters are empirical and increasing the number of parameters didnt yield any significant improvements. We also employ the front-end adaptation techniques of Linear Discriminant Analysis (LDA), Maximum Likelihood Linear Regression (MLLR) and Feature space MLLR (fMLLR) for speaker independent and speaker adaptive training.

Dictionary: We employ the CMU Pronunciation dictionary [45]. This dictionary corresponds to American-English pronunciations and that makes it compatible with our available children and adult data. To account for the out-of-vocabulary (OOV) words during training, a grapheme to phoneme converter was used to generate phoneme transcripts for OOV words.

Language Model: Two language models were interpolated, one trained on a subset of children's training data reference transcripts and the second generic English language model from CMU-Sphinx-4[1] [46]. The interpolation helps incorporating children's grammar which is beneficial for children's ASR along with the adult's grammar to facilitate the transfer learning process between adults and children. Since this work deals with evaluating acoustic models, we keep the language model fixed for all our experiments.

Vocal Tract Length Normalization (VTLN)—VTLN is a speaker dependent transform aimed to normalize the variability found in vocal tract structures and reduce inter-speaker

[1]Language model version: cmusphinx-5.0-en-us.lm

variability. It involves computation of speaker specific frequency warping factors derived using maximum likelihood estimation. We adopt linear VTLN system [47], in which each warping factor is associated with a linear transformation, i.e.,

$$\begin{aligned} x^{\alpha} &= A^{\alpha}x + b^{\alpha} = W^{\alpha}\chi \\ W^{\alpha} &= \left[A^{\alpha}; b^{\alpha}\right] \ \chi = [x; 1] \end{aligned} \tag{1}$$

where x is the feature vector, α is the warp factor (chosen using grid search), x^{α} is the transformed feature vector for warp factor α, A^{α} is the linear transformation matrix & b^{α} is the linear bias for warp factor α, W^{α} is the affine transformation matrix and χ is the extended feature set. Linear VTLN computation involves a total of 31 × (120 × 120 + 120) = 450, 120 parameters for a MFCC feature dimension of 40 (with Δ & Δ-Δ's) over a grid search of 31 warp factors.

Feature space maximum likelihood linear regression (fMLLR)—fMLLR, also known as constrained space maximum likelihood linear regression (CMLLR) [48] is used for speaker adaptive training in this study. fMLLR is a linear model-space transformation which is computed using expectation-maximization technique. The parameters of the transformation is found as follows:

$$\begin{aligned} Q(M, \widehat{M}) = K - \\ \frac{1}{2}\sum_{s=1}^{S}\sum_{m=1}^{M}\sum_{r=1}^{T^{(s)}}\gamma_m(\tau)[K^{(m)} + log(|\widehat{\Sigma}^{(s,m)}|) \\ + (o(\tau) - \widehat{\mu}^{(s,m)})^T \widehat{\Sigma}^{(s,m)-1}(o(\tau) - \widehat{\mu}^{(s,m)})] \end{aligned} \tag{2}$$

where $\widehat{\mu}^{(s,m)}$ and $\widehat{\Sigma}^{(s,m)}$ are the transformed mean and variance for speaker s and Gaussian component m, S is the number of speakers, M is the number of Gaussian components associated with the particular transform, $O_T = o(1), \ldots, o(T)$ is the adaptation data (training data for the transform), K & $K^{(m)}$ are the normalization constant for transition probabilities and Gaussian component m respectively. $\gamma_m(\tau)$ is the posterior probability given by

$$\gamma_m(\tau) = p(q_m(\tau)|M, O_T) \tag{3}$$

where $q_m(\tau)$ is the Gaussian component m at time τ. With fMLLR, a constraint is applied such that the transformation of the mean should correspond to the transformation of the variance, i.e.,

$$\begin{aligned} \widehat{\mu} &= A'\mu - b' \\ \widehat{\Sigma} &= A'\Sigma A'^{T} \end{aligned} \tag{4}$$

The total number of parameters involved in fMLLR estimation is $S \times M \times (D \times (D + 1))$ for feature dimension D.

i-Vector Setup: The i-vector extraction can be formalized as:

$$\widetilde{F} = u + Tx \tag{5}$$

where $\tilde{F}$ is the mean super-vector, i.e., the vector of component means of the GMM obtained after maximum a-posteriori (MAP) adaptation of a universal background model (UBM), which is decomposed into u, the mean super-vector of the UBM, T, the total variability matrix spanning a low dimensional subspace estimated in a maximum likelihood sense, and x, the i-vector computed as a latent variable with standard normal prior via MAP. We employ high-resolution, 40-dimensional MFCCs as front-end features for i-vector training. To introduce context, we used an LDA transform with a context of 3 left and 3 right. Both the universal background model (UBM) and the total-variability matrix for the i-vector were trained on adults speech data to allow transfer learning from these as well. We used 2048 Gaussian components to train the UBM, whereas the i-vector dimension was fixed to 100. We experiment with 2 versions of i-vectors: (i) online i-vector, computed on the fly on a sliding window of speech signal with no look-ahead, and (ii) offline i-vector, computed on all available speech data pertaining to a specific speaker.

Hybrid DNN-HMM System: We employed a hybrid DNN-HMM system, where the DNN is used to replace the posterior probabilities of a traditional GMM system. DNN architecture employed is a time delay neural network which uses sub-sampling for exploiting long contextual information [49]. The DNN consumes high resolution MFCC features with a context of 13 left and 9 right frames. The MFCC features were concatenated with the i-vector and were used to train the DNN. The DNN has 7 hidden layers with p-norm non-linearity, each of dimension 3500, consisting of approximately 12.2 million parameters. The choice of number of parameters are inspired from the TEDLIUM recipe in Kaldi, given the amount of children data (91.6 hours) is similar to that of TEDLIUM (118 hours). The output Softmax layer consists of 3976 units trained to predict the posterior. We used greedy layer-wise training to train the DNN [50]. An exponential decay function is applied to the learning rate. To regularize the training, we use two techniques: (i) a normalization layer following each hidden activation layer [51], which normalizes the vector of activations such that the sum-square of the vector is 1.0, and (ii) a max-change limit for each hidden layer [51], to stabilize the training process. The convergence of the DNN is confirmed using a small subset of held-out training data.

Evaluation: To confirm the effectiveness of the proposed transfer learning techniques, we perform statistical significance test. We perform two statistical significance tests for (i) word error rate, and (ii) sentence error rates [52]. All the results reported in this study are statistically significant with $p < 0.001$, both in terms of word error rates and sentence error rates.

4.2. Baseline System

4.2.1. Children's ASR—The Children's ASR was trained only on the children speech data (splits illustrated in table 1). In order to compare to the DNN and to relate to the previous work [7], we provide the result of the GMM-HMM systems. To asses the advantage of the proposed transfer learning, we also trained a hybrid DNN-HMM based baseline system on children-only speech data. To provide a range of baselines we also employ popular adaptation techniques such as VTLN, SAT, i-vector, which have been proven successful for children's speech, in conjunction with the Hybrid DNN-HMM.

4.2.2. Adult's ASR—An additional ASR was trained only on adults speech data from TED-LIUM. The performance of this system was evaluated by decoding on the test set of children speech to compare its performance to that of the baseline children ASR. This system is used for transfer learning to adapt to children speech.

5. Recognition Results and Discussions

5.1. Baseline Results

Table 2 shows the results of the baseline system. The GMM-HMM results are comparable to that of the previous study [7] although more data has been incorporated for training in the current system. We see that the SAT gives the best results among the GMM-HMM framework. The hybrid DNN-HMM system improves over its respective GMM counterpart by 1% absolute. We believe the reason for the minimal improvement is that DNN requires more data to generalize well for children speech.

We also compare different adaptation techniques for the DNN-HMM model. VTLN provides an absolute 3.25% improvement over the raw MFCC features. SAT performs much better and reduces the WER to 21.31% an absolute improvement of 14.66% over raw features. However, we find that a combination of VTLN and SAT doesn't provide any major improvement. Trials augmenting raw features with i-vectors suggest that the best performance is achieved by using the offline version of i-vectors calculated on the whole utterance. However, these still fail to surpass the performance of the SAT by 4.22% absolute, thereby confirming SAT is crucial for children speech adaptation irrespective of GMM or DNN acoustic modeling.

5.2. Hypothesis Verification

In this section, we perform carefully designed experiments to prove our initial hypothesis that (i) transfer learning from adult ASR is advantageous, and (ii) adapting bottom layers of a DNN helps to address acoustic variability.

5.2.1. Transfer learning from Adults to children—We perform transfer learning from two different base models: (i) Adult's model, (ii) Combined model (Adult + Children) and compare with (iii) Children's ASR to verify our hypothesis for the need for transfer learning from adults to children. Table 4 shows the best results obtained on children's test set with each of the model. Results confirm our hypothesis that transfer learning from Adult's model leads to a more robust system.

5.2.2. Acoustic variability modeling—We conduct experiments on TEDLIUM (Adult's speech) corpus by artificially creating acoustic feature level variability. We apply (i) pitch-shifting, (ii) time-stretching on the raw speech signals and (iii) random VTLN warping on the MFCC features. We choose the variabilities due to their close relevance to the ones found in children's speech. The results of these experiments are presented in Table 3. From the table, the main observation is that the bottom-layer adaptation helps the most for handling acoustic variability. A more detailed look for each of the variability, indicates that VTLN can only be handled by adapting bottom layer and not the top-layer. We don't

observe any improvements for time-stretching with adaptations, probably because of the HMM's ability to compensate for variable speaking rates.

5.3. Transfer Learning Results

Table 5 shows results of the proposed transfer learning technique. Even-though the best performing model from Table 2 is obtained with LDA+MLLT+SAT, we choose the best performing i-vector (offline, utterance level) model as our baseline. This is because, we hope to estimate a feature level transform similar to fMLLR transform (SAT) with adaptation of bottom layers of the DNN. Since the fMLLR transform is a linear transform on the feature space, there is a possibility that the DNN adaptation is limited by the fMLLR transform. By allowing the adaptation directly on the features, we believe the DNN is able to estimate a more effective non-linear transformation of the feature space rather than being constrained to linear feature transforms like fMLLR, VTLN, LDA.

The baseline adult's model is significantly worse than children's model, as expected and consistent with previous studies. We first conduct adaptation experiments by adapting a single layer at a time. This allows us to assess the types of variability present in children's speech relative to the adult-trained DNN. It also allows us to evaluate performance benefits through addressing specific variability types. Adapting bottom layers should help counter acoustic variability in kids. Adapting top layers should attempt to account for pronunciation variability.

We observe as hypothesized that with single-layer modifications addressing acoustic variability (24.26%) is more advantageous than accounting for pronunciation variability (26.97%). Both are providing big gains over both the original adult's baseline of 39.32%.

Often, in transfer learning the top layers, representing high level abstract information, are used for adaptation [35, 34]. However, our finding is in agreement with prior studies showing high variability in spectral characteristics of children speech [1, 2, 4, 5] that denotes the need for input-layer adaptation. This suggests that the transfer learning adaptation configuration is task dependent.

We also investigate letting both the top and bottom layers update, i.e., by modeling both the acoustic and pronunciation variability simultaneously. We observe a further boost in accuracy with the WER dropping to 19.63% giving a relative gain of 23.1% over the baseline children model and 50.1% over adults model. One interesting observation is that the bottom layer adaptation (24.26% WER) is mostly complementary to the top layer adaptation (26.97% WER) and vice-versa, thus benefits from simultaneous adaptation of a bottom layer with a top layer (19.63% WER). This suggests that *the acoustic variability and the pronunciation variability are fairly exclusive of each other in case of children*. Dis-joint training doesn't provide improvements, likely due to the sufficient amounts of data to simultaneously account for the degrees of freedom of joint training. It could however be beneficial in the case of less data as we show in the subsequent section (Section 6.1).

In our experiments we also found that using 2 layers to update instead of 1 gives further improvements. We achieve a word error rate (WER) of 17.8% which is a modest 9.3% gain

over using single layers for adaptation. Subsequent experiments with more layers did not provide any significant improvements. Adapting all the layers gives the same performance of 17.8% WER. This suggests that all the variability present between the children and the adult is concentrated at the top (pronunciation level) and bottom (acoustic level) layers of the DNN in agreement with the initial hypothesis made in this work. This indicates that the underlying middle hidden layers efficiently model the basic human speech structure.

Overall, the proposed transfer learning technique outperforms the best results obtained using the baseline model trained on children's speech with SAT by a relative 16.5% (relative 54.7% improvements over the baseline adult model). The results highlight the power of transfer learning in the DNN framework in outperforming SAT, the prior best performing recipe for children ASR [7].

Finally, we compare the proposed adaptation technique against a model trained on combined data of adult's and children's speech which was proposed in [27, 26, 28, 29]. Combining adults' and children's data provides modest improvements over the baseline systems trained only on children (5.18% absolute) and adult (18.97% absolute) data. However, our proposed adaptation technique proves to be superior with 2.55% absolute improvement over the model trained on adults and children.

Informed by the above results, for the rest of this work, we experiment with four different adaptation configurations:

1. 2 layers: (bottom-most + top-most)
2. 4 layers: (2-bottom-most + 2-top-most)
3. 6 layers: (3-bottom-most + 3-top-most)
4. all layers.

We always adapt even number of layers, thus maintaining symmetry in the structure in terms of top and bottom layers for maximum performance. Moreover, from our experiments we found that adapting a single layer never surpasses the adaptation using symmetric 2 layers and thus we skip presenting those results.

5.4. Transfer learning with fMLLR transform, (Speaker adaptive training)

We perform further experiments with speaker adaptive training with fMLLR transforms. Particularly, we compare transfer learning results for fMLLR transform trained on (i) adult's speech, and (ii) children's speech. Table 6 displays the results. We observe that fMLLR transforms trained on children's speech perform significantly better than the ones trained on adult's speech. There is a huge offset in performance with fMLLR transforms trained on adult's speech and the transfer learning is unable to compensate for performance degradation offset. The transfer learning on fMLLR transforms trained on children's speech do not prove to be useful, but instead cause performance degradation compared to the baseline children ASR. These results suggest that fMLLR transforms bring constraints and are less suited for transfer learning.

6. Analysis of Amount of Adaptation Data

Figure 4 shows the transfer learning adaptation performance curve over amount of adaptation data (in terms of WER and hours). Each curve represents different adaptation architectures of the DNN in terms of number of layers used for adaptation. The following inferences can be drawn from the plot:

- The WER decays *exponentially* with increase in amount of data.
- The curves are almost always monotonically decreasing, suggesting that more adaptation data always helps. We note that the graph has not converged, meaning more data could help the adaptation further, suggesting that the constraint is still the amount of children data available.
- Any amount of children data is helpful for adaptation, as in our experiments even as low as 35 minutes of children adaptation data was found to give improvements of up-to 9.1% (relative) over the adult model.
- Adapting less number of layers yields better results for low data scenario, i.e., we find that adapting only 2 layers consistently outperforms adapting with more layers until about 25 hours of adaptation data.
- With 25 hours of adaptation data, all of the 4 curves more or less intersect suggesting that all the four architectures gives approximately the same improvements.
- For more than 25 hours of data, we find that adapting 4, 6 and all layers converge to approximately same performance in agreement of the findings in section 5.3.

6.1. Transfer Learning for low resource scenarios

Table 7 represents three extreme low data adaptation scenarios. We apply dis-joint training to account for data sparsity as explained in section 2.3. Since earlier experiments indicated that 2 layers provided maximum benefits for low data, we present the effect of dis-joint training for 1 & 2 layers only. For 1-layer, the series of experiments involved first adapting bottom layer (layer-1) with other layers fixed, then on adapting layer-2 with the rest fixed and so on. For 2-layers, the series of experiments involved first training with top and bottom layer. Fixing those weights, we then continue training with layer-2 and layer-6 to update and so on. We find that the dis-joint training further improves the adaptation for small amounts of data i.e., 35 and 45 minutes. The improvements diminish when more data is used, as in the case of 2 hours and as seen earlier in table 5. For smaller amounts of data, dis-joint training of a single layer is more beneficial (6.4% and 4.5% relative improvements for 35 and 45 minutes). For 2-layer adaptation, approximately 1.9% and 2.3% relative reduction in WER is observed with dis-joint training for 35 and 45 minutes respectively.

7. Age dependent analysis

7.1. Age vs. Adaptation layer configurations

In this section, we analyze the effect of different layer adaptation configurations on the children's age. The model is adapted on all available children data independent of age (age-

independent acoustic model). The results are plotted as a bar graph in Figure 5. We observe the following:

- Overall performance increases with increase in age, irrespective of the adaptation configuration. The two peaks corresponding to ages 12 and 13 years is probably a consequence of the acoustic model mismatch posed by relatively less training data for elder children (11 - 14 years) (See Figure 3).
- Performance is worse for younger children, consistent with past work [7].
- The adaptation configuration affects more younger children. To demonstrate this, Figure 6 shows the WER variance between the 4 configurations plotted over age. It is evident from the plot that the variance for younger children is significantly higher and decreases with increase in age. Similar peaks found in Figure 5 for ages 12 and 13 years is also apparent in variance plot.
- Younger children benefit with adaptation of more layers than older children. This aligns with the expectation that younger children manifest higher acoustic complexity and hence more parameters (layers) are necessary to capture the increased complexity. For example, from Figure 5, if 2 layers are adapted rather than all layers we have significantly fewer gains for 6 year olds than 14 year olds. This is also justified to certain extent by looking at the variances in Figure 6. This also suggests that despite the acoustic and pronunciation variability, young-children speech encodes more variability that affects the whole network.

7.2. Amount of Adaptation Data vs. Age

We also investigate the amount of adaptation data and its effect on children's age. Figure 7 shows a 3-d plot of WER over the amount of adaptation data and the children's age. Adaptation data are chosen at random and hence follow the proportions in Figure 3. We make the following inferences from the figure:

- It is evident that more the adaptation data better is the performance irrespective of age of the children.
- We see that younger children need more data to reach the same level of performance as older children. The trend is in accordance with the age, I e., as the age of children increases, less amount of adaptation data is sufficient.
- In-spite of large amount of matched-adaptation data, we observe that the performance of younger children of age 6-8 years doesn't meet that of the elder children.
- Although the adaptation data for older children is mainly mismatched (see Figure 3 for distribution of training data), they need as low as 30 minutes of adaptation data to surpass the performance of the younger children adapted on all (90 hours) of data.

7.3. Layer configurations vs. Amount of Adaptation Data vs. Age

To gain insights into the optimal adaptation strategy in terms of 4 earlier mentioned adaptation layer configurations as a function of the amount adaptation data and age of children, we plot the difference of WER between different adaptation layer configurations. Figure 8 shows a 3-d plot for difference between the WER when adapting all layers and WER when adapting only 2 layers. Any positive values indicate that adapting with 2 layers to be superior than adapting all the layers and vice-versa. We can deduce the following by looking at Figure 8:

1. Adapting 2-layers is more beneficial when adaptation data available is low. When more adaptation data is available, it is advantageous to adapt more layers. The trend is consistent over all the children ages - 6 years to 14 years which is in accordance with the finding from Section 7.1 and Section 7.2.
2. For younger children, 6 years to 11 years, we find that it is better to use fewer adaptation layers when the adaptation data available is low. The performance of the system is significantly lower when adapting with all the layers. This is because of the increased variability affecting the overall performance of the system. This is especially true when a large amount of parameters are adapted with little data, due to noise introduced from high variability. The performance of the system eventually recovers and surpasses the 2-layer adaptation configuration when sufficient amount of adaptation data is available.
3. For younger children, with sufficiently high adaptation data, we find that the effective gains made between the layer configuration is much higher compared to elder children. Thereby asserting their sensitivity to adaptation data and layer configurations.
4. For older children, 12 years to 14 years, the system adapts rapidly with considerably less data compared to younger children.
5. For older children, the performance gains are comparable between 2-layer adaptation and all layer adaptation.

The analysis is only presented for differences between adapting all the layers and adapting only 2-layers. The particular plot was chosen to illustrate the differences as in an extreme case. Similar trends were observed for differences of other configurations, i.e., adapting more layers versus fewer layers.

The observations in the section make sense from an acoustic point of view, because as the children grow their vocal tracts mature and voices tend to sound more like that of an adult. Therefore, the adaptation is less important in this case.

8. Analysis of Age Dependent Transformations

In order to assess the validity of the transformations learnt by adapting the layers and its extensibility and relevance to children’s speech, we analyze age specific transformations resulting from age dependent transfer learning adaptation. The transformations would be

meaningful if there exists some level of meaningful portability between different ages. Note: these transformations are not equivalent and shouldn't be mistaken to age specific models.

Figure 9 shows the 3-d plot of WER from application of age dependent transformations on each age group, when adapting the model with all the layers. The following can be inferred from the plot:

1. For younger children, ages 6 years to 10 years, the matched models i.e., application of same aged transformations provide significant improvements.
2. For younger children, as the mismatch increases (in terms of age), the performance decreases.
3. For younger children, the rate of performance degradation is much more drastic as the mismatch (in terms of age) increases compared to older children.
4. For ages 11 years to 14 years, the surface is more or less plateaued, this is probably because of data scarcity for estimation of meaningful transformations (See Figure 3).
5. The overall surface is tilted towards the left, indicating that performance of elder children are significantly better irrespective of the applied transformation.

The above observations confirm the validity of the transformations and its portability across the ages. Although the transformations are not equivalent to age-dependent models, the above observations prove they exhibit similar trends (performance-wise) as reported in previous literature [13].

8.1. Age dependent transformations versus Adaptation layer configurations

Figure 10 illustrates the confusion matrix obtained by the application of age dependent transformations on each group for each of the 4 adaptation configurations. A quick inspection shows that all of the configurations exhibit similar trends observed in Section 8.

9. Age dependent transformations versus Age independent transformations

Figure 11 compares the performance of the age independent transformations (obtained by adapting on all the data) against the application of matched age dependent transformations. To keep the analysis consistent over different adaptation layer configurations, we consider only the configuration of adapting all the layers. We find that the age independent transformation trained on significantly more data outperforms the age dependent transformations consistently over all the ages. This finding suggests that DNN can exploit more data to offset and surpass the performance and effectively generalize over different ages due to its large parameter space. It does not lose its generalizability when exposed to different ages. This is in contrast to GMM models, that when adapted (e.g. via MLLR), to a wider diverse population with limited data underperform specific adaptations [53]. By examining the difference between the WER trajectories over age, we find a peak over the

ages 11 years to 14 years, highlighting the aforementioned effect of limited data for these age groups as in Figure 3.

However, by providing a correction factor to compensate for the difference between the amount of data between the age dependent and independent transforms, enables for a more fair comparison between the transforms. To enable such an analysis we adopt 2 different types of data correction factors for age independent transformation:

1. We train an age independent model by randomly sampling data equal to the average data (over ages - Figure 3) which in our case is approximately 10 hours. We refer to this as average age-independent transform (Blue line in Figure 12). We then compute the performance on test set for each age category.
2. We train multiple age independent transforms, each trained with the corresponding amount of data available in each age category (see Figure 3). This gives us one age-independent model for each age best matched to age-dependent transform in terms of adaptation data. We refer to this as matched age-independent transform (Green line in Figure 12).

Since the sampling of data in either case is random, this retains the original corpus proportions (with respect to age).

Figure 12 compares the data normalized age independent transform against the age dependent transformations. The following observations are apparent from the plot:

1. After normalizing the amount of data, we now see that the age dependent transformations outperform the age independent transformations for younger children (ages 6 years to 10 years) in both cases (average and matched versions).
2. We observe that the improvements from age dependent transforms gets more prominent as the age decreases, with maximum gains for 6 year old.
3. We observe a crossover for elder children (ages 11 years to 14 years) in both the cases of average and matched versions, i.e., the age independent transformations are better compared to that of age dependent. (For elder children, the average version of age independent transformation shows higher demarcation due to the heightened mismatch in adaptation data. Hence, the matched version is more representative.). This interesting finding could be attributed towards the higher similarity between the speech of adults and elder children. (Note: this is not a case of age-dependent acoustic modeling, but rather an adaptation from adult's speech).
4. Looking at the difference between the 'best performing' age-independent transformation and the age dependent transform, i.e., the potential gains from exploiting more data with age-independent transform increases with increase in age. This is expected, considering that the elder children exhibit relatively lower variability in acoustic and pronunciation constructs and hence exhibit much similar speech structure to that of the adult.

9.1. Effect of adaptation layer configurations

Figure 13 plots the difference between the age-dependent transform and the matched version of age-independent transform for different adaptation layer configurations. The takeout from the plot is, the age-dependent transforms outperform the age-independent transforms for younger children, whereas the age-independent transforms are beneficial for elder children. The trend observed earlier, with all the layers, remains apparent over all the layer configurations. Note the absolute values (trajectories) are a function of the amount of data present for each age and age itself, as supported by our earlier observations in Section 7.3. Hence, the inter-relations of different configuration trajectories is complex.

10. Conclusion & Future Work

In this study, we conduct an analysis of LVCSR adaptation and transfer learning for children's speech using multiple databases. We compare the advantages of DNN acoustic models over the GMM-HMM systems. We also compare adult and children DNN acoustic model performance for decoding children speech. Several transfer learning techniques are evaluated, on adult models, specifically to address the increased acoustic variability and pronunciation variability found in children. Extensive analysis is performed to study the effect of the amount of adaptation data, DNN transfer learning configurations and their impact on different age groups. In the case of severely limited in-domain (kids) data we proposed and analyzed disjoint adaptation. We also analyzed the amount of adaptation data required for children of different ages. We investigated various transfer learning configurations and their effect on different age groups and data sizes. Our work validated the benefits of age dependent transfer learning and examined the portability and extensibility of models over the different age groups. We also presented comparisons of age dependent and age independent transfer learning. These provide valuable insights towards future research directions in terms of persisting challenges and problems in children's speech recognition.

In future we would like to analyze the variability internal to the DNN, i.e. how the weights of the "adapted layers" change. Comparisons of such variability between adult and child models can inform on linguistic and structural aspects of kids speech. These can also help identify the aspects of non-linearities in normalization and adaptation techniques towards improved kids speech processing. Such models can provide insights in analyzing the effect on various speech parameters in regards to pitch, intensity, voice quality, duration, formant frequencies etc, which are valuable aspects in assessing the difficulties faced for children ASR.

Acknowledgments

Financial Support

The U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick MD 21702-5014 is the awarding and administering acquisition office. This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs through the Psychological Health and Traumatic Brain Injury Research Program under Award No. W81XWH-15-1-0632.

Appendices

Appendix

Analysis on Balanced-Age data

In this section we present results on age-balanced set of data. Although, ideally, we would want to perform analysis with age balanced data over the entire range of ages, i.e., 6 years to 14 years, we are restricted in terms of available speech data for children of each age category. From Figure 3, it is evident we have less than an hour of data available for balanced analysis. Thus, we perform analysis only over the range of 6 years to 10 years. This gives us approximately 11.5 hours of data in each of the 5 age categories.

Appendix A.: Age vs. Adaptation layer configurations

We repeat the experiments of Section 7.1 here with age-balanced data. The model is adapted on 11.5 hours of children's speech of each age category (from 6 years to 10 years) and tested on matched age category. Figure A.14 presents the analysis of adaptation layer configurations as a function of age. Following observations are made:

- Similar trend is evident as Figure 5, i.e., performance increases with increase in children age.
- Adapting all layers is beneficial for children of age 6, i.e., more parameters are required to capture high acoustic variabilities found in younger children.
- For children of ages 7 and 8, adapting 6 layers provides the best performance, hinting at decrease in variabilities and hence relatively lesser parameters required for better adaptation.
- For children of ages 9 and 10, adaptation with only 2 layers provide optimal performance (for 11.5 hours of speech), supporting previous observations.

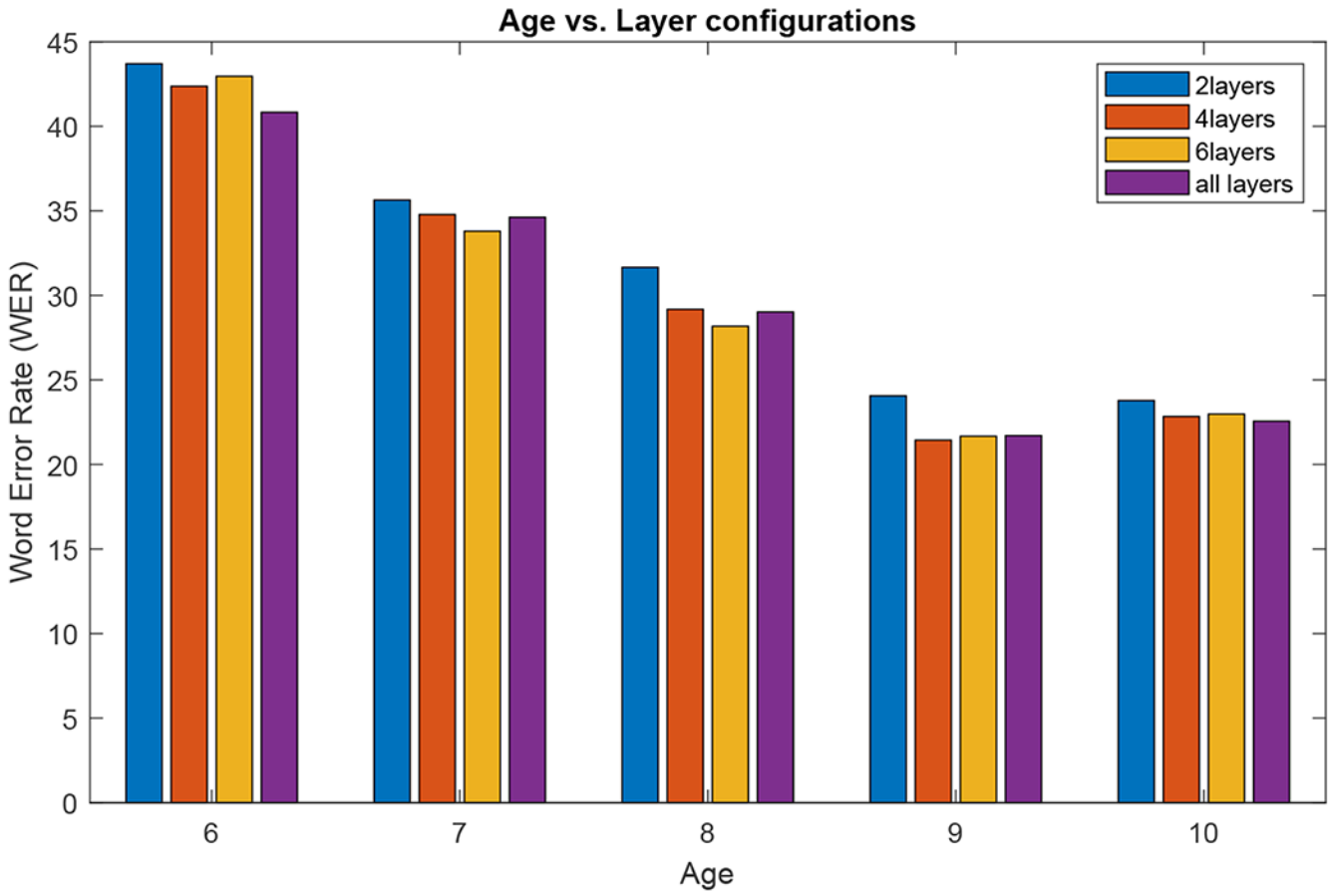

Figure A.14:
Children Age vs. Adaptation layer configurations (Age Balanced Data)

Appendix B.: Analysis of Age Dependent Transforms

In this section, we repeat the experiments from Section 8 but with balanced amount of data in each age category. Figure B.15 illustrates the WER 3-D plot for each of the age dependent models against each test categories. The following observations can be made:

- The observations are consistent with the ones made in Section 8 for the ages (6-10 years).
- The surface plot is much smoother, exhibits better linearity and is more predictable compared to Figure 9.
- Age dependent transforms are critical for younger children (6-8 years), higher slope in the surface plot along both axes, versus the elder ones (9-10 years) where the performance plateaus.

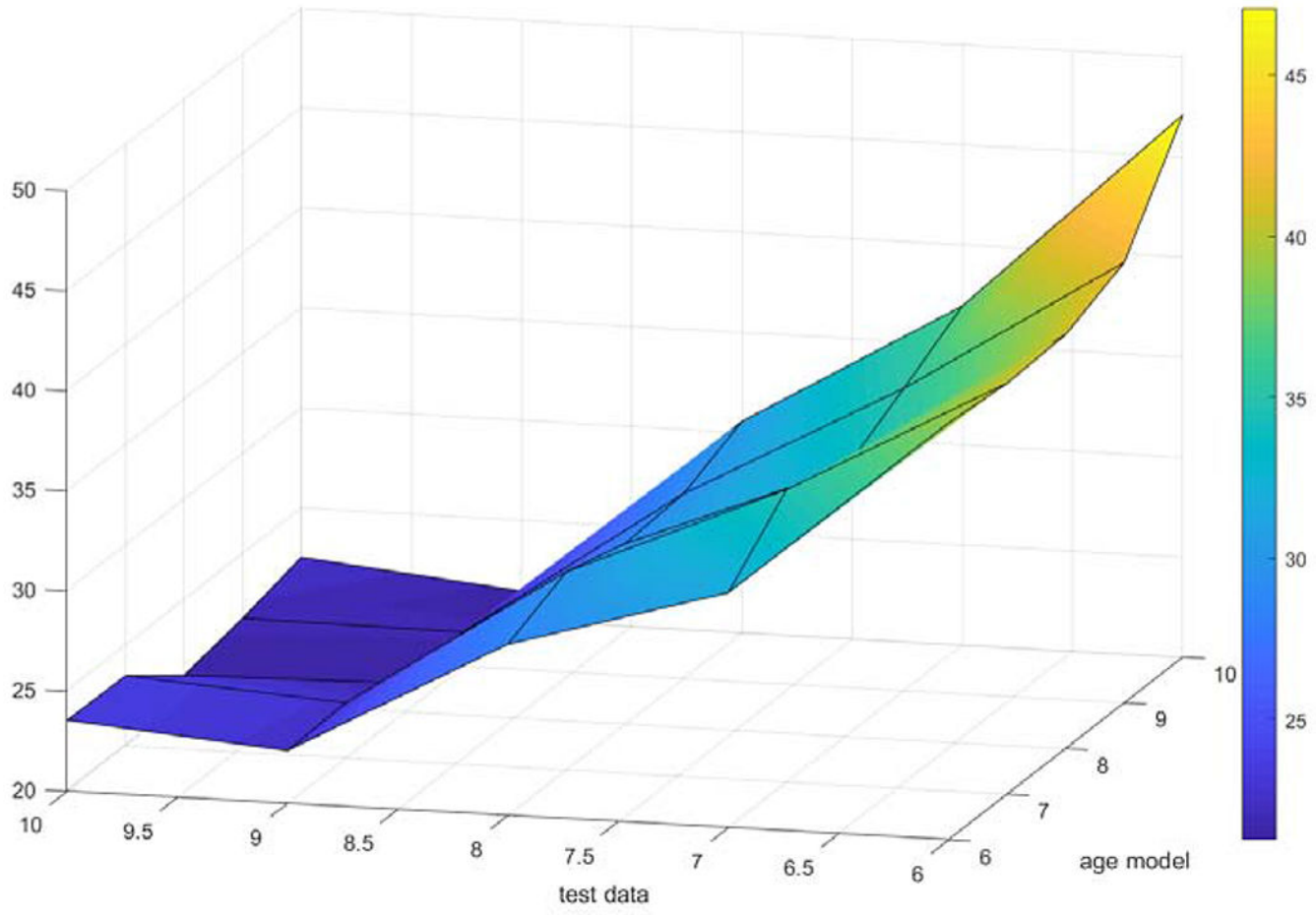

Figure B.15:
Age dependent model performance - Adapting all layers (Age balanced data)

Appendix C.: Age dependent transformations vs. Adaptation layer configurations

In this section, we repeat the experiments from Section 8.1 with age balanced data. The observations are similar to as in Section 8.1. In addition the following observations are made:

- The models adapted with all layers on younger children (age 6 & 7) tend to be less generalizable on elder children (age 9 & 10). We believe as per our earlier hypothesis adaptation with all layers increases the noise in the model and hence is less robust overall for other ages.
- The models adapted with only 2 layers on elder children (age 8-10 years) perform considerably worse on younger children compared to models adapted with more layers. This agrees with our earlier findings that younger children exhibit more complexity and hence need more parameter adaptation.

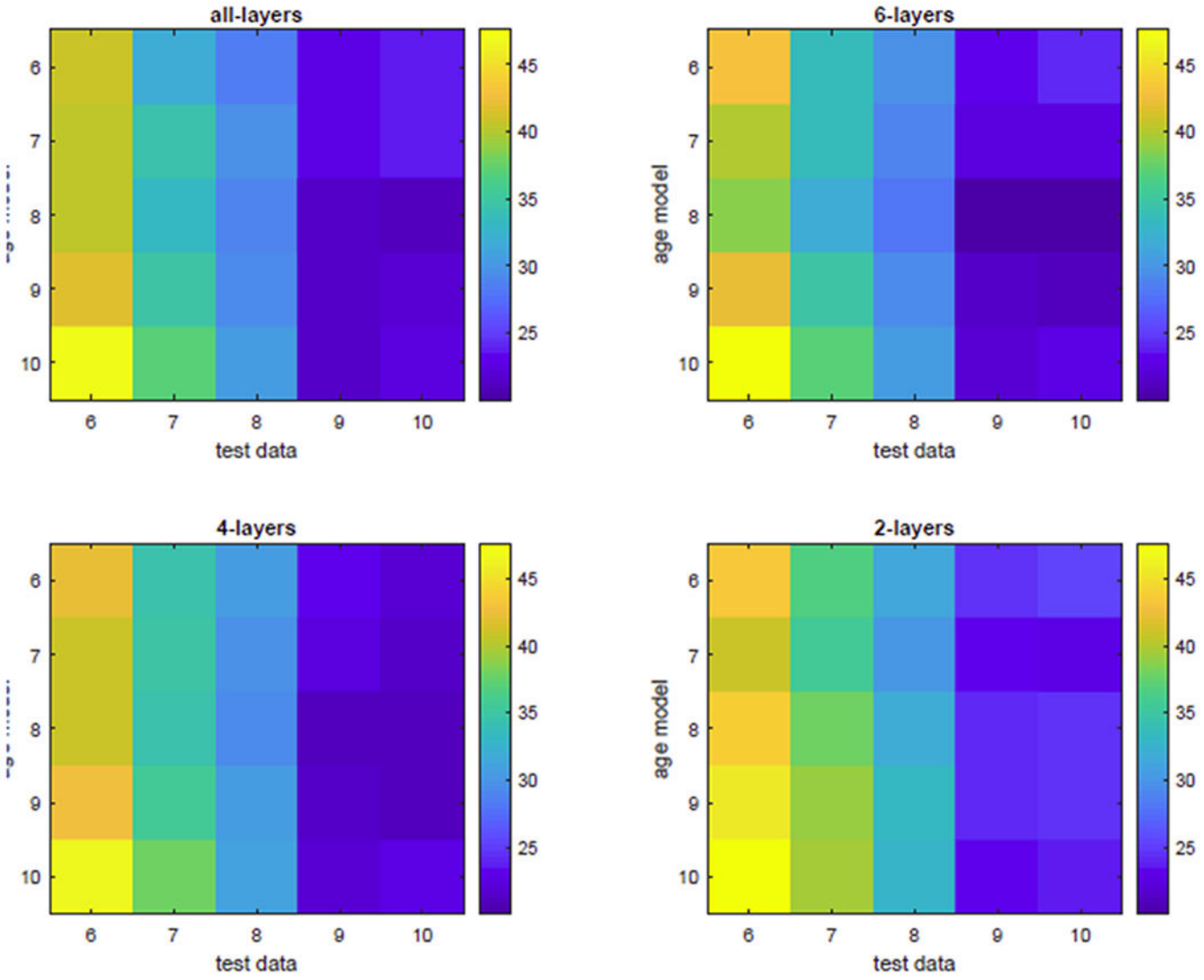

Figure C.16:
Age dependent model performances - All layer configurations (Balanced Age) Colorbar pertains to WER

References

[1]. Potamianos A, Narayanan S, Robust recognition of children's speech, IEEE Transactions on speech and audio processing 11 (6) (2003) 603–616.

[2]. Lee S, Potamianos A, Narayanan S, Acoustics of childrens speech: Developmental changes of temporal and spectral parameters, The Journal of the Acoustical Society of America 105 (3) (1999) 1455–1468. [PubMed: 10089598]

[3]. Gerosa M, Giuliani D, Narayanan S, Acoustic analysis and automatic recognition of spontaneous children's speech, in: Ninth International Conference on Spoken Language Processing, 2006.

[4]. Potamianos A, Narayanan S, Lee S, Automatic speech recognition for children., in: Eurospeech, 1997.

[5]. Li Q, Russell MJ, An analysis of the causes of increased error rates in children's speech recognition, in: Seventh International Conference on Spoken Language Processing, 2002.

[6]. Russell QLMJ, Why is automatic recognition of children's speech difficult? .

[7]. Shivakumar PG, Potamianos A, Lee S, Narayanan S, Improving speech recognition for children using acoustic adaptation and pronunciation modeling, in: Proc. Workshop on Child, Computer and Interaction (WOCCI), 2014.

[8]. Umesh S, Sinha R, A study of filter bank smoothing in MFCC features for recognition of children's speech, IEEE Transactions on audio, speech, and language processing 15 (8) (2007) 2418–2430.

[9]. Ghai S, Sinha R, Pitch adaptive MFCC features for improving childrens mismatched ASR, International Journal of Speech Technology 18 (3) (2015) 489–503.

[10]. Shahnawazuddin S, Dey A, Sinha R, Pitch-Adaptive Front-End Features for Robust Children's ASR., in: INTERSPEECH, 3459–3463, 2016.

[11]. Giuliani D, Gerosa M, Investigating recognition of children's speech, in: Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on, vol. 2, IEEE, II–137, 2003.

[12]. Stemmer G, Hacker C, Steidl S, Nöth E, Acoustic normalization of children's speech., in: INTERSPEECH, 2003.

[13]. Elenius D, Blomberg M, Adaptation and Normalization Experiments in Speech Recognition for 4 to 8 Year old Children., in: Interspeech, 2749–2752, 2005.

[14]. Gray SS, Willett D, Lu J, Pinto J, Maergner P, Bodenstab N, Child automatic speech recognition for US English: child interaction with living-room-electronic-devices, in: Proceedings of workshop on child computer interaction (WOCCI), 2014.

[15]. Gerosa M, Giuliani D, Narayanan S, Potamianos A, A review of ASR technologies for children's speech, in: Proceedings of the 2nd Workshop on Child, Computer and Interaction, ACM, 7, 2009.

[16]. Potamianos A, Narayanan S, Spoken dialog systems for children, in: Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, vol. 1, IEEE, 197–200, 1998.

[17]. Das S, Nix D, Picheny M, Improvements in children's speech recognition performance, in: Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, vol. 1, IEEE, 433–436, 1998.

[18]. Fringi E, Lehman JF, Russell M, Evidence of phonological processes in automatic recognition of children's speech, in: Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[19]. Tepperman J, Lee S, Narayanan SS, Alwan A, A generative student model for scoring word reading skills, IEEE Transactions on Audio, Speech, and Language Processing 19 (2) (2011) 348–360.

[20]. Tulsiani H, Swarup P, Rao P, Acoustic and language modeling for children's read speech assessment, in: Communications (NCC), 2017 Twenty-third National Conference on, IEEE, 1–6, 2017.

[21]. Tong R, Chen NF, Ma B, Multi-Task Learning for Mispronunciation Detection on Singapore Childrens Mandarin Speech, Proc. Interspeech 2017 (2017) 2193–2197.

[22]. Hagen A, Pellom B, Cole R, Highly accurate childrens speech recognition for interactive reading tutors using subword units, speech communication 49 (12) (2007) 861–873.

[23]. Giuliani D, BabaAli B, Large Vocabulary Children's Speech Recognition with DNN-HMM and SGMM Acoustic Modeling, in: Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[24]. Cosi P, A KALDI-DNN-based ASR system for Italian, in: 2015 International Joint Conference on Neural Networks (IJCNN), IEEE, 1–5, 2015.

[25]. Serizel R, Giuliani D, Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition, in: Spoken Language Technology Workshop (SLT), 2014 IEEE, IEEE, 135–140, 2014.

[26]. Liao H, Pundak G, Siohan O, Carroll M, Coccaro N, Jiang Q-M, Sainath TN, Senior A, Beaufays F, Bacchiani M, Large vocabulary automatic speech recognition for children .

[27]. Qian M, McLoughlin I, Quo W, Dai L, Mismatched training data enhancement for automatic recognition of children's speech using DNN-HMM, in: Chinese Spoken Language Processing (ISCSLP), 2016 10th International Symposium on, IEEE, 1–5, 2016.

[28]. Fainberg J, Bell P, Lincoln M, Renals S, Improving Children's Speech Recognition Through Out-of-Domain Data Augmentation., in: INTERSPEECH, 1598–1602, 2016.

[29]. Qian Y, Wang X, Evanini K, Suendermann-Oeft D, Improving DNN-Based Automatic Recognition of Non-native Children Speech with Adult Speech, in: Workshop on Child Computer Interaction, 40–44, 2017.

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

[31]. Serizel R, Giuliani D, Deep neural network adaptation for childrens and adults speech recognition, in: Proc. of the First Italian Computational Linguistics Conference, 2014.

[32]. Serizel R, Giuliani D, Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children, Natural Language Engineering 23 (3) (2017) 325–350.

[33]. Matassoni M, Gretter R, Falavigna GD, Daniele and, NON-NATIVE CHILDREN SPEECH RECOGNITION THROUGH TRANSFER LEARNING, Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on (2018) 6229–6233.

[34]. Heigold G, Vanhoucke V, Senior A, Nguyen P, Ranzato M, Devin M, Dean J, Multilingual acoustic models using distributed deep neural networks, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 8619–8623, 2013.

[35]. Huang J-T, Li J, Yu D, Deng L, Gong Y, Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 7304–7308, 2013.

[36]. Cireşan DC, Meier U, Schmidhuber J, Transfer learning for Latin and Chinese characters with deep neural networks, in: The 2012 International Joint Conference on Neural Networks (IJCNN), IEEE, 1–6, 2012.

[37]. Bengio Y, Courville A, Vincent P, Representation learning: A review and new perspectives, IEEE transactions on pattern analysis and machine intelligence 35 (8) (2013) 1798–1828. [PubMed: 23787338]

[38]. Dehak N, Kenny P, Dehak R, Dumouchel P, Ouellet P, Front-end factor analysis for speaker verification, Audio, Speech, and Language Processing, IEEE Transactions on 19 (4) (2011) 788–798.

[39]. Shivakumar PG, Li M, Dhandhania V, Narayanan SS, Simplified and supervised i-vector modeling for speaker age regression, in: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, IEEE, 4833–4837, 2014.

[40]. Saon G, Soltau H, Nahamoo D, Picheny M, Speaker adaptation of neural network acoustic models using i-vectors., in: ASRU, 55–59, 2013.

[41]. Cole R, Hosom P, Pellom B, University of colorado prompted and read childrens speech corpus, Tech. Rep., Technical Report TR-CSLR-2006-02, Center for Spoken Language Research, University of Colorado, Boulder, 2006.

[42]. Cole R, Pellom B, University of Colorado read and summarized story corpus, Tech. Rep., Technical Report TR-CSLR-2006-03, University of Colorado, 2006.

[43]. Shobaki K, Hosom J-P, Cole R, The OGI kids' speech corpus and recognizers, in: Proc. of ICSLP, 564–567, 2000.

[44]. Rousseau A, Deléglise P, Estève Y, Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks., in: LREC, 3935–3939, 2014.

[45]. Weide R, The CMU pronunciation dictionary, release 0.6, 1998.

[46]. Walker W, Lamere P, Kwok P, Raj B, Singh R, Gouvea E, Wolf P, Woelfel J, Sphinx-4: A flexible open source framework for speech recognition .

[47]. Kim D, Umesh S, Gales M, Hain T, Woodland P, Using VTLN for broadcast news transcription, in: Eighth International Conference on Spoken Language Processing, 2004.

[48]. Gales MJ, Maximum likelihood linear transformations for HMM-based speech recognition, Computer speech & language 12 (2) (1998) 75–98.

[49]. Peddinti V, Povey D, Khudanpur S, A time delay neural network architecture for efficient modeling of long temporal contexts, in: Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[50]. Bengio Y, Lamblin P, Popovici D, Larochelle H, et al., Greedy layer-wise training of deep networks, Advances in neural information processing systems 19 (2007) 153.

[51]. Povey D, Zhang X, Khudanpur S, Parallel training of dnns with natural gradient and parameter averaging, arXiv preprint arXiv:1410.7455 .

[52]. Gillick L, Cox SJ, Some statistical issues in the comparison of speech recognition algorithms, in: International Conference on Acoustics, Speech, and Signal Processing,, IEEE, 532–535, 1989.

[53]. Gales MJ, Woodland PC, Mean and variance adaptation within the MLLR framework, Computer Speech & Language 10 (4) (1996) 249–264.

Highlights

- In this work, we conduct Evaluations on large vocabulary continuous speech recognition (LVCSR) for children, to:
- We Compare older GMM-HMM models and newer DNN models.
- Investigate different transfer learning adaptation techniques.
- Assess effectiveness of different speaker normalization and adaptation techniques like VTLN, fMLLR, i-vector based adaptation versus the employed transfer learning technique.
- Further, we conduct Analysis over the following parameters in context of transfer learning:
- DNN model parameters.
- Amount of adaptation data.
- Effect of children's ages.
- Age dependent transformations obtained from transfer learning and their validity, portability over the children's age span.
- We finally provide Recommendations on:
- Favorable transfer learning adaptation strategies for low data and high data scenarios.
- Suggested transfer learning adaptation techniques for children of different ages.
- Amount of adaptation data required for efficient performance over children's ages.
- Potential future research directions and relevant challenges and problems persisting in children speech recognition.

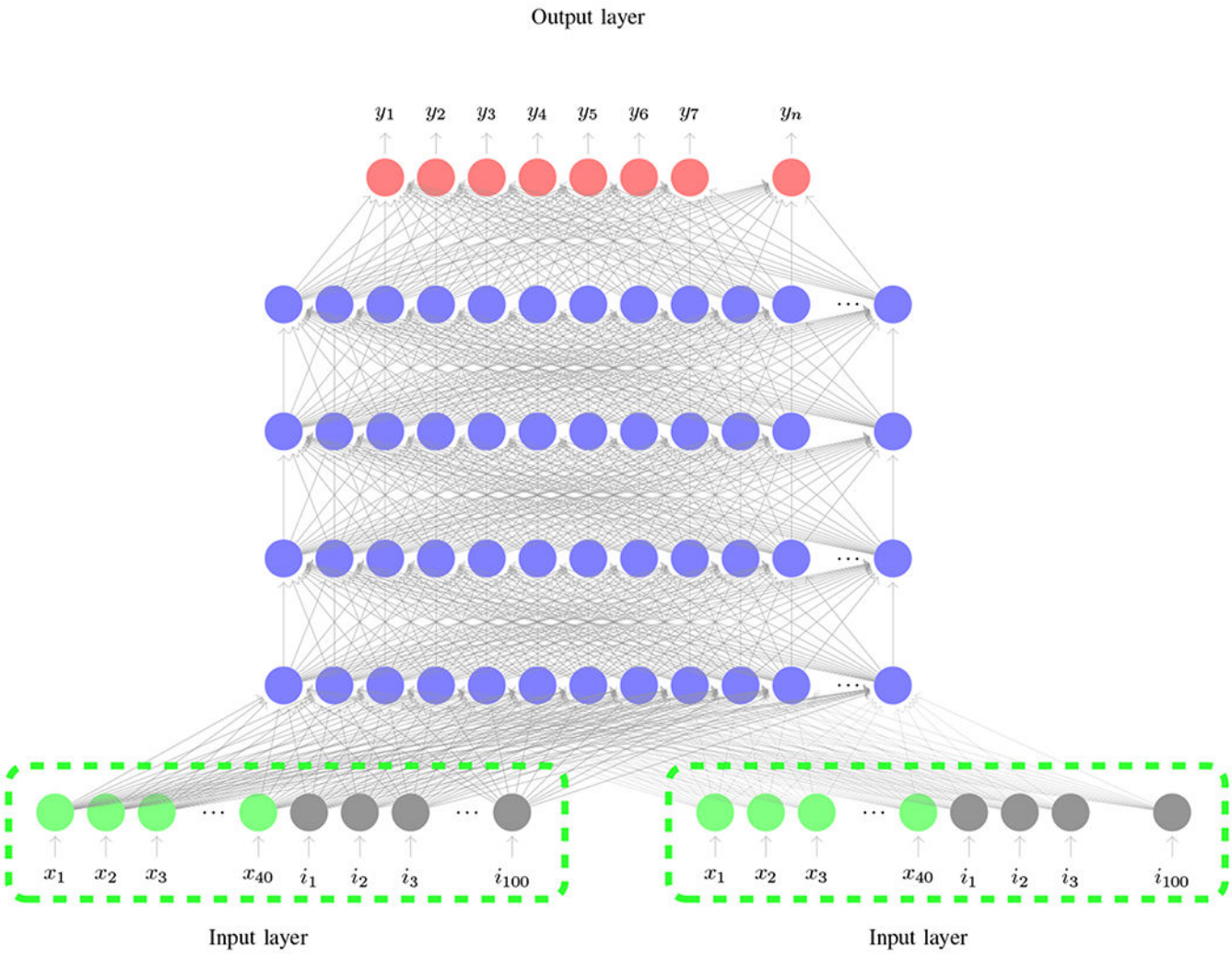

Figure 1:

Acoustic Variability Modeling

Neuron color scheme: Red-Output, Blue-Hidden, Gray-ivector input, Green-MFCC input

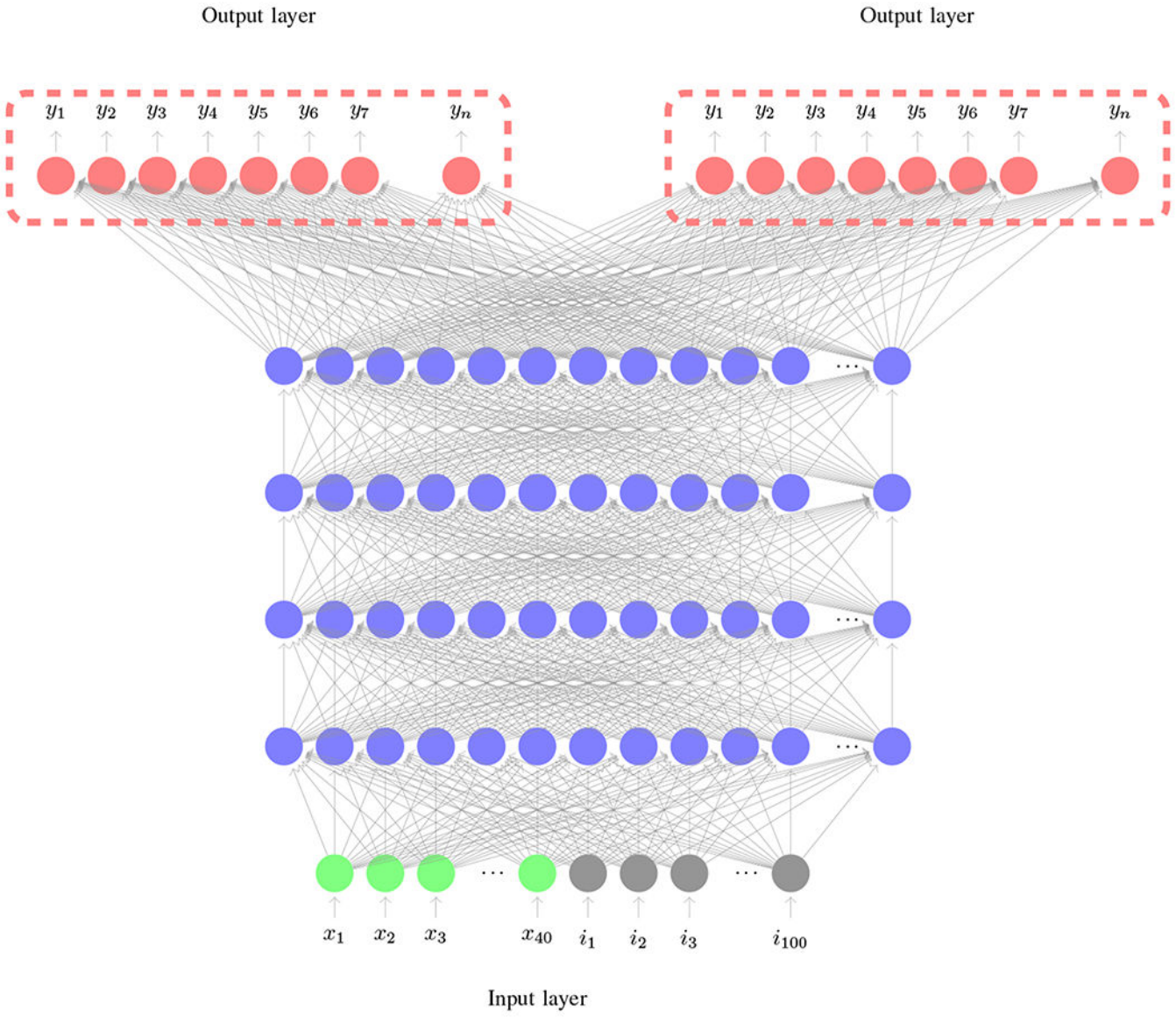

Figure 2:
Pronunciation Variability Modeling
Neuron color scheme: Red-Output, Blue-Hidden, Gray-ivector input, Green-MFCC input

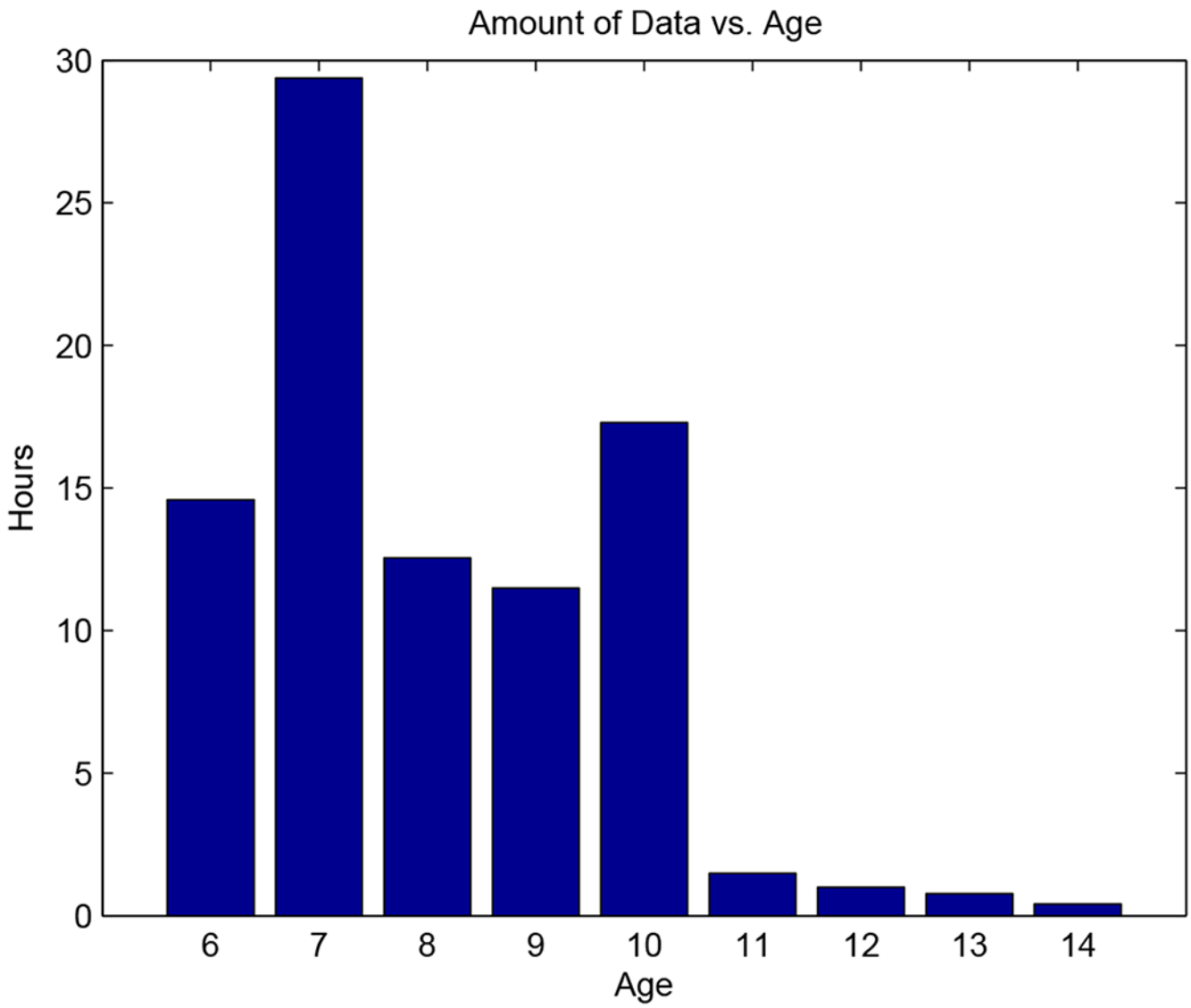

Figure 3:
Distribution of training data over Children Age

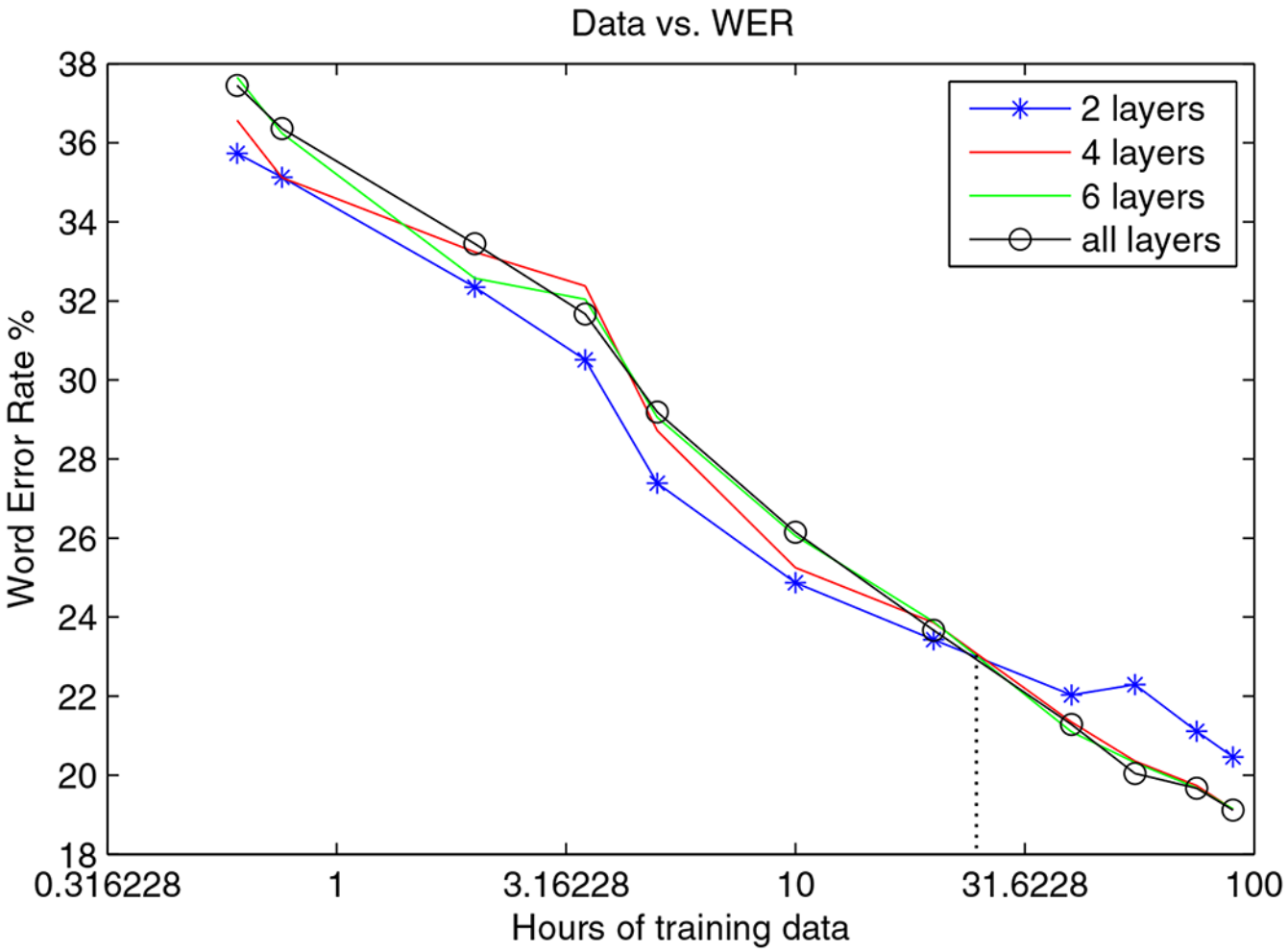

Figure 4:
Amount of Adaptation Data (Log-scale) versus Word Error Rate; Four Different DNN configurations

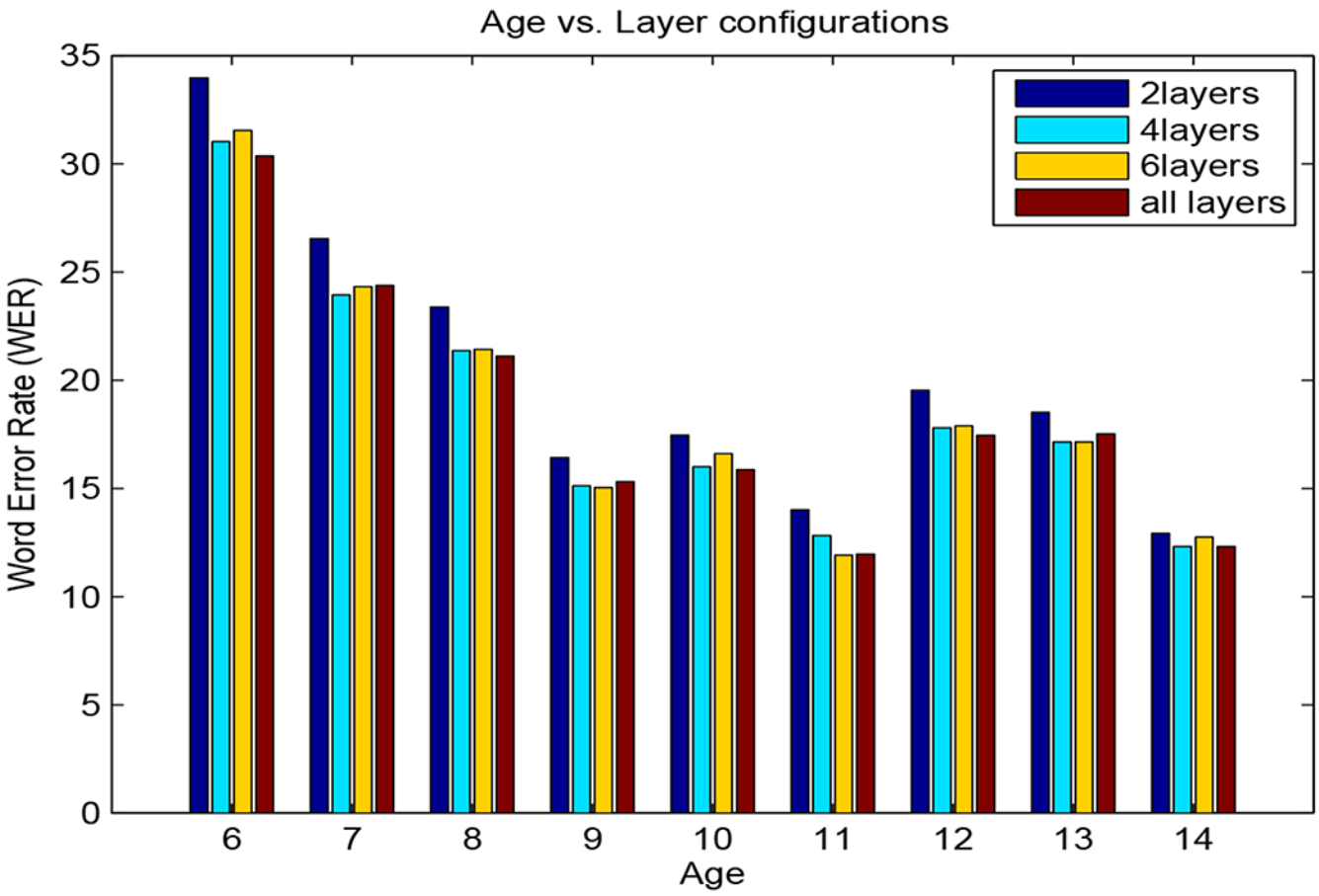

Figure 5:
Children Age vs. Adaptation layer configurations

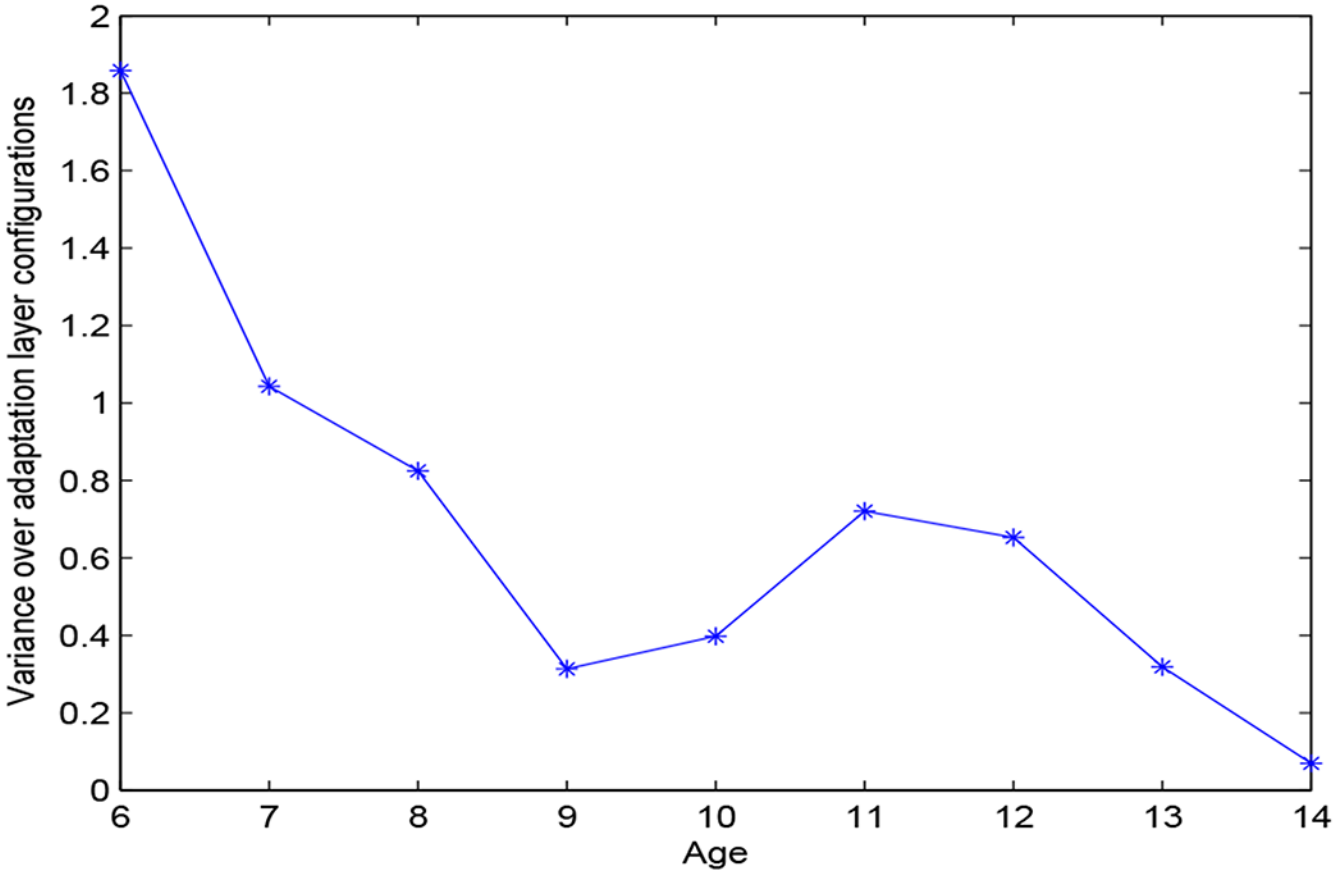

Figure 6:
WER Variance over Adaptation Layer Configurations across Children Age Groups

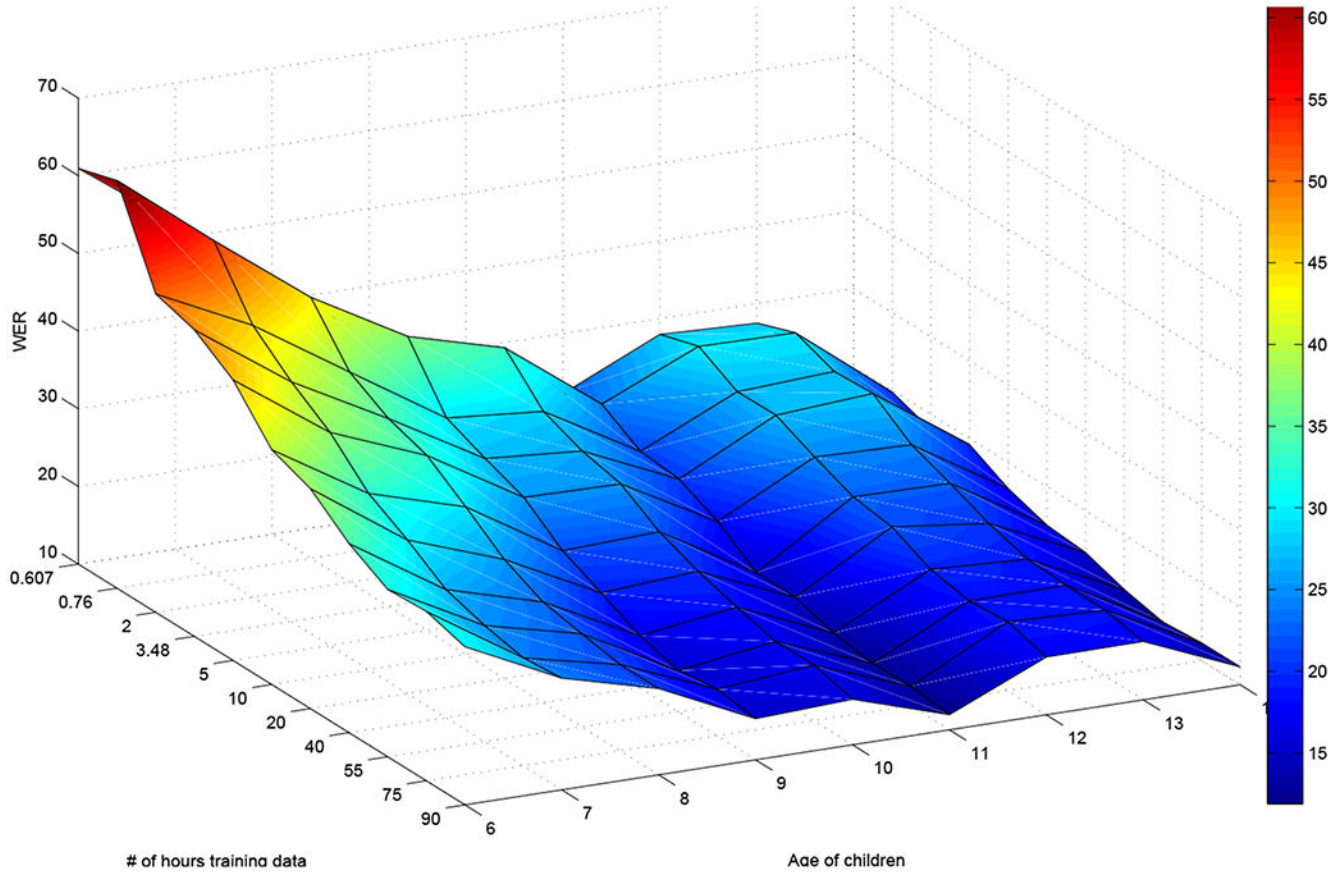

Figure 7:
Amount of training data vs. Children Age

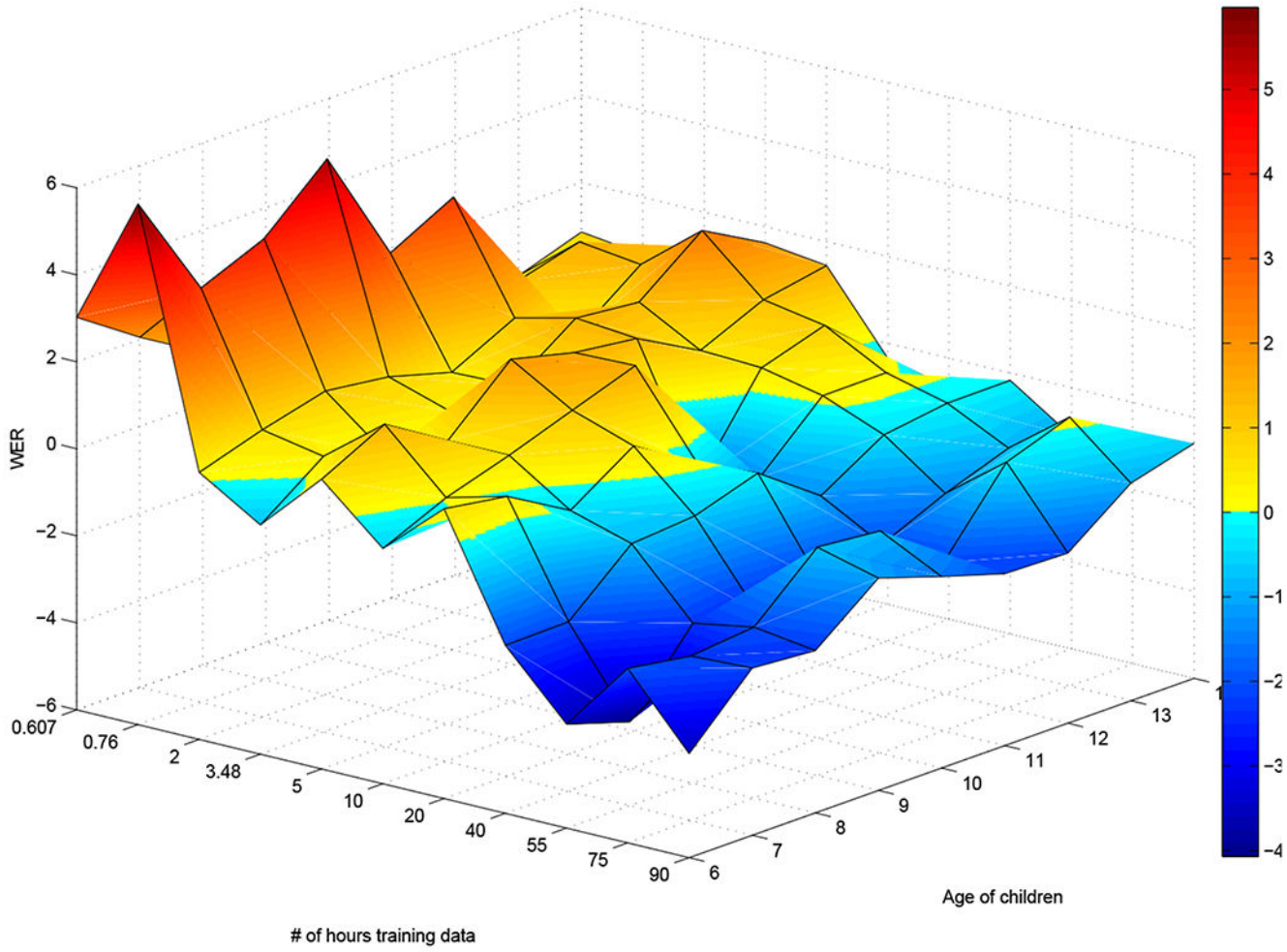

Figure 8:
Layer configurations vs. Amount of Adaptation Data vs. Age

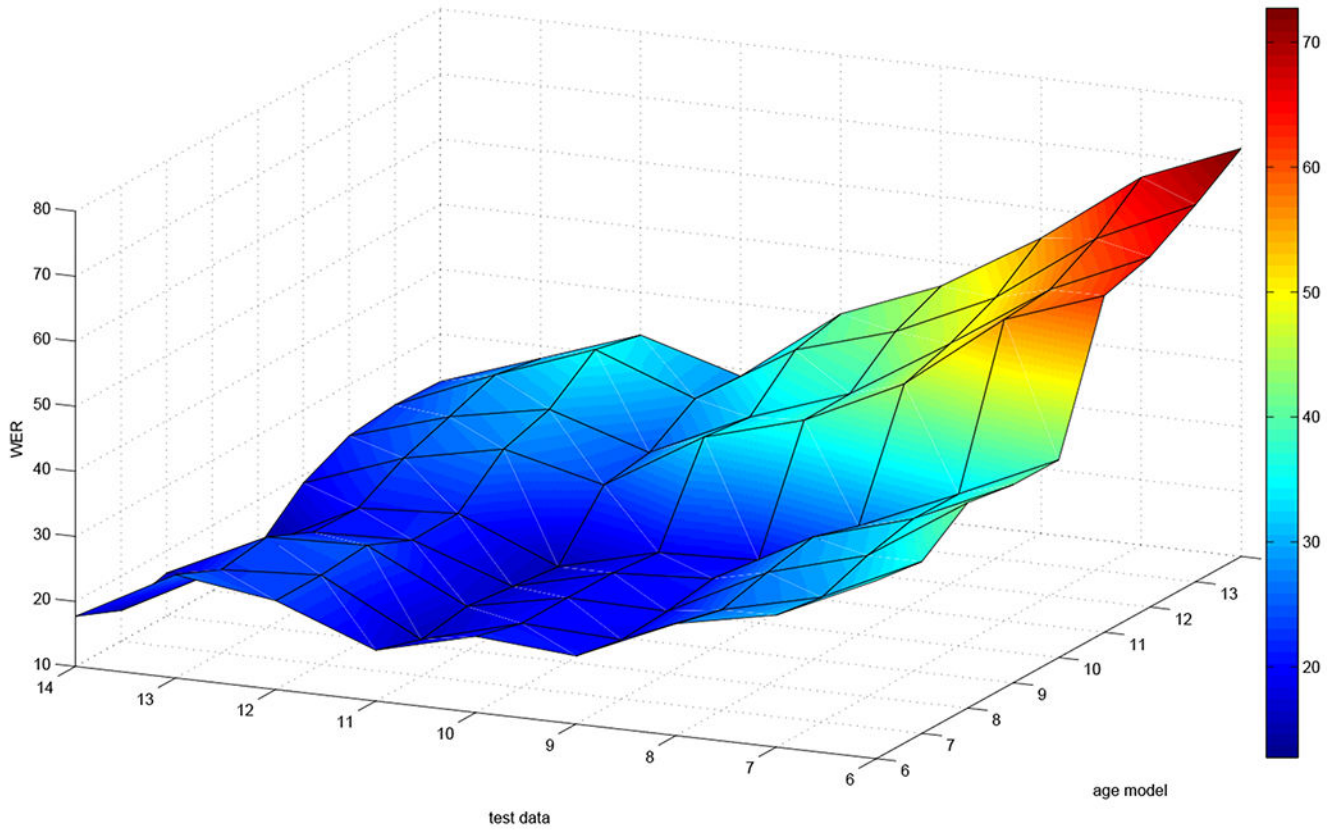

Figure 9:
Age dependent model performance - Adapting all layers

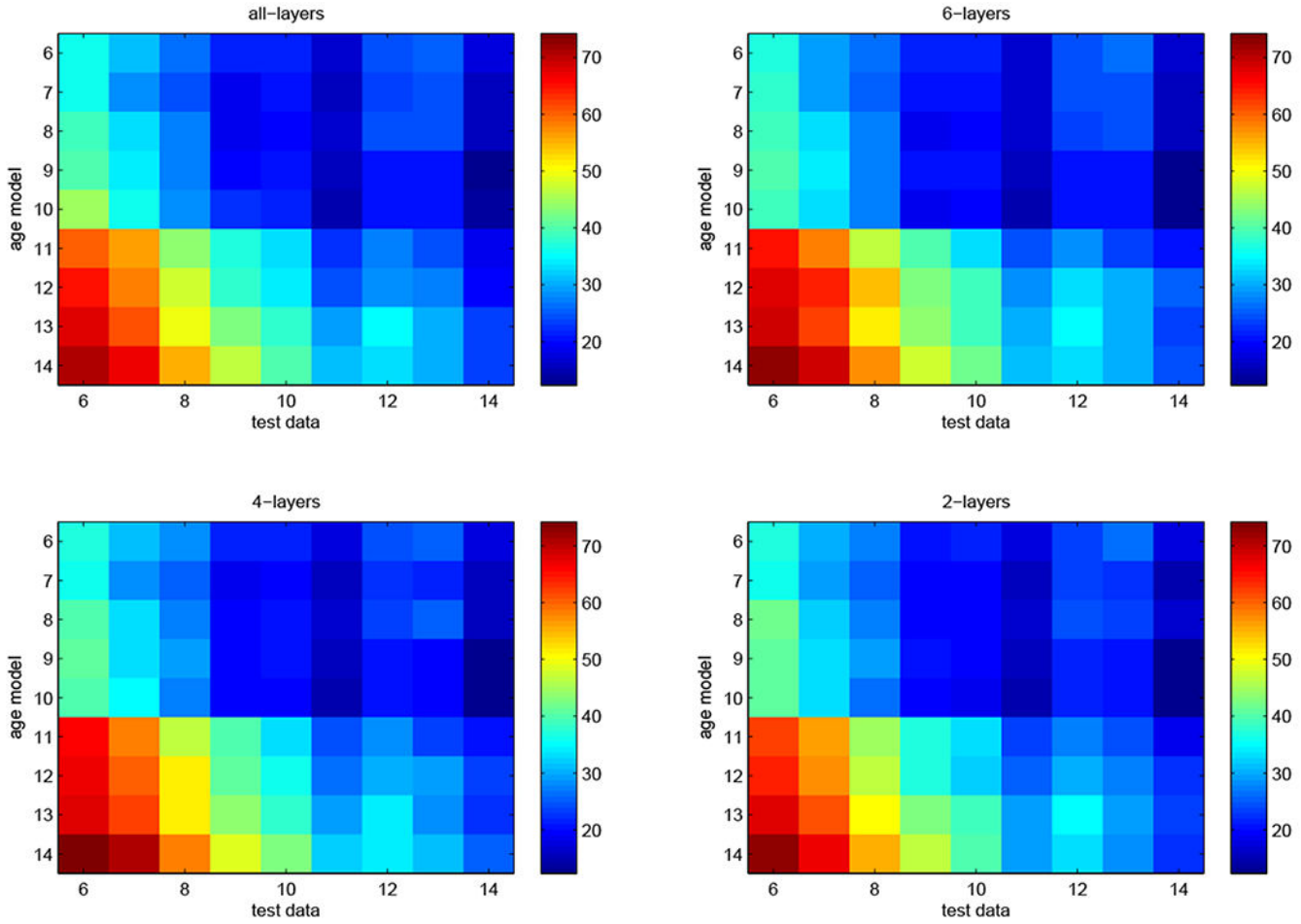

Figure 10:
Age dependent model performances - All layer configurations Colorbar pertains to WER

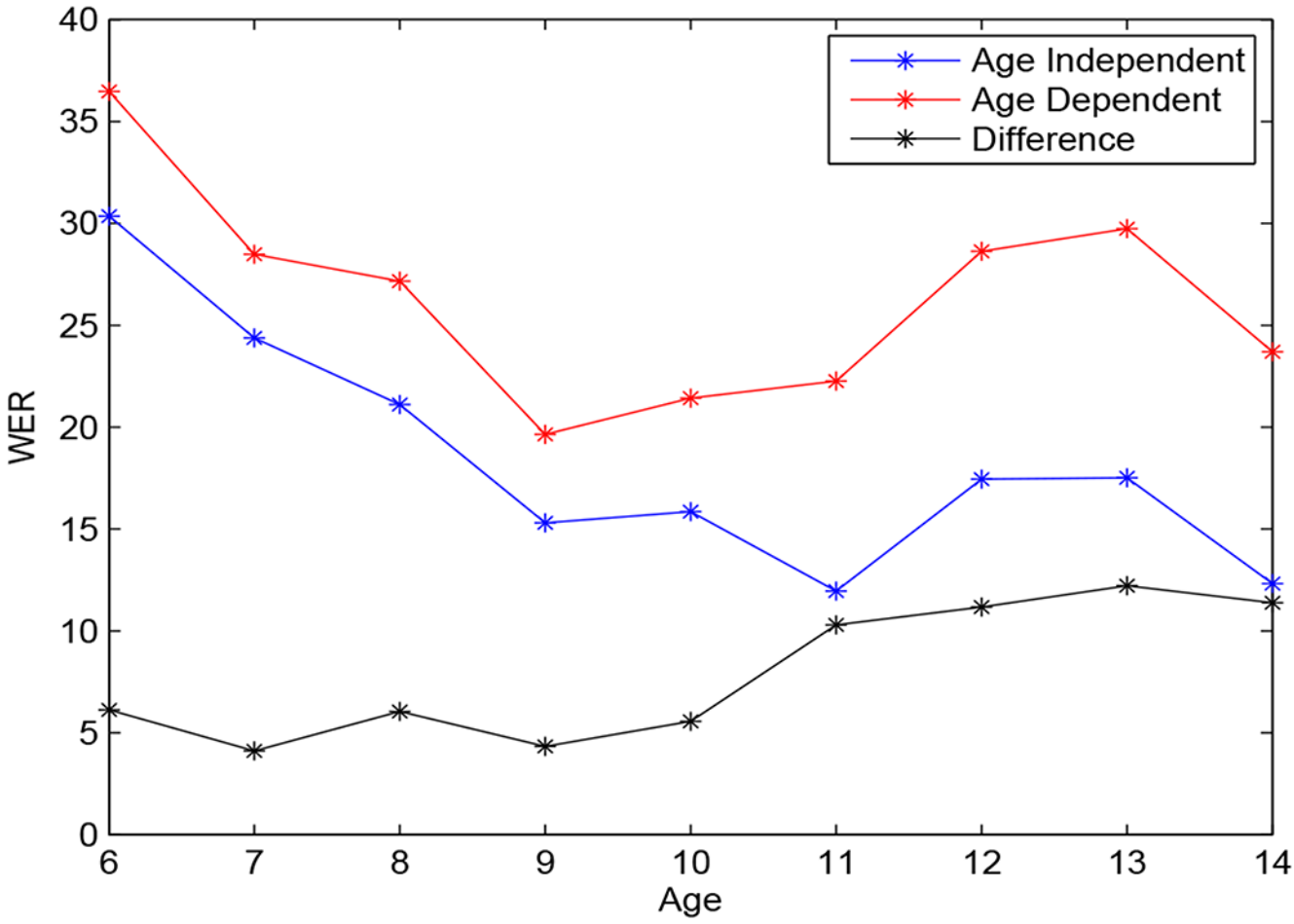

Figure 11:
Age dependent transformations versus Age independent transformations

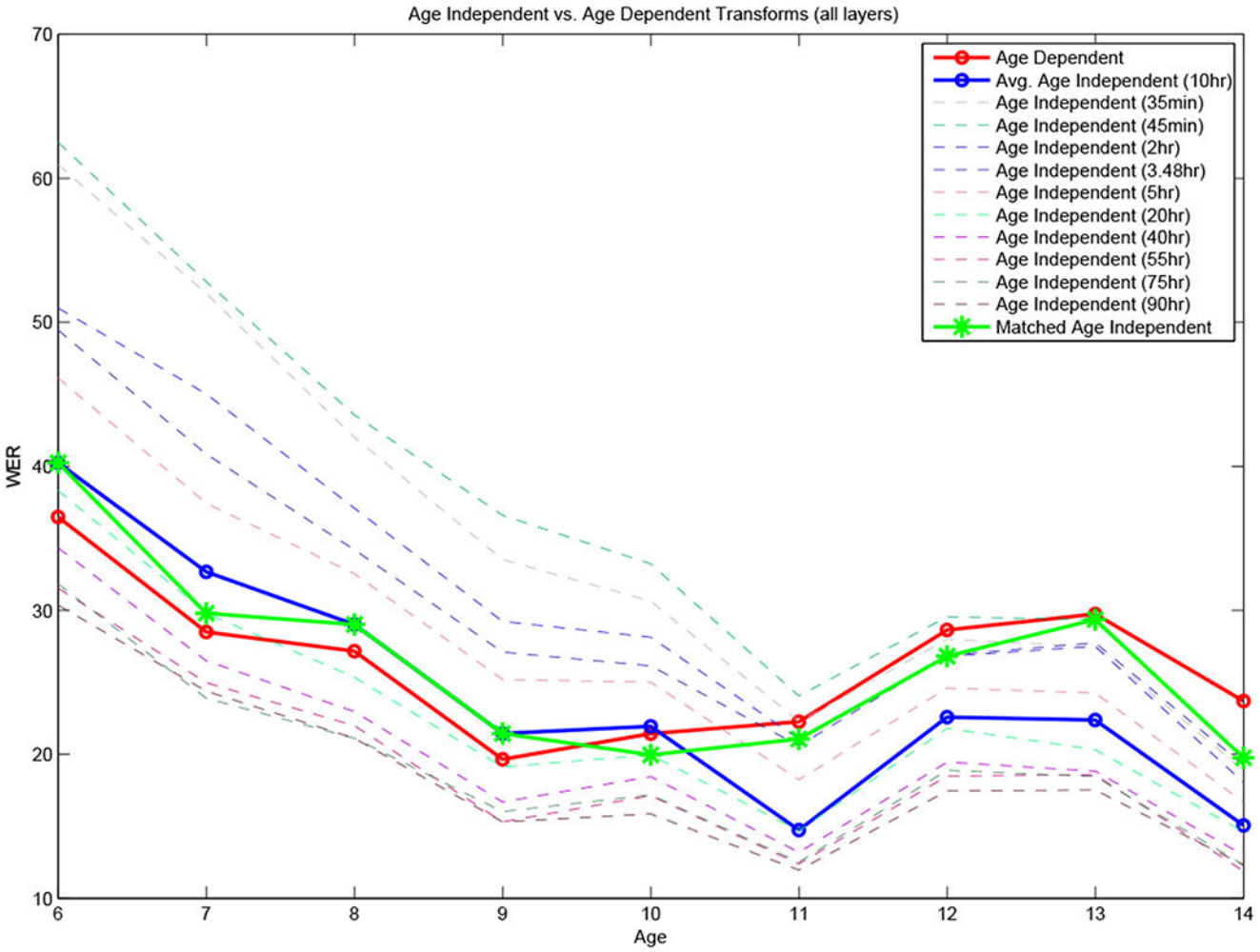

Figure 12:

Age dependent transformations versus data-normalized Age independent transformations (Adapting all layers)

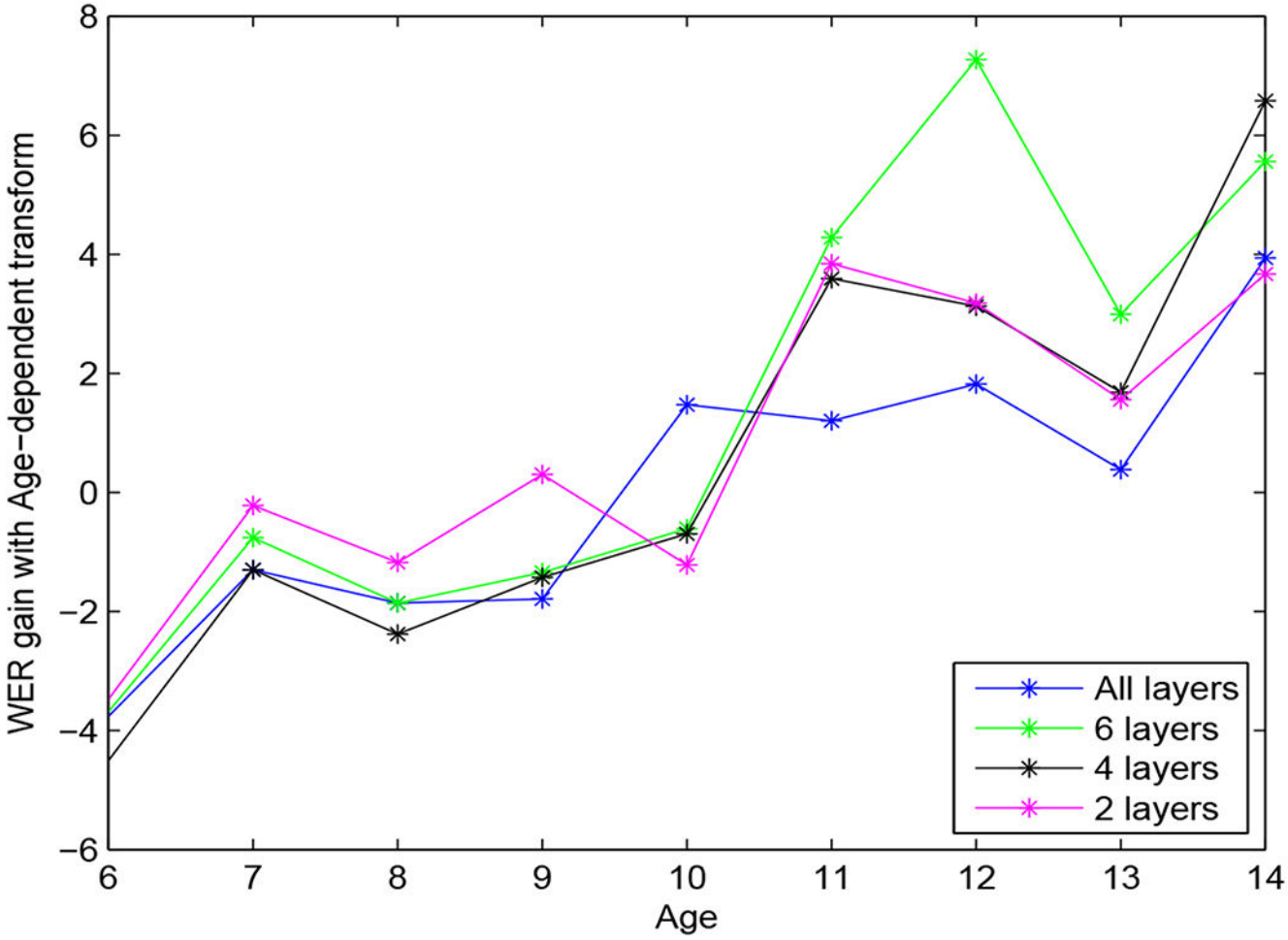

Figure 13:
Effect of adaptation layer configurations: Difference of WER between Age dependent transformations and matched age-independent transformations

Table 1:

Summary of Corpora and their training-testing splits

Corpus	# Hours	# Speakers	Age	Split
CU Prompted & Read	25.69	663	6-11	Train
CU Read & Summarized	33.11	320	6-11	Train
OGI	22.56	509	6-11	Train
ChIMP	10.25	97	6-14	Train
CID	2.26	324	6-14	Test
Total (Children-Train)	91.61	1589	6 - 14	Train
TED-LIUM (Adult)	205.82	774	NA	Train

Table 2:

Baseline results of ASR trained only on children's speech (91 hours).

Model	WER
GMM-HMM Monophone	54.53%
GMM-HMM Triphone	36.96%
GMM-HMM Triphone LDA+MLLT	32.79%
GMM-HMM Triphone LDA+MLLT+SAT	**24.55%**
GMM-HMM Triphone LDA+MLLT+SAT + VTLN	25.66%
Hybrid DNN-HMM	35.97%
Hybrid DNN-HMM + VTLN	32.72%
Hybrid DNN-HMM + LDA+MLLT+SAT	**21.31%**
Hybrid DNN-HMM + LDA+MLLT+SAT + VTLN	21.82%
Hybrid DNN-HMM + online i-vector (speaker)	28.03%
Hybrid DNN-HMM + online i-vector (utterance)	26.59%
Hybrid DNN-HMM + offline i-vector (utterance)	25.53%

Table 3:

WER results for artificially generated acoustic variability on TEDLIUM

	Time-stretching	Pitch-shifting	VTLN-warping	Global
No adaptation	16.08	37.14	13.52	20.05
Bottom layer	16.41	20.68	13.11	16.27
Top layer	16.88	26.77	13.95	18.37

Table 4:

Best results obtained for different base models

Model	(Proposed) Adult (TL)	Adult + Children (TL)	Children ASR
WER %	**17.8%**	19.06%	25.53%

Table 5:

Transfer Learning Results (DNN: Hybrid DNN-HMM + offline i-vector AV: Acoustic Variability Modeling, PV: Pronunciation Variability Modeling) - 91 hours

Model	AV	PV	Configuration	WER
DNN Children	✘	✘	Baseline	25.53%
DNN Adult	✘	✘	Baseline	39.32%
DNN Children + Adult	✘	✘	-	20.35%
DNN TL	✘	✓	1 layer	26.97%
DNN TL	✓	✘	1 layer	24.26%
DNN TL	✓	✓	1 layer each	**19.63%**
DNN TL	✓	✓	dis-joint 1 layer each	20.01%
DNN TL	✓	✓	2 layers each	**17.8%**
DNN TL	✓	✓	dis-joint 2 layers each	18.74%
DNN TL	-	-	all layers	**17.8%**

Table 6:

Transfer learning on fMLLR Transforms (SAT)

	fMLLR/Adult	fMLLR/Children
No transfer learning	59.59%	21.31%
1-middle & 1-top	36.76%	23.61%
2-middle & 2-top	36.09%	23.8%
3-middle & 3-top	35.16%	23.74%
all layers	34.38%	24.04%

Table 7:

Adaptation at extreme low data scenarios

Adaptation Data	Model (training)	WER
35 minutes	1 layer	36.47%
35 minutes	1 layer (dis-joint)	34.13%
35 minutes	2 layers (simultaneous)	35.73%
35 minutes	2 layers (dis-joint)	35.04%
45 minutes	1 layer	35.23%
45 minutes	1 layer (dis-joint)	33.62%
45 minutes	2 layers (simultaneous)	35.13%
45 minutes	2 layers (dis-joint)	34.33%
2 hours	1 layer	33.25%
2 hours	1 layer (dis-joint)	33.62%
2 hours	2 layers (simultaneous)	32.35%
2 hours	2 layers (dis-joint)	32.94%